

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA AGRÍCOLA

**PROCESSO DE DESCOBERTA DE CONHECIMENTO EM
BASES DE DADOS PARA A ANÁLISE E O ALERTA
DE DOENÇAS DE CULTURAS AGRÍCOLAS E SUA
APLICAÇÃO NA FERRUGEM DO CAFEIRO**

CARLOS ALBERTO ALVES MEIRA

CAMPINAS
JUNHO DE 2008

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA AGRÍCOLA

**PROCESSO DE DESCOBERTA DE CONHECIMENTO EM
BASES DE DADOS PARA A ANÁLISE E O ALERTA
DE DOENÇAS DE CULTURAS AGRÍCOLAS E SUA
APLICAÇÃO NA FERRUGEM DO CAFEEIRO**

Tese de Doutorado submetida à banca examinadora
para obtenção do título de Doutor em Engenharia
Agrícola, na área de concentração em Planejamento
e Desenvolvimento Rural Sustentável.

CARLOS ALBERTO ALVES MEIRA

Orientador: Prof. Dr. Luiz Henrique Antunes Rodrigues

CAMPINAS
JUNHO DE 2008

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

M478p Meira, Carlos Alberto Alves
Processo de descoberta de conhecimento em bases de dados para a análise e o alerta de doenças de culturas agrícolas e sua aplicação na ferrugem do cafeeiro / Carlos Alberto Alves Meira.--Campinas, SP: [s.n.], 2008.

Orientador: Luiz Henrique Antunes Rodrigues
Tese (Doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Mineração de dados (Computação). 2. Classificação. 3. Modelos. 4. Árvore de decisão. 5. *Hemileia vastatrix*. I. Rodrigues, Luiz Henrique Antunes. II. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. III. Título.

Título em Inglês: Process of knowledge discovery in databases for analysis and warning of crop diseases and its application on coffee rust.

Palavras-chave em Inglês: Data mining; Classification; Decision tree; Plant disease forecasting system; Predictive model; *Hemileia vastatrix*.

Área de concentração: Planejamento e Desenvolvimento Rural Sustentável

Titulação: Doutor em Engenharia Agrícola

Banca examinadora: Laércio Zambolim, Maria Carolina Monard, Sérgio Almeida de Moraes, Sílvia Maria Fonseca Silveira Massruhá.

Data da defesa: 13/06/2008

Programa de Pós-Graduação: Engenharia Agrícola

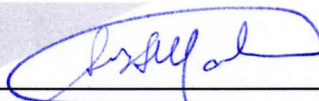
Este exemplar corresponde à versão final da **Tese de Doutorado** defendida por **Carlos Alberto Alves Meira**, aprovada pela Comissão Julgadora em 13 de junho de 2008, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.



Prof. Dr. Luiz Henrique Antunes Rodrigues - Orientador
FEAGRI/UNICAMP



Prof. Dr. Laércio Zambolim - Membro Titular
UFV/Viçosa-MG



Dr^a. Silvia Maria Fonseca Silveira Massruhá - Membro Titular
EMBRAPA/CNPTIA



Dr. Sérgio Almeida de Moraes - Membro Titular
IAC/Campinas-SP



Prof^a. Dr^a Maria Carolina Monard - Membro Titular
USP/São Carlos-SP

Dedico
à minha esposa LUCIANA,
ao nosso filho JOÃO VICTOR
e aos meus pais,
com muito carinho.

AGRADECIMENTOS

À Embrapa - Empresa Brasileira de Pesquisa Agropecuária, pela oportunidade de capacitação profissional.

Ao Prof. Dr. Luiz Henrique Antunes Rodrigues, pela orientação e pelo companheirismo.

Ao Dr. Sérgio Almeida de Moraes do Instituto Agronômico de Campinas, pela atenção e pela colaboração técnica na área de fitopatologia.

À Fundação Procafé, em nome de Antônio Wander Rafael Garcia e Leonardo Biscaro Japiassú, por ceder os dados utilizados no desenvolvimento do trabalho.

Ao SAS Brasil, pela concessão da licença de uso do SAS[®] *Education Analytical Suite* e do SAS[®] *Enterprise Miner*[™] por meio do seu Programa Acadêmico.

Ao Laboratório de Inteligência Computacional do ICMC/USP, campus de São Carlos, nas pessoas da Profa. Dra. Maria Carolina Monard e do Dr. Ronaldo Cristiano Prati, pela receptividade e pela colaboração técnica na área de inteligência artificial.

À Embrapa Informática Agropecuária, em nome do Dr. José Gilberto Jardine e do Dr. Eduardo Delgado Assad, pela oportunidade de utilizar as dependências físicas e a infraestrutura computacional durante o curso.

Às bibliotecárias da Embrapa Informática Agropecuária, em especial Leila Maria Lenk e Maria Goretti Gurgel Praxedes, pela presteza na obtenção de diversas referências bibliográficas utilizadas neste trabalho.

Ao Sistema Brasileiro de Informação do Café, pela disponibilidade na internet da Biblioteca do Café, por meio da qual foram obtidas referências bibliográficas utilizadas neste trabalho.

À amiga Marcia Izabel Fugisawa Souza, pela disposição de sempre em colaborar na elaboração das referências bibliográficas.

À Dra. Silvia Maria Fonseca Silveira Massruhá, pela amizade e pelo apoio, no âmbito da Embrapa, durante a realização do curso.

Ao pessoal do Laboratório de Informática da FEAGRI/UNICAMP, em especial Enzo Gomes Beato, pela amizade e pelo apoio em diversas atividades durante o curso.

Aos demais professores e funcionários da FEAGRI/UNICAMP, que direta ou indiretamente contribuíram para a realização do curso.

Aos meus pais, João Antonio e Conceição Aparecida (Cidinha), que sempre me incentivaram e me ensinaram a dar valor aos estudos e à contínua formação.

À minha esposa Luciana e ao nosso filho João Victor, pelo amor, pelo apoio incondicional e pela compreensão durante todo o curso, principalmente nos momentos mais difíceis.

E a Deus pela vida.

SUMÁRIO

LISTA DE FIGURAS	XI
LISTA DE TABELAS	XV
RESUMO	XIX
ABSTRACT	XXI
1 INTRODUÇÃO	1
2 REVISÃO BIBLIOGRÁFICA	5
2.1 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS - KDD	5
2.1.1 TAREFAS E TÉCNICAS DE MINERAÇÃO DE DADOS	7
2.1.2 ÁRVORES DE DECISÃO E REGRAS DE CLASSIFICAÇÃO	9
2.1.2.1 ÁRVORES DE DECISÃO	9
2.1.2.2 REGRAS DE CLASSIFICAÇÃO	11
2.1.2.3 INDUÇÃO DE ÁRVORES DE DECISÃO	12
2.1.3 AVALIAÇÃO DE MODELOS E DE REGRAS DE CLASSIFICAÇÃO	16
2.2 EPIDEMIOLOGIA DE DOENÇAS DE PLANTAS	22
2.2.1 INTRODUÇÃO E CONCEITOS	22
2.2.2 EPIDEMIOLOGIA DA FERRUGEM DO CAFEEIRO	25
2.3 ALERTA DE DOENÇAS DE PLANTAS	30
2.3.1 SISTEMAS DE ALERTA DE DOENÇAS DE PLANTAS	30
2.3.2 MODELOS DE PREVISÃO DE DOENÇAS DE PLANTAS	33
2.3.3 SISTEMAS E MODELOS DE ALERTA DA FERRUGEM DO CAFEEIRO	35
2.3.4 ÁRVORES DE DECISÃO COMO MODELOS DE PREVISÃO DE DOENÇAS DE PLANTAS	41
3 MATERIAL E MÉTODOS	45
3.1 OS DADOS BRUTOS	45
3.2 MODELO DO PROCESSO	46
3.3 ENTENDIMENTO DOS DADOS	48
3.3.1 COLEÇÃO DE DADOS INICIAL	48
3.3.2 DESCRIÇÃO DOS DADOS	49
3.3.3 EXPLORAÇÃO DOS DADOS	51
3.3.4 VERIFICAÇÃO DA QUALIDADE DOS DADOS	55
3.4 PREPARAÇÃO DOS DADOS	60
3.4.1 ESPECIFICAÇÃO DO ATRIBUTO META	60
3.4.2 ESPECIFICAÇÃO DOS ATRIBUTOS PREDITIVOS	61
3.4.3 ESPECIFICAÇÃO DOS ATRIBUTOS PREDITIVOS ESPECIAIS	63
3.4.4 PASSOS DA PREPARAÇÃO DOS DADOS	65
3.4.5 PREPARATIVOS FINAIS PARA A MODELAGEM	68
3.5 MODELAGEM	69
3.5.1 GERAÇÃO DOS MODELOS	69
3.5.2 AVALIAÇÃO DOS MODELOS E DAS REGRAS	73
3.5.3 USO E CONFIGURAÇÃO DAS FERRAMENTAS DE MODELAGEM	75
3.6 ESPECIALIZAÇÃO DO MODELO DO PROCESSO	81

4	<u>ANÁLISE DA EPIDEMIA DA FERRUGEM DO CAFEIEIRO COM ÁRVORE DE DECISÃO</u>	85
4.1	CONSIDERAÇÕES INICIAIS	85
4.2	RESULTADOS	85
4.3	DISCUSSÃO	90
4.4	CONSIDERAÇÕES FINAIS	94
5	<u>MODELOS DE ALERTA DA FERRUGEM DO CAFEIEIRO</u>	95
5.1	CONSIDERAÇÕES INICIAIS	95
5.2	MODELOS PARA LAVOURAS COM ALTA CARGA PENDENTE DE FRUTOS	96
5.2.1	ALERTA QUANDO A TAXA DE INFECÇÃO FOR ATINGIR OU ULTRAPASSAR 5 P.P.	96
5.2.1.1	RESULTADOS	96
5.2.1.2	DISCUSSÃO	109
5.2.2	ALERTA QUANDO A TAXA DE INFECÇÃO FOR ATINGIR OU ULTRAPASSAR 10 P.P.	113
5.2.2.1	RESULTADOS	113
5.2.2.2	DISCUSSÃO	125
5.3	MODELOS PARA LAVOURAS COM BAIXA CARGA PENDENTE DE FRUTOS	131
5.3.1	ALERTA QUANDO A TAXA DE INFECÇÃO FOR ATINGIR OU ULTRAPASSAR 5 P.P.	131
5.3.1.1	RESULTADOS	131
5.3.1.2	DISCUSSÃO	144
5.3.2	ALERTA QUANDO A TAXA DE INFECÇÃO FOR ATINGIR OU ULTRAPASSAR 10 P.P.	149
5.3.2.1	RESULTADOS	149
5.3.2.2	DISCUSSÃO	163
5.4	CONSIDERAÇÕES FINAIS	167
6	<u>CARACTERIZAÇÃO DO PROCESSO</u>	171
6.1	CONSIDERAÇÕES INICIAIS	171
6.2	CONTEXTO DE MINERAÇÃO DE DADOS	171
6.3	TAREFAS ESPECIALIZADAS	171
6.3.1	COMPREENSÃO DO DOMÍNIO	172
6.3.2	ENTENDIMENTO DOS DADOS	175
6.3.3	PREPARAÇÃO DOS DADOS	177
6.3.4	MODELAGEM	180
6.4	CONSIDERAÇÕES FINAIS	184
7	<u>CONCLUSÕES</u>	185
7.1	SUGESTÕES PARA A CONTINUIDADE DO TRABALHO	188
	<u>REFERÊNCIAS BIBLIOGRÁFICAS</u>	191

LISTA DE FIGURAS

Figura 1: Visão geral das fases do processo de KDD (FAYYAD et al., 1996a).	7
Figura 2: Objetivos e tarefas de mineração de dados (adaptada de REZENDE et al., 2002).	8
Figura 3: Árvore de decisão para um exemplo simples de viagem (MONARD e BARANAUSKAS, 2002b).	10
Figura 4: Exemplo das operações de poda <i>subtree replacement</i> e <i>subtree raising</i> .	15
Figura 5: Triângulo de doença de planta (AGRIOS, 1988).	23
Figura 6: Ciclo da doença - ferrugem do cafeeiro (adaptado de APSNET, 2008).	27
Figura 7: Fases do modelo de processo CRISP-DM (CHAPMAN et al., 2000).	46
Figura 8: Evolução mensal da incidência da ferrugem do cafeeiro em lavouras com diferentes espaçamentos e cargas pendentes de frutos – média de 1998/1999 a 2005/2006.	52
Figura 9: Evolução mensal da incidência da ferrugem do cafeeiro no ano agrícola 2003/2004 em lavouras com diferentes espaçamentos e cargas pendentes de frutos.	52
Figura 10: Distribuição dos valores de incidência da ferrugem do cafeeiro independente do espaçamento e da carga pendente de frutos da lavoura.	53
Figura 11: Distribuição dos valores de incidência da ferrugem do cafeeiro de acordo com o espaçamento da lavoura.	54
Figura 12: Distribuição dos valores de incidência da ferrugem do cafeeiro de acordo com a carga pendente de frutos da lavoura.	54
Figura 13: Representação dia-a-dia do esquema usado na preparação dos dados meteorológicos.	62
Figura 14: Esquema geral da preparação dos dados para a modelagem.	66
Figura 15: Relação entre a temperatura média durante o período de molhamento foliar (THUR95_PINF) e a temperatura média durante o período noturno de molhamento foliar (THNUR95_PINF).	69
Figura 16: Diagrama do projeto no SAS [®] Enterprise Miner [™] .	77

Figura 17: Configuração básica usada no <i>Tree node</i> do Enterprise Miner™ .	78
Figura 18: Configuração avançada usada no <i>Tree node</i> do Enterprise Miner™ .	79
Figura 19: Configuração do classificador J48 e da validação cruzada no Weka.	80
Figura 20: Visão hierárquica da metodologia CRISP-DM (adaptada de CHAPMAN et al., 2000).	81
Figura 21: Distribuição percentual das três classes de taxa de infecção da ferrugem do cafeeiro, para cada mês e para o total dos meses.	86
Figura 22: Árvore de decisão que auxilia na compreensão das epidemias da ferrugem do cafeeiro.	87
Figura 23: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 1; Geração: Enterprise Miner™ .	97
Figura 24: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 1; Geração: Weka.	98
Figura 25: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 2; Geração: Enterprise Miner™ .	102
Figura 26: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 2; Geração: Weka.	102
Figura 27: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 3; Geração: Enterprise Miner™ .	106
Figura 28: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 3; Geração: Weka.	106

- Figura 29: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 1; Geração: Enterprise Miner™. _____ 114
- Figura 30: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 1; Geração: Weka. _____ 115
- Figura 31: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 2; Geração: Enterprise Miner™. _____ 118
- Figura 32: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 3; Geração: Enterprise Miner™. _____ 122
- Figura 33: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 3; Geração: Weka. _____ 123
- Figura 34: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 1; Geração: Enterprise Miner™. _____ 132
- Figura 35: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 1; Geração: Weka. _____ 132
- Figura 36: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 2; Geração: Enterprise Miner™. _____ 136
- Figura 37: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 2; Geração: Weka. _____ 136

Figura 38: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 3; Geração: Enterprise Miner™.	140
Figura 39: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 3; Geração: Weka.	141
Figura 40: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 1; Geração: Enterprise Miner™.	150
Figura 41: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 1; Geração: Weka.	151
Figura 42: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 2; Geração: Enterprise Miner™.	155
Figura 43: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 2; Geração: Weka.	156
Figura 44: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 3; Geração: Enterprise Miner™.	159
Figura 45: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 3; Geração: Weka.	160
Figura 46: Histograma com distribuição de valores de pressão barométrica.	176

LISTA DE TABELAS

Tabela 1: Matriz de confusão para a classificação com duas classes. _____	18
Tabela 2: Matriz de contingência de uma regra $A \rightarrow C$ (A - antecedente; C - consequente). _	20
Tabela 3: Matriz para cálculo dos valores de severidade da ferrugem (VSF) do cafeeiro, com base no período de molhamento foliar e na temperatura média do período (GARÇON et al., 2004). _____	38
Tabela 4: Descrição dos atributos relevantes registrados pela estação meteorológica. _____	49
Tabela 5: Descrição dos atributos relevantes dos boletins de avisos. _____	51
Tabela 6: Exploração dos dados registrados pela estação meteorológica no mês de janeiro de 2003. _____	55
Tabela 7: Testes de consistência aplicados nos atributos registrados pela estação meteorológica. _____	57
Tabela 8: Atributos preditivos usados na modelagem. _____	62
Tabela 9: Matriz de condições diárias de infecção e seus respectivos índices numéricos. ____	64
Tabela 10: Atributos preditivos especiais usados na modelagem. _____	65
Tabela 11: Conjunto de treinamento usado na indução da árvore de decisão aplicada na análise da epidemia da ferrugem do cafeeiro. _____	70
Tabela 12: Atributos preditivos de cada opção de seleção de atributos para a geração dos modelos de alerta da ferrugem do cafeeiro. _____	73
Tabela 13: Avaliação da árvore de decisão da Figura 22. _____	89
Tabela 14: Matriz de confusão da árvore de decisão da Figura 22. _____	89
Tabela 15: Distribuição dos exemplos de lavouras com alta carga pendente entre as classes '1' e '0' dos atributos meta TAXA_INF_M5 e TAXA_INF_M10. _____	96
Tabela 16: Matrizes de confusão da árvore de decisão da Figura 24. _____	99
Tabela 17: Avaliação da árvore de decisão da Figura 24. _____	99

Tabela 18: Regras extraídas da árvore de decisão da Figura 24 e avaliação de cada regra individualmente. _____	100
Tabela 19: Matrizes de confusão da árvore de decisão da Figura 25. _____	103
Tabela 20: Avaliação da árvore de decisão da Figura 25. _____	104
Tabela 21: Regras extraídas da árvore de decisão da Figura 25 e avaliação de cada regra individualmente. _____	104
Tabela 22: Matrizes de confusão da árvore de decisão da Figura 27. _____	107
Tabela 23: Avaliação da árvore de decisão da Figura 27. _____	107
Tabela 24: Regras extraídas da árvore de decisão da Figura 27 e avaliação de cada regra individualmente. _____	107
Tabela 25: Matrizes de confusão da árvore de decisão da Figura 30. _____	115
Tabela 26: Avaliação da árvore de decisão da Figura 30. _____	115
Tabela 27: Regras extraídas da árvore de decisão da Figura 30 e avaliação de cada regra individualmente. _____	116
Tabela 28: Matrizes de confusão da árvore de decisão da Figura 31. _____	119
Tabela 29: Avaliação da árvore de decisão da Figura 31. _____	119
Tabela 30: Regras extraídas da árvore de decisão da Figura 31 e avaliação de cada regra individualmente. _____	120
Tabela 31: Matrizes de confusão da árvore de decisão da Figura 33. _____	123
Tabela 32: Avaliação da árvore de decisão da Figura 33. _____	124
Tabela 33: Regras extraídas da árvore de decisão da Figura 33 e avaliação de cada regra individualmente. _____	124
Tabela 34: Distribuição dos exemplos de lavouras com baixa carga pendente entre as classes ‘1’ e ‘0’ dos atributos meta TAXA_INF_M5 e TAXA_INF_M10. _____	131
Tabela 35: Matrizes de confusão da árvore de decisão da Figura 35. _____	133
Tabela 36: Avaliação da árvore de decisão da Figura 35. _____	133

Tabela 37: Regras extraídas da árvore de decisão da Figura 35 e avaliação de cada regra individualmente.	133
Tabela 38: Matrizes de confusão da árvore de decisão da Figura 36.	137
Tabela 39: Avaliação da árvore de decisão da Figura 36.	137
Tabela 40: Regras extraídas da árvore de decisão da Figura 36 e avaliação de cada regra individualmente.	138
Tabela 41: Matrizes de confusão da árvore de decisão da Figura 39.	142
Tabela 42: Avaliação da árvore de decisão da Figura 39.	142
Tabela 43: Regras extraídas da árvore de decisão da Figura 39 e avaliação de cada regra individualmente.	143
Tabela 44: Matrizes de confusão da árvore de decisão da Figura 40.	151
Tabela 45: Avaliação da árvore de decisão da Figura 40.	152
Tabela 46: Regras extraídas da árvore de decisão da Figura 40 e avaliação de cada regra individualmente.	152
Tabela 47: Matrizes de confusão da árvore de decisão da Figura 42.	156
Tabela 48: Avaliação da árvore de decisão da Figura 42.	156
Tabela 49: Regras extraídas da árvore de decisão da Figura 42 e avaliação de cada regra individualmente.	157
Tabela 50: Matrizes de confusão da árvore de decisão da Figura 44.	160
Tabela 51: Avaliação da árvore de decisão da Figura 44.	161
Tabela 52: Regras extraídas da árvore de decisão da Figura 44 e avaliação de cada regra individualmente.	161
Tabela 53: Contexto de mineração de dados usado na especialização do modelo do processo.	171

RESUMO

Sistemas de alerta de doenças de plantas permitem racionalizar o uso de agrotóxicos, mas são pouco utilizados na prática. Complexidade dos modelos, dificuldade de obtenção dos dados necessários e custos para o agricultor estão entre as razões que inibem o seu uso. Entretanto, o desenvolvimento tecnológico recente – estações meteorológicas automáticas, bancos de dados, monitoramento agrometeorológico na Web e técnicas avançadas de análise de dados – permite se pensar em um sistema de acesso simples e gratuito.

Uma instância do processo de descoberta de conhecimento em bases de dados foi realizada com o objetivo de avaliar o uso de classificação e de indução de árvores de decisão na análise e no alerta da ferrugem do cafeeiro causada por *Hemileia vastatrix*. Taxas de infecção calculadas a partir de avaliações mensais de incidência da ferrugem foram agrupadas em três classes: TX1 - redução ou estagnação; TX2 - crescimento moderado (até 5 p.p.); e TX3 - crescimento acelerado (acima de 5 p.p.). Dados meteorológicos, carga pendente de frutos do cafeeiro (*Coffea arabica*) e espaçamento entre plantas foram as variáveis independentes. O conjunto de treinamento totalizou 364 exemplos, preparados a partir de dados coletados em lavouras de café em produção, de outubro de 1998 a outubro de 2006.

Uma árvore de decisão foi desenvolvida para analisar a epidemia da ferrugem do cafeeiro. Ela demonstrou seu potencial como modelo simbólico e interpretável, permitindo a identificação das fronteiras de decisão e da lógica contidas nos dados, auxiliando na compreensão de quais variáveis e como as interações dessas variáveis condicionaram o progresso da doença no campo. As variáveis explicativas mais importantes foram a temperatura média nos períodos de molhamento foliar, a carga pendente de frutos, a média das temperaturas máximas diárias no período de incubação e a umidade relativa do ar.

Os modelos de alerta foram desenvolvidos considerando taxas de infecção binárias, segundo os limites de 5 p.p e 10 p.p. (classe '1' para taxas maiores ou iguais ao limite; classe '0', caso contrário). Os modelos são específicos para lavouras com alta carga pendente ou para lavouras com baixa carga. Os primeiros tiveram melhor desempenho na avaliação. A estimativa de acurácia, por validação cruzada, foi de até 83%, considerando o alerta a partir de 5 p.p. Houve ainda equilíbrio entre a acurácia e medidas importantes como sensibilidade, especificidade e confiabilidade positiva ou negativa. Considerando o alerta a partir de 10 p.p.,

a acurácia foi de 79%. Para lavouras com baixa carga pendente, os modelos considerando o alerta a partir de 5 p.p. tiveram acurácia de até 72%. Os modelos para a taxa de infecção mais elevada (a partir de 10 p.p.) tiveram desempenho fraco. Os modelos mais bem avaliados mostraram ter potencial para servir como apoio na tomada de decisão referente à adoção de medidas de controle da ferrugem do cafeeiro.

O processo de descoberta de conhecimento em bases de dados foi caracterizado, com a intenção de que possa vir a ser útil em aplicações semelhantes para outras culturas agrícolas ou para a própria cultura do café, no caso de outras doenças ou pragas.

PALAVRAS-CHAVE: mineração de dados; classificação; árvore de decisão; sistema de previsão de doenças de plantas; modelo de previsão; *Hemileia vastatrix*.

ABSTRACT

Plant disease warning systems can contribute for diminishing the use of chemicals in agriculture, but they have received limited acceptance in practice. Complexity of models, difficulties in obtaining the required data and costs for the growers are among the reasons that inhibit their use. However, recent technological advance – automatic weather stations, databases, Web based agrometeorological monitoring and advanced techniques of data analysis – allows the development of a system with simple and free access.

A process instance of knowledge discovery in databases has been realized to evaluate the use of classification and decision tree induction in the analysis and warning of coffee rust caused by *Hemileia vastatrix*. Infection rates calculated from monthly assessments of rust incidence were grouped into three classes: TX1 - reduction or stagnation; TX2 - moderate growth (up to 5 pp); and TX3 - accelerated growth (above 5 pp). Meteorological data, expected yield and space between plants were used as independent variables. The training data set contained 364 examples prepared from data collected in coffee-growing areas between October 1998 and October 2006.

A decision tree has been developed to analyse the coffee rust epidemics. The decision tree demonstrated its potential as a symbolic and interpretable model. Its model representation identified the existing decision boundaries in the data and the logic underlying them, helping to understand which variables, and interactions between these variables, led to coffee rust epidemics in the field. The most important explanatory variables were mean temperature during leaf wetness periods, expected yield, mean of maximum temperatures during the incubation period and relative air humidity.

The warning models have been developed considering binary infection rates, according to the 5 pp and 10 pp thresholds (class '1' for rates greater than or equal the threshold; class '0', otherwise). These models are specific for growing areas with high expected yield or areas with low expected yield. The former had best performance in the evaluation. The estimated accuracy by cross-validation was up to 83%, considering the warning for 5 pp and higher. There was yet equivalence between accuracy and such important measures like sensitivity, specificity and positive or negative reliability. Considering the warning for 10 pp and higher, the accuracy was 79%. For growing areas with low expected

yield, the accuracy of the models considering the warning for 5 pp and higher was up to 72%. The models for the higher infection rate (10 pp and higher) had low performance. The best evaluated models showed potential to be used in decision making about coffee rust disease control.

The process of knowledge discovery in databases was characterized in such a way it can be employed in similar problems of the application domain with other crops or other coffee diseases or pests.

KEYWORDS: data mining; classification; decision tree; plant disease forecasting system; predictive model; *Hemileia vastatrix*.

1 INTRODUÇÃO

A racionalidade no uso de agrotóxicos traz proveito aos produtores, por um lado, por reduzir gastos com defensivos e com mão de obra. É benéfica ao meio ambiente e à sociedade, por outro lado, pois diminui os riscos de contaminação do solo, da água e dos alimentos cultivados, em época que a população mundial se preocupa, cada vez mais, com o consumo de produtos saudáveis e com a proteção da natureza.

Um dos meios de promover o uso racional de agrotóxicos é a utilização de sistemas de previsão ou de alerta de doenças de culturas agrícolas, principalmente daquelas causadas por fungos fitopatogênicos (REIS, 2004). O conhecimento, o monitoramento e o aviso das condições que favorecem uma doença e aumentam o risco de uma epidemia permitem o emprego de medidas de controle apenas quando necessário (CAMPBELL e MADDEN, 1990). Portanto, quando as condições não são favoráveis à doença, o número de aplicações de fungicida pode diminuir, em comparação com os esquemas convencionais baseados em calendário fixo.

Os sistemas de alerta, no entanto, são pouco utilizados na prática pelos agricultores. As razões dessa baixa adoção passam pela complexidade do modelo de previsão, pela dificuldade de obtenção dos dados necessários e pelos custos de implementação e de manutenção para o agricultor (CAMPBELL e MADDEN, 1990).

Os avanços tecnológicos dos últimos anos – instalação, pelos órgãos públicos, de um grande número de estações meteorológicas automáticas, organização de bancos de dados meteorológicos, disponibilidade de sistemas de monitoramento agrometeorológico na Web e proliferação de técnicas avançadas de análise de dados – permitem se pensar em um sistema de alerta de doenças de alcance público, gratuito e simples de usar.

A hipótese deste trabalho foi que uma análise de dados meteorológicos junto com registros de intensidade de doenças de culturas agrícolas causadas por fungos, caracterizada como um processo de descoberta de conhecimento em bases de dados (FAYYAD et al., 1996a), indicaria a viabilidade de uso dos modelos obtidos, em termos de acurácia de predição e de outras medidas cabíveis, na emissão de alertas dessas doenças, como parte integrante de um sistema de monitoramento agrometeorológico de acesso público e gratuito.

O objetivo geral foi executar e avaliar uma instância do processo de descoberta de conhecimento em bases de dados no desenvolvimento de modelos de alerta da ferrugem do

cafeeiro, buscando estabelecer as bases para a definição de um procedimento de emissão de alertas quanto à intensidade dessa doença, apoiado, basicamente, em dados meteorológicos.

Fez parte, também, do objetivo mais geral, caracterizar o processo realizado de descoberta de conhecimento, para permitir sua reprodução e adaptação em problemas similares de outras culturas agrícolas ou mesmo da cultura do café, para outras doenças e pragas.

O objetivo específico foi produzir modelos confiáveis de alerta da ferrugem do cafeeiro, a partir de dados meteorológicos, da carga pendente de frutos do cafeeiro e do espaçamento entre plantas, por meio de classificação e de indução de árvores de decisão. Tais modelos podem vir a ser usados como apoio na tomada de decisão referente à adoção de medidas de controle da doença, em busca de racionalização no uso de fungicidas.

A ferrugem é a principal doença do cafeeiro em todo o mundo e pode ser encontrada em todas as lavouras de café cultivadas no Brasil (ZAMBOLIM et al., 1997). A doença, além da importância econômica, atende outros requisitos importantes, que justificam o desenvolvimento de um sistema de alerta, como a variação na sua intensidade entre cada estação de cultivo e a disponibilidade de medidas de controle químico economicamente viáveis.

As árvores de decisão são de interesse especial para a descoberta de conhecimento em bases de dados, pois utilizam representações simbólicas e interpretáveis. Essas representações permitem a compreensão das fronteiras de decisão que existem nos dados e também da lógica implícita neles (APTE e WEISS, 1997).

O propósito básico da indução de uma árvore de decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema. No último caso, a intenção é compreender quais variáveis e interações dessas variáveis conduzem o fenômeno estudado. Esses dois objetivos podem aparecer juntos em um mesmo estudo (BREIMAN et al., 1984).

Sendo assim, um outro objetivo específico foi aplicar e avaliar o potencial da indução de árvores de decisão na análise da epidemia da ferrugem do cafeeiro. O intuito foi obter uma árvore de decisão capaz de auxiliar na compreensão de como as condições do ambiente, a carga pendente de frutos do cafeeiro e o espaçamento entre as plantas na lavoura condicionaram a taxa de infecção da doença, identificando os fatores mais importantes no progresso da ferrugem do cafeeiro no campo.

Os capítulos seguintes estão organizados conforme descrito a seguir. O capítulo 2 contém a revisão bibliográfica referente à temática deste trabalho: a descoberta de conhecimento em bases de dados, com noções gerais e a respeito das tarefas e técnicas, especificamente as árvores de decisão e as regras de classificação (seção 2.1); a epidemiologia de doenças de plantas, com uma introdução, conceitos gerais e a epidemiologia da ferrugem do cafeeiro (seção 2.2); e o alerta de doenças de plantas, com sistemas de alerta e modelos de previsão, em geral, sistemas e modelos de alerta da ferrugem do cafeeiro e o uso de árvores de decisão como modelos de previsão de doenças de plantas (seção 2.3).

O capítulo 3 apresenta o material e os métodos utilizados no desenvolvimento do trabalho. São descritos: os dados brutos utilizados (seção 3.1); o modelo adotado para o processo realizado de descoberta de conhecimento em bases de dados (seção 3.2); e como as fases desse processo foram executadas, em especial o entendimento dos dados (seção 3.3), a preparação dos dados (seção 3.4) e a modelagem (seção 3.5). É descrito, também, como foi feita a caracterização do processo de descoberta de conhecimento, a partir da especialização do modelo do processo (seção 3.6).

Os resultados e a discussão a respeito deles estão contemplados em três capítulos. No capítulo 4, é feita a análise da epidemia da ferrugem do cafeeiro com árvore de decisão. O capítulo 5 trata dos modelos de alerta da ferrugem do cafeeiro. E a caracterização do processo de descoberta de conhecimento em bases de dados é abordada no capítulo 6.

O capítulo 7 apresenta as conclusões deste trabalho e algumas sugestões para a sua continuidade. Ao final, são relacionadas todas as referências bibliográficas consultadas.

Ainda como parte deste trabalho, há um CD-ROM, em anexo, que contém arquivos com informações complementares a respeito do processo realizado de descoberta de conhecimento em bases de dados.

2 REVISÃO BIBLIOGRÁFICA

2.1 Descoberta de conhecimento em bases de dados - KDD

Nas últimas décadas, a capacidade de gerar e armazenar dados aumentou rapidamente. Esse crescimento explosivo na quantidade de dados armazenados gerou a necessidade por novas técnicas e ferramentas automatizadas que pudessem auxiliar na transformação desses dados em informação útil e conhecimento. A abundância de dados, em conjunto com a necessidade por ferramentas de análise, é conhecida como uma situação “rica em dados, mas pobre em informação” (HAN e KAMBER, 2001).

Ao mesmo tempo em que se percebia uma desproporção entre a geração de dados e a sua compreensão, havia uma expectativa crescente de que os dados, analisados e apresentados inteligentemente, seriam um recurso valioso a ser usado como vantagem competitiva (FRAWLEY et al., 1992).

Nesse contexto, surgiu a Descoberta de Conhecimento em Bases de Dados ou KDD, do termo em inglês *Knowledge Discovery in Databases*. KDD foi definida inicialmente como a extração de informação implícita, previamente desconhecida e potencialmente útil a partir de dados (FRAWLEY et al., 1992).

Embora diferentes tipos de informação possam ser descobertos nos dados, o foco foi sobre padrões expressos em linguagem de alto nível. Linguagem natural é muitas vezes desejável, pela perspectiva humana, mas não é conveniente para a manipulação pelos algoritmos de descoberta. Representações lógicas são mais naturais para a computação e, se necessário, podem ser traduzidas para um formato em linguagem natural (FRAWLEY et al., 1992).

A definição inicial de KDD foi posteriormente revisada:

“Descoberta de conhecimento em bases de dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em repositórios de dados.” (FAYYAD et al., 1996a, p. 6).

Nesta definição, o termo ‘processo’ indica que KDD compreende várias fases; ‘dados’ são um conjunto de fatos (p.ex. registros de uma base de dados); ‘padrões’ podem ser expressões em alguma linguagem, descrevendo subconjuntos dos dados, ou podem ser modelos aplicáveis a subconjuntos dos dados; os padrões descobertos devem ser ‘válidos’ em

novos dados, com algum grau de certeza; é desejável que os padrões sejam ‘novos’, de preferência para o usuário, e sejam ‘potencialmente úteis’, podendo conduzir a algum benefício; por fim, os padrões devem ser ‘compreensíveis’, se não imediatamente, então, após algum pós-processamento (FAYYAD et al., 1996a, 1996b, 1996c).

A área de KDD evoluiu, e continua a evoluir, da intersecção de áreas de pesquisa tais como aprendizado de máquina, reconhecimento de padrões, banco de dados, estatística e visualização de dados. Pode ser imaginada como a confluência dessas disciplinas (FAYYAD et al., 1996b).

KDD se baseia fortemente em técnicas conhecidas de aprendizado de máquina, de reconhecimento de padrões e de estatística para encontrar os padrões nos dados. A estatística oferece também métodos de quantificação da incerteza inerente quando se procura inferir padrões gerais a partir de amostras de uma população. As técnicas de visualização de dados estimulam naturalmente a percepção e a inteligência humana, aumentando a capacidade de entendimento e de associação de novos padrões (REZENDE et al., 2002).

O termo “mineração de dados” é muitas vezes usado como um sinônimo de KDD. Alternativamente, a mineração de dados é considerada uma etapa essencial do processo de KDD (HAN e KAMBER, 2001). Segundo Fayyad et al. (1996b), KDD refere-se ao processo global de descoberta de conhecimento a partir de dados, enquanto a mineração de dados é uma fase desse processo. Por essa visão, a mineração de dados refere-se à aplicação de algoritmos específicos para extrair os padrões dos dados. As outras fases do processo são também importantes para se garantir que conhecimento útil seja derivado dos dados.

A mineração de dados está intimamente associada à noção de extração de conhecimento a partir de um grande volume de dados. Entretanto, o processo de KDD pode ser realizado independentemente da quantidade de dados disponível, em todas as suas fases.

A Figura 1 apresenta uma visão geral das fases do processo de KDD (FAYYAD et al., 1996a): primeiro, antes de se começar a mexer com os dados, é preciso compreender o domínio de aplicação e identificar a meta da descoberta de conhecimento pelo ponto de vista do usuário; em seguida, os dados de interesse são selecionados, é feito um pré-processamento nos dados (p.ex. eliminação de ruídos e tratamento de dados ausentes), os dados sofrem transformações (p. ex. conversão de dados e derivação de novos atributos) e é realizada a mineração de dados; ao final, é feita a interpretação e a avaliação dos resultados

obtidos, e o conhecimento descoberto é distribuído conforme se tenha definido no planejamento. Esse processo pode envolver várias iterações e quase sempre é necessário o retorno para fases anteriores.

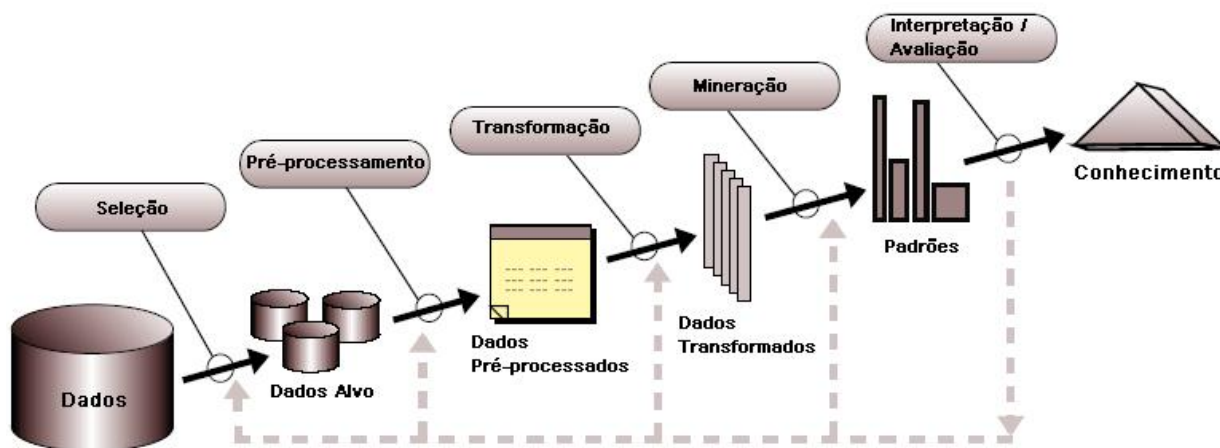


Figura 1: Visão geral das fases do processo de KDD (FAYYAD et al., 1996a).

Rezende et al. (2002) consideraram o processo de mineração de dados dividido em três grandes etapas: pré-processamento, extração de padrões e pós-processamento. Incluíram, ainda, uma fase anterior ao processo, referente ao conhecimento do domínio e à identificação do problema, e outra fase posterior, relacionada com a utilização do conhecimento obtido.

O processo de KDD é centrado na cooperação entre os seus diversos atores, e o seu sucesso depende, em parte, dessa cooperação. Os atores do processo podem ser divididos em três classes (REZENDE et al., 2002):

Especialista do domínio: pessoa que deve possuir amplo conhecimento do domínio de aplicação e deve fornecer apoio para a execução do processo.

Analista de dados: pessoa responsável pela execução do processo de KDD. Este usuário deve conhecer a fundo as etapas que compõem o processo.

Usuário final: representa a classe de usuários que vai utilizar o conhecimento extraído como auxílio em um processo de tomada de decisão.

2.1.1 Tarefas e técnicas de mineração de dados

Na prática, os dois objetivos principais da mineração de dados são a predição e a descrição. A predição envolve o uso de variáveis com valores conhecidos para prever um valor desconhecido ou futuro de outra variável (atributo meta). A descrição caracteriza propriedades gerais encontradas nos dados, com foco em padrões interpretáveis pelo ser

humano. Esses objetivos podem ser alcançados por meio de vários tipos de tarefa. A escolha de uma ou mais tarefas depende do problema em questão. As tarefas tradicionais de mineração de dados estão representadas na Figura 2 e são brevemente descritas a seguir (HAN e KAMBER, 2001; FAYYAD et al., 1996b).

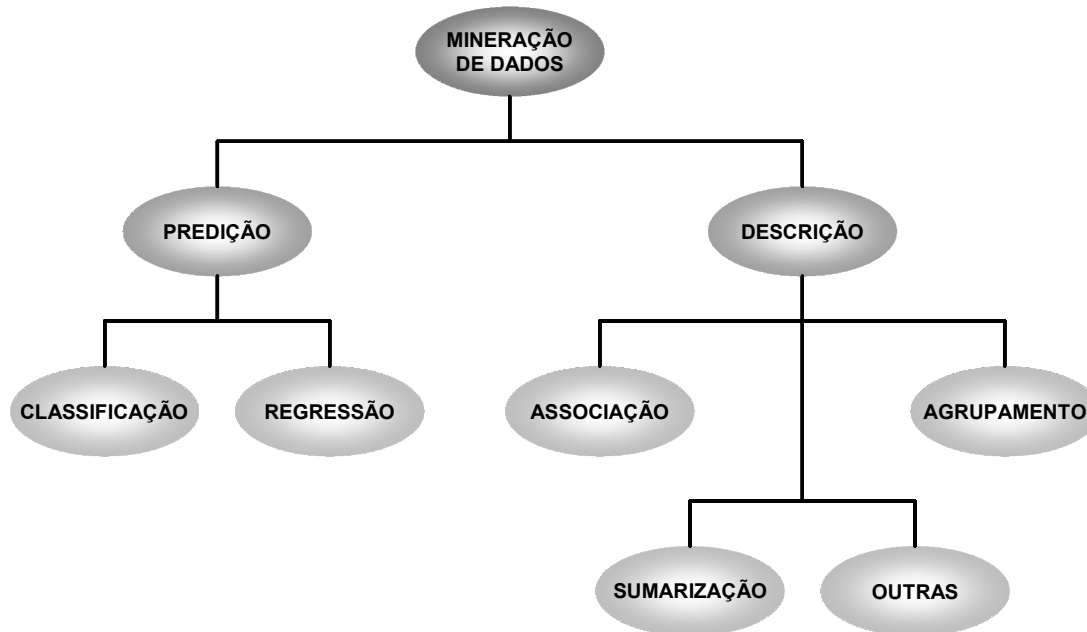


Figura 2: Objetivos e tarefas de mineração de dados (adaptada de REZENDE et al., 2002).

Classificação: consiste em descobrir uma função que mapeie (classifique) um item ou registro de dados para uma classe dentre algumas pré-definidas. A variável de predição é discreta ou categórica.

Regressão: consiste em descobrir uma função que mapeie um item de dados para uma variável de predição de valor numérico contínuo.

Associação: é a descoberta de regras de associação indicando condições de atributo-valor que ocorrem frequentemente juntas em um conjunto de dados.

Agrupamento (*clustering*): consiste em agrupar os dados em classes ou *clusters*, tal que os elementos de uma classe tenham alta similaridade entre si e sejam diferentes dos elementos das outras classes. Ao contrário da classificação, o rótulo de classe de cada elemento não é conhecido de antemão.

Sumarização: envolve meios de encontrar uma descrição compacta para um subconjunto de dados, como, por exemplo, derivação de regras resumidas ou visualização multivariada.

Cada tarefa de mineração de dados possui técnicas diferentes associadas. Dentre as mais populares estão (MICHALSKI et al., 1998; HAN e KAMBER, 2001): árvores de decisão, regras de classificação, redes neurais, análise de cesta de mercado, vizinhos mais próximos, regressão linear ou não linear e algoritmos genéticos. Há também abordagens híbridas, que aplicam duas ou mais técnicas em conjunto.

Não existe a melhor técnica, cada uma possui vantagens e desvantagens. A escolha de uma técnica requer uma análise mais detalhada do problema em mãos e a decisão de qual representação e estratégia de descoberta sejam mais adequadas. Melhor ainda é poder aplicar mais de uma técnica para resolver o mesmo problema e no final escolher o produto daquela que apresentar os melhores resultados.

2.1.2 Árvores de decisão e regras de classificação

As técnicas de indução¹ de árvores de decisão e de regras de classificação são consideradas técnicas de aprendizado orientadas a conhecimento, em que o interesse principal consiste em obter descrições simbólicas que sejam de fácil compreensão e utilização por meio de modelos mentais (MONARD e BARANAUSKAS, 2002b). São adequadas ao processo de KDD, pois, como visto, KDD dá ênfase especial em descobrir padrões compreensíveis que possam ser interpretados como conhecimento útil ou interessante.

Soluções simbólicas permitem a compreensão das fronteiras de decisão que existem nos dados e também da lógica implícita neles (APTE e WEISS, 1997). As redes neurais, por exemplo, embora possam ter alta precisão, são relativamente difíceis de compreender quando comparadas com as árvores de decisão (FAYYAD et al., 1996a).

2.1.2.1 Árvores de decisão

Uma árvore de decisão é um modelo representado graficamente por nós e ramos, parecido com uma árvore, mas no sentido invertido (HAN e KAMBER, 2001; MONARD e BARANAUSKAS, 2002b; WITTEN e FRANK, 2005).

O nó raiz é o primeiro nó da árvore, no topo da estrutura. Os nós internos, incluindo o nó raiz, são nós de decisão. Cada um contém um teste sobre um ou mais atributos (variáveis independentes) e os resultados desse teste formam os ramos da árvore. Geralmente, o teste em um nó compara o valor de um atributo com um valor constante. No entanto, algumas árvores

¹ A indução é a forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos (MONARD e BARANAUSKAS, 2002a).

podem comparar dois atributos entre si, ou utilizar alguma função envolvendo um ou mais atributos (WITTEN e FRANK, 2005).

Cada nó folha, nas extremidades da árvore, representa um valor de predição para o atributo meta (variável dependente) ou uma distribuição de probabilidade dos seus possíveis valores. As árvores de decisão são também chamadas de árvores de classificação ou de regressão, caso o atributo meta seja categórico ou numérico, respectivamente.

Se o atributo de teste em um nó é categórico (nominal), o número de ramos a partir do nó de decisão pode ser a quantidade dos possíveis valores do atributo. Outra maneira é dividir os valores do atributo em dois subconjuntos e, portanto, a decisão vai ficar entre duas opções (dois ramos). Se o atributo de teste é numérico, geralmente o nó de decisão se ramifica em dois, de acordo com o teste do valor do atributo em relação a um valor constante. Uma árvore de decisão é denominada binária quando ela possui somente dois ramos a partir de qualquer um dos seus nós internos de decisão.

Como exemplo simples, a Figura 3 apresenta a árvore de decisão para uma aplicação ilustrativa, em que o atributo meta que se deseja prever é a recomendação para viajar ou não, dadas certas condições do tempo.

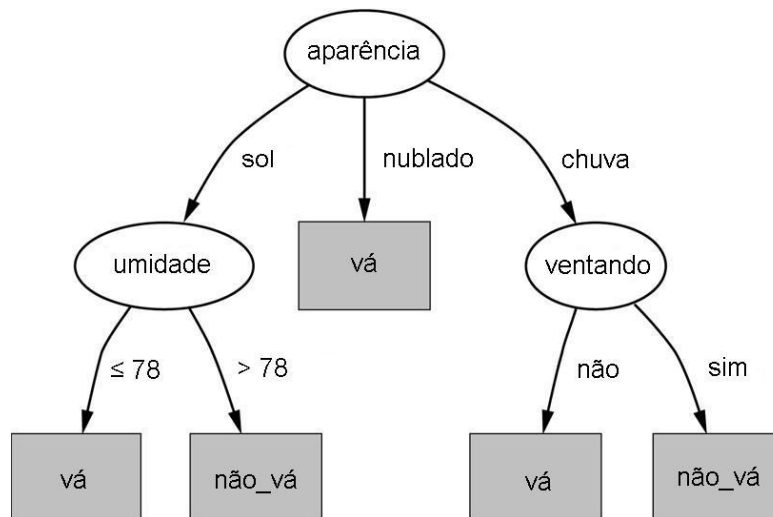


Figura 3: Árvore de decisão para um exemplo simples de viagem (MONARD e BARANAUSKAS, 2002b).

A árvore de decisão, depois de construída, pode ser utilizada para classificar exemplos cuja classe é desconhecida. Para classificar um exemplo, testam-se os valores de seus atributos segundo a árvore de decisão. Um caminho é traçado a partir do nó raiz, descendo pelos ramos de acordo com os resultados dos testes, até chegar em um nó folha, que representa a classe de predição do exemplo (HAN e KAMBER, 2001).

O propósito básico da indução de uma árvore de decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema (BREIMAN et al., 1984). No último caso, a intenção é compreender quais variáveis e interações dessas variáveis conduzem o fenômeno estudado. Esses dois propósitos não são excludentes, podendo aparecer juntos em um mesmo estudo.

2.1.2.2 Regras de classificação

Regras de classificação são uma alternativa às árvores de decisão. O antecedente de uma regra, ou pré-condição, é uma série de testes, como os testes nos nós de uma árvore de decisão. O conseqüente, ou conclusão, indica a classe que se aplica aos exemplos cobertos pela regra (WITTEN e FRANK, 2005).

O conhecimento representado em árvores de decisão pode ser extraído e representado na forma de regras de classificação SE-ENTÃO. Uma regra é criada para cada caminho entre a raiz e um nó folha. Os testes de valor de atributo ao longo do caminho formam uma conjunção no antecedente da regra e o nó folha transforma-se no conseqüente da regra. Essas regras podem ser mais fáceis de compreender, especialmente se a árvore de decisão for muito grande (HAN e KAMBER, 2001). A seguir, seguem as regras de classificação extraídas da árvore de decisão representada na Figura 3:

SE aparência = sol **E** umidade \leq 78 **ENTÃO** classe = vá

SE aparência = sol **E** umidade $>$ 78 **ENTÃO** classe = não_vá

SE aparência = nublado **ENTÃO** classe = vá

SE aparência = chuva **E** ventando = não **ENTÃO** classe = vá

SE aparência = chuva **E** ventando = sim **ENTÃO** classe = não_vá

Regras produzidas a partir de uma árvore de decisão são não ambíguas e disjuntas, ou seja, a ordem com que são executadas é irrelevante. Uma única regra é disparada quando um novo exemplo é classificado. Em geral, essas regras são mais complexas do que o necessário, sendo possível remover testes redundantes (MONARD e BARANAUSKAS, 2002b; WITTEN e FRANK, 2005).

Regras de classificação também podem ser induzidas diretamente do conjunto de exemplos. Basicamente, há duas formas de indução de regras: as ordenadas e as não ordenadas (MONARD e BARANAUSKAS, 2002b).

Na classificação com regras ordenadas, o classificador vai testando cada regra, na ordem definida, até encontrar uma cuja expressão seja satisfeita pelo novo exemplo. A classe de predição do novo exemplo é aquela indicada pelo conseqüente da regra disparada. Assim, a ordem das regras é fundamental e, exceto a primeira regra, as demais isoladamente não possuem validade própria. Se nenhuma regra é satisfeita, existe uma regra padrão que, em geral, atribui ao novo exemplo a classe mais comum (MONARD e BARANAUSKAS, 2002b).

No caso das regras não ordenadas, todas as regras são testadas na classificação de um novo exemplo. Sendo assim, mais de uma regra pode ser disparada, podendo ocorrer um conflito quando classes diferentes são preditas. Nesse caso, é comum atribuir ao novo exemplo a classe mais provável, considerando-se a soma das distribuições de probabilidade entre as classes associadas com cada regra disparada (MONARD e BARANAUSKAS, 2002b).

2.1.2.3 Indução de árvores de decisão

O algoritmo básico de indução de árvores de decisão constrói a árvore de forma recursiva, de cima para baixo, segundo a abordagem conhecida como “dividir-e-conquistar” (HAN e KAMBER, 2001; WITTEN e FRANK, 2005).

Inicia com um conjunto de exemplos de treinamento, que é dividido de acordo com um teste sobre um dos atributos preditivos, formando-se subconjuntos mais homogêneos em relação ao atributo meta. Esse procedimento é repetido até que se consiga conjuntos de exemplos bem homogêneos, para os quais seja possível atribuir um único valor para o atributo meta.

Critério de escolha do atributo de teste

O critério utilizado para escolher o atributo que divide o conjunto de exemplos em cada repetição é um dos aspectos principais do processo de indução, do qual depende o sucesso de um algoritmo de aprendizado por árvore de decisão (MONARD e BARANAUSKAS, 2002b). Entre os critérios mais conhecidos e usados estão: o ganho de informação e a razão de ganho, definidos com base na teoria da informação (QUINLAN, 1993); e o índice Gini (BREIMAN et al., 1984).

O ganho de informação, por exemplo, é uma medida usada para selecionar o atributo de teste em cada nó de decisão de uma árvore. O atributo com o maior ganho de informação

(ou maior redução de entropia²) é escolhido como o atributo de teste de cada nó, em cada iteração do processo de indução. Esse atributo minimiza a informação necessária para classificar os exemplos das partições resultantes da divisão, o que reflete a menor “impureza” dessas partições. Tal abordagem, ligada à teoria da informação, minimiza o número de testes esperados para classificar um exemplo e garante que uma árvore simples (não necessariamente a mais simples) seja encontrada (HAN e KAMBER, 2001).

Quando algum atributo preditivo tiver um grande número de valores possíveis, cuja escolha como atributo de teste em um nó vá resultar em uma divisão com vários ramos, surge um efeito indesejado com o cálculo do ganho de informação. O problema é que a medida de ganho de informação tende a dar preferência para atributos com grande quantidade de valores (WITTEN e FRANK, 2005). Para compensar esse efeito, foi feita uma alteração na medida do ganho de informação, resultando na medida denominada razão de ganho.

A razão de ganho leva também em consideração o número e o tamanho dos nós resultantes da divisão dos exemplos pelo atributo de teste, desconsiderando qualquer informação a respeito da classe. Quinlan (1993) a descreveu como uma medida robusta em uma grande variedade de circunstâncias. Embora uma solução prática, a razão de ganho sacrificou parte da elegância e da motivação teórica do critério de ganho de informação.

Valores ausentes (*missing values*)

Valores ausentes de atributos (*missing values*) incorporam uma questão clara na indução de árvores de decisão: qual ramo deve-se escolher quando um nó testa um atributo cujo valor está ausente?

Em alguns casos, o valor ausente pode ser tratado como um valor de atributo em si. Isso assume que a ausência de valor para o atributo tem um significado próprio. Se não for esse o caso, o valor ausente deve ser tratado de uma maneira especial, que não apenas ser considerado como mais um possível valor para o atributo em questão (WITTEN e FRANK, 2005).

Uma solução simples é registrar o número de exemplos do conjunto de treinamento associado com cada ramo de cada nó de decisão. Então, quando algum exemplo tiver valor

² Entropia: medida da variação ou desordem em um sistema; medida da desordem ou da imprevisibilidade da informação.

ausente para o atributo de teste de um nó, escolhe-se o ramo “mais popular”, ou seja, aquele com o maior número de exemplos.

Uma solução mais sofisticada é fracionar o exemplo com valor ausente e encaminhar as suas partes a cada um dos ramos da divisão do nó de decisão. Um peso é atribuído a cada ramo, entre 0 e 1, proporcional ao número de exemplos do conjunto de treinamento associado a esse ramo – os pesos de todos os ramos do nó de decisão somam 1. O exemplo fracionado pode sofrer novas divisões nos níveis seguintes da indução da árvore. Cada parte do exemplo, ao final do processo de indução, chega a um nó folha. Para a classificação desse exemplo, depois de induzida a árvore, as decisões em cada nó folha devem ser recombinadas, usando os pesos que percorreram o caminho até esses nós folhas, para que se decida a classe de predição do exemplo (WITTEN e FRANK, 2005).

Poda (*pruning*)

Após a construção da árvore de decisão, é possível que o modelo induzido seja muito específico para o conjunto de treinamento. Essa situação é conhecida como um super-ajuste aos dados ou *overfitting* (MONARD e BARANAUSKAS, 2002b). Alguns dos ramos da árvore de decisão podem também refletir anomalias nos dados devido a ruídos ou *outliers* (HAN e KAMBER, 2001). Para resolver esse problema, alguns indutores podam a árvore de decisão, reduzindo o número de nós internos e, conseqüentemente, diminuindo a complexidade da árvore (MONARD e BARANAUSKAS, 2002b).

Os métodos de poda geralmente usam medidas estatísticas para remover ramos menos confiáveis, resultando em uma classificação mais rápida e em um aumento da capacidade da árvore de classificar corretamente dados de teste independentes (HAN e KAMBER, 2001). Entre os métodos de pós-poda, que é a poda feita após o processo de indução da árvore de decisão, estão o erro pessimista (QUINLAN, 1993) e a complexidade do erro (BREIMAN et al., 1984).

Dois tipos diferentes de pós-poda foram considerados por Quinlan (1993) no algoritmo C4.5: substituição de subárvore (*subtree replacement*) e elevação de subárvore (*subtree raising*). *Subtree replacement* é a operação básica de poda, que visa selecionar alguma subárvore e substituí-la por um nó folha simples. *Subtree raising*, por sua vez, é uma operação mais complexa. Com ela, uma subárvore inteira é elevada e ocupa o lugar de outra subárvore, da qual a que foi elevada fazia parte.

Como exemplo ilustrativo, a árvore de decisão representada na Figura 4 (b) foi obtida a partir da árvore da Figura 4 (a) com os dois tipos de poda mencionados. Na subárvore à esquerda do nó raiz, a subárvore a partir do nó com o atributo de teste A_4 foi substituída por um nó folha (*subtree replacement*). Com isso, o nó folha F_1' da Figura 4 (b) passou a reunir todos os exemplos dos nós folhas F_1 , F_2 e F_3 da Figura 4 (a).

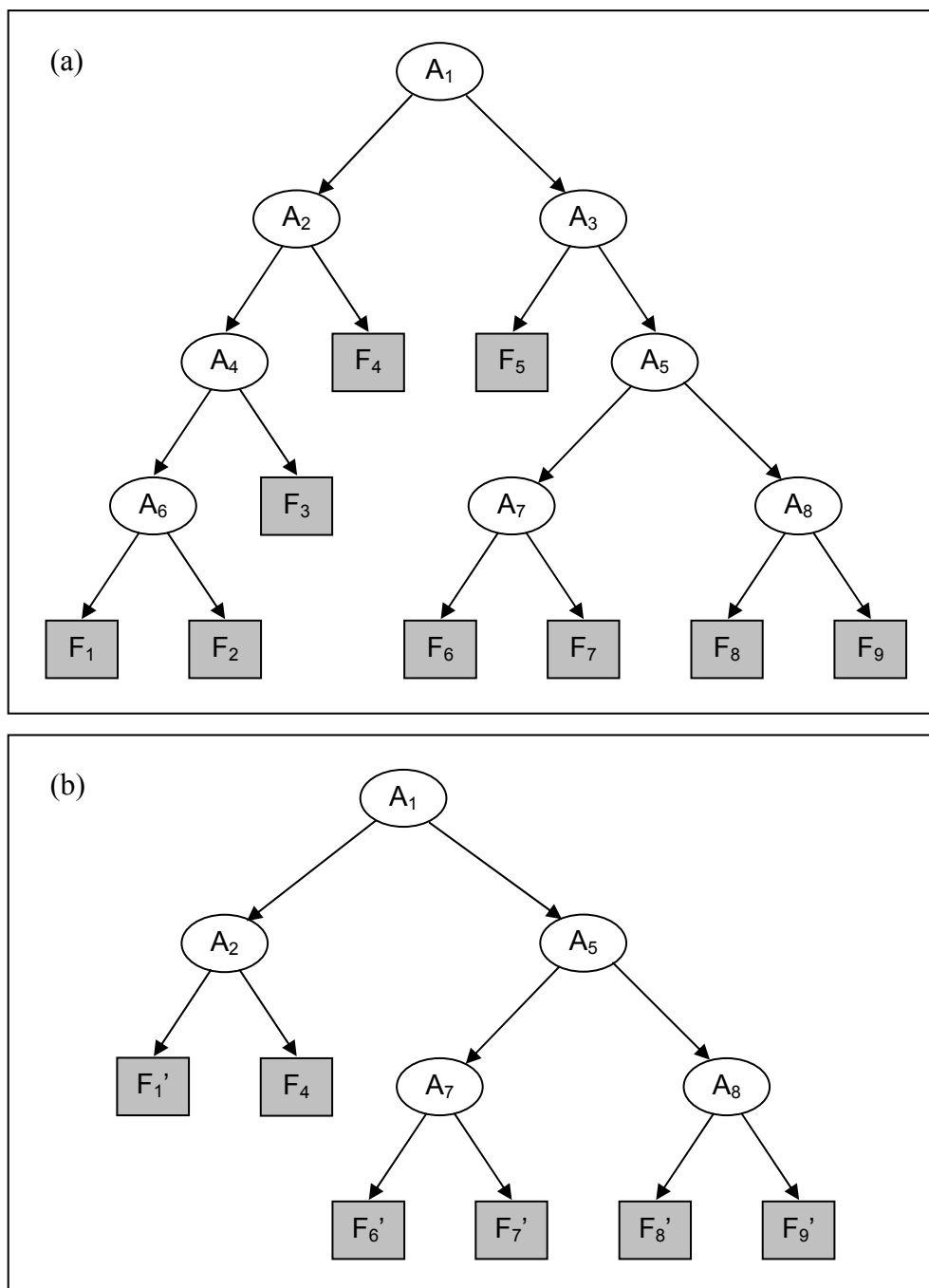


Figura 4: Exemplo das operações de poda *subtree replacement* e *subtree raising*.

Do outro lado, na subárvore à direita, a subárvore a partir do nó com o atributo de teste A_5 foi elevada e substituiu a subárvore a partir do nó com o atributo de teste A_3 (*subtree raising*). Com essa operação, os exemplos do nó folha F_5 foram reclassificados de acordo com a nova subárvore. Por isso, os nós folhas F_6' , F_7' , F_8' e F_9' da Figura 4 (b) estão identificados com apóstrofo, para indicar que não são os mesmos da árvore de decisão original representada na Figura 4 (a).

Na pré-poda, a abordagem é parar a construção da árvore de decisão assim que uma certa condição for satisfeita. Medidas como significância estatística, o próprio ganho de informação e outras, podem ser usadas durante a indução de uma árvore para avaliar a qualidade da divisão por um atributo de teste.

Se, particionando os exemplos em um nó, for resultar em uma divisão que fique abaixo de um limite especificado, segundo a medida adotada, decide-se por não criar o nó de decisão, deixando-o como um nó folha. A dificuldade, nesses casos, está em escolher um limite apropriado – valores altos podem resultar em árvores muito simplificadas, enquanto valores baixos podem resultar em pouca generalização (HAN e KAMBER, 2001).

A pré-poda, entretanto, pode ocasionar um efeito indesejado, quando dois atributos não contribuem para a decisão individualmente, mas juntos possuem alto poder preditivo ou descritivo. Uma conjunção de teste entre esses atributos seria a melhor forma de particionar os exemplos, mas a pré-poda pode evitar que tal conjunção se manifeste na árvore de decisão (MONARD e BARANAUSKAS, 2002b).

Outras possíveis regras de parada para a pré-poda são (SAS INSTITUTE INC., 2004c; WITTEN e FRANK, 2005): não deixar que a árvore de decisão ultrapasse um determinado nível de profundidade durante a indução; e não permitir que um nó folha seja criado sem um número mínimo de exemplos do conjunto de treinamento.

2.1.3 Avaliação de modelos e de regras de classificação

A taxa de erro e a acurácia, que é o complemento da taxa de erro, são as medidas de avaliação mais comuns para os modelos de classificação (HAN e KAMBER, 2001; WITTEN e FRANK, 2005). A taxa de erro é a proporção de erros de predição sobre um conjunto de exemplos em que se conhece o valor do atributo meta. A acurácia e a taxa de erro são estimativas do percentual de acertos e de erros do classificador, respectivamente, na predição da classe de novos exemplos.

Um conjunto de exemplos pode ser denotado por $\{(x_i, y_i), i = 1, 2, \dots, n\}$, onde x_i é um vetor de valores das variáveis independentes, y_i é o valor do atributo meta e n é a quantidade de exemplos. A taxa de erro de uma árvore de decisão h sobre este conjunto de exemplos é dada pela equação:

$$err(h) = \frac{1}{n} \sum_{i=1}^n D(h(x_i), y_i) \quad 2.1$$

onde $h(x_i)$ é o valor de predição para x_i e $D(a, b)$ é igual a 1, se a é diferente de b , ou igual a 0, caso contrário. A partir da taxa de erro, a acurácia é dada pela equação:

$$acc(h) = 1 - err(h) \quad 2.2$$

Um dos métodos de estimativa das medidas de avaliação, chamado de método de substituição, consiste em construir o classificador e testar o seu desempenho no mesmo conjunto de exemplos, ou seja, o conjunto de teste é idêntico ao conjunto de treinamento (MONARD e BARANAUSKAS, 2002a). Calcular a acurácia, por exemplo, sobre o conjunto de treinamento, normalmente resulta em uma estimativa altamente otimista, devido à especialização do modelo com respeito aos exemplos.

Uma das formas mais usadas de contornar esse problema é dividir aleatoriamente os exemplos em dois conjuntos independentes, um de treinamento e o outro de teste. Esse método é conhecido como *holdout* (HAN e KAMBER, 2001; WITTEN e FRANK, 2005). O conjunto de treinamento (tipicamente dois terços dos dados) é usado para induzir o modelo e a sua taxa de erro, bem como outra medida de avaliação qualquer, é estimada a partir do conjunto de teste.

Outro método bastante usado é a validação cruzada (*cross-validation*), particularmente quando a quantidade de dados para dividir entre treinamento e teste é limitada (WITTEN e FRANK, 2005). Na validação cruzada, os exemplos são aleatoriamente divididos em k partições mutuamente exclusivas (*folds*) de tamanho aproximadamente igual.

Uma das partições é reservada para teste, enquanto as demais, juntas, são usadas para treinamento. Este procedimento é executado k vezes, cada vez com uma partição diferente para teste. A taxa de erro, ao final, é calculada como a média das taxas de erro obtidas em cada uma das partições de teste. A vantagem da validação cruzada é usar cada um dos exemplos tanto para treinamento quanto para teste.

Testes extensivos em muitos e diferentes conjuntos de dados mostraram que dez é um número próximo do número exato de partições para se obter a melhor estimativa de erro pela validação cruzada, conhecida como *10-fold cross-validation* (WITTEN e FRANK, 2005).

A matriz de confusão de um classificador oferece meios efetivos para a avaliação do modelo com base em cada classe (MONARD e BARANAUSKAS, 2002a). Cada elemento da matriz mostra o número de exemplos para os quais a classe verdadeira é a linha e a classe predita é a coluna. A diagonal principal da matriz (elementos (i,j) , onde $i = j$) representa os acertos do modelo, enquanto os demais elementos representam os erros, discriminados para cada classe. Cada elemento da matriz (M) pode ser calculado segundo a equação:

$$M(C_i, C_j) = \sum_{\{x,y \in T: y=C_i\}} I(h(x), C_j) \quad 2.3$$

onde $i, j = 1, 2, \dots, k$; $\{C_1, C_2, \dots, C_k\}$ é o conjunto das classes para o atributo meta; (x, y) são os exemplos do conjunto de treinamento T ; e $I(a, b)$ é igual a 1, se a é igual a b , ou igual a 0, caso contrário.

A Tabela 1 ilustra a matriz de confusão para um problema com duas classes, denominadas C_+ (classe positiva) e C_- (classe negativa). Nesses casos, existem quatro possibilidades de acertos e de erros do classificador, identificadas como:

- Verdadeiros positivos (VP), quando os exemplos pertencem à classe C_+ e foram preditos como pertencentes a essa mesma classe.
- Falsos negativos (FN), quando os exemplos pertencem à classe C_+ e foram preditos como pertencentes à classe C_- .
- Verdadeiros negativos (VN), quando os exemplos pertencem à classe C_- e foram preditos como pertencentes a essa mesma classe.
- Falsos positivos (FP), quando os exemplos pertencem à classe C_- e foram preditos como pertencentes à classe C_+ .

Tabela 1: Matriz de confusão para a classificação com duas classes.

		Predita	
		C_+	C_-
Verdadeira	C_+	VP	FN
	C_-	FP	VN

Além da própria acurácia (equação 2.4) e da taxa de erro, outras medidas podem ser derivadas da matriz de confusão, tais como (MONARD e BARANAUSKAS, 2002a): **sensitividade** ou precisão da classe C_+ (*sensitivity* ou *recall* ou *true C_+ rate*); **especificidade** ou precisão da classe C_- (*specificity* ou *true C_- rate*); **confiabilidade positiva** ou confiabilidade de predição da classe C_+ (*positive reliability* ou *C_+ predictive value*); e **confiabilidade negativa** ou confiabilidade de predição da classe C_- (*negative reliability* ou *C_- predictive value*). Essas medidas são calculadas a partir das equações 2.5 a 2.8, respectivamente:

$$acc(h) = \frac{VP + VN}{n} \quad 2.4$$

$$sens(h) = \frac{VP}{VP + FN} \quad 2.5$$

$$spec(h) = \frac{VN}{VN + FP} \quad 2.6$$

$$prel(h) = \frac{VP}{VP + FP} \quad 2.7$$

$$nrel(h) = \frac{VN}{VN + FN} \quad 2.8$$

Em alguns casos, a acurácia ou precisão de um classificador pode não ser satisfatória, mas o conhecimento induzido, isto é, o conjunto de regras, pode conter regras que, individualmente, tenham boa precisão ou que possuam alguma outra propriedade interessante.

Lavraç et al. (1999) desenvolveram uma visão unificadora sobre algumas medidas de avaliação de regras, proporcionando uma terminologia e uma notação comuns, por meio da matriz de contingência. A matriz de contingência é uma generalização da matriz de confusão binária. Enquanto a matriz de confusão refere-se ao classificador como um todo, a matriz de contingência está associada a cada uma das regras que compõem o classificador.

Considerando cada regra no formato $A \rightarrow C$ (A - antecedente; C - conseqüente), sua correspondente matriz de contingência é mostrada na Tabela 2. Na tabela, A denota o conjunto de exemplos para os quais o antecedente da regra é verdadeiro e o seu complemento

\bar{A} denota o conjunto de exemplos para os quais o antecedente é falso. A interpretação de C e \bar{C} é análoga, em relação ao conseqüente da regra. Os demais elementos da tabela significam:

- ac é o número de exemplos para os quais A e C são verdadeiros.
- $a\bar{c}$ é o número de exemplos para os quais A é verdadeiro e C é falso.
- $\bar{a}c$ é o número de exemplos para os quais A é falso e C é verdadeiro.
- $\bar{a}\bar{c}$ é o número de exemplos para os quais A e C são falsos.
- a é o número de exemplos para os quais A é verdadeiro.
- \bar{a} é o número de exemplos para os quais A é falso.
- c é o número de exemplos para os quais C é verdadeiro.
- \bar{c} é o número de exemplos para os quais C é falso.
- n é o número total de exemplos.

Tabela 2: Matriz de contingência de uma regra $A \rightarrow C$ (A - antecedente; C - conseqüente).

	A	\bar{A}	
C	ac	$\bar{a}c$	c
\bar{C}	$a\bar{c}$	$\bar{a}\bar{c}$	\bar{c}
	a	\bar{a}	n

A **acurácia** ou **precisão** de uma regra $R = A \rightarrow C$ é definida como a probabilidade condicional de C ser verdadeiro dado que A é verdadeiro (equação 2.9).

$$acc(R) = \frac{ac}{a} \tag{2.9}$$

Essa forma de calcular a precisão não considera se o número de exemplos cobertos corretamente pela regra é alto ou baixo, dando maior valor a regras precisas que cubram poucos exemplos em comparação com regras não tão precisas, mas que cubram corretamente mais exemplos.

A **precisão de Laplace**, conforme a equação 2.10, procura corrigir esse problema, penalizando as regras que cubram poucos exemplos – na equação, N_{Cl} é o número de classes possíveis para o atributo meta.

$$Lacc(R) = \frac{ac + 1}{a + N_{Cl}} \quad 2.10$$

A **sensitividade** é a proporção de exemplos cobertos pela regra pertencentes à classe predita em C. É definida como a probabilidade condicional de A ser verdadeiro dado que C é verdadeiro (equação 2.11).

$$sens(R) = \frac{ac}{c} \quad 2.11$$

A **especificidade** é o correspondente à sensibilidade, mas para os exemplos que não são cobertos pela regra. É definida como a probabilidade condicional de A ser falso dado que C é falso (equação 2.12).

$$spec(R) = \frac{\overline{ac}}{c} \quad 2.12$$

A **cobertura** (*coverage*) é a proporção de exemplos cobertos pela regra (equação 2.13). Quanto maior a cobertura, maior a quantidade de exemplos cobertos pela regra. O **suporte** (*support*) é a proporção de exemplos cobertos corretamente pela regra (equação 2.14). Quanto maior o suporte, maior o número de exemplos pertencentes à classe em questão cobertos corretamente pela regra.

$$cov(R) = \frac{a}{n} \quad 2.13$$

$$sup(R) = \frac{ac}{n} \quad 2.14$$

A próxima medida procura avaliar a **novidade** (*novelty*), ou grau de interesse (*interestingness*), de uma regra. Uma regra é definida como nova se a probabilidade $P(AC)$ de A e C ocorrerem juntos não puder ser inferida das probabilidades de A e C isoladamente ou, em outras palavras, se A e C não são estatisticamente independentes. Isso pode ser obtido comparando-se o valor observado $P(AC)$ com o valor esperado sob a consideração de independência $\mu(AC) = P(A) \times P(C)$ (equação 2.15).

$$nov(R) = P(AC) - P(A) \times P(C) = \frac{ac}{n} - \frac{a \times c}{n^2} \quad 2.15$$

Quanto mais o valor observado diferir do valor esperado, maior é a chance de que exista uma associação verdadeira e inesperada entre A e C expressa pela regra. Pode ser demonstrado que $-0,25 \leq nov(R) \leq 0,25$ (LAVRAC et al., 1999); quanto maior um valor positivo (próximo de 0,25), mais forte é a associação entre A e C ; e quanto menor um valor negativo (próximo de -0,25), mais forte é a associação entre A e \bar{C} .

Lavrac et al. (1999) demonstraram que a novidade é igual à medida que chamaram de **acurácia relativa com peso** (*weighted relative accuracy*). Essa medida foi definida para promover um balanceamento entre a generalidade e a acurácia relativa de uma regra – a acurácia relativa é o ganho de acurácia da regra relativo à regra trivial “todos os exemplos pertencem à classe em questão”.

2.2 Epidemiologia de doenças de plantas

2.2.1 Introdução e conceitos

As plantas tornam-se doentes quando sofrem distúrbios que interferem, além do normal, em uma ou mais de suas funções, como a fotossíntese. As principais causas são organismos vivos (os patógenos) ou fatores do ambiente. De início, a reação das plantas ao agente causal é invisível, mas logo surgem mudanças que se manifestam visíveis e constituem os sintomas da doença. Doença de planta, então, pode ser definida como uma disfunção de suas células e tecidos resultante de irritação contínua por agente patogênico ou fator ambiental, que conduz ao desenvolvimento de sintomas (AGRIOS, 1988).

Doenças infecciosas são aquelas que resultam de infecção da planta por patógeno. São caracterizadas pela habilidade do patógeno de crescer e se multiplicar rapidamente nas plantas doentes e também de se propagar para plantas saudáveis. Além da planta (hospedeiro) e do patógeno, um conjunto de condições ambientais dentro de um intervalo favorável também deve ocorrer para que a doença se desenvolva.

As interações dos três componentes de uma doença podem ser representadas por um triângulo (Figura 5). Se as plantas são resistentes ao patógeno, por exemplo, o lado do triângulo do hospedeiro (e a quantidade de doença) é pequeno ou inexistente. Quanto mais virulento, abundante e ativo, maior é o lado do triângulo do patógeno e maior a quantidade potencial da doença. Também, quanto mais favoráveis as condições ambientais, que ajudam o patógeno ou que reduzem a resistência do hospedeiro, maior é o lado do triângulo referente ao ambiente.

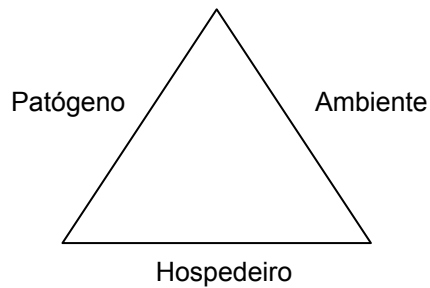


Figura 5: Triângulo de doença de planta (AGRIOS, 1988).

O desenvolvimento de doenças em plantas cultivadas também é influenciado pelo ser humano. As influências incluem, entre outras, o grau de resistência a certos patógenos do cultivar escolhido, a data e a densidade de plantio, as práticas culturais e os sistemas de produção.

Uma epidemia ocorre quando um patógeno se propaga e afeta muitos indivíduos de uma população de plantas sobre uma área relativamente grande e num espaço de tempo relativamente curto (AGRIOS, 1988). As epidemias ocorrem como resultado de combinações no tempo dos mesmos elementos que originam as doenças de plantas: hospedeiros suscetíveis cultivados em grande extensão, grande produção de patógeno virulento e condições ambientais favoráveis e duradouras. A epidemiologia é o estudo das epidemias e dos fatores que as influenciam.

Em toda doença infecciosa, uma série de eventos ocorre em seqüência e conduz ao desenvolvimento e perpetuação da doença e do patógeno. Esta corrente de eventos é chamada de ciclo da doença. Alguns patógenos completam apenas um ou parte de um ciclo da doença em uma estação de crescimento das plantas e são chamados monocíclicos. Na maioria das doenças, entretanto, o patógeno produz mais de uma geração por estação de crescimento e são chamados policíclicos. Eles completam vários ciclos da doença e são os responsáveis pela maior parte das epidemias de culturas agrícolas (AGRIOS, 1988).

Eventos ou fases importantes de um ciclo de doença são a inoculação, a germinação, a penetração, a infecção, a colonização, o aparecimento dos sintomas e lesões, a disseminação e a sobrevivência do patógeno. Segue breve descrição de alguns deles (AGRIOS, 1988):

- **Inoculação:** é a entrada em contato do patógeno com a planta. Inóculo é qualquer parte viva do patógeno capaz de iniciar uma infecção. Nos fungos pode ser conídio, esporo, esclerócio ou fragmento de micélio. Inóculo que sobrevive de uma estação para outra e causa as primeiras infecções é chamado de inóculo primário ou inicial. Inóculo produzido

a partir das infecções primárias é chamado de inóculo secundário e causa as infecções secundárias.

- **Germinação e penetração:** os patógenos penetram nas superfícies das plantas diretamente, por aberturas naturais ou por ferimentos. Os esporos de fungos precisam germinar primeiro e produzir o tubo germinativo capaz de penetrar na epiderme das folhas de plantas. Para isso requerem temperatura favorável e uma película de água na superfície da planta ou, pelo menos, alta umidade relativa do ar. As condições de umidade devem durar o tempo necessário para o patógeno penetrar, caso contrário ele seca e morre.
- **Infecção:** é o processo pelo qual os patógenos estabelecem contato com as células ou tecidos suscetíveis do hospedeiro em busca de nutrientes. Durante a infecção, os patógenos crescem e/ou se multiplicam dentro dos tecidos, invadindo e colonizando a planta. Os sintomas da doença resultam de infecções realizadas com sucesso.

A infecção também é referida na literatura como o processo que compreende os subprocessos de germinação, penetração e estabelecimento do parasitismo ou início da colonização (ZADOKS e SCHEIN, 1979).

O intervalo de tempo entre a inoculação e o surgimento dos sintomas é chamado de período de incubação. Período latente é o tempo decorrido desde a penetração do patógeno até a sua esporulação em pústulas ou lesões. A duração desses períodos, em várias doenças, depende da combinação patógeno-hospedeiro, do estágio de desenvolvimento do hospedeiro e da temperatura no ambiente da planta infectada.

- **Colonização:** é o mecanismo de invasão de células, tecidos e órgãos do hospedeiro pelo micélio ou haustórios do patógeno e a concomitante extração de nutrientes. A colonização tem início quando a primeira molécula de nutriente do hospedeiro é transferida e incorporada ao protoplasma do parasita.
- **Disseminação:** é a dispersão do patógeno a várias distâncias e em qualquer direção. A disseminação de patógenos responsáveis pelo aparecimento de doenças, mesmo daquelas de menor importância econômica, quase sempre é realizada passivamente por agentes como vento, chuva, insetos, animais, humanos e maquinário.

A presença em uma mesma área de plantas suscetíveis e de patógenos virulentos não garante muitas infecções, nem o desenvolvimento de uma epidemia, o que faz sobressair a

influência do ambiente. Os fatores ambientais mais importantes no desenvolvimento de epidemias são a umidade, a temperatura e o molhamento foliar.

Umidade abundante, prolongada e repetitiva, na forma de chuva, orvalho ou alta umidade relativa do ar, resultando em molhamento foliar, é o fator predominante no desenvolvimento da maioria das epidemias de doenças causadas por fungos e bactérias. E quando a temperatura permanece dentro de um intervalo favorável, para cada um dos estágios, um patógeno policíclico pode completar seu ciclo no espaço de tempo mais curto possível. Desde que a quantidade de inóculo é multiplicada com cada ciclo da doença, um maior número de ciclos resulta em mais plantas sendo infectadas por mais e mais patógenos, e assim uma epidemia se desenvolve.

A avaliação de doenças de plantas é uma das mais importantes e frequentemente mais difíceis tarefas da epidemiologia. A medição da quantidade de doença em um determinado momento é pedra fundamental da análise dos dados, dos esforços de modelagem e das interpretações dos patossistemas. A quantidade presente de doença pode ser referenciada como intensidade da doença e dentro dela distingue-se entre a incidência e a severidade (CAMPBELL e MADDEN, 1990).

Incidência da doença refere-se ao número de unidades da planta (planta inteira ou partes dela, como folhas, troncos ou frutos) visivelmente doentes, geralmente relativo ao número total de unidades avaliadas. Severidade da doença é a área ou volume do tecido da planta que está doente, geralmente relativo à área ou volume total. Agrios (1988) indica ainda a medida de dano de produção, como a proporção da produção que se deixa de colher porque a doença a destruiu ou impediu a planta de produzi-la.

Medir a incidência da doença é relativamente rápido e fácil. No entanto, ela pode ter pouca relação com a severidade da doença ou com os danos na produção. Embora a severidade e os danos sejam de maior importância para o produtor, suas medições são mais difíceis e, em alguns casos, possíveis apenas em fases adiantadas do desenvolvimento da epidemia (AGRIOS, 1988).

2.2.2 Epidemiologia da ferrugem do cafeeiro

O café é considerado um dos principais produtos agrícolas do mercado mundial. O Brasil é o maior produtor e o principal país exportador, tendo reconquistado a posição de fornecedor de quase 30% do mercado internacional. O consumo interno, recentemente, atingiu

um recorde, alcançando a marca de 16 milhões de sacos, cerca de 50% da produção (LIMA e SILVA, 2006).

Os cultivares comerciais de café recomendados para plantio são suscetíveis às principais doenças, o que causa grande preocupação aos cafeicultores, pois, dependendo da região, uma determinada doença pode ser o fator responsável pelas baixas produtividades da cultura. Como, na maioria dos casos, o patógeno está presente nos campos de cultivo, o ambiente e os fatores que predispõem ao ataque são os mais importantes a serem considerados no controle de doenças do cafeeiro (ZAMBOLIM et al., 1997).

A ferrugem do cafeeiro, cujo agente etiológico é *Hemileia vastatrix* Berk. & Br., foi constatada pela primeira vez no Brasil, e também na América do Sul, em janeiro de 1970, no estado da Bahia. Quatro meses depois foi encontrada em cafeeiros de quase todos os estados brasileiros. Atualmente, pode ser encontrada em todas as lavouras de café cultivadas e é considerada a principal doença da cultura. No país, os prejuízos na produção atingem cerca de 35%, em média, nas regiões onde as condições climáticas são favoráveis à doença (ZAMBOLIM et al., 1997; ZAMBOLIM et al., 2002).

Os danos provocados pela ferrugem surgem em consequência da queda precoce das folhas e da seca dos ramos, que reduzem a produção de frutos no ano seguinte. A desfolha, antes do florescimento, interfere no desenvolvimento dos botões florais e na frutificação e, durante o desenvolvimento dos frutos, leva à formação de grãos anormais, defeituosos, e frutos com lojas vazias, afetando sensivelmente a produção (ZAMBOLIM et al., 2002).

Os sintomas da ferrugem podem ser observados na face inferior das folhas, onde aparecem manchas de coloração amarelo-pálida, inicialmente pequenas, com 1 a 3 mm, que evoluem atingindo até 2 cm de diâmetro, quando apresentam aspecto pulverulento e coloração amarelo-alaranjada característica da doença. Na face superior das folhas observam-se manchas cloróticas amareladas, correspondendo aos limites da pústula na face inferior, que posteriormente necrosam (ZAMBOLIM et al., 1997).

O ciclo da doença pode ser representado como na Figura 6. Ele inicia-se pelos uredósporos do fungo, que, ao caírem na face inferior das folhas, na presença de água líquida, germinam, penetram e infectam, produzindo a urédia com os uredósporos. Esses uredósporos produzidos podem infectar novamente outras folhas da mesma planta ou de plantas diferentes. Em determinadas condições climáticas, a télia e os teliósporos formam-se nas lesões. Os

teliósporos, ao germinarem, formam o basídio (pró-micélio) e os basidiósporos, cuja função ainda é desconhecida (ZAMBOLIM et al., 1997).

O conhecimento dos fatores que determinam a maior taxa de progresso da ferrugem é de grande importância, uma vez que eles condicionam a distribuição da doença, a sua incidência e a severidade. O estudo das relações entre o patógeno, o hospedeiro e o ambiente pode auxiliar na compreensão da ocorrência de epidemias e, conseqüentemente, permitir a aplicação de medidas de controle mais adequadas (MONTROYA e CHAVES, 1974).

Variações regionais de incidência e os prejuízos causados pela ferrugem podem ser atribuídos principalmente aos fatores climáticos, através da coincidência de condições de temperatura e de umidade (intensidade, duração e freqüência das chuvas; orvalho; e período de molhamento) favoráveis à doença, no período de outubro a março (MORAES, 1983).

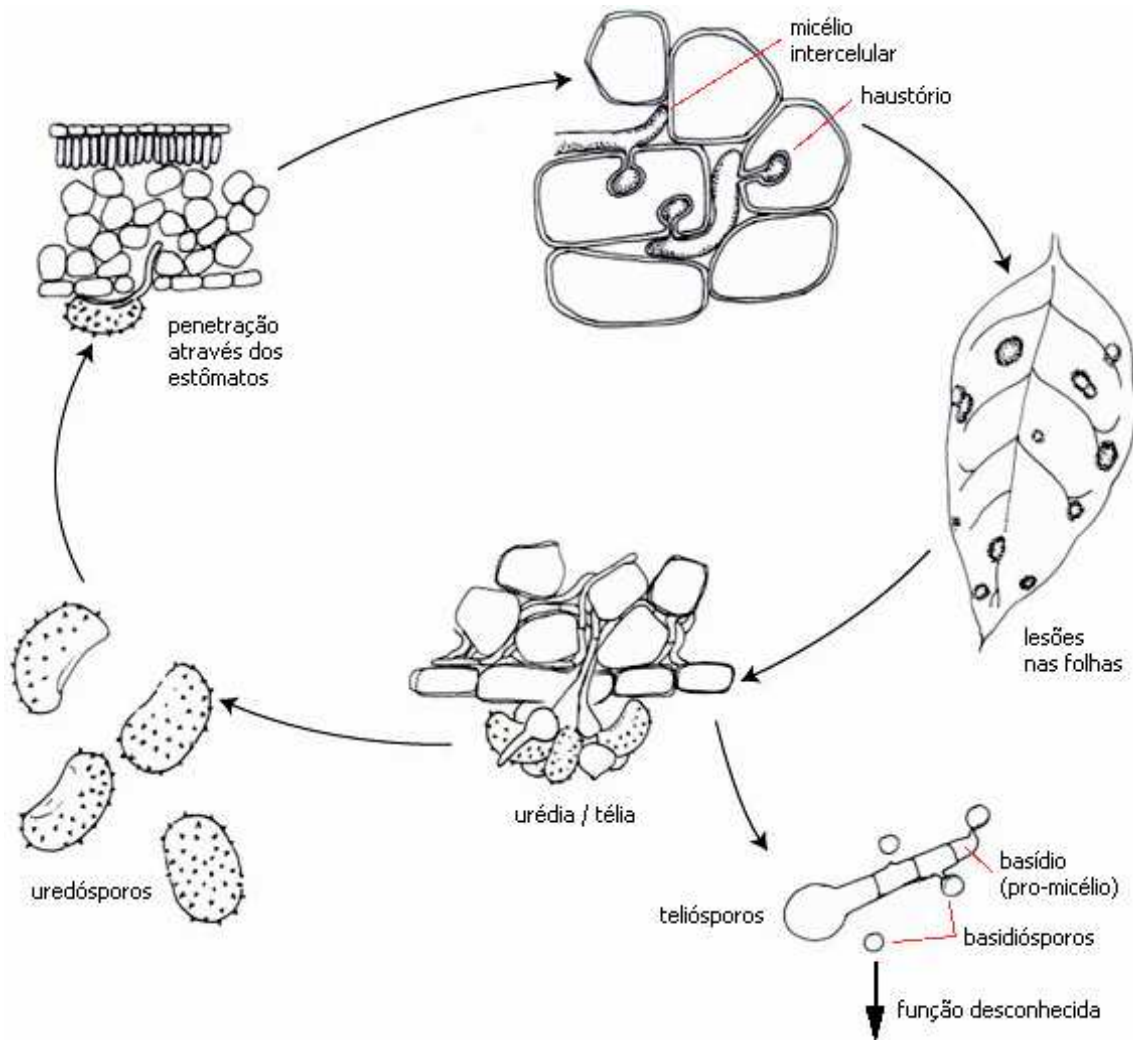


Figura 6: Ciclo da doença - ferrugem do cafeeiro (adaptado de APSNET, 2008).

A severidade da ferrugem do cafeeiro está relacionada também com a carga pendente de frutos das plantas. O cafeeiro é uma planta tipicamente bianual, ou seja, de dois em dois anos apresenta alta produção. Nesses anos, a ferrugem atinge alta severidade, normalmente iniciando em dezembro/janeiro, aumentando rapidamente de março a abril até atingir o pico em maio/junho. A partir daí decresce, devido às baixas temperaturas, à queda de folhas provocada pela colheita, à senescência natural e ao fato de que grande severidade da doença provoca intensa desfolha das plantas. No ano seguinte, de baixa produção, a doença não é severa, mesmo sob condições favoráveis do clima (ZAMBOLIM et al., 1997).

Outros fatores que influenciam o desenvolvimento da doença são a densidade de plantio, podendo afetar o microclima dentro da lavoura, e o nível de resistência do cultivar. As raças de *H. vastatrix* e a intensidade de inóculo são os principais fatores relacionados com o patógeno (VALE et al., 2000).

Por mais favoráveis que sejam as condições relacionadas com o hospedeiro e o patógeno, epidemias não ocorrem se as condições ambientais não forem favoráveis. Vários fatores do ambiente interferem na ocorrência da ferrugem do cafeeiro, atuando todos conjunta e simultaneamente (MORAES, 1983).

Os uredósporos de *H. vastatrix* germinam apenas na presença de água livre na superfície das folhas, principalmente no período noturno. Se a água secar antes da penetração, o processo é inibido (KUSHALAPPA, 1989a). Um período de seis horas de água livre na superfície da folha foi o tempo mínimo necessário para ocorrer infecção; a máxima infecção foi obtida após 24 horas de água livre na superfície da folha (KUSHALAPPA et al., 1983).

A temperatura, enquanto a superfície da folha está molhada, é um dos mais importantes fatores que determinam a quantidade de germinação dos esporos e de penetração (KUSHALAPPA, 1989a). A temperatura ótima para a germinação e a penetração do fungo nos estômatos de folhas de café varia de 22 a 24° C (ZAMBOLIM et al., 2002); temperaturas superiores a 30° C e inferiores a 14° C foram limitantes para a infecção (KUSHALAPPA et al., 1983). Estimou-se em 23,7° C a temperatura ótima de germinação sobre mudas de cafeeiro (MONTROYA e CHAVES, 1974).

O período de incubação (tempo em dias desde a germinação e penetração nos tecidos da planta até o aparecimento dos sintomas) pode variar de 18 a 45 dias; entretanto, está em

média entre 25 e 30 dias. Temperaturas altas, acima de 28° C, ou baixas, abaixo de 18° C, tendem a aumentar o período de incubação (ZAMBOLIM et al., 1997).

Moraes et al. (1976) observaram que o período de incubação – período decorrido até a formação de 50% de pústulas (o período latente foi referido como período de incubação) – tendeu a encurtar nos meses mais quentes (28 dias) e tornar-se mais longo nos meses mais frios (65 dias). Os autores sugeriram a utilização da seguinte equação para a estimativa do período de incubação:

$$y = 103,01 - 0,98 \times x_1 - 2,1 \times x_2 \quad 2.16$$

onde y é a estimativa do período de incubação em dias, x_1 é a temperatura média máxima e x_2 a temperatura média mínima durante o período.

O processo de esporulação é influenciado pelo ambiente e pelo nível de resistência do hospedeiro. Muito poucos trabalhos foram realizados com relação à influência do ambiente na esporulação de *H. vastatrix*. Sabe-se que a umidade relativa do ar melhora o processo de esporulação e a temperatura também influencia a esporulação, mas nenhum estudo detalhado a esse respeito foi realizado (KUSHALAPPA, 1989a).

O vento e a chuva são os principais agentes de disseminação dos esporos no campo, de uma região para outra. Dentro da planta, o respingo de chuva é o principal meio de dispersão (ZAMBOLIM et al., 2002).

Os estudos epidemiológicos da ferrugem do cafeeiro, como também para outras doenças de plantas, na grande maioria das vezes, utilizaram a regressão múltipla para ajustar os dados (KUSHALAPPA et al., 1983; KUSHALAPPA, 1989a; MONTOYA e CHAVES, 1974; MORAES et al., 1976; ZAMBOLIM et al., 2002). Trabalhos mais recentes empregaram técnicas alternativas, procurando, principalmente, contornar a limitação da análise de regressão com relação à multicolinearidade entre as variáveis independentes (BUTT e ROYLE, 1990), quando estas estão correlacionadas entre si.

Silva-Acuña et al. (1998), por meio da análise de trilha, estudaram a relação entre algumas variáveis climáticas e a taxa de infecção da ferrugem do cafeeiro durante três anos. Segundo a análise de trilha, as variáveis de efeito-causa, nos dois anos de alta carga pendente de frutos, foram a temperatura entre 21 e 26° C e o molhamento foliar noturno. No ano de baixa produção, apesar de não ter sido possível detectar relação de efeito-causa das variáveis

explicativas, foi constatado efeito direto do molhamento foliar noturno sobre a taxa de infecção da ferrugem.

Pinto et al. (2002) avaliaram o potencial das redes neurais para descrever epidemias da ferrugem do cafeeiro. Eles empregaram as redes neurais para estabelecer relações entre variáveis climáticas e produção com a incidência da ferrugem do cafeeiro.

As variáveis precipitação pluvial, número de dias com e sem precipitação, umidade relativa do ar, horas de insolação, temperaturas média, máxima e mínima, calculadas como médias ou somatórios para os 15, 30, 45 e 60 dias anteriores às avaliações da incidência da ferrugem, e a variável produção, a qual assumiu valor '0' para as plantas antes do início da produção e '1' para as plantas em fase de produção, foram utilizadas para construir as redes. Séries temporais da incidência da doença, isoladamente, também foram utilizadas na elaboração de redes neurais.

A rede que melhor descreveu a epidemia da ferrugem do cafeeiro incluiu as variáveis temperatura mínima, umidade relativa do ar, produção e insolação, referentes a 30 dias antes da data de avaliação da incidência da ferrugem. As redes elaboradas a partir das séries temporais também foram adequadas para descrever a epidemia da ferrugem, sendo que a melhor delas incluiu as observações da incidência da doença das quatro quinzenas anteriores à data de avaliação. A escolha dos melhores modelos baseou-se nos menores valores do quadrado médio do desvio e do erro médio de previsão avaliados para as redes.

2.3 Alerta de doenças de plantas

2.3.1 Sistemas de alerta de doenças de plantas

Um sistema de previsão de doença de planta é aquele que prevê o aparecimento ou um aumento na intensidade de uma doença baseado em informação sobre o ambiente, a cultura e/ou o patógeno (CAMPBELL e MADDEN, 1990). Esse aparecimento ou aumento futuro da doença é frequentemente baseado na observação de períodos críticos ocorridos, o que acaba causando certa confusão. Com respeito aos sintomas, a previsão é anterior ao fato, mas com respeito à infecção, a previsão é posterior (ZADOKS, 1984).

Zadoks (1984) sugeriu o termo aviso de doença ou alerta de doença, para evitar problemas de terminologia e enfatizar que a mensagem aos produtores é mais importante do que a sua origem técnica. Madden e Ellis (1988), entretanto, usaram ambos os termos, sistema

de previsão e sistema de alerta, sabendo-se que todos os sistemas estão “prevendo” aumento geral da doença baseados em componentes concluídos do seu ciclo.

Prever doenças de plantas é importante por duas razões principais: economia e segurança. A questão econômica é reduzir o custo de produção por meio de aplicações oportunas de medidas de controle, geralmente na forma de fungicidas. Segurança envolve não apenas a cultura, reduzindo efeitos tóxicos sobre as plantas, mas também o ambiente externo, reduzindo a exposição de agrotóxicos a outras espécies de plantas, aos trabalhadores e aos consumidores (HARDWICK, 1998).

Os alertas auxiliam os produtores a determinar a necessidade e o momento de aplicar técnicas de controle de doenças (CAMPBELL e MADDEN, 1990). Um alerta proporciona indicação de quando é provável que a doença vá se tornar crítica e, portanto, ter impacto econômico. Para algumas doenças, é importante ser capaz de prever a primeira ocorrência, enquanto para outras um certo nível de doença pode ser tolerado, particularmente em partes da planta que possuem pouca contribuição para a produção ou a qualidade (HARDWICK, 1998).

Sistemas de previsão de doenças de plantas podem ser classificados de várias maneiras, de acordo com o tipo de informação usada para fazer a previsão ou com a abordagem conceitual para a previsão. Especificamente, previsores podem ser classificados com respeito a (CAMPBELL e MADDEN, 1990): se informação da cultura, da doença, do patógeno ou do ambiente, ou uma combinação dessas, são usadas para fazer as previsões; se as previsões são pré-plantio ou pós-plantio; se informação empírica ou fundamental foi usada no desenvolvimento do sistema; e se características específicas das epidemias, tais como inóculo primário, inóculo secundário ou taxa de aumento da doença, servem de base para as previsões. Uma ampla revisão bibliográfica de sistemas de previsão de doenças de plantas, abrangendo importantes doenças de várias culturas agrícolas, encontra-se em Reis e Bresolin (2004).

Um sistema de alerta, para que tenha sucesso, precisa ser adotado e implementado pelos produtores, devendo haver a percepção de que é possível obter benefícios específicos e tangíveis com o seu uso. Atributos que asseguram o sucesso incluem (CAMPBELL e MADDEN, 1990): confiabilidade, simplicidade para implementar, importância da doença, utilidade do alerta, disponibilidade aos produtores, aplicabilidade a várias doenças/pragas e eficiência de custo.

Segundo Coakley (1988), para desenvolver um sistema de alerta é preciso que a doença satisfaça quatro requisitos: a doença causa perdas economicamente significativas na qualidade ou na quantidade da produção; a doença varia entre cada estação de cultivo; medidas de controle da doença estão disponíveis e são economicamente viáveis; e há informação suficiente a respeito da natureza da dependência da doença em relação às condições meteorológicas.

Muitos estudos a respeito do papel das condições meteorológicas nas doenças de plantas indicam que a doença é mais afetada pelas condições microclimáticas no dossel das plantas do que pelas condições macroclimáticas medidas a uma certa distância da cultura em estações meteorológicas padrão. Entretanto, condições macroclimáticas produzem o microclima – é possível usar regras para determinar relacionamentos entre o macro e o microclima – e existe um limite na extensão com que o microclima pode facilitar o desenvolvimento da doença sob condições macroclimáticas desfavoráveis (COAKLEY, 1988). Coakley (1988) chegou ainda a sugerir que tentativas de relacionar dados macroclimáticos com doenças podem ter alcançado resultados limitados, em parte, por causa da dificuldade de analisar grandes quantidades de dados sem o auxílio de computador.

Um tempo atrás, desvantagens de usar dados microclimáticos de dentro dos campos de cultivo no desenvolvimento e/ou implementação de sistemas de alerta incluíam períodos de tempo relativamente curtos e poucos locais para os quais estavam disponíveis os dados, alto custo de coleta desses dados e frequência considerável de perda de dados devido a falhas nos equipamentos (COAKLEY, 1988). Nos últimos anos, os instrumentos manuais deram lugar às estações meteorológicas automáticas, que podem ser instaladas próximas aos campos de cultivo. A cada geração, essas estações tornam-se mais sofisticadas, confiáveis e de menor preço. Mesmo assim, exigir de produtores a instalação e a manutenção dessas estações meteorológicas e o gerenciamento dos dados de cada campo não é prático nem economicamente viável.

Como alternativa a essa questão, esforços têm sido feitos para desenvolver e validar sistemas de alerta que utilizem dados de redes regionais de estações meteorológicas (GENT e SCHWARTZ, 2003; MADDEN et al., 2004) e tecnologias para a obtenção de dados de locais específicos sem sensores *in loco* (MAGAREY et al., 2001). Outra tendência é utilizar dados meteorológicos estimados a partir de modelos de previsão do tempo para antecipar ainda mais

os alertas, permitindo um tempo maior para a tomada de decisão e para a aplicação de medidas de controle (SHTIENBERG e ELAD, 1997; MADDEN et al., 2004).

2.3.2 Modelos de previsão de doenças de plantas

Modelos representam a percepção (ou imaginação) da realidade de forma simbólica e simplificada. Modelagem e simplificação são essenciais ao processo científico. A solução ideal é o modelo contemplar os aspectos essenciais do sistema real pertinentes ao problema em questão. A complexidade ou simplicidade do modelo deve estar em acordo com o seu propósito – enquanto a simplicidade facilita o entendimento do modelo, a complexidade pode permitir maior acurácia na descrição do sistema (CAMPBELL et al., 1988).

Modelos podem ser classificados em dois grupos, dependendo da abordagem de desenvolvimento. O primeiro tipo de modelo é chamado fundamental ou mecânico. O desenvolvimento desses modelos parte de um conceito, hipótese ou teoria, em vez de um conjunto de dados. Um modelo consistente com o conceito é elaborado e depois são realizados experimentos para testar a sua acurácia (CAMPBELL e MADDEN, 1990).

Esses modelos são derivados de tentativas de compreensão da realidade, sendo que essa compreensão pode ser obtida a partir de experimentos prévios – em laboratório, câmara de ambiente controlado, casa de vegetação ou campo – ou de princípios biológicos (CAMPBELL et al., 1988).

Modelos de previsão fundamentais são geralmente simples e baseados em um ou poucos componentes do ciclo da doença, nos quais a infecção é o componente que prevalece. Contudo, é possível considerar o ciclo completo da doença por meio de uma abordagem de análise sistêmica e modelos de simulação (MADDEN e ELLIS, 1988).

O segundo tipo de modelo é chamado empírico ou correlativo. Os modelos empíricos são desenvolvidos a partir da coleta e análise de dados atuais e históricos sobre níveis da doença e outros fatores bióticos e abióticos (MADDEN e ELLIS, 1988). Eles descrevem um relacionamento observado entre duas ou mais variáveis do conjunto de dados, normalmente derivados a partir do ajuste dos dados a um modelo aceitável (CAMPBELL e MADDEN, 1990). Conhecimento teórico relacionado a mecanismos básicos não é exigido (CAMPBELL et al., 1988).

Modelos empíricos estão relacionados com uma previsão apenas por estação de cultivo ou podem envolver múltiplas previsões. Os primeiros são úteis quando é importante

predizer o inóculo inicial ou o nível inicial da doença, e o seu desenvolvimento geralmente requer observações de vários anos e/ou locais. Aqueles que envolvem múltiplas previsões são úteis quando a doença pode aumentar rapidamente durante a estação de crescimento, ou quando o valor econômico da cultura justifica várias intervenções de controle. O seu desenvolvimento requer observações feitas sobre a doença, o ambiente e/ou outros fatores por toda a estação de cultivo. Em geral, mais de dois anos ou locais precisam ser estudados e anos adicionais são necessários para a validação (MADDEN e ELLIS, 1988).

Coakley (1988), baseado em sua experiência, sugeriu um mínimo de oito a doze anos de registro de dados, de campos com fontes naturais de inóculo, para identificar com segurança quais podem ser os fatores climáticos de influência no desenvolvimento de uma doença. Sugeriu também, quando se tem menos de oito anos de registro, que dados de diferentes localidades de uma região geográfica podem ser utilizados.

Madden e Ellis (1988) indicaram duas formas pelas quais modelos de previsão empíricos podem ser desenvolvidos. A primeira, chamada de qualitativa, envolve o desenvolvimento de critérios de previsão sem qualquer análise estatística formal. Alguns exemplos de modelos desenvolvidos dessa forma são: previsão da mancha preta do amendoim (JENSEN e BOYLE, 1966; PARVIN JR. et al., 1974; PEDRO JÚNIOR et al., 1994; MORAES, 1999), previsão da requeima da batateira (WALLIN, 1962; MICHEL et al., 1997; COSTA et al., 2002), previsão da pinta-preta do tomateiro (MADDEN et al., 1978) e previsão da queima das folhas da cenoura (SOUZA et al., 2002).

A segunda forma de desenvolvimento dos modelos de previsão, chamada de quantitativa, é baseada em análise estatística e modelagem dos dados observados. Os métodos e técnicas utilizados são variados. Em uma retrospectiva dos modelos apresentados na literatura, é possível perceber que o emprego desses métodos e técnicas acompanha a evolução nas disciplinas relacionadas com a análise de dados.

O predomínio é de métodos e técnicas estatísticos, sendo a análise de regressão a mais popular (MADDEN e ELLIS, 1988). Como exemplo, cita-se o uso de regressão linear múltipla no desenvolvimento de modelos para predizer a severidade de epidemias da ferrugem asiática da soja (DEL PONTE et al., 2006). Alguns métodos matemáticos e estatísticos para a modelagem de dados epidemiológicos foram revisados por Hau e Kranz (1990).

Mais recentemente, os trabalhos publicados relatam o uso de métodos e técnicas modernos e sofisticados, não necessariamente inovadores, mas que ganharam/recuperaram visibilidade e importância nos últimos anos, como redes neurais (PAUL e MUNKVOLD, 2005; BATCHELOR et al., 1997), regressão logística (DE WOLF et al., 2003), árvores de decisão (seção 2.3.4) e análise estatística de séries temporais (XU et al., 2000).

2.3.3 Sistemas e modelos de alerta da ferrugem do cafeeiro

Modelos empíricos – abordagem quantitativa

Equações de regressão desenvolvidas para a predição do período de incubação ou do período latente de *H. vastatrix*, baseadas nas temperaturas máximas e mínimas (média durante o período), como a equação 2.16 apresentada na seção 2.2.2, foram usadas para dar uma idéia de quão severa a ferrugem poderia ser durante certas estações ou meses do ano. Moraes (1983) sugeriu, com base em estimativas pela equação 2.16, para o período de outubro a março, os seguintes níveis de severidade de ataque da ferrugem do cafeeiro: risco alto de ataque severo, quando o período de incubação (PI) estimado for inferior a 35 dias; risco médio de ataque severo quando o PI for estimado entre 35 e 45 dias; e pequena probabilidade de risco de ataque severo, quando o PI estimado for superior a 45 dias.

Chaves et al. (1970 apud KUSHALAPPA, 1989b), logo que a ferrugem do cafeeiro surgiu no Brasil, consideraram desnecessárias aplicações de fungicida, de maio a agosto, devido aos períodos latentes mais longos. Kushalappa (1989b), entretanto, argumentou que a razão de não haver necessidade de aplicações de fungicida nos meses mais frios seria diferente: a temperatura durante o período de molhamento foliar geralmente fica abaixo de 15°C, o que é limitante para a infecção.

Alfonsi et al. (1974) estudaram a associação entre níveis de infecção (média de pústulas por folha), variáveis climáticas e área foliar das plantas. A média das temperaturas máximas, a média das temperaturas mínimas e o total de chuvas, registrados em períodos de 15, 30 e 45 dias, foram correlacionados com os níveis de infecção observados ao final dos respectivos períodos. Os coeficientes de determinação (R^2) obtidos entre o nível de infecção e as três variáveis climáticas, independente da área foliar, mostraram que a associação com o período de 45 dias expressou melhor a proporção de acréscimos de pústulas (cerca de 95% de explicação da variação na severidade da doença). Na literatura consultada, não há registro de sistema de alerta da ferrugem do cafeeiro baseado nas equações de regressão obtidas.

Também por análise de regressão, vários fatores meteorológicos e biológicos foram considerados para explicar a taxa de progresso da ferrugem (KUSHALAPPA e ESKES, 1989). Os fatores mais significativos foram identificados pelo critério de seleção *stepwise*. Como variáveis dependentes, foram consideradas a severidade da doença na data de previsão (DP) e a taxa de infecção da ferrugem para os intervalos de um a dois períodos latentes (28 dias) após DP. A equação que explicou a máxima variação (94%) na taxa de infecção foi:

$$k'' = 0,031 + 4,881 \times PAFE + 0,022 \times PNF - 0,001 \times MIN - 0,001 \times MAX - 0,001 \times CHUVA \quad 2.17$$

onde k'' é a taxa de infecção, corrigida para o crescimento do hospedeiro, para 56 dias depois de DP; $PAFE$ é a proporção de área foliar com esporos na DP; PNF é a proporção de novas folhas formadas durante 14 dias antes de DP; MIN é a média das mínimas e MAX é a média das máximas temperaturas (em °C) para 14 dias antes de DP; $CHUVA$ é o total de chuvas (em mm) entre 14 a 28 dias antes de DP. Não foi encontrado, na revisão bibliográfica, trabalho a respeito do uso das equações desenvolvidas em sistema de alerta da ferrugem do cafeeiro.

Correlações significativas foram observadas entre variáveis independentes usadas na formulação de equações de regressão para prever a taxa de progresso da ferrugem do cafeeiro (KUSHALAPPA et al., 1983). Como resultado, alguns parâmetros, que independentemente explicaram variação significativa na doença, foram eliminados devido à multicolinearidade. O sucesso preditivo de tais modelos depende da ocorrência futura dos diferentes parâmetros, incluindo aqueles não usados no modelo, em combinações semelhantes àquelas observadas. Isso torna esse tipo de modelo menos estável sob condições de campo, a menos que esteja baseado em vários anos de registro de dados (KUSHALAPPA e ESKES, 1989).

Na seção 2.2.2, foi descrito o uso de redes neurais para descrever epidemias da ferrugem do cafeeiro. A camada de entrada para as redes foi formada pelas variáveis climáticas mais a variável de produção (variáveis independentes) e a variável de saída foi a incidência da ferrugem (variável dependente). Os menores valores do erro médio de previsão (EMP = 1,17%) e do quadrado médio do desvio (QMD = 3,43) foram obtidos para a rede neural elaborada com as variáveis produção, umidade relativa, horas de insolação e temperatura mínima, relativas ao período de 30 dias anteriores à avaliação da incidência da doença. A melhor rede neural (EMP = 4,72% e QMD = 3,95) elaborada a partir das séries

temporais teve como variáveis de entrada as observações da incidência da doença de quatro quinzenas anteriores à data de avaliação (PINTO et al., 2002).

A partir destes resultados, Pinto et al. (2002) sugeriram que o emprego de séries temporais, baseado na incidência ou severidade da doença, poderia facilitar a previsão de epidemias da ferrugem do cafeeiro. Apesar do melhor desempenho do modelo que incluiu as variáveis climáticas, os autores consideraram que avaliar a intensidade da doença é mais fácil para o produtor ou o agente de extensão, quando comparada à coleta de variáveis climáticas. Não foi encontrado, na revisão da literatura, relato de uso e/ou validação de redes neurais como modelo de previsão de sistema de alerta da ferrugem do cafeeiro.

Modelo empírico – abordagem qualitativa

Modificações no clima, nos últimos anos, têm ocasionado alterações na severidade da ferrugem, bem como no início e no pico da doença em algumas regiões do Brasil (ZAMBOLIM et al., 2002). Diante dessas alterações, surgiu a dúvida sobre a aplicação de fungicidas sistêmicos por meio de duas pulverizações foliares ou aplicações via solo, para que se obtivesse controle racional e econômico da ferrugem.

Para se identificar, então, os períodos favoráveis à ferrugem, nos quais as plantas deveriam ser atomizadas, foi desenvolvido um sistema de previsão ou de aviso (GARÇON et al., 2004). O objetivo foi desenvolver um sistema simples – simplicidade é um atributo importante para a aceitação do sistema, pois maiores são as chances de adoção pelos agricultores – e confiável de prever o desenvolvimento da doença no campo, determinando o momento propício para iniciar o controle químico por meio de pulverizações com fungicida sistêmico, bem como o intervalo entre as aplicações.

As variáveis meteorológicas empregadas no sistema de previsão foram o molhamento foliar diário e a temperatura média durante esse período de molhamento, obtidas em estação meteorológica colocada no meio da área experimental. Com os dados diários dessas variáveis meteorológicas calculou-se o valor de severidade da ferrugem (VSF), a partir de uma matriz de valores de severidade semelhante à idealizada por Wallin (1962) para a requeima da batateira, modificada para a ferrugem do cafeeiro (Tabela 3).

O limiar de ação para indicação do momento da pulverização foi baseado no acúmulo dos valores diários de VSF. Os limites de VSF estipulados para o teste e a validação do sistema foram 29, 34, 39 e 44, para anos de alta carga pendente de frutos (alta intensidade da

ferrugem), e 49, 59, 69 e 79, para anos de baixa ou média carga pendente (baixa intensidade da ferrugem).

Tabela 3: Matriz para cálculo dos valores de severidade da ferrugem (VSF) do cafeeiro, com base no período de molhamento foliar e na temperatura média do período (GARÇON et al., 2004).

Molhamento foliar (h/diárias)	Temperatura (°C)						
	< 16	16-18	19-20	21-24	25-26	27-29	30
0	0*	0	0	0	0	0	0
0 < h ≤ 8	0	0	1	2	1	0	0
8 < h ≤ 17	0	1	2	3	2	1	0
17 < h ≤ 24	0	2	3	4	3	2	0
h = 24**	0	0	1	2	1	0	0

* Valor de severidade da ferrugem (VSF) diário.

** Molhamento foliar diário de 24 h, porém este dentro de um período de molhamento de mais de 48 h sem interrupção.

OBS.: se dentro de 30 dias não houver acumulado mais de 5 VSF, desconsideram os VSF's acumulados até o momento.

Em uma lavoura de alta carga pendente (101,5 sacas beneficiadas/ha), foram recomendadas duas pulverizações com fungicida sistêmico quando o valor acumulado de VSF atingiu 29-31, igualando-se as duas aplicações do tratamento com calendário fixo. Em uma lavoura de média carga pendente (22,4 sacas beneficiadas/ha), recomendou-se uma única pulverização quando o valor acumulado de VSF atingiu 49-51, enquanto todos os outros tratamentos demandaram duas aplicações. Portanto, o sistema baseado no VSF foi tão eficiente quanto o calendário no controle da ferrugem do cafeeiro, porém com economia de uma pulverização na lavoura com carga média de frutos (GARÇON et al., 2004).

Os resultados mostraram que, normalmente, em lavouras com baixa a média carga pendente de frutos, uma única aplicação de fungicida sistêmico, no momento oportuno, indicado com base no número de horas de molhamento foliar e na temperatura média durante o período de molhamento, como sugerido pelo modelo de aviso proposto, foi suficiente para se alcançar um controle eficiente e racional da ferrugem do cafeeiro.

Modelo fundamental

Considerando que os vários fatores que influenciam o progresso da ferrugem do cafeeiro no campo não poderiam ser identificados por um experimento ou por outros

procedimentos estatísticos clássicos, como a análise de regressão, Kushalappa e Eskes (1989) justificaram o desenvolvimento de um modelo mais compreensivo, capaz de explicar o curso de ação biológica do patógeno e integrar os vários fatores que influenciam o sistema.

O sistema epidêmico da ferrugem do cafeeiro é composto de processos epidemiológicos policíclicos, que consistem de uma série de processos monocíclicos. O início de uma epidemia começa com um inóculo inicial e cada ciclo da doença (processo monocíclico) é formado pelos macroprocessos de esporulação, disseminação e infecção. Todos constituem os componentes estruturais do sistema epidêmico (KUSHALAPPA, 1994).

Baseado nesses aspectos, foi desenvolvido um modelo de previsão da taxa de progresso da ferrugem, considerando o inóculo inicial e fatores significativos do ambiente e do hospedeiro que influenciam o processo monocíclico de *H. vastatrix*. Denominado de “razão de sobrevivência líquida para o processo monocíclico” (*RSLPM*), o modelo foi formado pela integração de modelos fundamentais e empíricos desenvolvidos para cada um dos componentes estruturais da doença (KUSHALAPPA et al., 1983; KUSHALAPPA et al., 1984).

Na obtenção do modelo *RSLPM*, vários fatores que influenciam o progresso da ferrugem do cafeeiro, relacionados com o hospedeiro, o patógeno e o ambiente – ou, especificamente, os micro e mesoprocessos componentes dos macroprocessos – foram transformados em “equivalentes de processo” para o ambiente e o hospedeiro. Depois, os produtos multiplicativos dos equivalentes de micro e mesoprocessos foram derivados, designados de “equivalentes de processo monocíclico” para o ambiente e para o hospedeiro. Estes e o nível de inóculo foram então transformados em outro parâmetro multiplicativo, a razão de sobrevivência líquida para o processo monocíclico de *H. vastatrix* (KUSHALAPPA, 1989a). A incorporação dos três componentes do triângulo de doenças de plantas no modelo foi baseada na atividade biológica do fungo, o que fez os criadores do modelo o considerarem do tipo fundamental.

A influência do inóculo, denominada de razão de sobrevivência básica (*RSB*), foi quantificada com base na incidência (proporção de folhas com ferrugem) ou na severidade (proporção de área foliar com ferrugem).

A influência do ambiente, ou equivalente de processo monocíclico para o ambiente (*EPMA*), foi calculada pela multiplicação de sua influência nos processos de disseminação e

de infecção. O equivalente de disseminação foi determinado em função da velocidade diária do vento, da quantidade de chuva diária e da densidade de plantas, enquanto o equivalente de infecção foi determinado em função da duração do molhamento foliar (em horas) e da temperatura durante esse período.

O equivalente de processo monocíclico para o hospedeiro (*EPMH*) foi determinado pelo equivalente de processo devido à predisposição do hospedeiro ao ataque da ferrugem por causa de alta produção.

Por fim, a razão de sobrevivência líquida para processo monocíclico foi definida como o produto final da multiplicação das influências do inóculo, do ambiente e do hospedeiro, conforme a equação 2.18 (KUSHALAPPA et al., 1983; KUSHALAPPA, 1989a).

$$RSLPM = RSB \times EPMA \times EPMH \quad 2.18$$

Valores de *RSLPM* a partir de dados observados no campo, para um intervalo de 28 dias antes da data de predição (DP), foram relacionados, por análise de regressão, com taxas de infecção da ferrugem observadas 28 dias após DP, corrigidas para o crescimento do hospedeiro (KUSHALAPPA et al., 1984). Esse intervalo de predição foi escolhido por causa da média observada do período latente do fungo, de outubro a março, que foi de aproximadamente 28 dias.

Várias equações foram desenvolvidas para predizer a taxa de infecção, considerando diferentes parâmetros de área de produção de inóculo no cálculo dos valores de *RSLPM* (KUSHALAPPA et al., 1984). As equações que obtiveram os melhores coeficientes de determinação (R^2) foram:

$$k'' = 0,00044 + 14,766 \times RSLPM - 2511,21 \times RSLPM^2 \quad 2.19$$

$$k'' = 0,023 + 14,026 \times RSLPM - 87,382 \times RSLPM^2 \quad 2.20$$

onde a equação 2.19 é para predizer a severidade da doença, considerando a proporção de área foliar com ferrugem como o parâmetro *RSB*; a equação 2.20 é para predizer a incidência da doença, considerando a proporção de folhas com ferrugem como o parâmetro *RSB*; k'' é a taxa de infecção para 28 dias após DP, corrigida para o crescimento do hospedeiro; *RSLPM* é dada pela média diária da razão de sobrevivência líquida para o processo monocíclico durante 28 dias antes de DP. As equações 2.19 e 2.20 explicaram 76% e 64%, respectivamente, da variação em k'' .

Considerando que uma incidência de ferrugem de cerca de 10% justificaria uma aplicação de fungicida, um limite do valor de *RSLPM* para recomendar aplicações de fungicida foi derivado pela substituição de $k'' = 0,1$ na equação 2.20 (KUSHALAPPA et al., 1984; KUSHALAPPA, 1989b). Esse limite foi $RSLPM = 0,0057$, considerando a proporção de folhas com ferrugem como inóculo (*RSB*). Substituições semelhantes em outras equações permitiram se chegar no limite $RSLPM = 0,00015$, considerando a proporção de área foliar com ferrugem como inóculo.

A partir desses limites, um sistema de alerta simples e outro mais complexo foram desenvolvidos para recomendar aplicações de fungicida para o controle da ferrugem do cafeeiro. O sistema simples se resumiu em uma tabela, que foi formada pelo agrupamento dos valores de inóculo (*RSB*), de produção (*EPMH*) e de condições do ambiente (*EPMA*), observados no estado de Minas Gerais, em certos intervalos convenientes.

No caso do sistema simples, em intervalos quinzenais, deve-se quantificar a incidência (percentual de folhas atacadas) ou a severidade (área foliar atacada) da ferrugem e indicar a produção como alta ou baixa; depois, basta consultar a tabela sobre a recomendação ou não de aplicação de fungicida.

No caso do sistema complexo, deve-se quantificar *RSLPM*, em intervalos de 14 dias, e se recomenda a aplicação de fungicida quando o limite pré-estabelecido é igualado ou superado – $RSLPM \geq 0,0057$ ou $RSLPM \geq 0,00015$, caso se tenha quantificado a incidência ou a severidade da doença, respectivamente. Ambos os sistemas foram validados em condições de campo e foram considerados eficientes na determinação das épocas oportunas de aplicação de fungicidas (KUSHALAPPA et al., 1986).

2.3.4 Árvores de decisão como modelos de previsão de doenças de plantas

Algumas experiências foram relatadas na literatura com respeito à aplicação de árvores de decisão como modelo de previsão de doenças de plantas. Paul e Munkvold (2004) usaram este tipo de modelagem para avaliação de risco da cercosporiose do milho (*Cercospora zae-maydis* Tehon & E.Y. Daniels).

Árvores de decisão e modelos de regressão logística foram usados para prever a severidade da doença em estágio avançado do cultivo, a partir de dados obtidos no pré-plantio e de características do genótipo do milho (*Zea mays* L.).

Classes (categorias) de severidade da cercosporiose foram definidas e serviram como os valores da variável resposta (atributo meta). A abordagem de modelagem CART, de “Classification and Regression Trees” (BREIMAN et al., 1984), e diferentes abordagens de regressão logística foram empregadas para prever as classes de severidade em função da data do plantio, da quantidade de resíduo de milho no solo, da seqüência de plantio, da maturidade do genótipo, do seu nível de resistência à cercosporiose do milho e da longitude do local.

A severidade da cercosporiose do milho foi considerada em um estágio específico do crescimento da planta, no qual a doença proporciona o melhor relacionamento com a perda de produtividade. Os valores de severidade foram categorizados em uma variável resposta com cinco classes (‘1’: <20%; ‘2’: ≥ 20 e <40%; ‘3’: ≥ 40 e <60%; ‘4’: ≥ 60 e <80%; e ‘5’: $\geq 80\%$) e em outra variável resposta binária (‘0’: <20% e ‘1’: $\geq 20\%$). Esta estrutura de classes foi decidida após uma análise gráfica preliminar dos dados brutos e com base em limites críticos de severidade da doença em relação à perda de produtividade do milho.

Modelos foram desenvolvidos tanto para a variável resposta com cinco classes quanto para a variável resposta binária. Um total de 332 casos foi usado no desenvolvimento dos modelos e 30 casos independentes foram usados para avaliar a acurácia de predição. Para os modelos com resposta binária, calculou-se também a sensibilidade (capacidade de classificar corretamente casos com severidade $\geq 20\%$) e a especificidade (capacidade de classificar corretamente casos com severidade < 20%).

Os modelos de regressão logística classificaram corretamente de 60% (modelo para a variável resposta com cinco classes) a 70% (modelo para a variável resposta binária) dos casos de validação, enquanto as árvores de decisão classificaram corretamente de 57% (modelo para a variável resposta com cinco classes) a 77% (modelo para a variável resposta binária) desses mesmos casos.

A árvore de decisão para prever a variável resposta binária apresentou a maior acurácia (23 casos de 30 classificados corretamente). Ela também classificou corretamente 100% dos casos de validação com severidade $\geq 20\%$ (sensibilidade) e 61% dos casos com severidade < 20% (especificidade). Este nível máximo de sensibilidade foi considerado de suma importância, uma vez que perdas de produtividade do milho foram relatadas na literatura, associadas a valores de severidade da cercosporiose maiores ou iguais a 20%.

Os modelos em árvore de decisão, por meio da representação diagramática, proporcionaram uma compreensão clara do relacionamento entre as variáveis preditivas e a variável dependente. Tanto as árvores de decisão quanto os modelos de regressão logística desempenharam bem, apesar de terem usado basicamente dados de pré-plantio para fazer as predições da cercosporiose do milho. Ambas as abordagens mostraram potencial como ferramentas de tomada de decisão no gerenciamento da doença.

Árvores de decisão foram também usadas por Molineros et al. (2005) para modelar epidemias de giberela do trigo [*Gibberella zeae* (Schwein.) Petch]. A doença foi codificada como uma variável binária, com valor '1' atribuído aos casos com severidade maior ou igual a 10% e valor '0', caso contrário.

Cada caso consistiu de variáveis meteorológicas horárias, incluindo temperatura, umidade relativa e chuva, sintetizadas para sete dias antes da data de florescimento da cultura. Um total de 154 casos foi usado, 70% para modelagem (108 casos) e os 30% restantes para validação (46 casos). O relacionamento entre as condições do ambiente e a doença foi modelado com regressão logística, redes neurais, K-vizinhos mais próximos e árvores de decisão.

Os modelos obtidos foram avaliados quanto à habilidade de classificar corretamente os casos, bem como quanto à sensibilidade (capacidade de classificar corretamente casos com severidade $\geq 10\%$) e à especificidade (capacidade de classificar corretamente casos com severidade $< 10\%$). Os resultados preliminares indicaram que a acurácia de todos os modelos variou de 50 a 79%. As árvores de decisão e os modelos de regressão logística tiveram os melhores desempenhos.

Baker et al. (1993) desenvolveram uma árvore de decisão para prever a extensão de mortalidade de pínus (*Pinus elliotii* e *Pinus taeda*) em decorrência de podridão das raízes causada por *Heterobasidion annosum*. Os dados foram obtidos de plantações inoculadas com *H. annosum*. Após cinco anos, contou-se a quantidade de árvores mortas mais a de árvores com infecções letais, em cada um dos 152 talhões de 16 locais selecionados em quatro estados norte-americanos.

O risco de mortalidade foi categorizado em baixo ou alto antes da modelagem. O risco foi definido como baixo, quando seis ou menos árvores de um talhão estavam mortas ou tinham infecções letais; e foi definido como alto, quando sete ou mais árvores estavam mortas

ou infectadas. Segundo esse critério, os 152 exemplos se dividiram entre 94 de risco baixo e 58 de risco alto. Vinte e duas medidas das características físicas e químicas do solo foram usadas como as variáveis preditivas.

Análise discriminante linear *stepwise* foi inicialmente usada, mas nenhum relacionamento linear forte foi encontrado. Usou-se, então, a modelagem por meio de árvores de decisão, segundo a abordagem CART. Uma árvore com apenas dois atributos de teste classificou corretamente o risco de mortalidade para 85% dos casos (129 dos 152 exemplos). Esse modelo revelou que o percentual de silte e o pH, ambos do horizonte A do solo, foram as variáveis importantes na predição da classe de risco, indicando que solos com baixo teor de silte ou alto pH são, geralmente, de alto risco. Por meio de validação cruzada (*10-fold cross-validation*), a estimativa de acurácia da árvore de decisão foi de 80%.

A indução de árvores de decisão foi usada ainda para se obter modelo de estimativa da duração do período de molhamento foliar, variável bastante importante na epidemiologia de muitas doenças e em diversos sistemas de alerta de doenças de plantas. Gleason et al. (1994) desenvolveram um modelo empírico não paramétrico para a estimativa da duração do molhamento foliar devido ao orvalho, usando árvores de classificação e regressão (CART) em conjunto com análise discriminante linear. Esse modelo foi usado para melhorar a estimativa de duração de períodos de molhamento foliar em comparação com um modelo proprietário (KIM et al., 2002).

3 MATERIAL E MÉTODOS

3.1 Os dados brutos

Os dados utilizados foram coletados por Japiassú et al. (2007) e se referem ao acompanhamento mensal da incidência da ferrugem do cafeeiro, de outubro de 1998 a outubro de 2006, na fazenda experimental da Fundação Procafé, localizada em Varginha, MG, latitude sul de 21° 34' 00'', longitude oeste de 45° 24' 22'' e altitude de 940 m.

A cada ano, no mês de setembro, foram selecionadas oito lavouras de café adultas em produção, com idade entre 6 e 20 anos, quatro em espaçamento largo (por volta de 3,5 m entre linhas e 0,7 m entre plantas – densidade média de 4.000 plantas/ha) e quatro adensadas (por volta de 2,5 m entre linhas e 0,5 m entre plantas – densidade média de 8.000 plantas/ha).

Para cada espaçamento, foram escolhidas duas lavouras com alta carga pendente de frutos (acima de 30 sacas beneficiadas/ha) e duas com baixa carga (abaixo de 10 sacas beneficiadas/ha). Em cada par de lavouras, uma foi da cultivar Catuaí e a outra da cultivar Mundo Novo. Não houve controle da doença durante o ano agrícola nos talhões escolhidos. O período de colheita foi entre junho e agosto.

O processo de amostragem, realizado no final de cada mês, foi o recomendado por Chalfoun (1997): coleta de 100 folhas do terço médio das plantas em cada talhão, entre o terceiro e o quarto par de folhas; contagem do número de folhas com lesões de ferrugem; e determinação da incidência (percentual de folhas atacadas) para cada uma das quatro combinações de espaçamento e produção das lavouras.

Dados meteorológicos, como temperatura (média, máxima e mínima), precipitação pluviométrica, umidade relativa do ar, entre outros, foram registrados a cada 30 min por uma estação meteorológica automática (marca Davis, modelo Groweather Industrial) instalada próximo dos locais de avaliação da incidência da ferrugem. Estes dados foram recebidos desde setembro de 1998.

O conjunto de dados recebido é composto por três tipos de arquivo, para cada mês do ano:

- Um arquivo texto (.txt) com os valores dos atributos meteorológicos registrados a cada meia hora pela estação meteorológica.

- Uma planilha (.xls) com valores diários de alguns dos atributos meteorológicos, calculados a partir do arquivo texto da estação meteorológica.
- Um documento (.doc) referente ao boletim de avisos mensal emitido pela Fundação Procafé, com informações climáticas e fenológicas relacionadas com a cultura do café e informações sobre a ocorrência de doenças e pragas, dentre elas a ferrugem do cafeeiro. Em cada boletim, é divulgado um valor de percentual de ataque para cada uma das quatro combinações de espaçamento e produção das lavouras, de cada doença ou praga.

3.2 Modelo do processo

O planejamento e a execução do processo de descoberta de conhecimento em bases de dados, também referido como processo de KDD, foram baseados no modelo de processo de mineração de dados da metodologia CRISP-DM (*CRoss Industry Standard Process for Data Mining*; CHAPMAN et al., 2000), a qual divide o ciclo de vida de um projeto em seis fases (Figura 7): compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição. CRISP-DM é atualmente a metodologia mais usada em iniciativas de mineração de dados ao redor do mundo (KDNUGGETS, 2008).

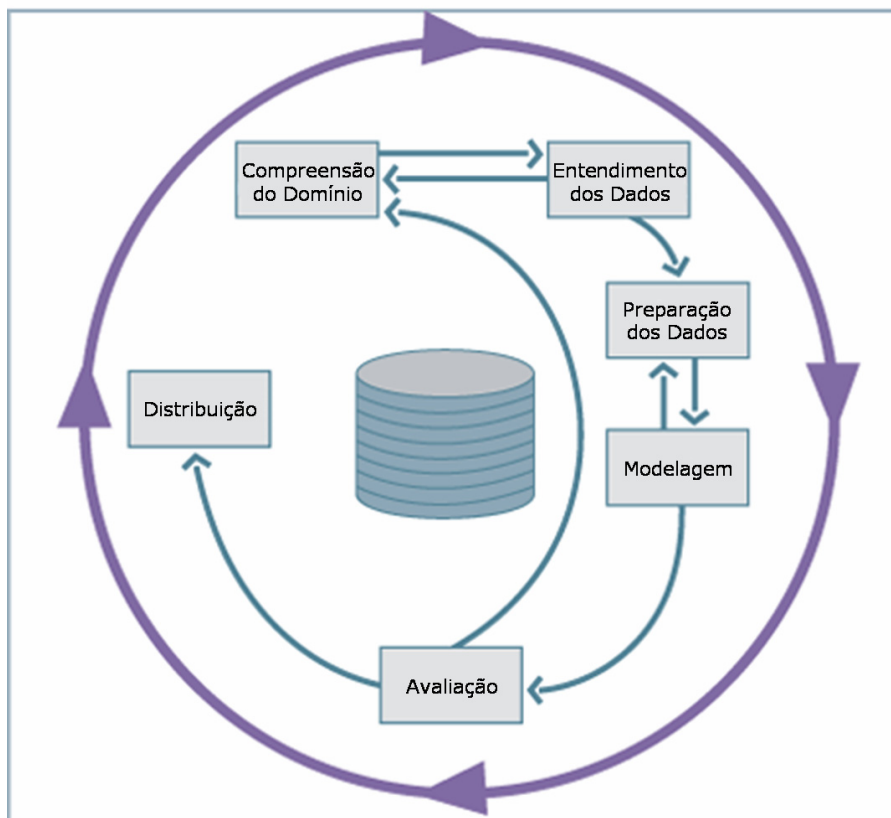


Figura 7: Fases do modelo de processo CRISP-DM (CHAPMAN et al., 2000).

A seqüência lógica entre as fases não é rígida, sendo comum, e quase sempre necessário, voltar e avançar entre diferentes fases. Qual fase, ou tarefa específica de uma fase, deve ser executada na seqüência depende do resultado da fase anterior. As setas na Figura 7 indicam as mais importantes e mais freqüentes dependências entre as fases.

Segue uma descrição sucinta e um breve comentário sobre cada fase:

- **Compreensão do domínio:** fase inicial para entender os objetivos e requisitos do projeto, pela perspectiva do domínio de aplicação, e depois converter esse conhecimento na definição do problema e no plano para alcançar os objetivos.

Esta fase consistiu da elaboração do plano de pesquisa para o exame de qualificação de doutorado e foi tema de trabalho apresentado em evento científico (MEIRA e RODRIGUES, 2005).

- **Entendimento dos dados:** começa com uma coleção de dados inicial e prossegue com atividades para se familiarizar com os dados, para identificar problemas de qualidade nesses dados e para buscar as primeiras compreensões (*insights*) a partir deles.

A seção 3.3 trata especificamente desta fase.

- **Preparação dos dados:** cobre todas as atividades para construir, a partir dos dados iniciais brutos, o conjunto de dados final para a modelagem. É uma fase muito importante do processo, que normalmente consome a maior parte do tempo e, em muitos projetos, não recebe a devida atenção. O desafio é preparar os dados de forma que a informação contida neles seja exposta da melhor maneira para as ferramentas de modelagem (PYLE, 1999). A seção 3.4 descreve em detalhes esta fase.

- **Modelagem:** técnicas de mineração de dados são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ótimos. Tipicamente, existem várias técnicas que podem ser utilizadas em um mesmo tipo de problema.

Desde a fase de compreensão do domínio, no plano de pesquisa, ficou decidido empregar a tarefa de classificação e, referente a ela, utilizar as técnicas de indução de árvore de decisão e de regras de classificação. A escolha se baseou no fato dessas técnicas permitirem a extração de regras compreensíveis, facilitando a participação de especialistas do domínio na interpretação e na avaliação dos modelos e tornando possível que algum padrão novo e interessante fosse percebido e posteriormente investigado. A seção 3.5 fornece mais informações sobre a fase de modelagem.

- **Avaliação:** nesse estágio do projeto tem-se um modelo, ou modelos, que apresentam boa qualidade na perspectiva da análise de dados. Antes de prosseguir com a sua distribuição, é importante avaliar cada modelo de forma mais completa e rever os passos executados na sua construção, para se ter a certeza de que atendem aos objetivos traçados. A avaliação deve cobrir tanto os modelos obtidos como também quaisquer outras descobertas não diretamente relacionadas, que possam vir a desvendar novos desafios ou idéias para futuras investigações.

As discussões sobre a fase de avaliação encontram-se nos capítulos 4 e 5.

- **Distribuição:** a criação de modelos geralmente não finaliza um projeto. É preciso colocá-los disponíveis para o uso – mesmo que o propósito de um modelo seja apenas aumentar o conhecimento sobre os dados, esse conhecimento adquirido deve ser organizado e apresentado de modo que o usuário possa aproveitá-lo; e ainda que o analista de dados não vá realizar o esforço de colocar os modelos em uso, é importante que o usuário tenha a noção de quais ações devem ser realizadas para se ter o proveito dos modelos criados. A discussão e as orientações a respeito desta fase estão no capítulo 5.

Durante a execução das fases deste projeto, os resultados parciais foram divulgados em eventos acadêmicos e científicos (MEIRA e RODRIGUES, 2006a; MEIRA e RODRIGUES, 2006b; MEIRA e RODRIGUES, 2007a; MEIRA e RODRIGUES, 2007b).

3.3 Entendimento dos dados

3.3.1 Coleção de dados inicial

Os dados foram recebidos da Fundação Procafé em remessas incrementais, por meio de mensagens eletrônicas com arquivos anexados. Houve uma vez em que foram recebidos em CD-ROM, este obtido em uma viagem a Varginha.

Para facilitar o processamento na fase de preparação dos dados, foi estabelecida uma estrutura base de diretórios para o processo de KDD e uma nomenclatura padrão para os arquivos. Criou-se um diretório para os arquivos da estação meteorológica, outro para as planilhas com dados meteorológicos diários e mais um para os dados de incidência da ferrugem.

Os arquivos da estação foram nomeados no formato ‘AAAA-MM.txt’, onde ‘AAAA’ é o ano e ‘MM’ é o número correspondente ao mês (p.ex. ‘2000-01’ refere-se ao mês de janeiro de 2000). Da mesma maneira, os arquivos das planilhas foram nomeados no formato

‘AAAA-MM.xls’ e os arquivos dos boletins no formato ‘BoletimAAAA-MM.doc’. Cada arquivo devidamente nomeado foi colocado no seu local correspondente da estrutura de diretórios criada.

3.3.2 Descrição dos dados

A descrição dos dados brutos foi feita para cada tipo de fonte de dados recebida. Foram elaborados um relatório de descrição de dados referente aos arquivos da estação meteorológica, um relatório referente às planilhas com dados meteorológicos diários e um relatório referente aos boletins mensais de avisos.

A Tabela 4 relaciona os atributos considerados relevantes para a pesquisa registrados pela estação meteorológica. A relevância dos atributos foi determinada com base na importância de cada atributo observada na revisão bibliográfica e com base na opinião de especialista do domínio.

Tabela 4: Descrição dos atributos relevantes registrados pela estação meteorológica.

Atributo	Tipo	Unidade de medida
DATA	Alfanumérico (DD/MM/AAAA)	-
Significado: Data em que foram obtidos os valores dos atributos medidos pelos sensores da estação meteorológica.		
HORA	Alfanumérico ([0-23]:[00 30])	h:min
Significado: Hora em que os dados foram obtidos no dia correspondente. A estação meteorológica da Fundação Procafé esteve programada para realizar as leituras a cada 30 min, de 0:00 hora até 23:30.		
TMED	Numérico	°C
Significado: Temperatura do ar (média dos últimos 30 min) medida através de um sensor de temperatura, dada em graus Celsius (°C). Precisão: +/- 0,5°C. Nome original do atributo no arquivo da estação: ‘Ar Temp’.		
TMAX	Numérico	°C
Significado: Temperatura máxima do ar, dada em °C, obtida a cada intervalo de tempo definido pelo usuário (30 min). Precisão: +/- 0,5°C. Nome original do atributo no arquivo da estação: ‘Ar Max’.		

TMIN	Numérico	°C
<p>Significado: Temperatura mínima do ar, dada em °C, obtida a cada intervalo de tempo definido pelo usuário (30 min). Precisão: +/- 0,5°C. Nome original do atributo no arquivo da estação: 'Ar Min'.</p>		
VVENTO	Numérico (≥0)	km/h
<p>Significado: Velocidade do vento (média dos últimos 30 min). É medida em quilômetros por hora (km/h). Nome original do atributo no arquivo da estação: 'Vento Vel.'.</p>		
PRECIP	Numérico (≥0, múltiplo de 0,2)	mm
<p>Significado: Medida por um coletor, registra os dados de precipitação pluviométrica durante o último período (30 min.). É dada em milímetros (mm). Precisão de pluviosidade: +/- 2% para taxas de 0,2 mm a 50 mm por hora, +/- 3% para taxas de 50 mm a 100 mm por hora. Nome original do atributo no arquivo da estação: 'Prec'.</p>		
INDPLUVMAX	Numérico (≥0)	mm/h
<p>Significado: O índice pluviométrico máximo (ou taxa de pluviosidade) é calculado através da determinação do intervalo de tempo entre cada aumento de 0,2 mm na precipitação. É dado em milímetros por hora (mm/h). É uma medida de intensidade das chuvas. Precisão: +/- 5%. Nome original do atributo no arquivo da estação: 'Max Índ.'.</p>		
UR	Numérico inteiro (≥0)	%
<p>Significado: Umidade relativa do ar no momento do registro. A umidade relativa fornece uma leitura de umidade que reflete a porcentagem de vapor de água que o ar tem capacidade de armazenar. A umidade relativa não é a quantidade de vapor de água no ar, mas sim a proporção de vapor de água do ar para a sua capacidade. Precisão: +/- 13%. Nome original do atributo no arquivo da estação: 'Umid'.</p>		

Apesar da estação meteorológica contar com sensor de molhamento foliar, o atributo não foi escolhido por duas razões: o sensor entrou em operação apenas no dia 21 de julho de 1999; e não existe um padrão de medida de molhamento foliar nos diversos tipos de sensores disponíveis no mercado – um modelo que dependesse de um atributo com este grau de especificidade iria requerer medidas de molhamento foliar compatíveis com a do sensor em questão, o que restringiria o seu uso.

Os atributos de interesse nas planilhas com dados diários meteorológicos são equivalentes aos registrados pela estação, diferentes apenas no nível de granularidade: temperatura média diária, temperatura máxima diária, temperatura mínima diária, velocidade média diária do vento, precipitação acumulada no dia e umidade relativa média diária.

Por que usar dados das planilhas com valores diários dos atributos meteorológicos se eles podem ser calculados a partir dos registros da estação? De fato, apenas em casos específicos foram usados os dados meteorológicos diários presentes nas planilhas. As razões para isso encontram-se na seção 3.3.4, no item “Verificação da qualidade das planilhas com dados meteorológicos diários”.

Os atributos considerados relevantes extraídos dos boletins de avisos estão descritos na Tabela 5. A escolha desses atributos também foi feita com base na literatura e com o apoio de especialista do domínio.

Tabela 5: Descrição dos atributos relevantes dos boletins de avisos.

Atributo	Tipo	Unidade de medida
LAVOURA	Categórico	Adensada ou Larga
Significado: Condição da lavoura quanto ao espaçamento entre plantas: adensada ou larga.		
CARGA	Categórico	Alta ou Baixa
Significado: Nível de produção da lavoura (carga pendente de frutos): alta ou baixa. Nome original do atributo no boletim de avisos: ‘Produção’.		
INCIDENCIA	Numérico	%
Significado: Incidência da ferrugem do cafeeiro, ou seja, percentual de folhas com lesões de ferrugem. Nome original do atributo no boletim de avisos: ‘% Folhas/Frutos ³ Atacados - Ferrugem’.		

3.3.3 Exploração dos dados

A exploração dos dados permeou todo o pré-processamento dos dados, desde o estado inicial, com os dados brutos, até os dados preparados para as ferramentas de modelagem. Esta seção trata da exploração dos dados brutos. Em seções mais adiante, alguns resultados da exploração dos dados preparados são exibidos e discutidos.

³ No boletim, esse atributo indica também o ataque de outras doenças e pragas, que podem atacar os frutos.

Exploração dos dados de incidência da ferrugem do cafeeiro

Os dados de incidência da ferrugem do cafeeiro foram explorados por meio de gráficos de evolução da doença ao longo do tempo e também por gráficos para análise de sua distribuição. Os primeiros foram gerados com o editor de planilhas Microsoft® Excel, enquanto os últimos foram gerados com a ferramenta SAS/INSIGHT® (SAS INSTITUTE INC., 2004a).

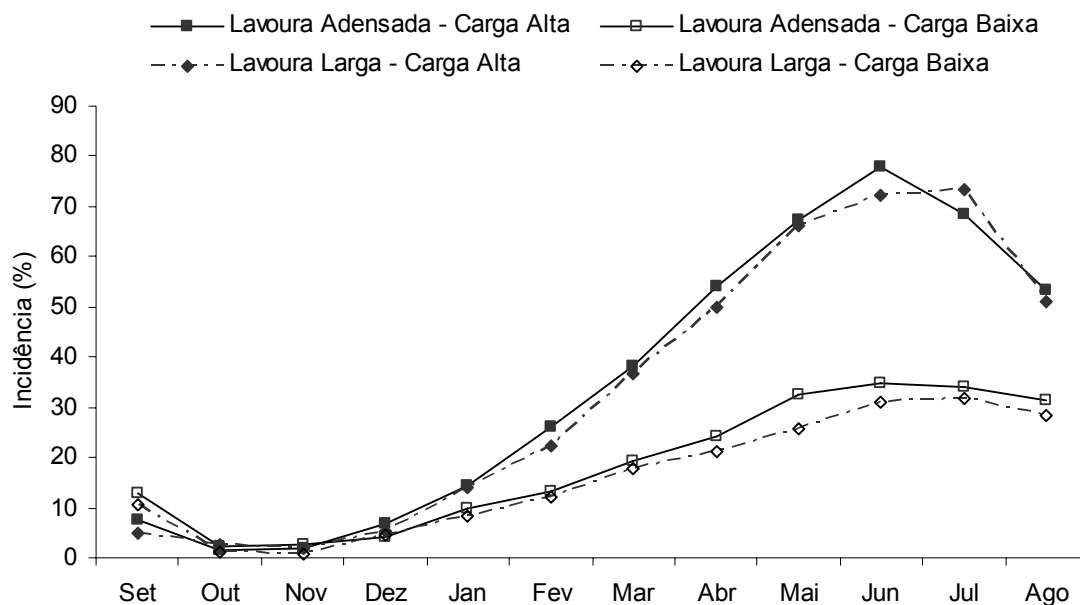


Figura 8: Evolução mensal da incidência da ferrugem do cafeeiro em lavouras com diferentes espaçamentos e cargas pendentes de frutos – média de 1998/1999 a 2005/2006.

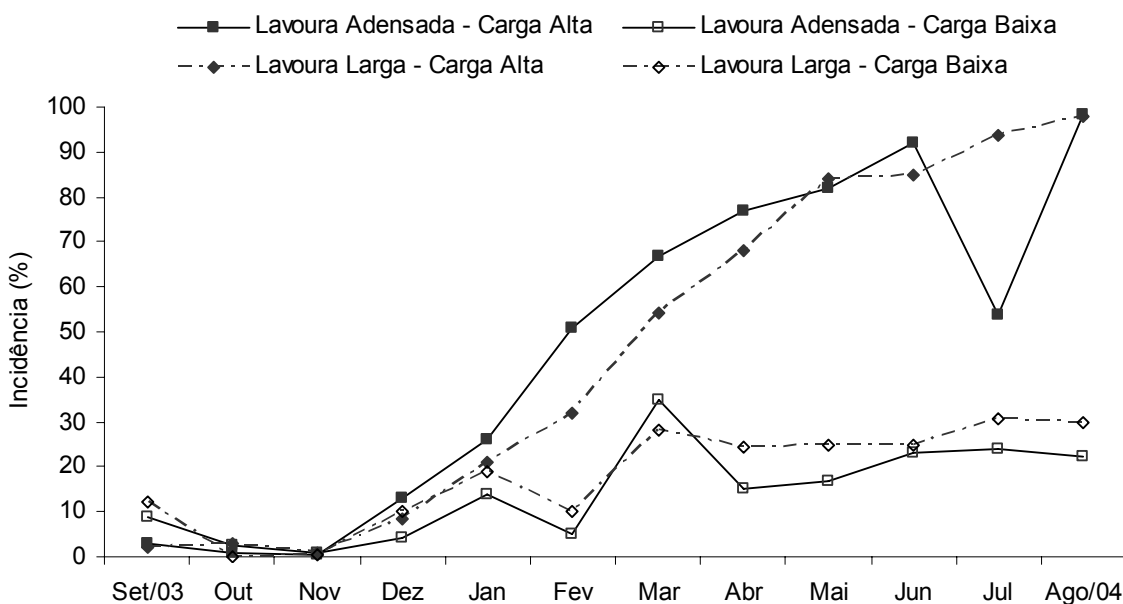


Figura 9: Evolução mensal da incidência da ferrugem do cafeeiro no ano agrícola 2003/2004 em lavouras com diferentes espaçamentos e cargas pendentes de frutos.

A Figura 8 representa a evolução média mensal da incidência da ferrugem do cafeeiro no período analisado, para as quatro combinações de espaçamento-carga. Gráficos do mesmo tipo também foram usados para explorar a evolução da doença em cada ano agrícola, como, por exemplo, de setembro de 2003 a agosto de 2004 (Figura 9).

Um aspecto importante da exploração dos dados, além de ter proporcionado um melhor entendimento deles, foi o auxílio na tarefa de verificar a sua qualidade. Por exemplo, no gráfico da Figura 9, é possível visualizar uma queda suspeita da incidência da ferrugem no mês de julho de 2004 (lavoura adensada - carga alta). A discussão sobre este comportamento estranho da doença e a conclusão a que se chegou está adiante, no item 3.3.4.

A distribuição dos valores dos atributos foi analisada por meio de histogramas e gráficos do tipo *box plot*. A Figura 10 é o gráfico *box plot* que representa a distribuição dos valores de incidência da ferrugem do cafeeiro, independente do espaçamento e da carga pendente da lavoura.

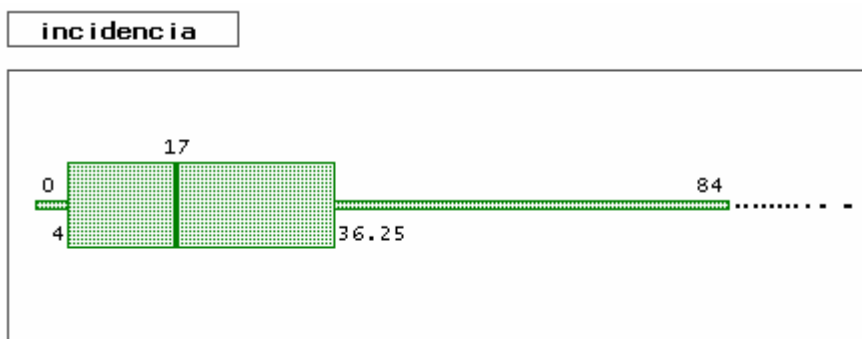


Figura 10: Distribuição dos valores de incidência da ferrugem do cafeeiro independente do espaçamento e da carga pendente de frutos da lavoura.

A linha de dentro da caixa representa a mediana (17%). As laterais da caixa representam os quartis Q1 (4%) e Q3 (36,25%). As caixas menores que se estendem a partir desses quartis são chamadas de *whiskers*. Elas se estendem a partir dos quartis até a observação mais distante que não tenha distância maior do que 1,5 vezes a diferença entre os quartis Q3 e Q1 ($Q3 - Q1 = 32,25$). Valores além dos *whiskers* são representados por marcas individuais. Outras estatísticas a respeito da distribuição dos valores de incidência: valor máximo = 98,5%; valor mínimo = 0% e valor médio = 25,8%.

Gráficos do tipo *box plot* foram usados também para comparar a distribuição dos valores de incidência de acordo com o espaçamento (Figura 11) ou com a carga pendente de frutos (Figura 12). Pela Figura 11, nota-se que a distribuição dos valores de incidência da

ferrugem para lavouras adensadas é muito parecida com a distribuição para lavouras largas. Quando a comparação é entre lavouras de carga alta ou baixa (Figura 12), percebe-se que as lavouras com alta carga atingiram valores de incidência bem mais elevados do que as lavouras com baixa carga.

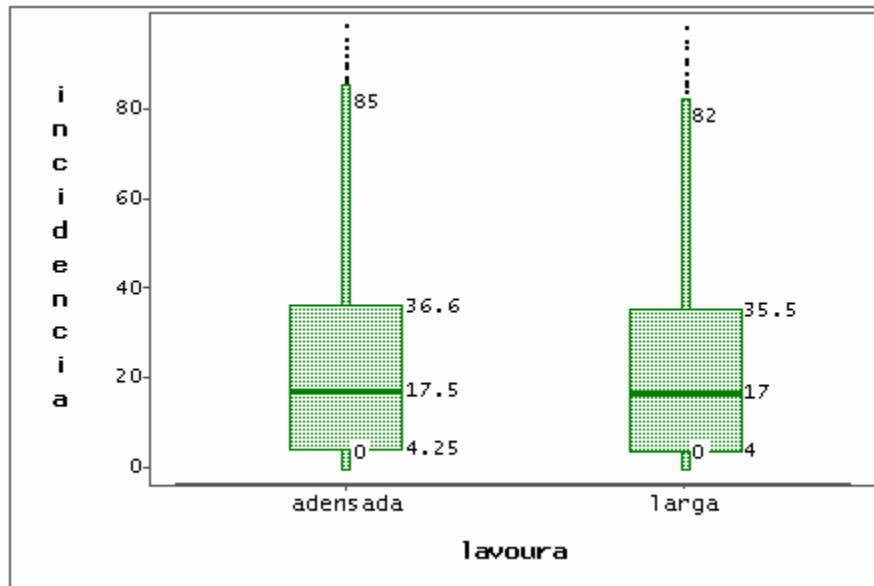


Figura 11: Distribuição dos valores de incidência da ferrugem do cafeeiro de acordo com o espaçamento da lavoura.

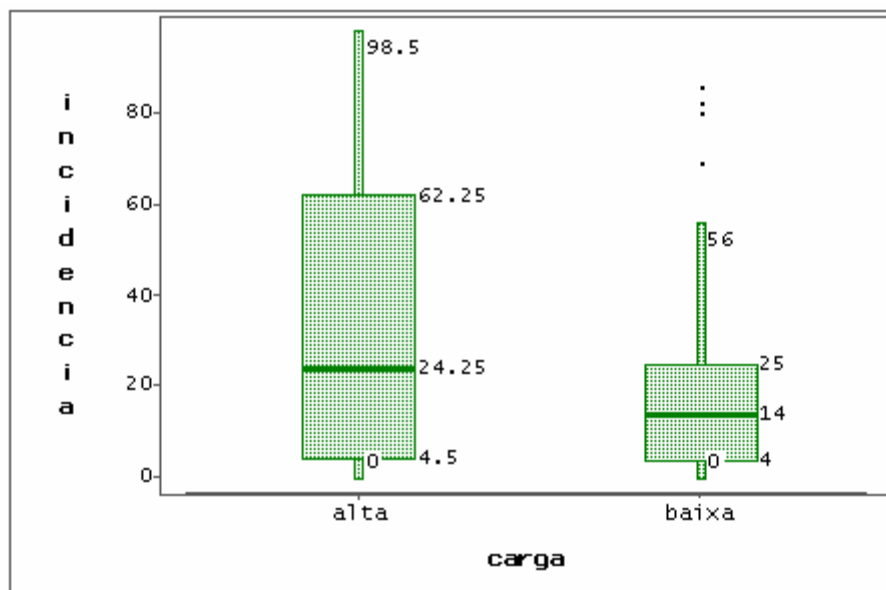


Figura 12: Distribuição dos valores de incidência da ferrugem do cafeeiro de acordo com a carga pendente de frutos da lavoura.

Exploração dos dados da estação meteorológica

Os dados meteorológicos de cada mês foram explorados com a ferramenta DP, disponível no CD-ROM que acompanha o livro *Data Preparation for Data Mining* (PYLE, 1999). Os relatórios de saída da ferramenta foram usados na elaboração de relatórios de exploração para cada ano contemplado no conjunto de dados.

Esses relatórios apresentam os atributos avaliados e seu tipo, os valores mínimo e máximo de cada atributo, o número de valores distintos que foram registrados, quantos valores vazios (nulos) foram encontrados e a quantidade de registros. A Tabela 6 corresponde a uma parte da saída da ferramenta DP para o arquivo de dados meteorológicos de janeiro de 2003.

Tabela 6: Exploração dos dados registrados pela estação meteorológica no mês de janeiro de 2003.

Atributo	Tipo	Min	Max	Distintos	Vazios	Reg
data	C	-	-	19	0	853
hora	C	-	-	48	0	853
tmed	N	16.4	31.6	136	0	853
tmax	N	16.7	32.2	135	0	853
tmin	N	16.2	31.1	129	0	853
vvento	N	0	14.5	10	0	853
precip	N	0	11	26	0	853
indpluvmax	N	0	15.2	28	1	852
ur	N	2	100	74	5	848

Durante a exploração, observou-se que o atributo velocidade do vento teve sempre poucos valores distintos registrados. Analisando os arquivos, descobriu-se que a velocidade do vento não foi medida em uma escala contínua de valores. Seus valores possuem uma diferença de 1,6 km/h (exceto entre os valores 8,0 e 9,7): 0,0; 1,6; 3,2; 4,8; 6,4; 8,0; 9,7; 11,3; 12,9; 14,5 etc.

Também com os relatórios de exploração foi possível identificar problemas de qualidade nos dados meteorológicos registrados pela estação, conforme apresentado na próxima seção.

3.3.4 Verificação da qualidade dos dados

Verificação inicial da qualidade dos dados

A tarefa de verificar a qualidade dos dados iniciou-se antes mesmo de se colocar cada arquivo de dados recebido no seu respectivo local na estrutura de diretórios criada para o processo de KDD (seção 3.3.1).

Primeiro, uma verificação da quantidade de arquivos e do conteúdo deles permitiu identificar alguns problemas: ausência dos boletins de avisos de setembro e dezembro de 2000 (foram enviados arquivos por engano como sendo dos boletins); ausência da planilha com os dados meteorológicos diários de dezembro de 2001; e ausência dos arquivos da estação meteorológica dos meses de outubro e novembro de 2000.

Depois, por meio de uma inspeção visual nos arquivos, foi possível identificar outros tipos de problema com os dados, principalmente nos arquivos da estação meteorológica: registros ausentes, em decorrência de paradas no funcionamento da estação; registros repetidos e/ou não pertencentes ao mês correspondente do arquivo da estação; e registros para o dia 29 de fevereiro de 2006, sendo que 2006 não foi ano bissexto.

Ao final desse esforço de verificação da qualidade dos dados, foi produzido um documento com várias observações e dúvidas sobre os dados e a forma como foram obtidos. Tudo foi discutido e resolvido com o pessoal técnico da Fundação Procafé em uma viagem a Varginha - MG, em maio de 2006. Pôde-se apurar, por exemplo, que as paradas no funcionamento da estação meteorológica ocorreram por problemas na bateria da estação ou em períodos em que a estação esteve desligada para manutenção.

Os arquivos de dados foram colocados na estrutura base de diretórios com os problemas resolvidos, exceto, é claro, os registros ausentes devido às paradas no funcionamento da estação meteorológica.

Verificação da qualidade dos dados de incidência da ferrugem do cafeeiro

A qualidade dos dados de incidência da ferrugem do cafeeiro foi verificada visualmente por meio de gráficos de evolução da doença. Comportamentos considerados suspeitos, observados nas curvas de progresso da ferrugem, foram anotados e, depois, discutidos com especialistas do domínio. Um único caso foi considerado como erro: queda acentuada da incidência da doença no mês de julho de 2004 (lavoura adensada - carga alta) e alta expressiva no mês seguinte, como pode ser observado no gráfico da Figura 9.

A conclusão a que se chegou é que deve ter ocorrido algum problema na amostragem que determinou aquele valor de incidência da ferrugem do cafeeiro em julho de 2004. A solução adotada foi trocar o valor com problema pela média aritmética dos valores de incidência de junho e agosto de 2004.

Verificação da qualidade dos dados da estação meteorológica

A qualidade dos dados da estação meteorológica foi verificada de três maneiras: mediante inspeção visual; com o auxílio dos resultados da exploração dos dados (seção 3.3.3); e por meio de testes de consistência dos dados.

Além da inspeção visual geral nos arquivos, descrita no início desta seção, foi realizada uma inspeção visual mais detalhada em cada atributo dos arquivos de registro da estação meteorológica. Esta inspeção tornou possível, por exemplo, a identificação de valores inconsistentes de umidade relativa do ar (UR). Um desses casos foi a identificação de seqüências de valores com UR = 33% entre valores com UR = 100%, em ocasiões com chuva e/ou molhamento foliar. Decidiu-se, então, que cada um desses valores 33 seria trocado pelo valor 100 no início da preparação dos dados.

A exploração dos dados meteorológicos, assim como com os dados de incidência da ferrugem do cafeeiro, auxiliou na identificação de problemas. Por exemplo, no mês de janeiro de 2003, a estação meteorológica deveria ter realizado 1488 registros (31 dias x 48 registros diários). Na Tabela 6, a coluna 'Reg' indica que foram realizados apenas 853 registros, o que significa que houve uma parada considerável no funcionamento da estação naquele mês. Ainda, na Tabela 6, é possível observar outros problemas: um valor ausente (vazio) para o atributo INDPLUVMAX; valor mínimo de 2% para a umidade relativa do ar (UR), o que é certamente um erro; e cinco valores ausentes para o atributo UR.

A verificação da qualidade dos dados meteorológicos se completou com vários testes de consistência sobre os valores de cada um dos atributos. A Tabela 7 apresenta os testes realizados em cada atributo e os resultados desses testes.

Tabela 7: Testes de consistência aplicados nos atributos registrados pela estação meteorológica.

Atributo DATA	Teste: DATA = DD/MM/AAAA
Resultado: Para os meses de fevereiro a setembro de 2004, as datas foram registradas no padrão DD-MM-AAAA.	
Atributo HORA	Teste: HORA = HH: [00 30]
Resultado: Foram encontrados 09 registros não realizados em hora inteira ou meia hora: <ul style="list-style-type: none">▪ 09/09/1998: 01 (17h17)▪ 14/09/1998: 08 (15h41, 15h45, 15h46, 15h47, 15h48, 15h49, 15h50, 15h51)	

Atributo TEMP	Teste: $TEMP > 0$
Resultado: Todos os valores de temperatura obedeceram a regra verificada.	
Atributo TMAX	Testes: $TMAX > 0$ $TMAX \geq TEMP$
Resultado: Todos os valores de temperatura máxima obedeceram as regras verificadas.	
Atributo TMIN	Testes: $TMIN > 0$ $TMIN \leq TEMP$
Resultado: Todos os valores de temperatura mínima obedeceram as regras verificadas.	
Atributo VVENTO	Testes: $VVENTO \geq 0$ $VVENTO \leq 150$
Resultado: Todos os valores de velocidade do vento obedeceram as regras verificadas.	
Atributo PRECIP	Testes: $PRECIP \geq 0$ $PRECIP$ múltiplo de $0,2^4$
Resultado: Todos os valores de precipitação obedeceram as regras verificadas.	
Atributo INDPLUVMAX	Teste: $INDPLUVMAX \geq 0$
Resultado: Todos os valores de índice pluviométrico obedeceram a regra verificada.	
Atributo UR	Testes: 1. $UR \geq 0$ 2. $UR \geq 20$
Resultado: <ol style="list-style-type: none"> 1. Todos os valores de umidade relativa obedeceram a regra verificada. 2. Alguns registros apresentaram umidade relativa abaixo de 20%. Em alguns casos, em épocas bastante secas, constatou-se que os registros estavam consistentes. Os demais casos ocorreram entre valores com $UR = 100\%$ e em situações com chuva e/ou molhamento foliar. Então, decidiu-se que cada valor com problema seria trocado pelo valor 100 no início da preparação dos dados. 	

⁴ O coletor da estação meteorológica mediu a precipitação em incrementos de 0,2 mm.

Verificação da qualidade das planilhas com dados meteorológicos diários

As planilhas mensais com dados meteorológicos diários foram elaboradas pelo pessoal da Fundação Procafé, a partir dos arquivos de registro da estação meteorológica. A verificação da qualidade desses dados foi feita mediante inspeção visual das planilhas.

O que chamou a atenção nas planilhas foi que todos os valores diários estavam preenchidos, mesmo para os dias em que houve ausência considerável de registro pela estação. Nesses casos, os valores diários dos atributos meteorológicos foram obtidos de três formas:

1. Valores diários de temperatura e de precipitação foram coletados de outros equipamentos (termômetro e pluviômetro) no mesmo local da estação meteorológica, enquanto os valores dos demais atributos foram copiados de dias com condições de tempo parecidas.

Os anos e dias em que isso ocorreu estão indicados no quadro abaixo:

2000	2003	2004
01 a 06/01/2000	11 a 24/01/2003	16 a 30/11/2004
03 a 24/10/2000		
01 a 24/11/2000		

2. Todos os valores diários foram copiados de dias com condições de tempo parecidas. Tanto nesses casos, como nos do item anterior, a determinação do dia com condições de tempo parecidas foi feita com base em avaliação feita pelo técnico responsável pelos dados na Fundação Procafé. Os anos e dias em que isso ocorreu estão indicados no quadro abaixo:

1998	1999	2002	2005	2006
17 a 21/12/1998	19 a 22/02/1999	06/06/2002	01 a 04/01/2005	01 e 02/03/2006
	13 a 15/12/1999		31/01/2005	
	27 a 31/12/1999		08 a 10/11/2005	

3. Os valores diários foram calculados a partir de registros realizados por uma estação meteorológica nova instalada no mesmo local, no ano de 2006. Nesses casos, usou-se o registro do mês inteiro da nova estação e não apenas dos dias e horários em que houve parada no funcionamento da estação antiga. Os meses em que isso ocorreu foram julho e setembro de 2006.

Esses valores diários mencionados das planilhas foram usados para complementar a geração dos atributos meteorológicos no nível diário durante a fase de preparação dos dados (passo 2 da preparação dos dados na seção 3.4.4).

3.4 Preparação dos dados

3.4.1 Especificação do atributo meta

O atributo meta (variável dependente) foi obtido por meio de transformações nos valores de incidência da ferrugem do cafeeiro. A evolução da doença, entre uma avaliação e a outra, foi definida como a variável de interesse. Sendo assim, foram calculadas as taxas de infecção da ferrugem de cada mês, segundo a equação 3.1:

$$t = y_i - y_{i-1} \quad 3.1$$

onde t é a taxa de infecção, y_i é a incidência da ferrugem do cafeeiro no mês em questão e y_{i-1} é a incidência da ferrugem no mês anterior.

Em problemas de classificação, o atributo meta é categórico. Por isso, os valores numéricos das taxas de infecção foram mapeados para três categorias ou classes: ‘TX1(≤ 0)’ - redução ou estagnação, para taxas de infecção negativas ou nulas; ‘TX2($> 0 \leq 5$)’ - crescimento moderado, para taxas de infecção positivas, menores ou iguais a 5 pontos percentuais (p.p.); e ‘TX3(> 5)’ - crescimento acelerado, para taxas de infecção maiores que 5 p.p. As classes foram escolhidas com base nas faixas de valores de incidência da ferrugem do cafeeiro recomendadas por Zambolim et al. (1997) para o controle da doença via foliar. Esta taxa de infecção categórica com três classes foi tomada como o atributo meta, nomeado TAXA_INF3N.

Após seguir em frente no processo de KDD, tendo-se obtido e avaliado modelos para o atributo meta TAXA_INF3N, decidiu-se também por gerar modelos em que a taxa de infecção da ferrugem fosse considerada como um atributo binário, com apenas dois valores. O intuito foi procurar obter modelos menos complexos (árvores de decisão com menor quantidade de nós) e com melhores valores das medidas de avaliação do que os modelos obtidos para o atributo com três classes.

A primeira opção de taxa de infecção binária foi estabelecida com a criação do atributo TAXA_INF_M5, com valor ‘1’ para taxas de infecção maiores ou iguais a 5 p.p. e valor ‘0’, caso contrário. Como opção adicional, foi criado também o atributo TAXA_INF_M10, com valor ‘1’ para taxas de infecção maiores ou iguais a 10 p.p. e valor ‘0’, caso contrário.

A fronteira em 5 p.p., como para TAXA_INF3N, foi baseada em Zambolim et al. (1997). Já a fronteira em 10 p.p. foi baseada em Kushalappa et al. (1984), que propuseram o

limite de risco de 10% na proporção de folhas com ferrugem para recomendar a aplicação de fungicida. Também, este valor está próximo do limite máximo de 12% de folhas doentes em que se recomenda a aplicação de fungicidas sistêmicos para o controle da doença (ZAMBOLIM et al., 1997).

3.4.2 Especificação dos atributos preditivos

Os atributos preditivos (variáveis independentes ou explicativas) meteorológicos foram construídos a partir do nível horário (registros da estação), passando pelo nível diário, até um nível que permitisse a análise de seu relacionamento com o atributo meta.

No nível diário, além de médias e de somatórios das variáveis meteorológicas, foram calculados valores estimados de molhamento foliar prolongado (mínimo de 6 h), uma vez que a germinação só ocorre se a folha estiver molhada, e seis horas de água livre na superfície da folha foi avaliado como o tempo mínimo necessário para ocorrer infecção (KUSHALAPPA et al., 1983).

O número de horas com alta umidade relativa do ar ($\geq 95\%$) foi utilizado como medida indireta de molhamento foliar contínuo (SUTTON et al., 1984). Em dias com períodos de molhamento disjuntos, foi considerado o maior, com tolerância de até uma hora entre eles para juntar em um único período.

Os períodos de molhamento foliar foram analisados tanto na sua extensão total como na sua fração noturna (das 20:00 h às 8:00 h), já que a infecção ocorre preferencialmente na ausência de ou com pouca luminosidade (MONTOYA e CHAVES, 1974). O dia foi considerado de 12:00 h de um dia comum até 12:00 h do dia seguinte (denominado “dia epidemiológico”), pois os períodos de molhamento ocorrem geralmente entre um dia e o outro.

As temperaturas médias durante os períodos total e noturno de molhamento foliar contínuo também foram calculadas para cada dia, uma vez que, enquanto a superfície da folha está molhada, a temperatura é o fator principal que determina o percentual de germinação dos esporos e de penetração (KUSHALAPPA et al., 1983).

Os dias com precipitação maior ou igual a 1 mm foram considerados chuvosos. Este critério foi o mesmo usado por Kushalappa et al. (1983) ao considerar um dia como chuvoso.

Na seqüência da preparação dos dados meteorológicos, cada dia foi tratado como um eventual dia de infecção. Considerando um período de incubação estimado, cada dia foi

associado ao mês correspondente de avaliação da incidência da ferrugem (Figura 13). O período de incubação para cada dia foi estimado pela equação 2.16 (seção 2.2.2) proposta por Moraes et al. (1976): Dessa forma, cada dia foi associado a uma taxa de infecção, para a qual possivelmente teve parcela de contribuição.

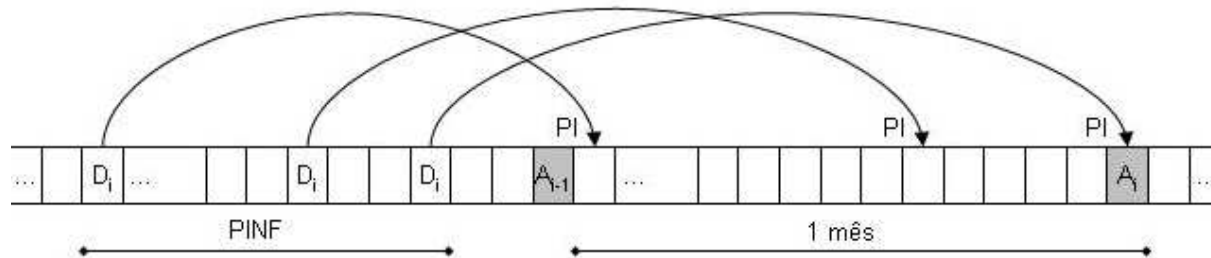


Figura 13: Representação dia-a-dia do esquema usado na preparação dos dados meteorológicos.
D_i - dia de infecção; A_i - avaliação da incidência da ferrugem do cafeeiro; A_{i-1} - avaliação da incidência no mês anterior; PI - período de incubação; PINF - período de infecção.

O conjunto de dias associado a uma taxa de infecção foi denominado de período de infecção (PINF), conforme representado na Figura 13. Para o PINF é que foram criados os atributos preditivos meteorológicos usados na modelagem, relacionados e descritos na Tabela 8. Também estão incluídos na Tabela 8 os dois atributos preditivos não relacionados com as condições meteorológicas: o espaçamento da lavoura (lavoura adensada ou larga) e a carga pendente de frutos (carga alta ou baixa).

Tabela 8: Atributos preditivos usados na modelagem.

Nome	Descrição		
	Tipo	Medida	Significado
CARGA	binário	-	Carga pendente de frutos: ALTA ou BAIXA.
DCHUV_PINF	numérico	dias	Número de dias chuvosos (precipitação ≥ 1 mm) no PINF (período de infecção).
LAVOURA	binário	-	Espaçamento: lavoura ADENSADA ou LARGA.
MED_INDPLUVMAX_PINF	numérico	mm/h	Média do índice pluviométrico máximo diário no PINF.
MED_PRECIP_PINF	numérico	mm	Média das precipitações pluviais diárias no PINF.
NHNUR95_PINF	numérico	h	Média diária do número de horas noturnas com umidade relativa do ar $\geq 95\%$ no PINF.
NHUR95_PINF	numérico	h	Média diária do número de horas com umidade relativa do ar $\geq 95\%$ no PINF.
PRECIP_PINF	numérico	mm	Precipitação pluvial acumulada no PINF.
SMT_NHNUR95_PINF	numérico	h	Somatório de NHNUR95 no PINF.

SMT_NHUR95_PINF	numérico	h	Somatório de NHUR95 no PINF.
SMT_VVENTO_PINF	numérico	km/h	Média do somatório da velocidade do vento de cada dia do PINF.
THNUR95_PINF	numérico	°C	Temperatura média diária durante as horas noturnas com umidade relativa $\geq 95\%$ no PINF.
THUR95_PINF	numérico	°C	Temperatura média diária durante as horas com umidade relativa $\geq 95\%$ no PINF.
TMAX_PINF	numérico	°C	Média das temperaturas máximas diárias no PINF.
TMAX_PI_PINF	numérico	°C	Média das temperaturas máximas diárias no período de incubação para os dias do PINF.
TMED_PINF	numérico	°C	Média das temperaturas médias diárias no PINF.
TMED_PI_PINF	numérico	°C	Média das temperaturas médias diárias no período de incubação para os dias do PINF.
TMIN_PINF	numérico	°C	Média das temperaturas mínimas diárias no PINF.
TMIN_PI_PINF	numérico	°C	Média das temperaturas mínimas diárias no período de incubação para os dias do PINF.
UR_PINF	numérico	%	Umidade relativa do ar média diária no PINF.
VVENTO_PINF	numérico	km/h	Velocidade média diária do vento no PINF.

3.4.3 Especificação dos atributos preditivos especiais

Alguns atributos meteorológicos foram criados de maneira especial. A intenção foi aglutinar e embutir nesses atributos, por meio de alguma representação, resultados e aspectos conhecidos da epidemiologia da ferrugem do cafeeiro encontrados na literatura.

De início, foi definida uma classificação para as condições diárias de molhamento foliar, e de luminosidade e temperatura durante o período de molhamento, com relação ao processo de infecção. A classificação, baseada nos trabalhos de Montoya e Chaves (1974) e Kushalappa et al. (1983), foi a seguinte:

- Condições diárias de molhamento foliar e luminosidade durante o período de molhamento com relação à infecção (M-L):
 - **Desfavorável:** molhamento foliar noturno inferior a 4 horas ($\text{NHNUR95} < 4$) ou molhamento foliar total inferior a 6 horas ($\text{NHUR95} < 6$).
 - **Pouco favorável:** molhamento foliar noturno maior ou igual a 4 horas ($\text{NHNUR95} \geq 4$) e molhamento foliar total maior ou igual a 6 horas ($\text{NHUR95} \geq 6$).

- **Favorável:** molhamento foliar noturno maior ou igual a 8 horas ($NHNUR95 \geq 8$) e molhamento foliar total maior ou igual a 12 horas ($NHUR95 \geq 12$).
- **Muito favorável:** molhamento foliar noturno maior ou igual a 8 horas ($NHNUR95 \geq 8$) e molhamento foliar total maior ou igual a 18 horas ($NHUR95 \geq 18$).
- Condições diárias de temperatura durante o período de molhamento foliar com relação à infecção (T):
 - **Desfavorável:** temperatura média durante o período de molhamento foliar inferior a 15 °C ($THUR95 < 15$) ou superior a 29 °C ($THUR95 > 29$).
 - **Pouco favorável:** temperatura média durante o período de molhamento foliar maior ou igual a 15 °C e menor que 18 °C ($15 \leq THUR95 < 18$) ou maior que 27 °C e menor ou igual a 29 °C ($27 < THUR95 \leq 29$).
 - **Favorável:** temperatura média durante o período de molhamento foliar maior ou igual a 18 °C e menor que 21 °C ($18 \leq THUR95 < 21$) ou maior que 24 °C e menor ou igual a 27 °C ($24 < THUR95 \leq 27$).
 - **Muito favorável:** temperatura média durante o período de molhamento foliar maior ou igual a 21 °C e menor ou igual a 24 °C ($21 \leq THUR95 \leq 24$).

Tabela 9: Matriz de condições diárias de infecção e seus respectivos índices numéricos.

M-L	T	Desfavorável ($T < 15$ ou $T > 29$)	Pouco favorável ($15 \leq T < 18$ ou $27 < T \leq 29$)	Favorável ($18 \leq T < 21$ ou $24 < T \leq 27$)	Muito favorável ($21 \leq T \leq 24$)
Desfavorável ($NHNUR95 < 4$ ou $NHUR95 < 6$)		Desfavorável 0	Desfavorável 0	Desfavorável 0	Desfavorável 0
Pouco favorável ($NHNUR95 \geq 4$ e $NHUR95 \geq 6$)		Desfavorável 0	Desfavorável 0*	Favorável 1*	Favorável 2*
Favorável ($NHNUR95 \geq 8$ e $NHUR95 \geq 12$)		Desfavorável 0	Favorável 1	Favorável 2	Muito favorável 3
Muito favorável ($NHNUR95 \geq 8$ e $NHUR95 \geq 18$)		Desfavorável 0	Favorável 2	Muito favorável 3	Muito favorável 4

* Caso $NHNUR95 \geq 8$ ou $NHUR95 \geq 12$, incrementa-se 0,5 no índice.

Em seguida, definiu-se mais uma classificação, para o que se denominou de condição diária de infecção, considerando a matriz das combinações de molhamento foliar e luminosidade (M-L) e de temperatura (T) definidas pela classificação anterior. Além da classificação, definiu-se também um índice numérico para essa condição diária de infecção. A Tabela 9 representa a matriz com cada opção de condição diária de infecção e o seu respectivo índice numérico.

A matriz da Tabela 9, considerando apenas os índices de condição diária de infecção, é parecida com a matriz de valores diários de severidade da ferrugem do cafeeiro usada por Garçon et al. (2004) como base de um sistema de previsão para o controle da doença (seção 2.3.3).

Por fim, a partir da matriz da Tabela 9, foram criados os atributos preditivos especiais, que estão relacionados e descritos na Tabela 10.

Tabela 10: Atributos preditivos especiais usados na modelagem.

Nome	Descrição		
	Tipo	Medida	Significado
ACDINF_PINF	numérico	-	Acumulado da condição diária de infecção no período de infecção (PINF), isto é, o somatório dos índices de condição diária de infecção no PINF.
DDI_PINF	numérico	dias	Número de dias desfavoráveis à infecção no PINF.
DFMFI_PINF	numérico	dias	Número de dias favoráveis e muito favoráveis à infecção no PINF.
DMFI_PINF	numérico	dias	Número de dias muito favoráveis à infecção no PINF.

3.4.4 Passos da preparação dos dados

Esquema geral da preparação dos dados

A preparação dos dados para a fase de modelagem foi realizada em quatro passos. A Figura 14 ilustra o esquema geral adotado, partindo da coleção de dados inicial e chegando nos dados preparados para a modelagem, ao término da seqüência de passos.

A execução de cada passo da preparação dos dados foi feita por meio de programas de computador escritos na linguagem de programação Perl (ActivePerl[®] versão 5.8.7, ActiveState Corp.). Perl é apontada, em pesquisas realizadas na internet (KDNUGGETS,

2008), como uma das principais linguagens de manipulação de dados adotadas em projetos de mineração de dados.

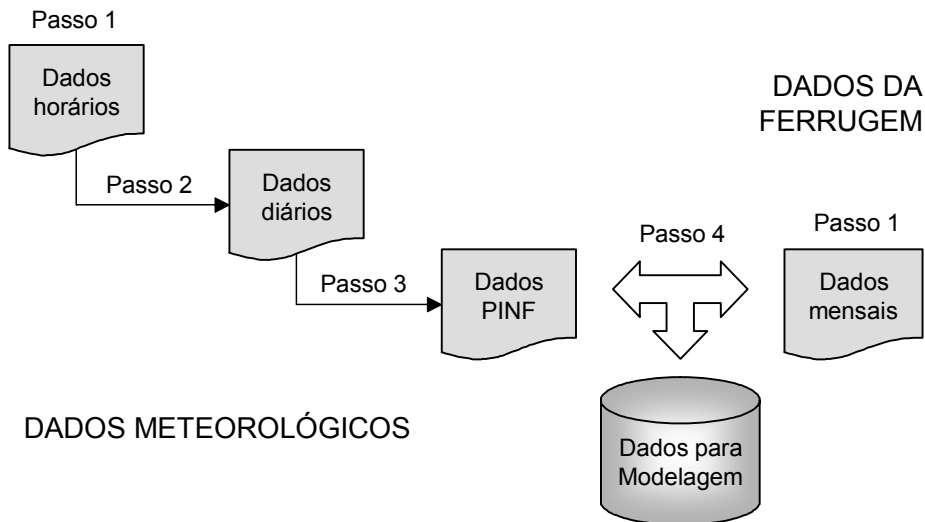


Figura 14: Esquema geral da preparação dos dados para a modelagem.

Na implementação dos programas em Perl, além dos recursos da própria linguagem, foram usados os seguintes módulos disponíveis no repositório de acesso livre CPAN (*Comprehensive Perl Active Network* – www.cpan.org): ‘Date::Calc’, para operações com datas; ‘Statistics::Descriptive’, para cálculo de estatísticas descritivas, como médias e somatórios; e ‘DBD-CSV’, para operações com arquivos no formato CSV (*Comma Separated Values*), adotado como formato padrão para os arquivos de dados.

Passo 1 - Reunião dos dados brutos e criação do atributo meta

O primeiro passo da preparação dos dados foi reunir os dados brutos da coleção de dados inicial. Os dados extraídos dos boletins de avisos foram reunidos em um arquivo e os dados meteorológicos em outro.

Os atributos relevantes dos boletins de avisos (LAVOURA, CARGA e INCIDENCIA, descritos na Tabela 5) tiveram seus valores digitados em um arquivo único do tipo planilha eletrônica. O atributo meta correspondente à taxa de infecção categórica com três classes (TAXA_INF3N) foi criado neste mesmo arquivo. Já os atributos meta referentes às taxas de infecção binárias (TAXA_INF_M5 e TAXA_INF_M10) foram criados depois, diretamente no arquivo de dados pronto para a modelagem (ver seção 3.5.3).

Os arquivos da estação meteorológica foram concatenados um ao outro. Alguns dos problemas encontrados na verificação da qualidade dos dados (ver seção 3.3.4) foram

resolvidos: substituição do padrão ‘DD-MM-AAAA’ de algumas datas pelo padrão ‘DD/MM/AAAA’; descarte de registros não realizados em hora inteira ou meia hora; e substituição de valores inconsistentes de umidade relativa do ar.

Foi criado, também, o dia epidemiológico (atributos DATA_EPID e HORA_EPID), com início às 12:00 h de um determinado dia (0:00 h do dia epidemiológico) e término às 12:00 h do dia seguinte (24:00 h do dia epidemiológico).

Passo 2 - Criação de atributos no nível diário

O segundo passo da preparação dos dados, e o mais trabalhoso, foi criar os atributos no nível diário. A execução deste passo foi subdividida:

1. **Cálculo de estatísticas descritivas diárias.** O arquivo gerado no passo anterior serviu de base para cálculo de estatísticas descritivas diárias dos atributos meteorológicos: temperaturas média, máxima e mínima do dia, velocidade média do vento do dia, somatório da velocidade do vento do dia, precipitação acumulada do dia, índice pluviométrico máximo do dia e umidade relativa média do dia.
O arquivo gerado foi depois complementado com valores extraídos das planilhas recebidas com dados meteorológicos diários. Isso ocorreu apenas para os dias em que houve ausência significativa de registro pela estação meteorológica (ver seção 3.3.4, no item “Verificação da qualidade das planilhas com dados meteorológicos diários”).
2. **Criação de atributos relacionados com molhamento foliar.** A partir do arquivo gerado no passo anterior, foram criados atributos no nível diário relacionados com molhamento foliar contínuo: período total e noturno de horas com umidade relativa $\geq 95\%$, temperatura média durante esses períodos e classe e índice da condição diária de infecção.
3. **Criação de atributos relacionados com período de incubação.** A partir do arquivo gerado no sub-passo 1, foram gerados atributos relacionados com período de incubação, tais como: período de incubação estimado do eventual dia de infecção e mês e ano prováveis de incidência da ferrugem. Estes atributos permitiram relacionar todos os dias (eventuais dias de infecção) com as taxas mensais de infecção da ferrugem do cafeeiro.
4. **Junção dos arquivos gerados.** Os arquivos gerados nos sub-passos anteriores foram juntados em um só arquivo para o passo seguinte da preparação dos dados.

Passo 3 - Criação de atributos para o PINF

O terceiro passo da preparação dos dados foi criar os atributos de cada período de infecção, a partir do arquivo de dados final gerado no passo 2. A especificação desses atributos encontra-se nas seções 3.4.2 e 3.4.3.

Passo 4 - Integração dos dados

O quarto passo da preparação dos dados foi integrar os dados preparados a partir dos boletins de avisos com os dados preparados desde os arquivos da estação meteorológica. A integração foi feita com base nos valores de mês e ano – o mês e o ano de avaliação da incidência da ferrugem, pelo lado dos dados vindos dos boletins, com o mês e o ano do período de infecção correspondente, pelo lado dos dados meteorológicos.

3.4.5 Preparativos finais para a modelagem

O conjunto de dados preparado totalizou 384 registros (8 anos x 12 meses x 4 combinações espaçamento-carga), referente ao período de outubro de 1998 a outubro de 2006. No entanto, os meses de novembro e dezembro de 2000 e fevereiro de 2000 e 2003 foram bastante comprometidos pelos períodos de parada no funcionamento da estação meteorológica. Os dados meteorológicos disponíveis também não foram suficientes para formar o PINF completo correspondente à incidência da ferrugem observada em outubro de 1998. Sendo assim, os registros dos referidos meses (5 meses x 4 combinações = 20) foram eliminados, resultando em um conjunto com 364 registros (exemplos ou casos) para a fase de modelagem.

Ao explorar os atributos preparados, notou-se redundância entre a temperatura média diária durante o período de molhamento foliar (THUR95_PINF) e a temperatura média diária durante o período noturno de molhamento foliar (THNUR95_PINF), com valores praticamente iguais para esses atributos (Figura 15). Decidiu-se, então, eliminar o atributo THNUR95_PINF do conjunto de dados.

Ainda como parte da preparação final para a modelagem, experimentou-se diferentes opções de seleção de atributos, começando pela seleção do conjunto completo dos atributos preparados e passando pela seleção de alguns subconjuntos deste. No caso das taxas de infecção binárias, decidiu-se também pela seleção de registros segundo a carga pendente de

frutos. A seção 3.5.1 fornece mais informações a respeito dessas seleções de atributos e de registros.

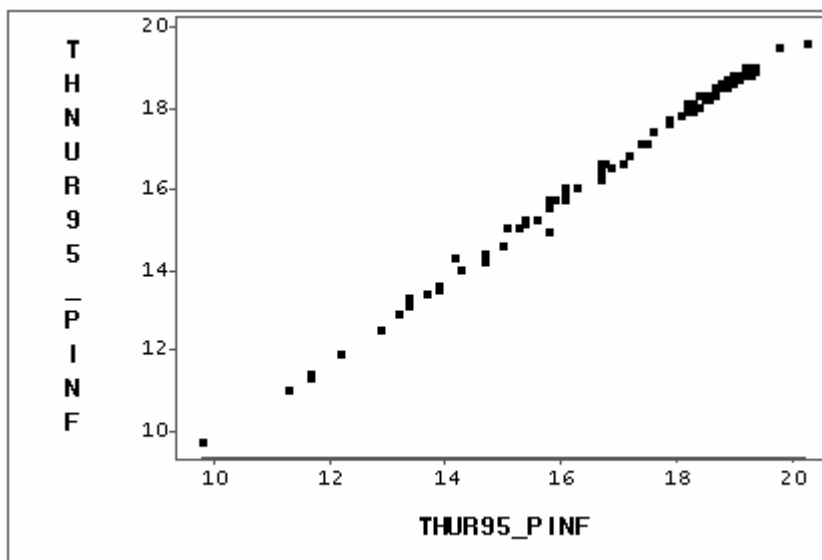


Figura 15: Relação entre a temperatura média durante o período de molhamento foliar (THUR95_PINF) e a temperatura média durante o período noturno de molhamento foliar (THNUR95_PINF).

3.5 Modelagem

Em resumo, após a fase de preparação dos dados, o conjunto de dados completo disponível para a fase de modelagem ficou composto da seguinte forma:

- 3 opções de atributo meta (seção 3.4.1):
 - taxa de infecção categórica com três classes (TAXA_INF3N).
 - taxas de infecção binárias (TAXA_INF_M5 e TAXA_INF_M10).
- 24 atributos preditivos ou explicativos (seções 3.4.2, 3.4.3 e 3.4.5).
- 364 registros ou exemplos (seção 3.4.5).

3.5.1 Geração dos modelos

Atributo meta: taxa de infecção categórica com três classes (TAXA_INF3N)

A geração dos modelos de alerta da ferrugem do cafeeiro foi iniciada considerando a taxa de infecção categórica com três classes (TAXA_INF3N) como o atributo meta (seção 3.4.1). Algumas opções de seleção de atributos foram experimentadas e os modelos gerados. Na avaliação desses modelos, constatou-se que os seus desempenhos não eram muito expressivos.

No entanto, pela interpretação dos modelos, percebeu-se que muitas das regras estavam bastante condizentes com os resultados de estudos epidemiológicos da ferrugem do cafeeiro encontrados na literatura. Daí surgiu a idéia de investir na técnica de indução de árvores de decisão aplicada na epidemiologia da doença.

Tabela 11: Conjunto de treinamento usado na indução da árvore de decisão aplicada na análise da epidemia da ferrugem do cafeeiro.

Nome	Descrição		
	Tipo	Medida	Significado
Atributo meta			
TAXA_INF3N	nominal	-	Taxa de infecção em três níveis categóricos: TX1(≤ 0), TX2($> 0 \leq 5$) ou TX3(> 5).
Atributos explicativos*			
LAVOURA	binário	-	Espaçamento: lavoura ADENSADA ou LARGA.
CARGA	binário	-	Carga pendente de frutos: ALTA ou BAIXA.
TMAX_PINF	numérico	°C	Média das temperaturas máximas diárias no PINF (período de infecção).
TMIN_PINF	numérico	°C	Média das temperaturas mínimas diárias no PINF.
TMED_PINF	numérico	°C	Média das temperaturas médias diárias no PINF.
UR_PINF	numérico	%	Umidade relativa do ar média diária no PINF.
MED_PRECIP_PINF	numérico	mm	Média das precipitações pluviiais diárias no PINF.
PRECIP_PINF	numérico	mm	Precipitação pluvial acumulada no PINF.
NHUR95_PINF	numérico	h	Média diária do número de horas com umidade relativa do ar $\geq 95\%$ no PINF.
THUR95_PINF	numérico	°C	Temperatura média diária durante as horas com umidade relativa do ar $\geq 95\%$ no PINF.
NHNUR95_PINF	numérico	h	Média diária do número de horas noturnas com umidade relativa do ar $\geq 95\%$ no PINF.
TMAX_PI_PINF	numérico	°C	Média das temperaturas máximas diárias no período de incubação para os dias do PINF.
TMIN_PI_PINF	numérico	°C	Média das temperaturas mínimas diárias no período de incubação para os dias do PINF.

* Os atributos estão relacionados na ordem em que aparecem no conjunto de treinamento.

Para isso, decidiu-se adotar uma seleção de atributos particular. Foram excluídos os atributos DCHUV_PINF, MED_INDPLUVMAX_PINF, SMT_NHNUR95_PINF, SMT_NHUR95_PINF, SMT_VVENTO_PINF, VVENTO_PINF e TMED_PI_PINF (seção 3.4.2), mais os 4 atributos especiais ACDINF_PINF, DDI_PINF, DFMEI_PINF e DMFI_PINF (seção 3.4.3). O conjunto de treinamento com 364 exemplos ficou composto conforme apresentado na Tabela 11.

Embora a velocidade do vento apareça com destaque na literatura, favorecendo a disseminação dos esporos do fungo causador da ferrugem do cafeeiro, decidiu-se por não incluí-la na análise (atributos VVENTO_PINF e SMT_VVENTO_PINF). Chegou-se a experimentar sua inclusão no conjunto de treinamento, mas sua importância foi marginal no modelo gerado.

Nos trabalhos pesquisados sobre o período de incubação do fungo, o destaque foi para as médias das temperaturas máximas e mínimas no período. Por isso, o atributo TMED_PI_PINF foi excluído da seleção, para se poder analisar exclusivamente o comportamento dos atributos TMAX_PI_PINF e TMIN_PI_PINF.

Os demais atributos excluídos não foram considerados importantes no caso particular da análise da epidemia da ferrugem do cafeeiro com árvore de decisão.

Atributo meta: taxa de infecção binária (TAXA_INF_M5 ou TAXA_INF_M10)

Com a intenção de melhorar o desempenho dos modelos de alerta da ferrugem do cafeeiro, foram criadas as taxas de infecção binárias, representadas pelos atributos meta TAXA_INF_M5 e TAXA_INF_M10 (seção 3.4.1).

Decidiu-se também separar a geração dos modelos por carga pendente de frutos, com base nos seguintes pontos:

- Conhecimento prévio sobre a epidemiologia da ferrugem do cafeeiro: a evolução da doença é bem mais acentuada nos anos de alta carga pendente de frutos; os próprios dados analisados confirmaram isso, conforme pode ser visto na Figura 8 e na Figura 12.
- Simplicidade dos modelos: um modelo geral seria mais complexo (com maior número de regras e de condições) do que dois modelos separados, um para alta carga pendente e outro para baixa carga pendente.

- Utilidade dos modelos: pela característica bianual dos cafezais, com anos de alta carga pendente seguidos de anos com baixa carga, o usuário não teria dificuldades para identificar qual modelo utilizar em cada ano agrícola.

Portanto, a geração dos modelos de alerta contemplou quatro combinações de atributo meta e carga pendente: TAXA_INF_M5 - carga alta; TAXA_INF_M5 - carga baixa; TAXA_INF_M10 - carga alta; e TAXA_INF_M10 - carga baixa.

O conjunto de treinamento, para cada uma dessas combinações, ficou com 182 exemplos (364 exemplos divididos entre carga pendente alta e baixa).

O atributo CARGA foi eliminado dos conjuntos de treinamento. A identificação da carga pendente referente ao conjunto foi feita no nome do arquivo.

Foram escolhidas três opções de seleção dos atributos preditivos para a modelagem. A Tabela 12 indica os atributos selecionados em cada uma destas opções. Segue a descrição de cada uma das opções de seleção e a razão de sua escolha:

- **Modelagem 1:** seleção de todos os atributos preditivos. Esta opção visou avaliar modelos gerados com todos os atributos especificados e preparados.
- **Modelagem 2:** seleção dos atributos derivados de elementos meteorológicos com disponibilidade ampla, como temperatura, precipitação e umidade relativa, mais alguns atributos relacionados com molhamento foliar, que foram derivados da umidade relativa do ar. Esta opção visou avaliar modelos gerados a partir de dados que são mais comuns e, portanto, independentes de registros meteorológicos detalhados.
- **Modelagem 3:** seleção dos atributos da Modelagem 2, exceto os atributos relacionados com molhamento foliar, que foram bem mais trabalhosos de preparar. Além da razão mencionada no item anterior, a escolha desta opção visou avaliar modelos gerados a partir de dados menos custosos de preparar e que não dependessem de registros meteorológicos no nível horário.

Tabela 12: Atributos preditivos de cada opção de seleção de atributos para a geração dos modelos de alerta da ferrugem do cafeeiro.

Atributos*	Opção de seleção dos atributos		
	Modelagem 1	Modelagem 2	Modelagem 3
LAVOURA	*	*	*
TMAX_PINF	*	*	*
TMIN_PINF	*	*	*
TMED_PINF	*	*	*
UR_PINF	*	*	*
MED_PRECIP_PINF	*	*	*
PRECIP_PINF	*	*	*
DCHUV_PINF	*	*	*
MED_INDPLUVMAX_PINF	*		
ACDINF_PINF	*		
DMFI_PINF	*		
DFMFI_PINF	*		
DDI_PINF	*		
NHUR95_PINF	*	*	
SMT_NHUR95_PINF	*		
THUR95_PINF	*	*	
NHNUR95_PINF	*	*	
SMT_NHNUR95_PINF	*		
TMAX_PI_PINF	*	*	*
TMIN_PI_PINF	*	*	*
TMED_PI_PINF	*	*	*
VVENTO_PINF	*		
SMT_VVENTO_PINF	*		

* Os atributos estão relacionados na ordem em que aparecem no conjunto de treinamento.

3.5.2 Avaliação dos modelos e das regras

A base para a avaliação dos modelos de alerta foi a matriz de confusão, formada pelos acertos e erros na predição das taxas de infecção. A partir da matriz de confusão, foram calculadas a acurácia e a taxa de erro do modelo e a precisão para cada classe de taxa de infecção (seção 2.1.3).

No caso das taxas de infecção binárias, as medidas de precisão para as classes correspondem à sensibilidade (precisão para a classe ‘1’) e à especificidade (precisão para a classe ‘0’). Além da sensibilidade e da especificidade, foram calculadas a confiabilidade positiva e a confiabilidade negativa (seção 2.1.3).

A matriz de confusão e as medidas de avaliação de cada modelo foram obtidas pelos métodos de resubstituição e de validação cruzada (seção 2.1.3). A validação cruzada foi realizada por meio de 10 partições aleatórias e estratificadas (distribuição dos exemplos de cada classe parecida com a do conjunto completo) do conjunto de treinamento.

Os valores finais das medidas de avaliação foram calculados pela média dos valores obtidos nas 10 repetições da validação cruzada. Sendo assim, o desvio padrão da média (*standard error of the mean* - SEM) foi também calculado, para dar uma noção da variação dos valores obtidos em cada repetição.

O cálculo do desvio padrão da média (*DP*) é dado pela equação 3.2, onde σ é o desvio padrão da amostra e n é o tamanho da amostra. O desvio padrão é dado pela equação 3.3, onde x_i é cada valor da amostra e \bar{x} é a média aritmética desses valores.

$$DP = \frac{\sigma}{\sqrt{n}} \quad 3.2$$

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad 3.3$$

Conforme visto na seção 2.1.3, um classificador pode não ter uma boa precisão, mas pode conter regras que, individualmente, sejam precisas ou possuam algum outro tipo de propriedade interessante. Então, além da avaliação dos modelos (classificadores), foi feita também uma avaliação individualizada das regras de classificação extraídas de cada modelo.

As regras correspondentes aos nós folhas de cada árvore de decisão foram extraídas. Para cada uma dessas regras, foi elaborada a matriz de contingência (seção 2.1.3) e, a partir dela, foram calculados valores de medidas de avaliação de regras. As medidas de avaliação de regras adotadas foram: precisão (razão de Laplace), sensibilidade, especificidade, novidade, cobertura e suporte (seção 2.1.3).

Esta avaliação das regras foi feita exclusivamente para os modelos de alerta com as taxas de infecção binárias como o atributo meta. No caso da árvore de decisão para a análise

da epidemia da ferrugem do cafeeiro, não foram calculadas e analisadas medidas de avaliação de regras individualmente.

Ainda como parte da avaliação, foram realizadas reuniões com um fitopatologista, especialista em epidemiologia de doenças de plantas, que realizou pesquisas relevantes na epidemiologia da ferrugem do cafeeiro. Essas reuniões foram de fundamental importância. Os modelos puderam ser avaliados pela perspectiva do domínio de aplicação e o contato com o especialista auxiliou em muito na análise e na interpretação dos modelos e de suas regras.

3.5.3 Uso e configuração das ferramentas de modelagem

Para a geração dos modelos, foram utilizadas duas ferramentas ou *software* de modelagem: o SAS[®] Enterprise Miner[™] (versão 4.3, SAS Institute Inc.) e o Weka (versão 3.4.11, Universidade de Waikato, Nova Zelândia). A execução de ambos foi feita em microcomputador com plataforma Windows.

O Enterprise Miner[™] é a solução SAS para o processo de mineração de dados (SAS INSTITUTE INC., 2004b). Ele proporciona uma infra-estrutura integrada de ferramentas que dá suporte a todo o processo, de acordo com a abordagem exclusiva SEMMA, que corresponde a cinco passos principais: amostragem (*Sampling*), exploração (*Exploration*), modificação (*Modification*), modelagem (*Modeling*) e avaliação (*Assessment*). Foi escolhido pela reconhecida qualidade dos produtos SAS e pela oportunidade de uso por meio de um convênio estabelecido entre a FEAGRI/UNICAMP e o SAS Brasil e de um contrato de licença de uso acadêmico diretamente com o doutorando.

O Weka é uma coleção de algoritmos de aprendizado de máquina e de ferramentas relacionadas, que também oferece suporte ao processo completo de mineração de dados (WITTEN e FRANK, 2005). Ele é um *software* livre, gratuito (www.cs.waikato.ac.nz/ml/weka), distribuído sob a licença de uso GNU (*General Public License*). Foi escolhido por ter sido apontado, em pesquisas realizadas na internet (KDNUGGETS, 2008), como uma das ferramentas mais utilizadas em iniciativas de mineração de dados que usam software livre. A escolha teve o intuito também de permitir a sua comparação com o software comercial do SAS.

Ambos, Enterprise Miner[™] e Weka, foram utilizados na modelagem com as taxas de infecção binárias como o atributo meta. No caso da modelagem aplicada à epidemiologia da

ferrugem do cafeeiro, foi utilizado apenas o Enterprise Miner™, cujas características se adequaram melhor ao tipo de aplicação.

Uso e configuração do SAS® Enterprise Miner™

A modelagem no Enterprise Miner™ foi realizada com a criação de projetos (a maior entidade de manipulação no software) e, para cada projeto, com a construção do diagrama de fluxo dos dados e de operações sobre esses dados (SAS INSTITUTE INC., 2004b; SAS INSTITUTE INC., 2004c).

Foram criados dois projetos, um para a indução da árvore de decisão para analisar a epidemia da ferrugem do cafeeiro e outro para a geração de modelos de alerta da ferrugem, usando as taxas de infecção binárias como o atributo meta.

Os diagramas foram elaborados pela composição de nós (*nodes*), que representam as ferramentas de operação sobre os dados, com setas de fluxo de dados. Também, os diagramas dos projetos foram organizados em níveis, dividindo-os em subdiagramas.

A Figura 16 ilustra como isso foi feito no caso da modelagem com as taxas de infecção binárias. Os quadros da figura representam:

- A. Diagrama geral.** Diagrama do primeiro nível, que representa o carregamento do conjunto de dados produzido na fase de preparação, um pré-processamento nos dados e a modelagem. Os subdiagramas de modelagem correspondem às três opções de seleção de atributos empregadas na geração dos modelos de alerta (seção 3.5.1).
- B. Subdiagrama de pré-processamento.** O pré-processamento foi incluído para a criação dos atributos meta TAXA_INF_M5 e TAXA_INF_M10, que representam as taxas de infecção binárias. Esses atributos foram criados dentro do Enterprise Miner™, pois a decisão de sua criação foi tomada no andamento do processo (seções 3.4.1 e 3.5.1). Em seguida, foram filtrados/eliminados os períodos de infecção mais comprometidos pelas paradas no funcionamento da estação meteorológica (3.4.5). Neste ponto, também, foram feitas algumas configurações dos atributos, próprias do Enterprise Miner™. Nós das ferramentas *Insight* (SAS/INSIGHT®) e *Distribution Explorer* foram usados para exploração dos dados finais disponíveis para a modelagem.
- C. Subdiagramas de modelagem.** Os subdiagramas de modelagem possuem a mesma estrutura. Primeiro, a seleção dos atributos segundo a opção de modelagem e, em seguida,

a modelagem por carga pendente de frutos. Foram incluídos, também, um subdiagrama para validação cruzada e outro para exportação dos dados para o Weka.

D. Subdiagramas de modelagem por CARGA. Os subdiagramas de modelagem por carga pendente de frutos, dentro de cada subdiagrama de modelagem, também possuem a mesma estrutura. De início, a filtragem entre os exemplos de carga alta ou baixa e, depois, a indução das árvores de decisão com o atributo meta sendo TAXA_INF_M5 ou TAXA_INF_M10.

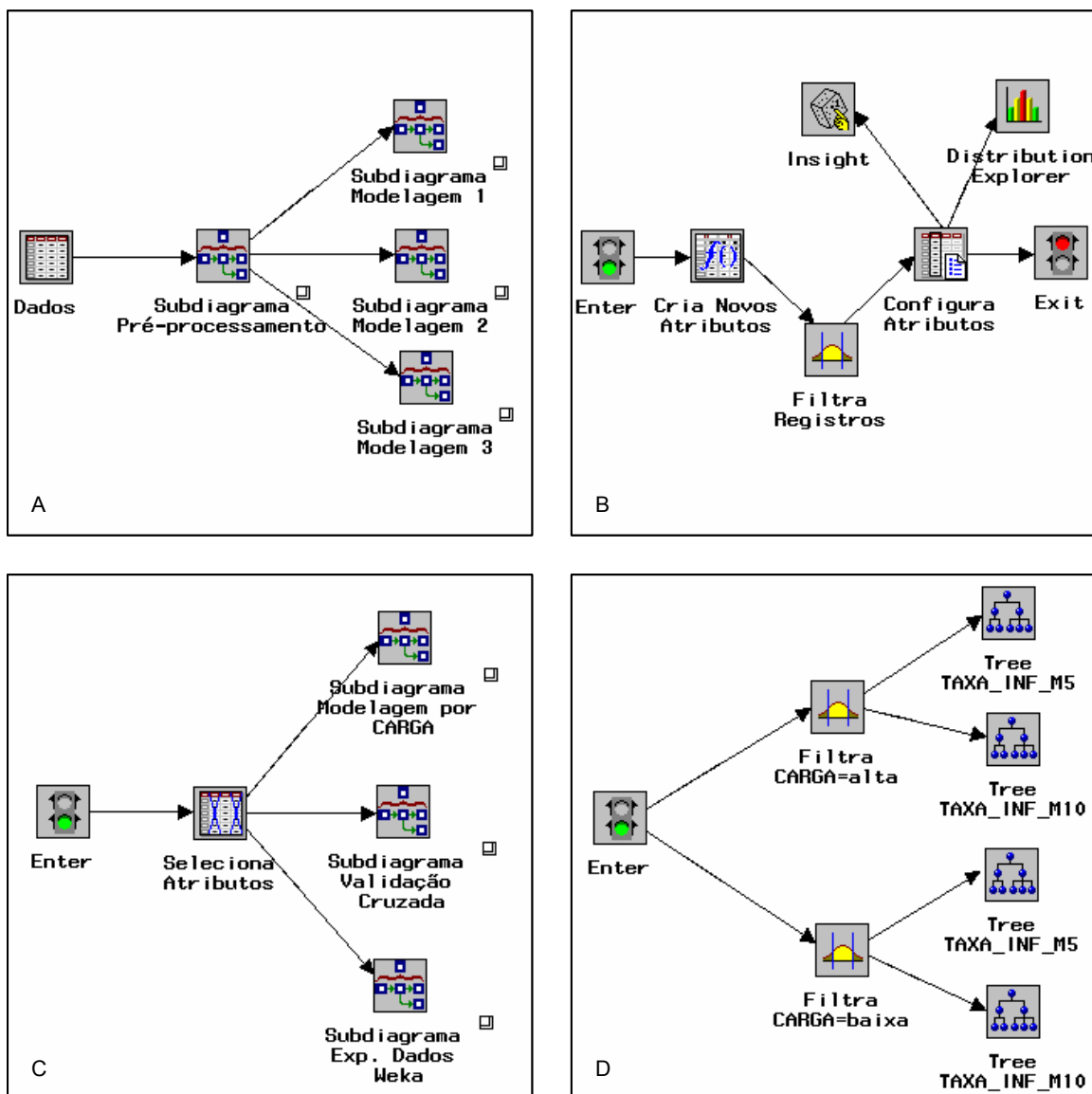


Figura 16: Diagrama do projeto no SAS® Enterprise Miner™.
 A. Diagrama geral; B. Subdiagrama de pré-processamento;
 C. Subdiagramas de modelagem; D. Subdiagramas de modelagem por CARGA.

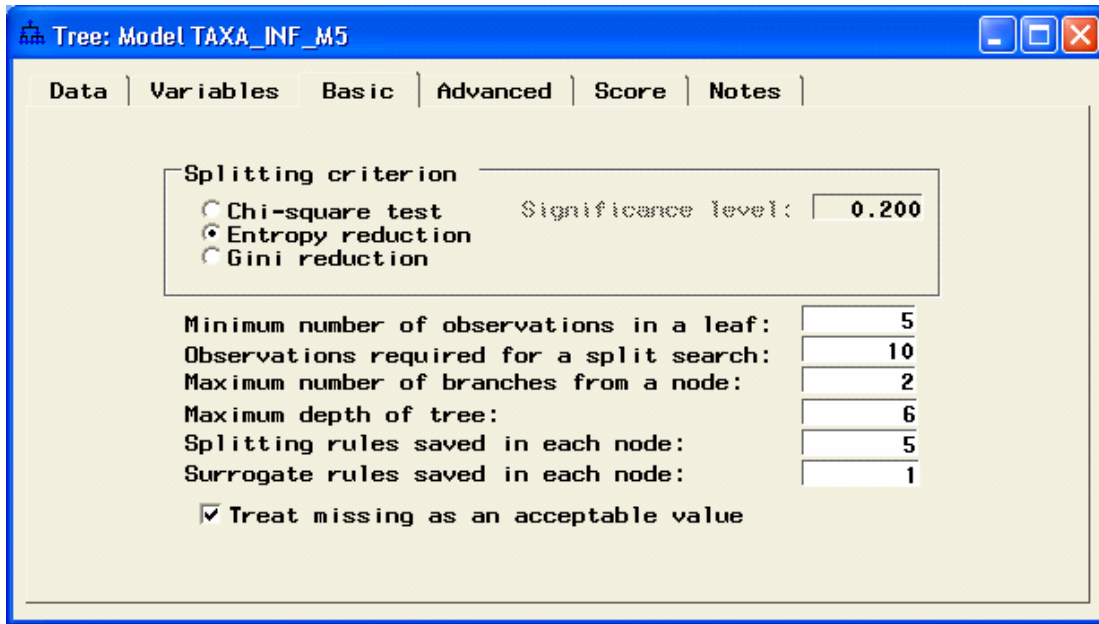


Figura 17: Configuração básica usada no *Tree node* do Enterprise Miner™.

O Enterprise Miner™ não possui uma ferramenta própria (*node*) para realizar validação cruzada. Foi necessário implementá-la com programas escritos na linguagem de programação do SAS®. Esses programas foram incluídos nos diagramas por intermédio de *nodes* do tipo *SAS Code*, que compuseram os subdiagramas de validação cruzada dentro de cada subdiagrama de modelagem (Figura 16 - C).

A indução das árvores de decisão foi realizada com a ferramenta *Tree* (Figura 16 - D). A configuração usada no *Tree node* está apresentada na Figura 17 (configuração básica) e na Figura 18 (configuração avançada).

O ganho de informação (seção 2.1.2.3), ou redução de entropia, foi o critério utilizado para escolher o atributo preditivo que dividiu o conjunto de exemplos em cada repetição do processo de indução da árvore de decisão (*'Splitting criterion'* na Figura 17).

A árvore de decisão foi escolhida para ser binária, com dois ramos a partir de cada nó interno (*'Maximum number of branches from a node'* na Figura 17).

Para evitar que o modelo ficasse muito específico para o conjunto de treinamento (*overfitting*), o que comprometeria a sua generalização e o desempenho com novos exemplos, foram adotadas duas regras de parada do algoritmo de indução. A primeira regra limitou a profundidade da árvore, permitindo-a ter no máximo seis níveis (*'Maximum depth of tree'* na Figura 17). A segunda regra limitou a fragmentação do conjunto de treinamento, requerendo um mínimo de dez exemplos em cada nó para a busca de uma nova divisão (*'Observations*

required for a split search' na Figura 17) e pelo menos cinco exemplos em cada nó folha (*Minimum number of observations in a leaf*' na Figura 17).

Foram salvas cinco regras de divisão em cada nó interno durante a indução (*Splitting rules saved in each node*' na Figura 17). Isso foi interessante para investigar e comparar regras e atributos concorrentes com a regra e o atributo escolhidos em cada nó.

Também, foi permitido que regras equivalentes fossem incluídas no modelo gerado, uma para cada nó (*Surrogate rules saved in each node*' na Figura 17), baseado na concordância com a regra principal do nó. A concordância é medida pela proporção de exemplos que a regra equivalente e a regra principal atribuem ao mesmo ramo da árvore. Uma regra desse tipo é interessante quando um exemplo sendo classificado possui valor nulo para o atributo escolhido na regra principal.

Além das regras de parada, denominadas de pré-poda, foi realizado um procedimento de pós-poda, após a indução da árvore de decisão completa. Junto com essa árvore completa, foram avaliadas todas as suas possíveis subárvores e escolhida a menor subárvore com o melhor valor de avaliação (*Sub-tree*' na Figura 18), no caso a menor taxa de erro (*Model assessment measure*' na Figura 18) sobre o conjunto de treinamento.

As opções de configuração restantes na Figura 18 não tiveram efeito na modelagem. São opções que se aplicam quando o conjunto de treinamento é muito grande, o que é comum em mineração de dados, mas não foi o caso neste projeto.

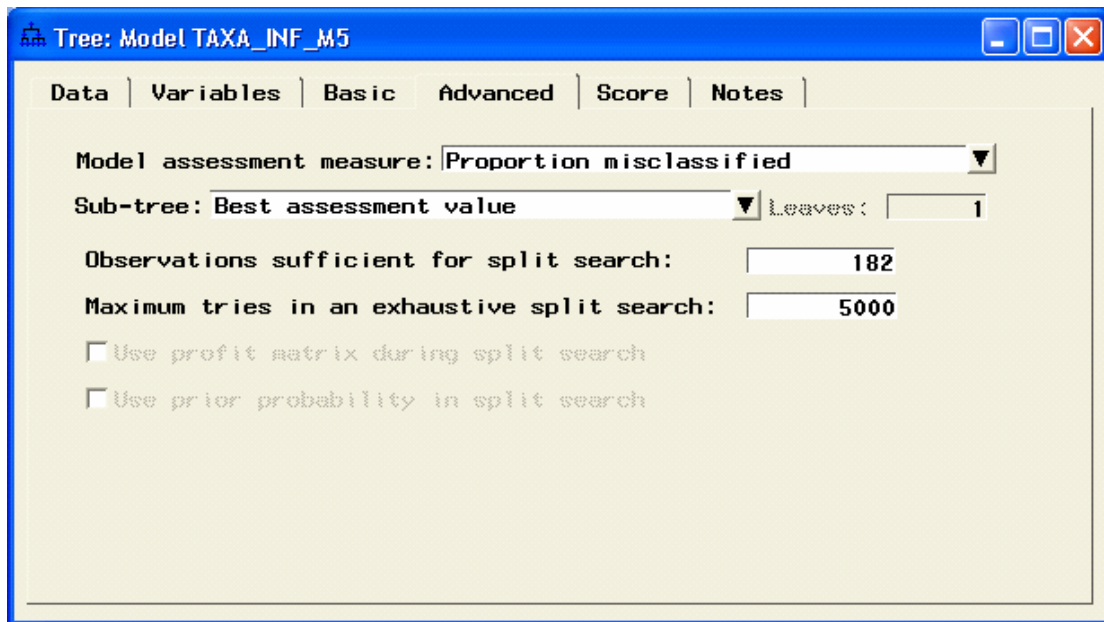


Figura 18: Configuração avançada usada no *Tree node* do Enterprise Miner™.

Depois de gerados, os modelos foram também visualizados e analisados com o auxílio da ferramenta SAS[®] Enterprise Miner[™] Tree Desktop Application (versão 9.1.32). As figuras das árvores de decisão apresentadas nos capítulos seguintes, referentes aos resultados obtidos com o Enterprise Miner[™], foram produzidas com o Tree Desktop Application.

Uso e configuração do Weka

O algoritmo J4.8 (WITTEN e FRANK, 2005) foi utilizado no Weka para a geração dos modelos. Ele é a implementação própria do Weka do indutor de árvores de decisão C4.5 (QUINLAN, 1993). Mais precisamente, J4.8 implementa uma versão posterior e ligeiramente melhorada, chamada C4.5 revisão 8, que foi a última versão de domínio público lançada antes da versão comercial C5.0.

A razão de ganho (seção 2.1.2.3) é o critério usado pelo algoritmo J4.8 na escolha do atributo preditivo que divide o conjunto de exemplos em cada repetição do processo de indução de uma árvore de decisão. As árvores geradas são binárias.

O classificador J48, que identifica o algoritmo J4.8 na interface do Weka (Figura 19), foi também configurado para deixar pelo menos cinco exemplos em cada nó folha (opção -M na linha de comando do classificador J48 exibida na Figura 19, equivalente ao parâmetro ‘minNumObj’ na caixa de diálogo de configuração do classificador).

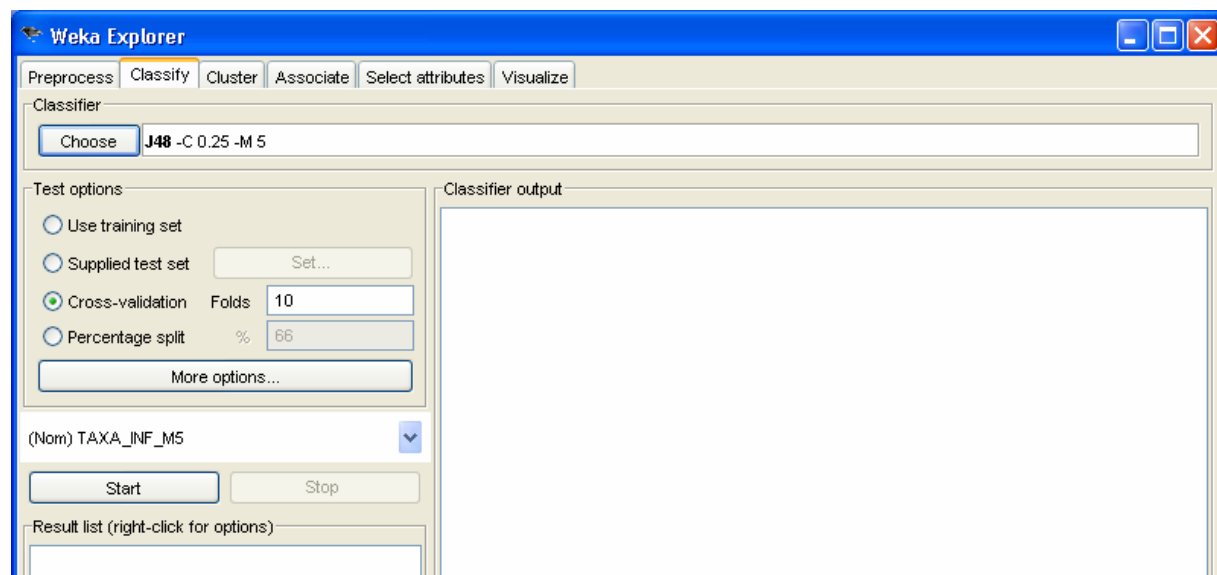


Figura 19: Configuração do classificador J48 e da validação cruzada no Weka.

O fator de confiança (opção -C na linha de comando exibida na Figura 19, equivalente ao parâmetro ‘confidenceFactor’), usado no mecanismo de poda do algoritmo, foi

mantido em 25%, que é o valor padrão (*default*) de configuração e funciona bem na maioria das situações (WITTEN e FRANK, 2005), como foi o caso neste projeto.

Os demais parâmetros de configuração do classificador tiveram seu valor padrão mantido. Dentre eles, destaca-se a ativação da operação de poda *subtree raising* (parâmetro ‘subtreeRaising’ na caixa de diálogo de configuração do classificador; seção 2.1.2.3).

O exemplo da Figura 19 mostra ainda a opção de validação cruzada selecionada (*10-fold cross-validation*) e o atributo meta escolhido, no caso TAXA_INF_M5.

3.6 Especialização do modelo do processo

Como última atividade, foi feita uma caracterização do processo de descoberta de conhecimento em bases de dados aplicado na obtenção dos modelos de alerta da ferrugem do cafeeiro, tal que permitisse sua reprodução e adaptação em problemas similares do domínio de aplicação.

A metodologia CRISP-DM (CHAPMAN et al., 2000) é descrita em termos de um modelo de processo hierárquico em quatro níveis de abstração (do geral para o específico): fases, tarefas genéricas, tarefas especializadas e instâncias de processo (Figura 20).

No nível mais alto, o processo de mineração de dados é organizado em seis fases (Figura 7, pg. 46); cada fase consiste de algumas tarefas genéricas do segundo nível. Este segundo nível é chamado genérico, pois a intenção é ser suficientemente geral para cobrir todas as possíveis situações de mineração de dados.

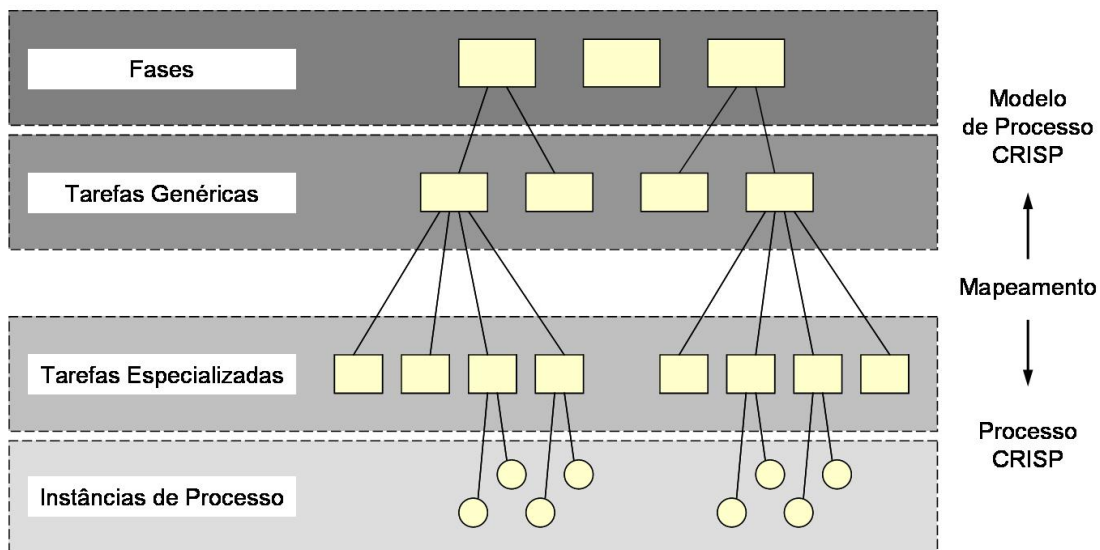


Figura 20: Visão hierárquica da metodologia CRISP-DM (adaptada de CHAPMAN et al., 2000).

O terceiro nível, das tarefas especializadas, é o local para descrever como as ações nas tarefas genéricas devem ser conduzidas em certas situações específicas. O quarto nível, da instância do processo, é o registro das ações, decisões e resultados de uma iniciativa real de mineração de dados. Uma instância do processo é organizada de acordo com as tarefas definidas nos níveis superiores, mas representa o que realmente aconteceu em um projeto posto em prática, em vez do que acontece em geral.

O mapeamento entre os níveis genérico e especializado é direcionado por um contexto de mineração de dados. Quatro diferentes dimensões podem estar discriminadas em um contexto de mineração de dados. Um contexto específico é um valor concreto para uma ou mais destas dimensões:

- **Domínio de aplicação:** é a área específica em que o projeto de mineração de dados se enquadra.
- **Tipo de problema:** indica a classe de objetivos específicos que o projeto de mineração de dados trata. Por exemplo: classificação, regressão, associação etc.
- **Aspecto técnico:** cobre questões específicas que representam diferentes desafios técnicos que geralmente ocorrem durante a mineração de dados. Por exemplo: valores ausentes, *outliers* etc.
- **Ferramenta e técnica:** especifica as ferramentas e/ou técnicas usadas no projeto de mineração de dados.

Um modelo de processo especializado, para uso futuro em contextos parecidos, pode ser obtido quando se especializa sistematicamente o modelo de processo genérico segundo um contexto de mineração de dados pré-definido ou, de maneira similar, quando se analisa e se consolida sistematicamente experiências de um projeto de mineração de dados em particular.

A caracterização ou especialização do processo realizado de descoberta de conhecimento em bases de dados, objeto deste documento, foi feita da segunda maneira, com base na instância do processo para a obtenção dos modelos de alerta da ferrugem do cafeeiro, ou seja, baseada numa análise sistemática das características do projeto e do registro das ações, decisões, resultados e dificuldades durante a sua execução.

A estratégia indicada na metodologia CRISP-DM para o mapeamento do modelo de processo, do nível genérico para o nível especializado, envolve (CHAPMAN et al., 2000):

- Analisar o contexto específico.

- Eliminar qualquer detalhe não aplicável ao contexto.
- Adicionar qualquer detalhe específico do contexto.
- Especializar os conteúdos genéricos de acordo com as características do contexto.
- Renomear conteúdos genéricos, se for o caso, provendo o modelo de significados mais explícitos e proporcionando maior clareza.

4 ANÁLISE DA EPIDEMIA DA FERRUGEM DO CAFEEIRO COM ÁRVORE DE DECISÃO

4.1 Considerações iniciais

A epidemiologia da ferrugem do cafeeiro já foi tema de diversos trabalhos (seção 2.2.2). A maioria desses estudos utilizou a regressão múltipla para ajustar os dados. Estudos mais recentes procuraram empregar outras técnicas, como a análise de trilha e as redes neurais.

No transcorrer deste trabalho, vislumbrou-se a possibilidade de utilizar uma árvore de decisão para analisar a epidemia da ferrugem do cafeeiro. As árvores de decisão, por utilizarem representações simbólicas e interpretáveis, permitem a compreensão das fronteiras de decisão que existem nos dados e da lógica implícita neles.

Redes neurais, por exemplo, embora possam ter alta precisão, são relativamente difíceis de compreender quando comparadas com as árvores de decisão. Multicolinearidade entre as variáveis independentes não afeta o desempenho das árvores de decisão, diferentemente das técnicas de regressão. Além disso, diversas variáveis, numéricas ou categóricas, podem ser analisadas ao mesmo tempo, sendo que o próprio algoritmo de indução se encarrega de selecionar as de maior importância.

O propósito básico da indução de uma árvore de decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema. No último caso, a intenção é compreender quais variáveis e interações dessas variáveis conduzem o fenômeno estudado.

Considerando tudo isso, resolveu-se aplicar e avaliar o potencial da técnica de indução de árvores de decisão na epidemiologia da ferrugem do cafeeiro. O intuito foi obter uma árvore de decisão capaz de auxiliar na compreensão de como as condições do ambiente, a carga pendente de frutos e o espaçamento entre as plantas na lavoura condicionaram a taxa de infecção da doença, identificando, dentre estes, os fatores mais importantes no progresso da ferrugem no campo.

4.2 Resultados

O início da epidemia da ferrugem do cafeeiro, na média de todos os anos, foi no mês de dezembro e atingiu o pico no mês de junho, independente da combinação de espaçamento e de carga pendente de frutos da lavoura (Figura 8, pg. 52).

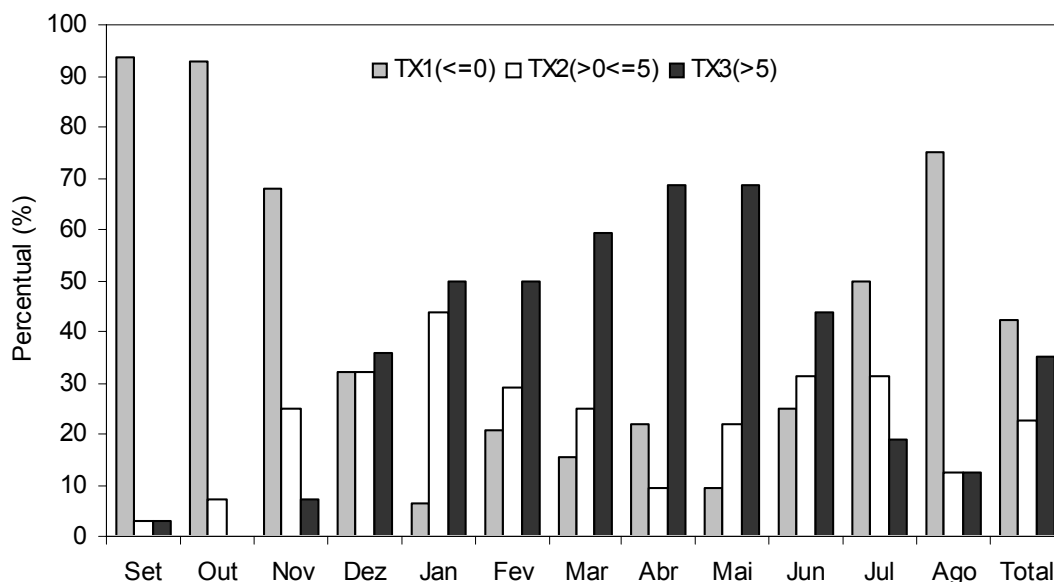


Figura 21: Distribuição percentual das três classes de taxa de infecção da ferrugem do cafeeiro, para cada mês e para o total dos meses.

A partir de dezembro, as taxas de infecção atingiram níveis mais elevados, com o percentual de distribuição da classe de taxa de infecção ‘TX3(>5)’, para taxas de infecção maiores que 5 p.p., ultrapassando o das classes de menor nível (Figura 21). O percentual de distribuição das três classes de taxa de infecção da ferrugem do cafeeiro, no conjunto de todos os meses do período analisado, foi 42% (154 exemplos) da classe ‘TX1(≤ 0)’, 23% (82 exemplos) da classe ‘TX2($> 0 \leq 5$)’ e 35% (128 exemplos) da classe ‘TX3(>5)’ (Figura 21).

A árvore de decisão que auxilia na compreensão das epidemias da ferrugem do cafeeiro é apresentada na Figura 22. As informações em cada nó da árvore representam, de cima para baixo, o número identificador do nó, a classe de taxa de infecção predominante e a distribuição percentual de cada classe no nó, na ordem ‘TX1(≤ 0)’, ‘TX2($> 0 \leq 5$)’ e ‘TX3(>5)’. Os nós estão coloridos em tons de cinza com base na proporção de exemplos da classe ‘TX3(>5)’ – quanto mais escuro, maior a proporção. Os números identificadores ausentes na seqüência numérica de 1 até 51 foram os nós eliminados na poda após a indução da árvore (pós-poda).

Para se compreender o relacionamento entre as variáveis explicativas e a taxa de infecção da ferrugem do cafeeiro, deve-se partir do topo da árvore de decisão (nó raiz) e descer pelos seus ramos, de acordo com os testes nas variáveis explicativas, até se chegar nos nós folhas.

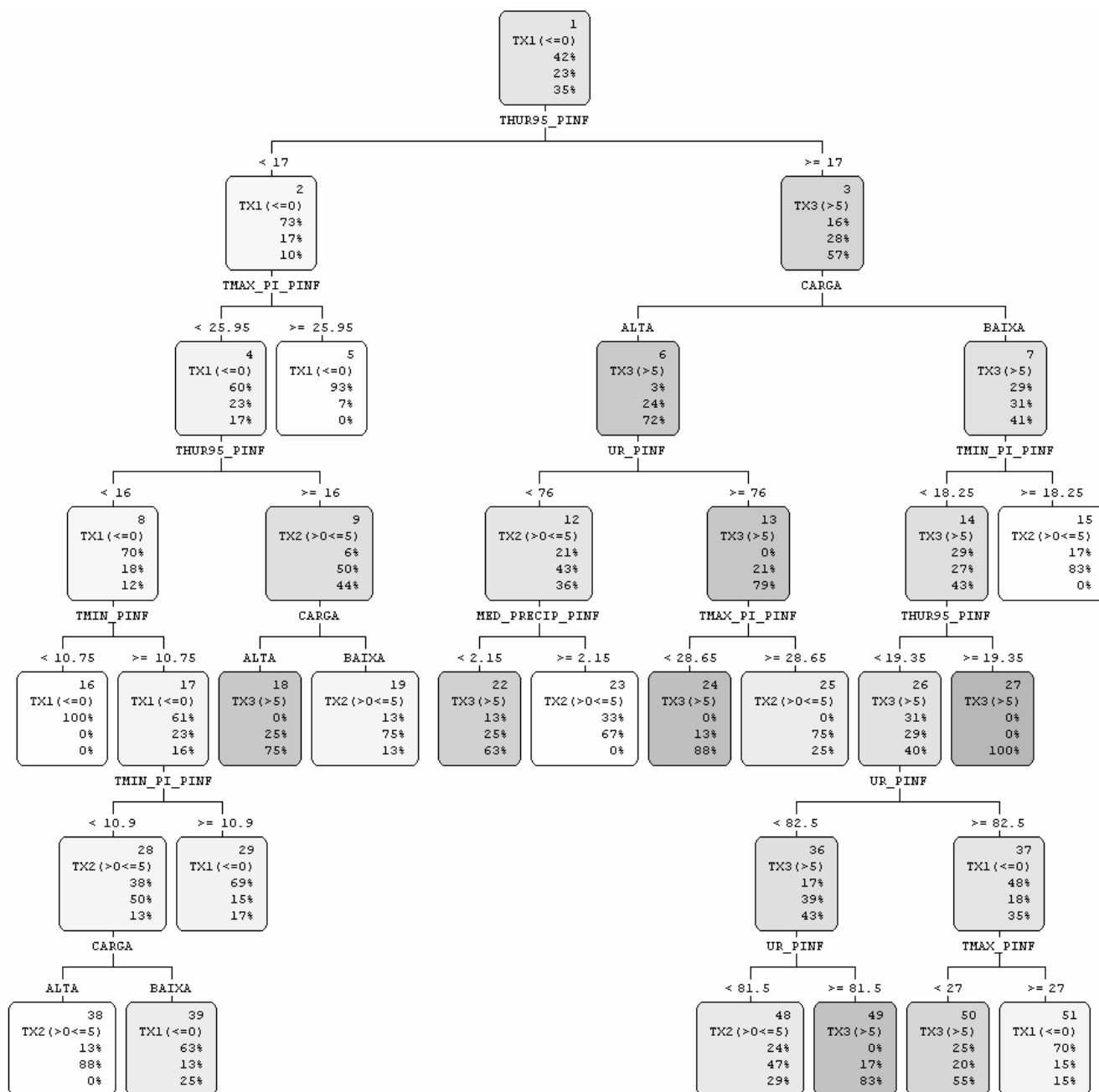


Figura 22: Árvore de decisão que auxilia na compreensão das epidemias da ferrugem do cafeeiro.

A primeira variável usada na decisão da classe da taxa de infecção (Figura 22, nó 1) foi a temperatura média nos períodos de alta umidade relativa do ar (THUR95_PINF). Temperaturas inferiores a 17 °C produziram taxas de infecção negativas ou nulas na maioria dos casos (73%), enquanto temperaturas maiores ou iguais a 17 °C resultaram em taxas de infecção positivas na maior parte das vezes (28% de ‘TX2(>0<=5)’ e 57% de ‘TX3(>5)’).

O próximo teste, descendo pelo ramo à esquerda do nó raiz (Figura 22, nó 2), foi escolhido sobre a média das temperaturas máximas diárias no período de incubação (TMAX_PI_PINF). Temperaturas maiores ou iguais a 25,95 °C resultaram em taxas de infecção negativas ou nulas (93% dos casos).

Neste ponto chega-se a um nó folha (Figura 22, nó 5) e, caso se estivesse classificando novos exemplos, para os quais não se conhece o valor da variável dependente, a árvore de decisão indicaria a taxa de infecção provável desses exemplos como da classe 'TX1(<=0)'.

O caminho de decisão entre o nó raiz e o nó folha pode ser traduzido para uma regra na forma 'SE <condição> ENTÃO <decisão>'. Assim, o caminho até o nó 5 se traduz na regra 'SE (THUR95_PINF < 17) e (TMAX_PI_PINF >= 25,95) ENTÃO TAXA_INF3N = TX1(<=0)'.

Em relação ainda a THUR95_PINF, a árvore de decisão indica que temperaturas abaixo de 16 °C foram desfavoráveis à infecção (Figura 22, nó 8) e que temperaturas maiores ou iguais a 19,35 °C foram bastante favoráveis à infecção (Figura 22, nó 27).

Segundo as decisões com base na carga pendente de frutos (Figura 22, nós 3, 9 e 28), taxas de infecção em níveis mais elevados ocorreram em cafeeiros com alta carga pendente, em comparação com os de baixa carga pendente.

Médias elevadas de temperatura máxima diária no período de incubação (TMAX_PI_PINF) tiveram efeito depressivo sobre a taxa de infecção (Figura 22, nó 25), semelhante ao efeito encontrado no nó 5, apesar da diferença dos limiares de decisão e das proporções de cada classe nos dois nós. Médias altas de temperatura mínima diária no período de incubação (TMIN_PI_PINF) também tiveram efeito depressivo sobre a taxa de infecção (Figura 22, nó 15). O teste sobre TMIN_PI_PINF no nó 17 parece não ter estabelecido um limite entre uma condição favorável e outra menos favorável, sendo que a decisão final da classe da taxa de infecção dependeu ainda de outras variáveis.

Valores médios diários mais elevados de umidade relativa do ar (UR_PINF) corresponderam a níveis mais elevados da taxa de infecção (Figura 22, nós 13 e 49). A decisão com base em UR_PINF no nó 26 parece também não ter dividido entre uma condição favorável e outra desfavorável.

Médias baixas de temperatura mínima diária (TMIN_PINF) exerceram influência negativa nas taxas de infecção (Figura 22, nó 16). Valores médios elevados de temperatura máxima diária (TMAX_PINF) também exerceram influência negativa nas taxas de infecção (Figura 22, nó 51).

A precipitação pluvial média diária (MED_PRECIP_PINF) foi escolhida no teste para o nó 12 (Figura 22), com efeito negativo nas taxas de infecção para valores maiores ou iguais a 2,15 mm (Figura 22, nó 23). A precipitação acumulada no período de infecção (PRECIP_PINF) poderia ter sido escolhida no teste para o mesmo nó 12. Nesse caso, o limiar de decisão, em vez de 2,15 mm para MED_PRECIP_PINF, seria 73 mm para PRECIP_PINF.

A árvore de decisão apresentou acurácia de 78% sobre o conjunto de treinamento e a acurácia obtida utilizando validação cruzada (*10-fold cross-validation*) foi de 73% (Tabela 13). Em relação aos acertos para cada classe de taxa de infecção, 88% (135 exemplos) da classe ‘TX1(≤ 0)’, 57% (47 exemplos) da classe ‘TX2($> 0 \leq 5$)’ e 79% (101 exemplos) da classe ‘TX3(> 5)’ foram corretamente classificados (Tabela 14). Quanto aos erros, por exemplo, 20% (16 exemplos) da classe ‘TX2($> 0 \leq 5$)’ foram classificados como da classe ‘TX1(≤ 0)’ e 23% (19 exemplos) classificados como da classe ‘TX3(> 5)’ (Tabela 14).

Tabela 13: Avaliação da árvore de decisão da Figura 22.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	78%	73%
Taxa de erro	22%	27%

Tabela 14: Matriz de confusão da árvore de decisão da Figura 22.

TAXA_INF3N		Preditá			
		TX1(≤ 0)	TX2($> 0 \leq 5$)	TX3(> 5)	TOTAL
Verdadeira	TX1(≤ 0)	135	13	6	154
	TX2($> 0 \leq 5$)	16	47	19	82
	TX3(> 5)	13	14	101	128
	TOTAL	164	74	126	364

4.3 Discussão

A importância da temperatura durante o período de molhamento foliar no progresso da ferrugem do cafeeiro é reconhecida na literatura (KUSHALAPPA, 1989a; MORAES, 1983; ZAMBOLIM et al., 1997; ZAMBOLIM et al., 2002). Enquanto a superfície da folha está molhada, a temperatura é o fator principal que determina o percentual de germinação dos esporos e de penetração do agente etiológico (KUSHALAPPA et al., 1983).

Na árvore de decisão gerada, a temperatura durante o molhamento foliar, medida indiretamente pela temperatura média nos períodos de alta umidade relativa do ar (THUR95_PINF), foi a variável mais importante na determinação da classe de taxa de infecção da ferrugem. Foi escolhida para o primeiro teste no nó raiz e para outros dois testes nos níveis intermediários da árvore de decisão.

THUR95_PINF inferiores a 17 e 16 °C cada vez mais desfavoráveis à infecção e superiores a 19 °C bastante favoráveis ficaram de acordo com os resultados obtidos por Montoya e Chaves (1974) e por Kushalappa et al. (1983). Os primeiros autores indicaram que o ponto mínimo de germinação seria encontrado em temperaturas inferiores a 18 °C e o ponto máximo na temperatura estimada de 23,7 °C. Kushalappa et al. (1983) consideraram 14 °C como limite mínimo de atividade do patógeno.

A árvore de decisão não identificou efeito negativo de temperaturas acima da ótima no poder germinativo de *H. vastatrix*, como observaram os autores citados. A razão disso pode ser atribuída ao valor máximo de THUR95_PINF (20,3 °C) ter ficado abaixo da temperatura ótima de germinação em todo o período analisado.

Os testes na árvore de decisão baseados na carga pendente de frutos confirmaram a predisposição das plantas à infecção de *H. vastatrix* devido à alta produção (KUSHALAPPA, 1989a). Segundo Zambolim et al. (2002), quanto maior a produção, maiores a incidência e a severidade da ferrugem. As decisões na árvore refletiram as diferenças nos níveis de taxa de infecção observadas entre as lavouras com carga pendente alta e com carga pendente baixa (Figura 8, pg. 52).

O espaçamento entre as plantas é considerado um fator de interferência no progresso da ferrugem do cafeeiro, provavelmente influenciando as condições microclimáticas dentro da lavoura (KUSHALAPPA, 1989a). Entretanto, o espaçamento nas lavouras de café não foi significativo na determinação da classe da taxa de infecção da ferrugem. A variável

LAVOURA não ter aparecido em nenhum teste na árvore de decisão refletiu também o comportamento do progresso da doença (Figura 8, pg. 52), que não exibiu nenhuma distinção evidente devido ao espaçamento.

Temperaturas elevadas no período de incubação exerceram efeito negativo nas taxas de infecção da ferrugem do cafeeiro, efeito esse observado pelos testes em TMAX_PI_PINF e TMIN_PI_PINF na árvore de decisão. Moraes et al. (1976) observaram que temperaturas médias máximas microclimáticas acima de 31 °C ocasionaram efeito depressivo sobre o desenvolvimento de *H. vastatrix*, fazendo com que o período de incubação aumentasse. Essas temperaturas corresponderam a temperaturas médias máximas macroclimáticas, obtidas em posto meteorológico, por volta de 28 °C, bem próximo dos 28,65 °C estabelecidos pela árvore de decisão no teste sobre TMAX_PI_PINF (Figura 22, nó 13).

Quando THUR95_PINF não foi favorável à infecção (menor que 17 °C), o limiar que determinou o efeito inibidor da temperatura no período de incubação mostrou-se menor: TMAX_PI_PINF \geq 25,95 °C (Figura 22, nó 5). Isto parece indicar que germinações ocorridas em condições menos favoráveis são mais sensíveis ao efeito da temperatura no período de incubação.

Montoya e Chaves (1974) observaram que temperaturas menos favoráveis à germinação (18 e 26 °C) prolongaram o período de incubação (referido como período de geração) e diminuíram o nível de infecção da ferrugem, enquanto temperaturas mais favoráveis à germinação (20, 22 e 24 °C) proporcionaram períodos de incubação mais curtos e maior número de ciclos de infecção. Segundo esses autores, as condições que afetaram o processo germinativo, além de terem determinado uma maior ou menor porcentagem de germinação, também exerceram influência na colonização do fungo no tecido vegetal.

Os resultados do presente trabalho indicam que, dependendo das condições de germinação, a colonização do fungo no tecido vegetal sofreu influência diferenciada das condições do ambiente, especificamente da temperatura, acrescentando-se à hipótese de Montoya e Chaves (1974).

A água da chuva é importante para a germinação dos esporos. A chuva, normalmente associada com o vento, é o principal agente de disseminação dos uredósporos (KUSHALAPPA, 1989a; ZAMBOLIM et al., 1997). As chuvas de baixa intensidade e o orvalho que umedecem as folhas durante várias horas são especialmente propícios

(MONTROYA e CHAVES, 1974). Por outro lado, chuvas fortes em um curto período de tempo podem conduzir a maior parte dos esporos para o chão (KUSHALAPPA, 1989a).

Esse comportamento irregular pode ter sido a razão das variáveis relacionadas com a precipitação (MED_PRECIP_PINF e PRECIP_PINF) não terem aparecido com grande importância na árvore de decisão. A umidade relativa média diária (UR_PINF) parece ter expressado melhor a importância das chuvas. As estações chuvosas estão frequentemente associadas com alta umidade relativa do ar (KUSHALAPPA, 1989a).

O único teste na árvore de decisão com base em MED_PRECIP_PINF (Figura 22, nó 12) pode ter expressado o efeito mencionado das fortes chuvas. Valores de MED_PRECIP_PINF maiores ou iguais a 2,15 mm (ou PRECIP_PINF \geq 73 mm) causaram efeito depressivo nas taxas de infecção da ferrugem do cafeeiro.

Médias baixas de temperatura mínima diária e médias altas de temperatura máxima diária (TMIN_PINF e TMAX_PINF, respectivamente) exerceram influência negativa nas taxas de infecção da ferrugem do cafeeiro. Isso mostrou que não só a temperatura durante o molhamento foliar foi importante; as máximas e mínimas de temperatura nos dias do período de infecção também exibiram importância, embora em grau bem menor.

No Brasil, é freqüente a presença de água livre na superfície das folhas do cafeeiro, mesmo no inverno, estação seca, devido principalmente ao orvalho; as baixas temperaturas enquanto a folha está molhada torna-se o fator limitante para a germinação e a penetração do fungo (KUSHALAPPA, 1989a). É o que parece ter sido capturado pela árvore de decisão. Os períodos de molhamento foliar prolongado (NHUR95_PINF e NHNUR95_PINF), presentes em praticamente todos os períodos de infecção, não serviram à árvore de decisão para identificar aqueles com maiores ou menores taxas de infecção. A temperatura média nos períodos de molhamento foliar (THUR95_PINF) é que foi escolhida como o fator determinante das classes de taxa de infecção.

Na avaliação geral, a árvore de decisão classificou corretamente 283 exemplos de um total de 364 do conjunto de treinamento, ou seja, 78% dos exemplos a partir dos quais a árvore de decisão foi gerada tiveram a classe de taxa de infecção predita igual à classe verdadeira.

Por classe de taxa de infecção, o desempenho da árvore para as classes 'TX1(\leq 0)' e 'TX3($>$ 5)' foi acima da média. Foram classificados corretamente 88% dos exemplos da classe 'TX1(\leq 0)' e 79% dos exemplos da classe 'TX3($>$ 5)'. O menor desempenho para a classe

'TX2(>0<=5)', com 57% dos exemplos classificados corretamente, talvez esteja relacionado com o menor número de exemplos desta classe no conjunto de treinamento (Figura 21), permitindo que as outras duas classes prevalecessem na distribuição final dos exemplos nos nós folhas.

O desempenho da árvore de decisão para classificação de outros exemplos, nos quais a árvore não foi treinada, foi estimado em 73% de acurácia. Esse valor de acurácia é uma estimativa de desempenho, caso se quisesse avaliar o potencial de uso da árvore de decisão como modelo de alerta da ferrugem do cafeeiro. Um eventual uso como modelo de alerta, entretanto, deveria estar restrito à região onde os dados analisados foram obtidos ou a regiões com características parecidas do ambiente estudado.

Cabe mencionar que a forma com que foi calculada a acurácia na validação cruzada incorporou um viés (*bias*) otimista. No processo de indução das árvores de decisão, em cada iteração da validação cruzada, foi mantida a opção no Enterprise Miner™ de escolher a menor subárvore com a menor taxa de erro (seção 3.5.3). Nas circunstâncias em que o processo foi realizado, porém, a melhor subárvore foi considerada com base nos conjuntos de teste da validação cruzada. Portanto, o processo de indução não foi totalmente independente dos conjuntos de teste, o que resultou no *bias* otimista.

No caso da árvore de decisão objeto deste capítulo, esse *bias* otimista não foi eliminado, uma vez que o objetivo principal não foi obter um modelo de alerta da ferrugem do cafeeiro, para o qual seria necessário uma estimativa de acurácia mais confiável. Além disso, o presente capítulo foi tema de um artigo científico (MEIRA et al., 2008), que foi submetido antes da descoberta do *bias* otimista mencionado. Sendo assim, mais um motivo para manter os resultados como foram apresentados.

Na avaliação dos modelos de alerta da ferrugem do cafeeiro, que são o tema do próximo capítulo, o procedimento de validação cruzada foi modificado e eliminou-se o *bias* otimista. Sendo assim, o processo de indução dos modelos de alerta foi totalmente independente dos conjuntos de teste da validação cruzada e, portanto, as medidas de avaliação apresentadas no próximo capítulo são bem mais confiáveis.

4.4 Considerações finais

A técnica de indução de árvores de decisão se mostrou uma ferramenta adequada para o estudo das epidemias da ferrugem do cafeeiro. A árvore de decisão apresentada demonstrou todo o seu potencial como modelo de representação simbólica e interpretável.

Permitiu a identificação das fronteiras de decisão presentes nos dados e da lógica contida neles, auxiliando na compreensão de quais variáveis explicativas, e de como as interações dessas variáveis, conduziram a taxa de progresso da ferrugem no campo. As variáveis explicativas mais importantes foram a temperatura média nos períodos de molhamento foliar, a carga pendente de frutos, a média das temperaturas máximas diárias no período de incubação e a umidade relativa do ar.

A indução de árvores de decisão, então, se apresenta como uma boa opção metodológica para a epidemiologia de doenças de plantas, em geral. As suas principais vantagens são:

- A facilidade de interpretação e compreensão dos padrões. A representação de uma árvore de decisão dá a impressão do entendimento das causas do comportamento observado da variável dependente. Modelos em árvore de decisão não provam causalidade, mas ajudam a explicar como determinaram a probabilidade estimada, em casos de classificação (REFAAT, 2006).
- O próprio algoritmo de indução descobre as fronteiras de decisão a partir dos dados. Nos experimentos tradicionais, geralmente, são estabelecidos alguns poucos valores para a pesquisa (p.ex. valores ou faixas de valores determinados de temperatura) e as conclusões são feitas com base nos resultados obtidos com esses valores.
- A possibilidade de construção da árvore de decisão interativamente. A indução interativa proporciona flexibilidade em analisar e escolher as variáveis explicativas em cada nó de decisão e liberdade para atribuir à árvore de decisão a configuração, ou estruturação, final que melhor atenda aos objetivos desejados.
- As árvores de decisão não impõem restrições ou requisitos especiais quanto aos procedimentos de preparação dos dados. Elas podem lidar com todos os tipos de variáveis e, até mesmo, com valores ausentes (REFAAT, 2006).

5 MODELOS DE ALERTA DA FERRUGEM DO CAFEIEIRO

5.1 Considerações iniciais

Este capítulo trata dos modelos de alerta da ferrugem do cafeeiro. Os alertas são considerados quando a taxa de infecção da ferrugem for esperada atingir ou ultrapassar os limites de 5 p.p. e 10 p.p. no prazo de um mês.

Este tipo de alerta pode ser útil na tomada de decisão referente à melhor época e à melhor maneira de se adotar medidas de controle da doença. Alertas emitidos corretamente (verdadeiros positivos) e situações acertadas em que os alertas não são emitidos (verdadeiros negativos) permitem identificar, por exemplo, os momentos mais oportunos para a aplicação de fungicidas.

A estimativa da acurácia de cada modelo é importante, pois ela dá a noção da proporção de acertos que o modelo pode ter caso venha a ser aplicado no problema real.

A sensibilidade, a confiabilidade positiva, a especificidade e a confiabilidade negativa são também medidas importantes para os modelos de alerta da doença.

A sensibilidade estima a capacidade do modelo de acertar nas situações em que se deve emitir um alerta. Em outras palavras, responde à seguinte pergunta: Qual percentual de alertas requeridos o modelo é esperado acertar?

A confiabilidade positiva, por sua vez, estima a capacidade do modelo de emitir corretamente os seus alertas. É a resposta à seguinte pergunta: Qual percentual de alertas emitidos pelo modelo é de se esperar que esteja correto?

A especificidade e a confiabilidade negativa são correspondentes à sensibilidade e à confiabilidade positiva, respectivamente, para as situações em que o alerta não é necessário e/ou não é emitido.

A especificidade estima a capacidade do modelo de acertar nas situações em que não se deve emitir um alerta; e a confiabilidade negativa estima a capacidade do modelo em acertar quando não emitir os alertas.

Sendo assim, além de procurar obter a melhor acurácia possível, boas avaliações em relação às outras quatro medidas, e um equilíbrio entre os seus valores, foram considerados para cada modelo de alerta da ferrugem do cafeeiro.

Também, como um modelo pode não ter boa acurácia, mas pode conter regras que, individualmente, sejam precisas e tenham outras características importantes, todas as regras componentes de cada modelo de alerta foram avaliadas. A precisão e outras medidas próprias de avaliação de regras foram utilizadas.

Além disso, procurou-se avaliar o desempenho das regras do mês de dezembro ao mês de abril, período crítico de evolução da ferrugem do cafeeiro e para o qual, normalmente, são recomendadas e realizadas as aplicações de fungicidas (ZAMBOLIM et al., 2002).

Na próxima seção, são apresentados e discutidos os modelos de alerta da ferrugem do cafeeiro para lavouras com alta carga pendente de frutos. Na seção 5.3, são apresentados e discutidos os modelos para lavouras com baixa carga pendente. Em seguida, na seção 5.4, são feitas as considerações finais deste capítulo.

5.2 Modelos para lavouras com alta carga pendente de frutos

Antes da apresentação dos modelos, é interessante observar as distribuições percentual e absoluta dos exemplos do conjunto de treinamento em relação às classes ‘1’ e ‘0’ dos atributos meta TAXA_INF_M5 e TAXA_INF_M10 (Tabela 15). Relembrando, para TAXA_INF_M5, ‘1’ significa que a taxa de infecção da ferrugem do cafeeiro foi maior ou igual a 5 p.p. e ‘0’ significa o contrário. Para TAXA_INF_M10, a interpretação é semelhante, considerando o limite de 10 p.p.

Tabela 15: Distribuição dos exemplos de lavouras com alta carga pendente entre as classes ‘1’ e ‘0’ dos atributos meta TAXA_INF_M5 e TAXA_INF_M10.

Atributo meta	Classe	
	1	0
TAXA_INF_M5	46% (84 exemplos)	54% (98 exemplos)
TAXA_INF_M10	29% (53 exemplos)	71% (129 exemplos)

5.2.1 Alerta quando a taxa de infecção for atingir ou ultrapassar 5 p.p.

5.2.1.1 Resultados

A Figura 23 apresenta a árvore de decisão, gerada no SAS[®] Enterprise Miner[™] (EM), para alertas da ferrugem do cafeeiro em lavouras com alta carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M5 (seção 3.4.1) e os atributos preditivos foram escolhidos conforme a opção de seleção de atributos Modelagem 1 (seção 3.5.1).

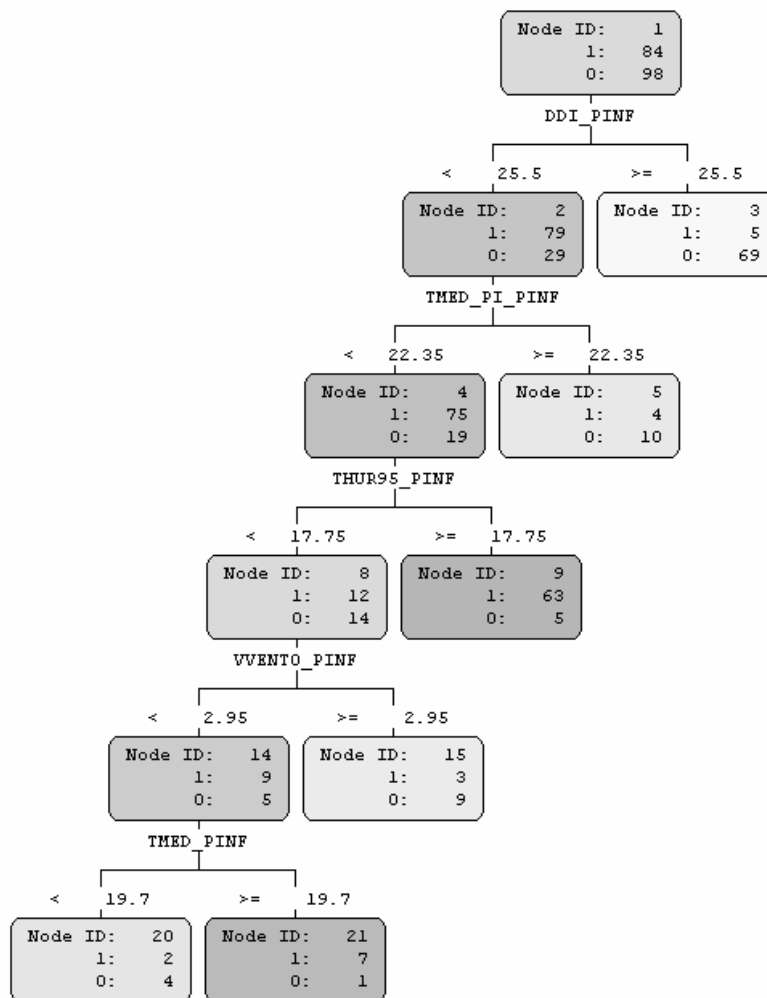


Figura 23: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 1; Geração: Enterprise Miner™.

Diferentemente da árvore de decisão apresentada na seção 4.2, os nós da Figura 23 indicam, além do seu número identificador (‘Node ID’), os valores absolutos de distribuição dos exemplos entre as classes ‘1’ e ‘0’, em vez da distribuição percentual. Desta forma, ajuda a dar a noção do número real de exemplos de cada classe, em cada nó. Além disso, evita uma confusão que poderia ocorrer entre os valores percentuais de distribuição nos nós folhas e a medida de precisão (razão de Laplace) adotada para as regras de classificação correspondentes.

Os nós da árvore de decisão são coloridos em tons de cinza com base na proporção de exemplos da classe ‘1’ – quanto mais escuro, maior a proporção. Os números identificadores ausentes na seqüência numérica de 1 até 21 foram os nós eliminados na poda após a indução

da árvore (pós-poda). A classe de predição em cada nó folha é a que apresenta o maior número de exemplos associado a ela.

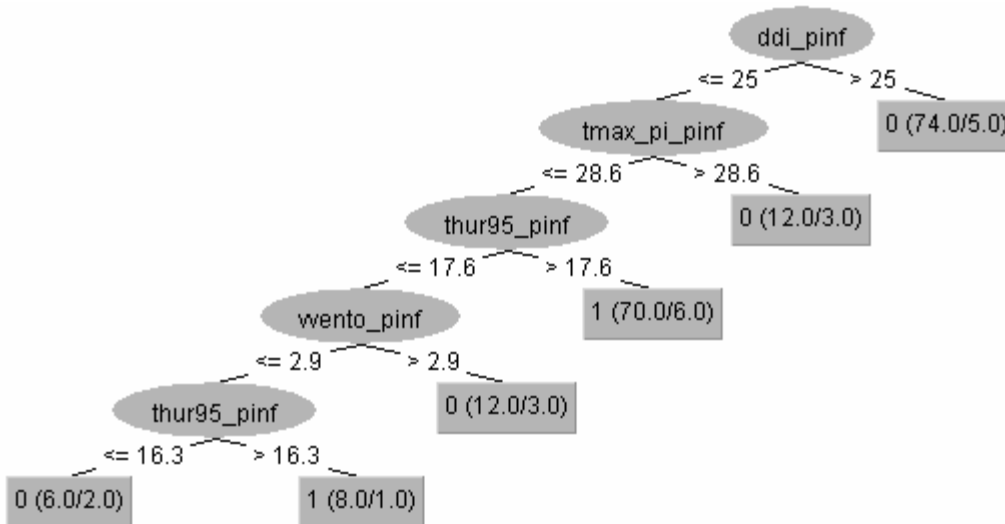


Figura 24: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 1; Geração: Weka.

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 24. A representação visual da árvore é diferente da do EM: os nós internos são denotados por elementos gráficos ovais, onde estão indicados os atributos de teste das regras; e os nós folhas são denotados por retângulos, onde há informação sobre a classe de predição e sobre a quantidade de exemplos classificados. Por exemplo, no nó folha à direita do nó raiz na Figura 24, ‘0 (74,0/5,0)’ significa que a classe ‘0’ foi atribuída aos exemplos classificados naquele nó e que 74 exemplos do conjunto de treinamento foram ali classificados, 5 dos quais de maneira incorreta – a classe verdadeira desses exemplos era ‘1’.

Os dois modelos possuem a mesma estrutura de árvore, com praticamente os mesmos atributos de teste e os mesmos valores para as fronteiras de decisão. Houve apenas trocas entre TMED_PI_PINF (EM – divisão do nó 2) e TMAX_PI_PINF (Weka) e entre TMED_PINF (EM – divisão do nó 14) e THUR95_PINF (Weka).

Com relação a essas trocas, verificou-se no EM que: a regra de divisão do nó 2 teve regra concorrente baseada em TMAX_PI_PINF (divisão em 28,65 °C) com quase o mesmo ganho de informação; e a regra de divisão do nó 14 teve regra concorrente baseada em THUR95_PINF (divisão em 16,3 °C) com ganho de informação igual.

Verificou-se também, pela análise da classificação dos exemplos do conjunto de treinamento por cada modelo, que a troca de TMED_PI_PINF por TMAX_PI_PINF significou

a troca de um par VN (verdadeiro negativo) / FN (falso negativo) por um par VP (verdadeiro positivo) / FP (falso positivo), respectivamente, referente a dois registros do mês de janeiro. Esta foi a razão da melhor confiabilidade positiva do modelo do EM (92,1% contra 91%) e da melhor sensibilidade do modelo do Weka (84,5% contra 83,3%) na ressubstituição. A troca de TMED_PINF por THUR95_PINF não alterou em nada a classificação entre os dois modelos.

A avaliação dos dois modelos foi parecida na validação cruzada, com pequena vantagem para o modelo do EM (acurácia = 82,4% contra 80,8%). Esta diferença esteve dentro do desvio padrão da média.

Optou-se pelo modelo gerado no Weka para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão, pela vantagem de acertar um alerta (VP) e de não errar outro que deveria ser dado (FN do modelo do EM) no mês de janeiro, mês crítico na evolução da ferrugem do cafeeiro.

Tabela 16: Matrizes de confusão da árvore de decisão da Figura 24.

Ressubstituição				Validação cruzada			
TAXA_INF_M5	Predita			TAXA_INF_M5	Predita		
	1	0			1	0	
Verdadeira	1	71	13	Verdadeira	1	66	18
	0	7	91		0	17	81

Tabela 17: Avaliação da árvore de decisão da Figura 24.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	89,0%	80,8% (2,1%)*
Taxa de erro	11,0%	19,2%
Sensibilidade	84,5%	78,5% (4,9%)
Especificidade	92,9%	82,6% (4,6%)
Confiabilidade positiva	91,0%	82,1% (3,8%)
Confiabilidade negativa	87,5%	83,8% (3,4%)

* Desvio padrão da média.

As matrizes de confusão da árvore de decisão da Figura 24, obtidas pelos métodos de ressubstituição e de validação cruzada, são apresentadas na Tabela 16. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 17.

As regras de classificação extraídas da árvore de decisão da Figura 24 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 18.

Tabela 18: Regras extraídas da árvore de decisão da Figura 24 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1	
SE DDI_PINF > 25 ENTÃO TAXA_INF_M5 = 0	Precisão: 92,1% Novidade: 0,16 Sensitividade: 70,4% Cobertura: 40,7% Especificidade: 94,0% Suporte: 37,9%
Distribuição*: JAN(2VN); JUL(11VN;1FN); AGO(14VN;2FN); SET(16VN); OUT(14VN); NOV(8VN); DEZ(4VN;2FN).	
Regra 2	
SE DDI_PINF ≤ 25 E TMAX_PI_PINF > 28,6 ENTÃO TAXA_INF_M5 = 0	Precisão: 71,4% Novidade: 0,01 Sensitividade: 9,2% Cobertura: 6,6% Especificidade: 96,4% Suporte: 4,9%
Distribuição: FEV(4VN;2FN); MAR(3VN;1FN); DEZ(2VN).	
Regra 3	
SE DDI_PINF ≤ 25 E TMAX_PI_PINF ≤ 28,6 E THUR95_PINF > 17,6 ENTÃO TAXA_INF_M5 = 1	Precisão: 90,3% Novidade: 0,17 Sensitividade: 76,2% Cobertura: 38,5% Especificidade: 93,9% Suporte: 35,2%
Distribuição: JAN(11VP;3FP); FEV(6VP); MAR(10VP;2FP); ABR(16VP); MAI(13VP;1FP); JUN(2VP); DEZ(6VP).	
Regra 4	
SE DDI_PINF ≤ 25 E TMAX_PI_PINF ≤ 28,6 E THUR95_PINF ≤ 17,6 E VVENTO_PINF > 2,9 ENTÃO TAXA_INF_M5 = 0	Precisão: 71,4% Novidade: 0,01 Sensitividade: 9,2% Cobertura: 6,6% Especificidade: 96,4% Suporte: 4,9%
Distribuição: JUN(2VN;2FN); JUL(2VN); NOV(5VN;1FN).	
Regra 5	
SE DDI_PINF ≤ 25 E TMAX_PI_PINF ≤ 28,6 E THUR95_PINF ≤ 16,3 E VVENTO_PINF ≤ 2,9 ENTÃO TAXA_INF_M5 = 0	Precisão: 62,5% Novidade: 0,00 Sensitividade: 4,1% Cobertura: 3,3% Especificidade: 97,6% Suporte: 2,2%
Distribuição: JUN(3VN;1FN); JUL(1VN;1FN).	
Regra 6	
SE DDI_PINF ≤ 25	Precisão: 80,0% Novidade: 0,02

E $TMAX_PI_PINF \leq 28,6$
E $16,3 < THUR95_PINF \leq 17,6$
E $VVENTO_PINF \leq 2,9$

Sensitividade: 8,3% **Cobertura:** 4,4%
Especificidade: 99,0% **Suporte:** 3,8%

ENTÃO $TAXA_INF_M5 = 1$

Distribuição: MAI(2VP); JUN(5VP;1FP).

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

Junto com cada regra, também é apresentada a distribuição mensal das predições do modelo em relação aos exemplos do conjunto de treinamento (Tabela 18): VP (verdadeiros positivos) foram os alertas preditos corretamente ($TAXA_INF_M5 = 1$ no conjunto de treinamento e predito certo); FP (falsos positivos) foram os alertas preditos incorretamente ($TAXA_INF_M5 = 0$ e predito errado); VN (verdadeiros negativos) foram os casos acertados em que não se predisse o alerta ($TAXA_INF_M5 = 0$ e predito correto); e FN (falsos negativos) foram os alertas que deveriam ter sido preditos, mas não o foram ($TAXA_INF_M5 = 1$ e predito errado).

As matrizes de confusão e a avaliação do modelo gerado no EM, bem como as regras de classificação extraídas da árvore de decisão da Figura 23 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

A Figura 25 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com alta carga pendente, em que o atributo meta foi a taxa de infecção binária $TAXA_INF_M5$ (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 2 (seção 3.5.1).

A regra equivalente (*surrogate rule*) à regra do nó raiz foi baseada no atributo $TMIN_PINF$: $TMIN_PINF < 14,2$ equivalendo a $THUR95_PINF < 16,5$ e $TMIN_PINF \geq 14,2$ equivalendo a $THUR95_PINF \geq 16,5$. Esta regra foi utilizada para classificar os dois exemplos de agosto de 2000 em que não houve período de molhamento foliar, com um mínimo de seis horas, no período de infecção correspondente ($NHUR95_PINF = 0$) e, portanto, os valores do atributo $THUR95_PINF$ foram nulos. Os dois exemplos foram classificados no nó 2 ($TMIN_PINF = 10,4$ °C) como da classe '0'.

A regra de divisão do nó 6, baseada originalmente no atributo MED_PRECIP_PINF (divisão em 1,75 mm), teve regra concorrente com mesmo ganho de informação baseada em $TMED_PI_PINF$ (divisão em 19,75 °C). Este último foi escolhido como o atributo de teste

para o modelo final (Figura 25), devido à melhor distribuição mensal dos exemplos classificados.

A regra de divisão original do nó 6 juntava exemplos de junho e de novembro de um lado e, do outro lado, juntava exemplos de maio com exemplos de novembro, dezembro e janeiro. O atributo de teste TMED_PI_PINF trouxe os exemplos de maio para junto dos de junho e levou os exemplos de novembro para o outro lado, junto com os outros exemplos de novembro. A troca do atributo de teste não alterou a distribuição numérica original dos exemplos entre as classes.

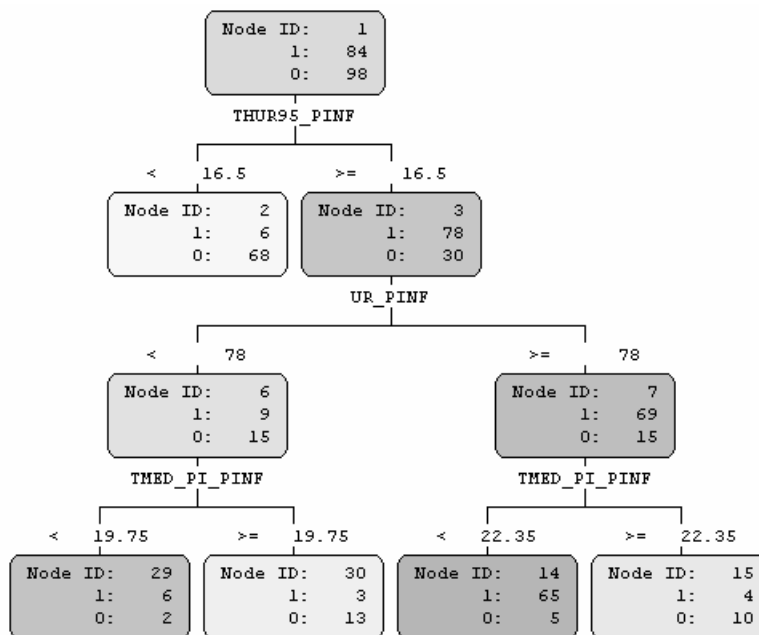


Figura 25: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 2; Geração: Enterprise Miner™.

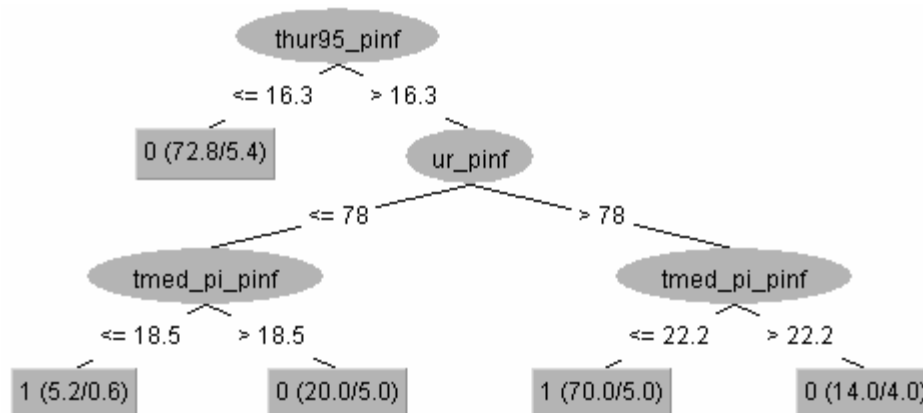


Figura 26: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 2; Geração: Weka.

Esta alteração na regra de divisão do nó 6 deixou a árvore de decisão gerada no EM muito parecida com a árvore de decisão correspondente gerada no Weka (Figura 26).

Analisando-se atentamente os dois modelos, percebe-se que houve uma diferença na distribuição dos exemplos do conjunto de treinamento classificados nos nós folhas. Essa diferença decorreu do tratamento diferenciado na classificação de exemplos com valor nulo para o atributo de teste de um nó.

No Weka, como não havia a opção de *surrogate rule*, os exemplos com valor nulo para THUR95_PINF foram fracionados, entre cada ramo do nó raiz, com base na proporção dos demais exemplos separados de acordo com a sua regra de divisão – os números decimais apresentados na Figura 26 (‘72,8/5,4’ e ‘5,2/0,6’) indicam esse fracionamento. Por isso, os dois exemplos de agosto de 2000 foram classificados como da classe ‘1’ (peso 0,4 para classificar como ‘0’ e peso 0,6 para classificar como ‘1’), diferente da classificação com o modelo do EM (classe ‘0’).

A distribuição desigual dos exemplos, a partir do nó raiz, ocasionou também uma pequena diferença na determinação da regra de divisão do nó cujo atributo de teste foi TMED_PI_PINF (divisão em 18,5 °C na Figura 26). Com isso, a distribuição dos exemplos ficou ainda mais diferenciada entre os modelos do EM e do Weka.

Essa diferença na distribuição dos exemplos acarretou diferença também na avaliação dos modelos com base na ressubstituição. A acurácia foi a mesma (89%), mas o modelo do EM teve melhor sensibilidade (84,5% contra 83,8%) e o modelo do Weka melhor confiabilidade positiva (92,1% contra 91%). A avaliação pela validação cruzada foi bem parecida para os dois modelos (acurácia = 81,3% para o modelo do EM e 81,9% para o modelo do Weka).

Optou-se pelo modelo gerado no EM para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão, considerando que a sua classificação dos exemplos de agosto de 2000, com a *surrogate rule*, seria a mais adequada.

Tabela 19: Matrizes de confusão da árvore de decisão da Figura 25.

Ressubstituição			Validação cruzada				
TAXA_INF_M5	Preditada		TAXA_INF_M5	Preditada			
	1	0		1	0		
Verdadeira	1	71	13	Verdadeira	1	67	17
	0	7	91		0	17	81

Tabela 20: Avaliação da árvore de decisão da Figura 25.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	89,0%	81,3% (4,8%)*
Taxa de erro	11,0%	18,7%
Sensitividade	84,5%	79,9% (6,9%)
Especificidade	92,9%	82,6% (4,3%)
Confiabilidade positiva	91,0%	79,4% (6,1%)
Confiabilidade negativa	87,5%	83,9% (4,7%)

* Desvio padrão da média.

As matrizes de confusão da árvore de decisão da Figura 25, obtidas pelos métodos de ressubstituição e de validação cruzada, são apresentadas na Tabela 19. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 20.

As regras de classificação extraídas da árvore de decisão da Figura 25 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 21. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

As matrizes de confusão e a avaliação do modelo gerado no Weka, bem como as regras de classificação extraídas da árvore de decisão da Figura 26 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

Tabela 21: Regras extraídas da árvore de decisão da Figura 25 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1 - Nó 2	
SE THUR95_PINF < 16,5 ENTÃO TAXA_INF_M5 = 0	Precisão: 90,8% Novidade: 0,15 Sensitividade: 69,4% Cobertura: 40,7% Especificidade: 92,9% Suporte: 37,4%
Distribuição*: JUN(4VN;2FN); JUL(14VN;2FN); AGO(14VN;2FN); SET(16VN); OUT(14VN); NOV(4VN); DEZ(2VN).	
Regra 2 - Nó 29	
SE THUR95_PINF ≥ 16,5 E UR_PINF < 78 E TMED_PI_PINF < 19,75 ENTÃO TAXA_INF_M5 = 1	Precisão: 70,0% Novidade: 0,01 Sensitividade: 7,1% Cobertura: 4,4% Especificidade: 98,0% Suporte: 3,3%
Distribuição: MAI(1VP;1FP); JUN(5VP;1FP).	

Regra 3 - Nó 30

SE THUR95_PINF \geq 16,5
E UR_PINF $<$ 78
E TMED_PI_PINF \geq 19,75
ENTÃO TAXA_INF_M5 = 0

Distribuição: JAN(2VN); NOV(9VN;1FN); DEZ(2VN;2FN).

Precisão: 77,8% **Novidade:** 0,02
Sensitividade: 13,3% **Cobertura:** 8,8%
Especificidade: 96,4% **Suporte:** 7,1%

Regra 4 - Nó 14

SE THUR95_PINF \geq 16,5
E UR_PINF \geq 78
E TMED_PI_PINF $<$ 22,35
ENTÃO TAXA_INF_M5 = 1

Distribuição: JAN(10VP;2FP); FEV(6VP); MAR(10VP;2FP); ABR(16VP); MAI(14VP); JUN(3VP;1FP); DEZ(6VP).

Precisão: 91,7% **Novidade:** 0,18
Sensitividade: 77,4% **Cobertura:** 38,5%
Especificidade: 94,9% **Suporte:** 35,7%

Regra 5 - Nó 15

SE THUR95_PINF \geq 16,5
E UR_PINF \geq 78
E TMED_PI_PINF \geq 22,35
ENTÃO TAXA_INF_M5 = 0

Distribuição: JAN(1VN;1FN); FEV(4VN;2FN); MAR(3VN;1FN); DEZ(2VN).

Precisão: 68,8% **Novidade:** 0,01
Sensitividade: 10,2% **Cobertura:** 7,7%
Especificidade: 95,2% **Suporte:** 5,5%

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

A Figura 27 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com alta carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M5 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 3 (seção 3.5.1).

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 28. Ela pode ser vista como uma subárvore do modelo do EM, com poda dos ramos a partir do nó 7 da Figura 27. Os atributos de teste correspondentes foram exatamente os mesmos e os valores para as fronteiras de decisão foram muito parecidos, proporcionando a mesma distribuição dos exemplos na classificação do conjunto de treinamento.

A avaliação dos dois modelos foi parecida na validação cruzada. A árvore sem a poda (EM) teve acurácia e sensibilidade um pouco melhores (83% e 80% contra 81,8% e 75,3%, respectivamente), enquanto a árvore podada (Weka) teve confiabilidade positiva pouca coisa melhor (85% contra 82,9%). O mesmo comportamento foi observado na resubstituição. O modelo do EM apresentou melhor equilíbrio entre as medidas na validação cruzada: 83%,

80%, 85,6%, 82,9% e 83,6% em comparação com 81,8%, 75,3%, 87,8%, 83,9% e 81%, na seqüência das medidas apresentadas na Tabela 23, desconsiderando a taxa de erro.

Por ter obtido a melhor acurácia, na resubstituição e na validação cruzada, e pelo melhor equilíbrio das medidas, optou-se pelo modelo gerado no EM para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão.

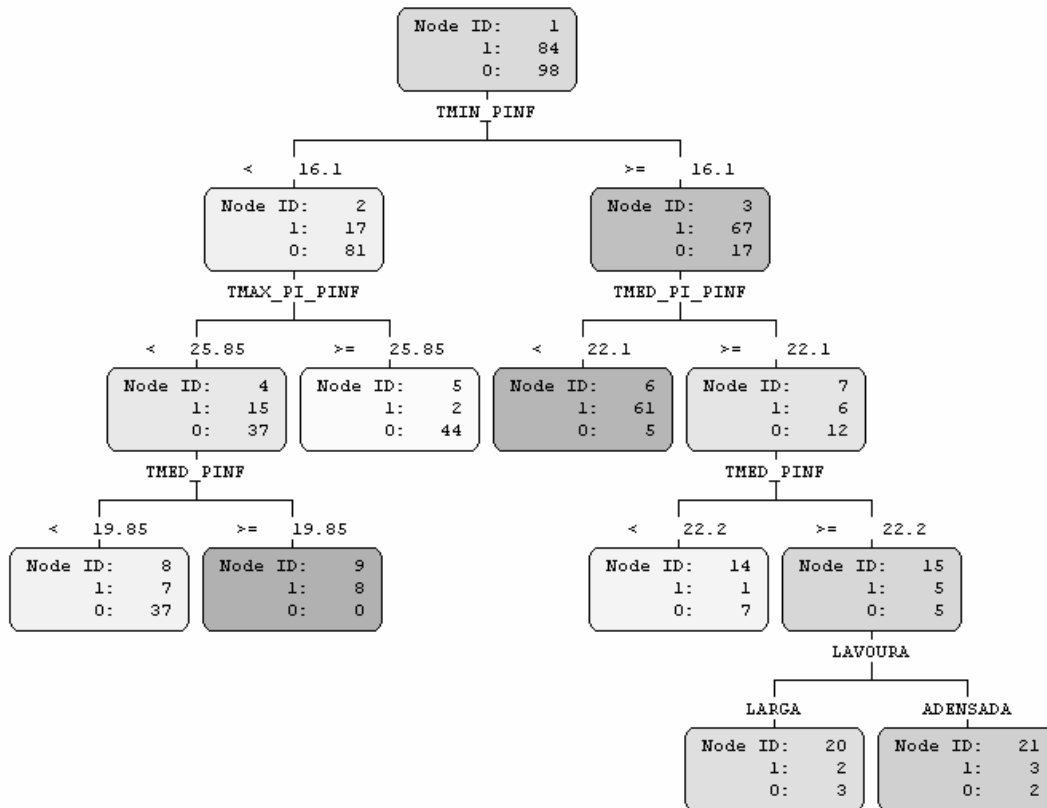


Figura 27: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 3; Geração: Enterprise Miner™.

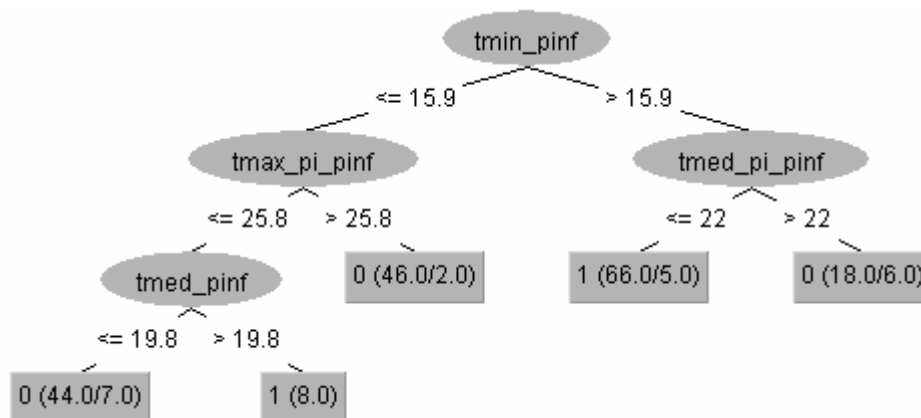


Figura 28: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 3; Geração: Weka.

As matrizes de confusão da árvore de decisão da Figura 27, obtidas pelos métodos de resubstituição e de validação cruzada, são apresentadas na Tabela 22. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 23.

Tabela 22: Matrizes de confusão da árvore de decisão da Figura 27.

Resubstituição			Validação cruzada				
TAXA_INF_M5	Pedita		TAXA_INF_M5	Pedita			
	1	0		1	0		
Verdadeira	1	72	12	Verdadeira	1	67	17
	0	7	91		0	14	84

Tabela 23: Avaliação da árvore de decisão da Figura 27.

Medida de avaliação	Método de estimativa	
	Resubstituição	Validação cruzada
Acurácia	89,6%	83,0% (3,0%)*
Taxa de erro	10,4%	17,0%
Sensitividade	85,7%	80,0% (4,4%)
Especificidade	92,9%	85,6% (2,8%)
Confiabilidade positiva	91,1%	82,9% (3,3%)
Confiabilidade negativa	88,3%	83,6% (3,1%)

* Desvio padrão da média.

As regras de classificação extraídas da árvore de decisão da Figura 27 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 24. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

Tabela 24: Regras extraídas da árvore de decisão da Figura 27 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1 - Nó 5	
SE TMIN_PINF < 16,1	Precisão: 93,8% Novidade: 0,11
E TMAX_PI_PINF ≥ 25,85	Sensitividade: 44,9% Cobertura: 25,3%
ENTÃO TAXA_INF_M5 = 0	Especificidade: 97,6% Suporte: 24,2%
Distribuição* : JUN(1VN;1FN); SET(12VN); OUT(14VN); NOV(13VN;1FN); DEZ(4VN).	
Regra 2 - Nó 6	
SE TMIN_PINF ≥ 16,1	Precisão: 91,2% Novidade: 0,17

E TMED_PI_PINF < 22,1
ENTÃO TAXA_INF_M5 = 1

Distribuição: JAN(10VP;2FP); FEV(6VP); MAR(8VP;2FP); ABR(16VP); MAI(13VP;1FP); DEZ(8VP).

Regra 3 - Nó 8

SE TMIN_PINF < 16,1
E TMAX_PI_PINF < 25,85
E TMED_PINF < 19,85
ENTÃO TAXA_INF_M5 = 0

Distribuição: JUN(5VN;3FN); JUL(14VN;2FN); AGO(14VN;2FN); SET(4VN).

Regra 4 - Nó 9

SE TMIN_PINF < 16,1
E TMAX_PI_PINF < 25,85
E TMED_PINF ≥ 19,85
ENTÃO TAXA_INF_M5 = 1

Distribuição: MAI(2VP); JUN (6VP).

Regra 5 - Nó 14

SE TMIN_PINF ≥ 16,1
E TMED_PI_PINF ≥ 22,1
E TMED_PINF < 22,2
ENTÃO TAXA_INF_M5 = 0

Distribuição: JAN(2VN); FEV(3VN;1FN); DEZ(2VN).

Regra 6 - Nó 20

SE TMIN_PINF ≥ 16,1
E TMED_PI_PINF ≥ 22,1
E TMED_PINF ≥ 22,2
E LAVOURA = LARGA
ENTÃO TAXA_INF_M5 = 0

Distribuição: JAN(1VN); FEV(1VN); MAR(1VN;2FN).

Regra 7 - Nó 21

SE TMIN_PINF ≥ 16,1
E TMED_PI_PINF ≥ 22,1
E TMED_PINF ≥ 22,2
E LAVOURA = ADENSADA
ENTÃO TAXA_INF_M5 = 1

Distribuição: JAN(1VP); FEV(1VP); MAR(1VP;2FP).

Sensitividade: 72,6% **Cobertura:** 36,3%
Especificidade: 94,9% **Suporte:** 33,5%

Precisão: 82,6% **Novidade:** 0,07
Sensitividade: 37,8% **Cobertura:** 24,2%
Especificidade: 91,7% **Suporte:** 20,3%

Precisão: 90,0% **Novidade:** 0,02
Sensitividade: 9,5% **Cobertura:** 4,4%
Especificidade: 100% **Suporte:** 4,4%

Precisão: 80,0% **Novidade:** 0,01
Sensitividade: 7,1% **Cobertura:** 4,4%
Especificidade: 98,8% **Suporte:** 3,8%

Precisão: 57,1% **Novidade:** 0,00
Sensitividade: 3,1% **Cobertura:** 2,7%
Especificidade: 97,6% **Suporte:** 1,6%

Precisão: 57,1% **Novidade:** 0,00
Sensitividade: 3,6% **Cobertura:** 2,7%
Especificidade: 98,0% **Suporte:** 1,6%

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

As matrizes de confusão e a avaliação do modelo gerado no Weka, bem como as regras de classificação extraídas da árvore de decisão da Figura 28 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

5.2.1.2 Discussão

No modelo gerado com a opção de seleção de atributos Modelagem 1 (Figura 24), em adição ao que já foi discutido com relação à análise da epidemia da ferrugem do cafeeiro com árvore de decisão (capítulo 4), pode-se comentar a inclusão dos dias desfavoráveis à infecção (atributo DDI_PINF) e da velocidade média diária do vento (atributo VVENTO_PINF).

A ocorrência de mais de 25 dias desfavoráveis à infecção ($DDI_PINF > 25$) correspondeu, na maioria dos casos, a taxas de infecção menores do que 5 p.p.

Ventos com velocidade média diária acima de 2,9 km/h exerceram efeito depressivo nas taxas de infecção. Na literatura, encontrou-se que ventos com velocidade de 4 km/h foram suficientes para a disseminação de esporos de *H. vastatrix* (KUSHALAPPA, 1989a). Entretanto, aumentos na velocidade acima de 4 km/h não estiveram associados com incrementos na quantidade de esporos coletados no ar. Esse relacionamento se mostrou irregular.

O modelo é simples e compacto, com apenas seis regras e baseado em quatro atributos de teste. No geral, o modelo foi bem avaliado, com altos valores para todas as medidas de avaliação (Tabela 17).

As predições para o período de dezembro a abril ficaram restritas a três regras (Tabela 18):

- **Regra 1:** precisão (92,1%); sensibilidade (70,4%); especificidade (94%); novidade (0,16); cobertura (40,7%) e suporte (37,9%); cobriu corretamente exemplos de dezembro e janeiro, na maior parte dos casos: 6 VN (verdadeiros negativos) e 2 FN (falsos negativos).
- **Regra 2:** precisão (71,4%); especificidade (96,4%); as demais medidas foram baixas; o período coberto se referiu aos meses de dezembro, fevereiro e março, com a maior parte dos casos classificados corretamente: 9 VN e 3 FN.
- **Regra 3:** precisão (90,3%); sensibilidade (76,2%); especificidade (93,9%); novidade (0,17); cobertura (38,5%) e suporte (35,2%); cobriu corretamente vários exemplos no período de dezembro a abril: 49 VP (verdadeiros positivos) e apenas 5 FP (falsos positivos).

Os altos valores de novidade (máximo igual a 0,25) indicaram alta probabilidade de correlação verdadeira entre o antecedente e o conseqüente da regra. Nenhuma regra, de

qualquer modelo, que tenha apresentado alto valor de novidade, revelou algo que pudesse ser considerado inesperado ou fora do comum.

No modelo gerado com a opção Modelagem 2 (Figura 25), se revelaram apenas as influências da temperatura e da umidade relativa do ar. A participação da temperatura dividiu-se na sua influência no processo de germinação (THUR95_PINF) e na sua influência durante o período de incubação (TMED_PI_PINF).

As influências desses fatores meteorológicos nas taxas de infecção da ferrugem do cafeeiro se assemelham ao que foi discutido no capítulo referente à análise da epidemia dessa doença com árvore de decisão (capítulo 4). Temperaturas médias mais baixas durante o molhamento foliar (THUR95_PINF < 16,5°C) foram desfavoráveis às taxas de infecção maiores ou iguais a 5 p.p. (Figura 25, nó 2), assim como também o foram temperaturas médias mais elevadas no período de incubação (Figura 25, nós 15 e 30). Umidade relativa média diária mais alta (UR_PINF ≥ 78%) foi mais favorável às taxas de infecção maiores ou iguais a 5 p.p. (Figura 25, nó 7).

O modelo é simples e compacto, com apenas cinco regras e baseado em três atributos de teste. No geral, o modelo foi bem avaliado, com altos valores para todas as medidas de avaliação (Tabela 20).

As previsões para o período de dezembro a abril foram distribuídas entre quatro das cinco regras (Tabela 21):

- **Regra 1:** precisão (90,8%); sensibilidade (69,4%); especificidade (92,9%); novidade (0,15); cobertura (40,7%) e suporte (37,4%); cobriu corretamente exemplos do mês de dezembro: 2 VN.
- **Regra 3:** precisão (77,8%); especificidade (96,4%); as demais medidas foram baixas; o período coberto incluiu os meses de dezembro (2 VN e 2 FN) e janeiro (2 VN).
- **Regra 4:** precisão (91,7%); sensibilidade (77,4%); especificidade (94,9%); novidade (0,18); cobertura (38,5%) e suporte (35,7%); cobriu corretamente vários exemplos no período de dezembro a abril: 48 VP e apenas 4 FP.
- **Regra 5:** precisão (68,8%); especificidade (95,2%); as demais medidas foram baixas; o período coberto foi de dezembro a março, com a maior parte dos exemplos classificados corretamente: 10 VN e 4 FN.

No modelo gerado com a opção Modelagem 3 (Figura 27), destacou-se a influência da temperatura, tanto no período de infecção quanto no período de incubação. Houve também a influência do espaçamento da lavoura (lavoura adensada ou larga), em grau bem menor.

A média das temperaturas mínimas diárias (TMIN_PINF) substituiu a temperatura média diária durante os períodos de molhamento foliar (THUR95_PINF) como o atributo de teste no nó raiz, em comparação com o modelo da opção Modelagem 2 (Figura 25).

Uma possível explicação para isso é que o molhamento foliar geralmente ocorre no período noturno, quando normalmente também ocorre a temperatura mínima do dia. Pedro Júnior et al. (1994) utilizaram a temperatura mínima do dia em um modelo de previsão da mancha preta do amendoim, em substituição à temperatura média durante o período de molhamento foliar, originalmente proposta para ser usada no modelo (JENSEN e BOYLE, 1966).

O teste sobre TMIN_PINF (divisão em 16,1 °C) foi diferente do teste sobre este mesmo atributo (divisão em 14,2 °C) na regra equivalente (*surrogate rule*) do nó raiz do modelo gerado com a opção Modelagem 2. Sendo assim, houve diferença na distribuição dos exemplos a partir do nó raiz entre os dois modelos.

Com essa nova distribuição dos exemplos a partir do nó raiz, as condições limitantes de temperatura devem ter prevalecido sobre as condições de umidade, fazendo com que o atributo UR_PINF, que entrou como atributo de teste na árvore de decisão da Figura 25, não entrasse no modelo da Figura 27. Cabe ressaltar que isso pode ter acontecido, ou que tenha sido válido, pelas características próprias do ambiente do local estudado.

A influência dos demais atributos sobre as taxas de infecção da ferrugem do cafeeiro é parecida com o que já foi discutido no capítulo referente à análise da epidemia dessa doença com árvore de decisão (capítulo 4).

O modelo também pode ser considerado simples e compacto, com sete regras e baseado em cinco atributos de teste. No geral, o modelo foi bem avaliado, com altos valores para todas as medidas de avaliação (Tabela 23).

As predições para o período de dezembro a abril foram distribuídas entre cinco das sete regras (Tabela 24):

- **Regra 1:** precisão (93,8%); sensibilidade (44,9%); especificidade (97,6%); novidade (0,11); cobertura (25,3%) e suporte (24,2%); classificou corretamente casos do mês de dezembro: 4 VN.
- **Regra 2:** precisão (91,2%); sensibilidade (72,6%); especificidade (94,9%); novidade (0,17); cobertura (36,3%) e suporte (33,5%); cobriu corretamente vários exemplos no período de dezembro a abril: 48 VP e apenas 4 FP.
- **Regra 5:** precisão (80%); especificidade (98,8%); as demais medidas foram baixas; cobriu corretamente o período de dezembro a fevereiro, na maioria dos casos: 7 VN e apenas 1 FN.
- **Regra 6:** precisão (57,1%); especificidade (97,6%); as demais medidas foram baixas; o período coberto foi de janeiro a março: 3 VN e 2 FN (os dois em março).
- **Regra 7:** precisão (57,1%); especificidade (98%); as demais medidas foram baixas; o período coberto foi de janeiro a março: 3 VP e 2 FP (os dois em março).

O conjunto das cinco regras acima, em relação às regras equivalentes dos modelos anteriores (Modelagem 1 e 2), melhorou o desempenho para os meses de dezembro (8 VP e 6 VN contra 6 VP, 6 VN e 2 FN), janeiro (11 VP, 2 FP e 3 VN contra 10 VP, 2 FP, 3 VN e 1 FN) e fevereiro (7 VP, 4 VN e 1 FN contra 6 VP, 4 VN e 2 FN), mas piorou o desempenho para o mês de março (9 VP, 4 FP, 1 VN e 2 FN contra 10 VP, 2 FP, 3 VN e 1 FN).

Comparando-se os modelos obtidos com as três opções de seleção de atributos preditivos, a árvore de decisão da Figura 27, gerada a partir do conjunto de treinamento mais reduzido (Modelagem 3), foi o modelo com a melhor avaliação.

Na resubstituição, foi o modelo com a melhor matriz de confusão (Tabela 22) e, conseqüentemente, com os melhores valores para todas as medidas de avaliação (Tabela 23).

Na validação cruzada, também apresentou a melhor matriz de confusão (Tabela 22) e, embora os três modelos tenham tido desempenhos semelhantes, considerando-se os desvios padrões das médias, foi o modelo que obteve os maiores valores, com os menores desvios padrões, para a maioria das medidas de avaliação (Tabela 23).

Outra boa opção como modelo de alerta da ferrugem do cafeeiro é a árvore de decisão gerada com a opção Modelagem 2 (Figura 25). Também teve bom desempenho na avaliação, tanto na resubstituição quanto na validação cruzada. As vantagens desse modelo são a menor

quantidade de regras, o número reduzido de atributos de teste e a dependência a apenas dois fatores meteorológicos, a temperatura e a umidade relativa do ar.

Os dois modelos destacados obtiveram uma boa relação de equilíbrio entre a acurácia, a sensibilidade, a especificidade e as confiabilidades positiva e negativa. Na validação cruzada, o mínimo de acerto para todas essas medidas ficou em torno de 80%. Na resubstituição, esse valor mínimo ficou em torno de 85%.

Em suma, as árvores de decisão da Figura 25 e da Figura 27 foram consideradas as melhores opções como modelos de alerta da ferrugem do cafeeiro para predizer quando a taxa de infecção da doença for atingir ou ultrapassar 5 p.p. em lavouras de café com alta carga pendente de frutos.

As melhores regras de cada modelo, considerando os verdadeiros positivos e os verdadeiros negativos para o período de dezembro a abril, além das outras medidas de avaliação de regras, foram:

- Modelagem 1 - VP: Regra 3; VN: Regras 1 e 2.
- Modelagem 2 - VP: Regra 4; VN: Regras 1, 3 e 5.
- Modelagem 3 - VP: Regra 2; VN: Regras 1 e 5.

5.2.2 Alerta quando a taxa de infecção for atingir ou ultrapassar 10 p.p.

5.2.2.1 Resultados

A Figura 29 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com alta carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M10 (seção 3.4.1) e os atributos preditivos foram escolhidos conforme a opção de seleção de atributos Modelagem 1 (seção 3.5.1).

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 30. O atributo de teste no nó raiz foi o mesmo, com o mesmo resultado na distribuição dos exemplos. A partir daí, alguns atributos de teste também coincidiram, mas em posições diferentes. A ordem dos atributos de teste no modelo do Weka proporcionou um menor número de regras e evitou testes repetidos sobre um mesmo atributo (TMIN_PINF).

A avaliação dos modelos foi parecida na validação cruzada, exceto com relação à sensibilidade. O modelo do EM teve acurácia de 78,2% contra 78,7% do modelo do Weka. Houve diferença de mais de 7 p.p. a favor do modelo do Weka na sensibilidade (64,7% contra 57%), mas há de se considerar o desvio padrão da média (6,8%). O modelo do Weka

apresentou melhor equilíbrio entre as medidas na validação cruzada: 78,7%, 64,7%, 84,6%, 66,8% e 85,6% em comparação com 78,2%, 57%, 86,7%, 67,2% e 82,9%, na seqüência das medidas apresentadas na Tabela 26, desconsiderando a taxa de erro.

No geral, pelo exposto, o modelo gerado no Weka teve melhor desempenho. Optou-se, então, por ele para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão.

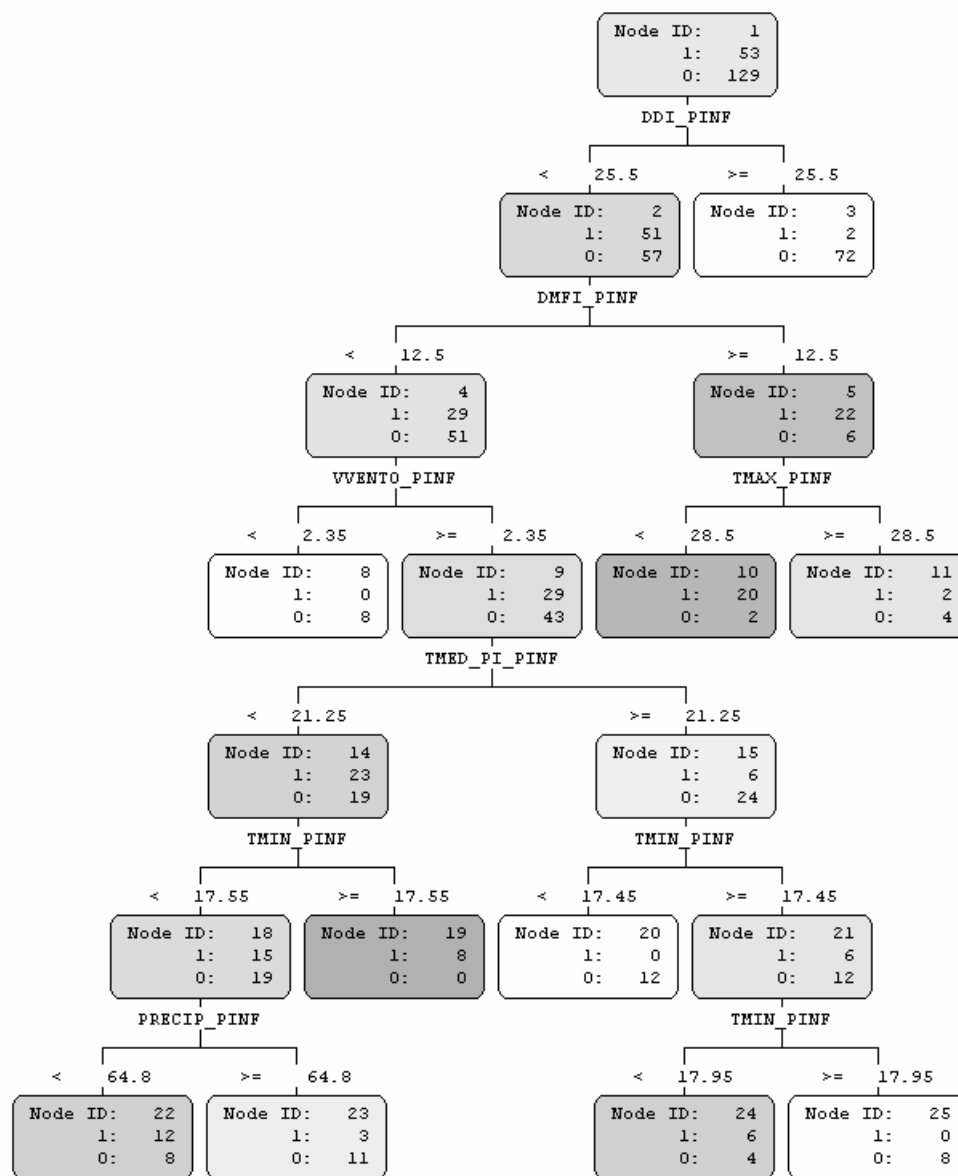


Figura 29: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 1; Geração: Enterprise Miner™.

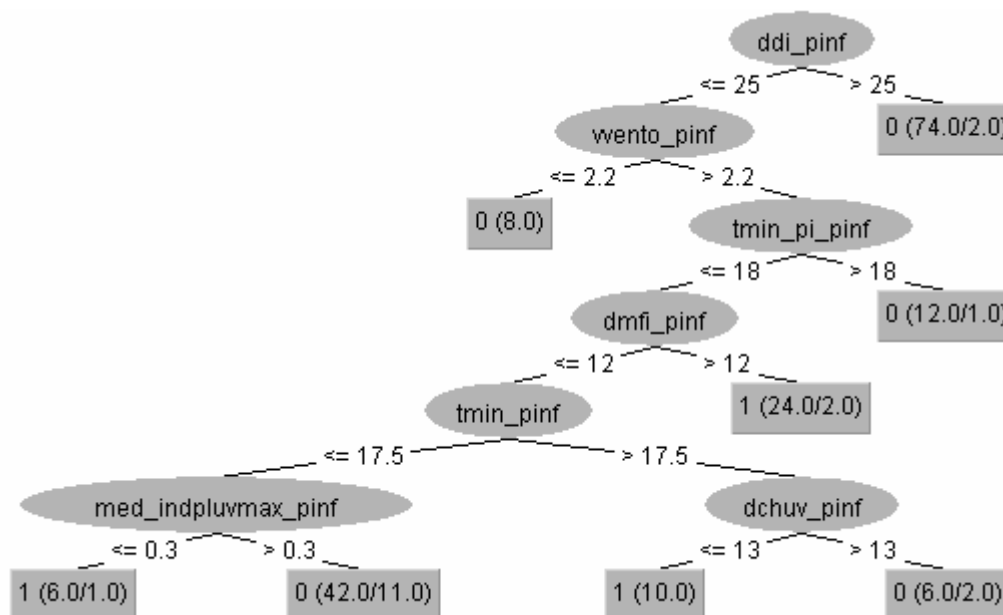


Figura 30: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 1; Geração: Weka.

As matrizes de confusão da árvore de decisão da Figura 30, obtidas pelos métodos de ressubstituição e de validação cruzada, são apresentadas na Tabela 25. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 26.

Tabela 25: Matrizes de confusão da árvore de decisão da Figura 30.

Ressubstituição			Validação cruzada				
TAXA_INF_M10	Predita		TAXA_INF_M10	Predita			
	1	0		1	0		
Verdadeira	1	37	16	Verdadeira	1	34	19
	0	3	126		0	20	109

Tabela 26: Avaliação da árvore de decisão da Figura 30.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	89,6%	78,7% (3,1%)*
Taxa de erro	10,4%	21,3%
Sensitividade	69,8%	64,7% (6,8%)
Especificidade	97,7%	84,6% (4,0%)
Confiabilidade positiva	92,5%	66,8% (6,1%)
Confiabilidade negativa	88,7%	85,6% (2,4%)

* Desvio padrão da média.

As regras de classificação extraídas da árvore de decisão da Figura 30 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 27. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

As matrizes de confusão e a avaliação do modelo gerado no EM, bem como as regras de classificação extraídas da árvore de decisão da Figura 29 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

Tabela 27: Regras extraídas da árvore de decisão da Figura 30 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1	
SE DDI_PINF > 25 ENTÃO TAXA_INF_M10 = 0	Precisão: 96,1% Novidade: 0,11 Sensitividade: 55,8% Cobertura: 40,7% Especificidade: 96,2% Suporte: 39,6%
Distribuição*: JAN(2VN); JUL(12VN); AGO(15VN;1FN); SET(16VN); OUT(14VN); NOV(8VN); DEZ(5VN;1FN).	
Regra 2	
SE DDI_PINF ≤ 25 E VVENTO_PINF ≤ 2,2 ENTÃO TAXA_INF_M10 = 0	Precisão: 90,0% Novidade: 0,01 Sensitividade: 6,2% Cobertura: 4,4% Especificidade: 100% Suporte: 4,4%
Distribuição: MAR(2VN); MAI(2VN); JUN(2VN); JUL(2VN).	
Regra 3	
SE DDI_PINF ≤ 25 E VVENTO_PINF > 2,2 E TMIN_PI_PINF > 18 ENTÃO TAXA_INF_M10 = 0	Precisão: 85,7% Novidade: 0,01 Sensitividade: 8,5% Cobertura: 6,6% Especificidade: 98,1% Suporte: 6,0%
Distribuição: JAN(4VN); FEV(3VN;1FN); MAR(2VN); DEZ(2VN).	
Regra 4	
SE DDI_PINF ≤ 25 E VVENTO_PINF > 2,2 E TMIN_PI_PINF ≤ 18 E DMFI_PINF > 12 ENTÃO TAXA_INF_M10 = 1	Precisão: 88,5% Novidade: 0,08 Sensitividade: 41,5% Cobertura: 13,2% Especificidade: 98,4% Suporte: 12,1%
Distribuição: JAN(5VP;1FP); FEV(4VP); MAR(7VP;1FP); ABR(6VP).	
Regra 5	
SE DDI_PINF ≤ 25 E VVENTO_PINF > 2,2 E TMIN_PI_PINF ≤ 18	Precisão: 75,0% Novidade: 0,02 Sensitividade: 9,4% Cobertura: 3,3% Especificidade: 99,2% Suporte: 2,7%

E DMFI_PINF \leq 12
E TMIN_PINF \leq 17,5
E MED_INDPLUVMAX_PINF \leq 0,3
ENTÃO TAXA_INF_M10 = 1

Distribuição: MAI(2VP); JUN(3VP;1FP).

Regra 6

SE DDI_PINF \leq 25
E VVENTO_PINF $>$ 2,2
E TMIN_PI_PINF \leq 18
E DMFI_PINF \leq 12
E TMIN_PINF \leq 17,5
E MED_INDPLUVMAX_PINF $>$ 0,3
ENTÃO TAXA_INF_M10 = 0

Distribuição: JAN(2VN); FEV(1VN;1FN); ABR(4VN); MAI(6VN;4FN); JUN(5VN;5FN); JUL(2VN); NOV(6VN); DEZ(5VN;1FN).

Regra 7

SE DDI_PINF \leq 25
E VVENTO_PINF $>$ 2,2
E TMIN_PI_PINF \leq 18
E DMFI_PINF \leq 12
E TMIN_PINF $>$ 17,5
E DCHUV_PINF \leq 13
ENTÃO TAXA_INF_M10 = 1

Distribuição: MAR(2VP); ABR(6VP); MAI(2VP).

Regra 8

SE DDI_PINF \leq 25
E VVENTO_PINF $>$ 2,2
E TMIN_PI_PINF \leq 18
E DMFI_PINF \leq 12
E TMIN_PINF $>$ 17,5
E DCHUV_PINF $>$ 13
ENTÃO TAXA_INF_M10 = 0

Distribuição: JAN(1VN;1FN); FEV(1VN;1FN); MAR(2VN).

Precisão: 72,7% **Novidade:** 0,01
Sensitividade: 24,0% **Cobertura:** 23,1%
Especificidade: 79,2% **Suporte:** 17,0%

Precisão: 91,7% **Novidade:** 0,04
Sensitividade: 18,9% **Cobertura:** 5,5%
Especificidade: 100% **Suporte:** 5,5%

Precisão: 62,5% **Novidade:** 0,00
Sensitividade: 3,1% **Cobertura:** 3,3%
Especificidade: 96,2% **Suporte:** 2,2%

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

A Figura 31 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com alta carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M10 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 2 (seção 3.5.1).

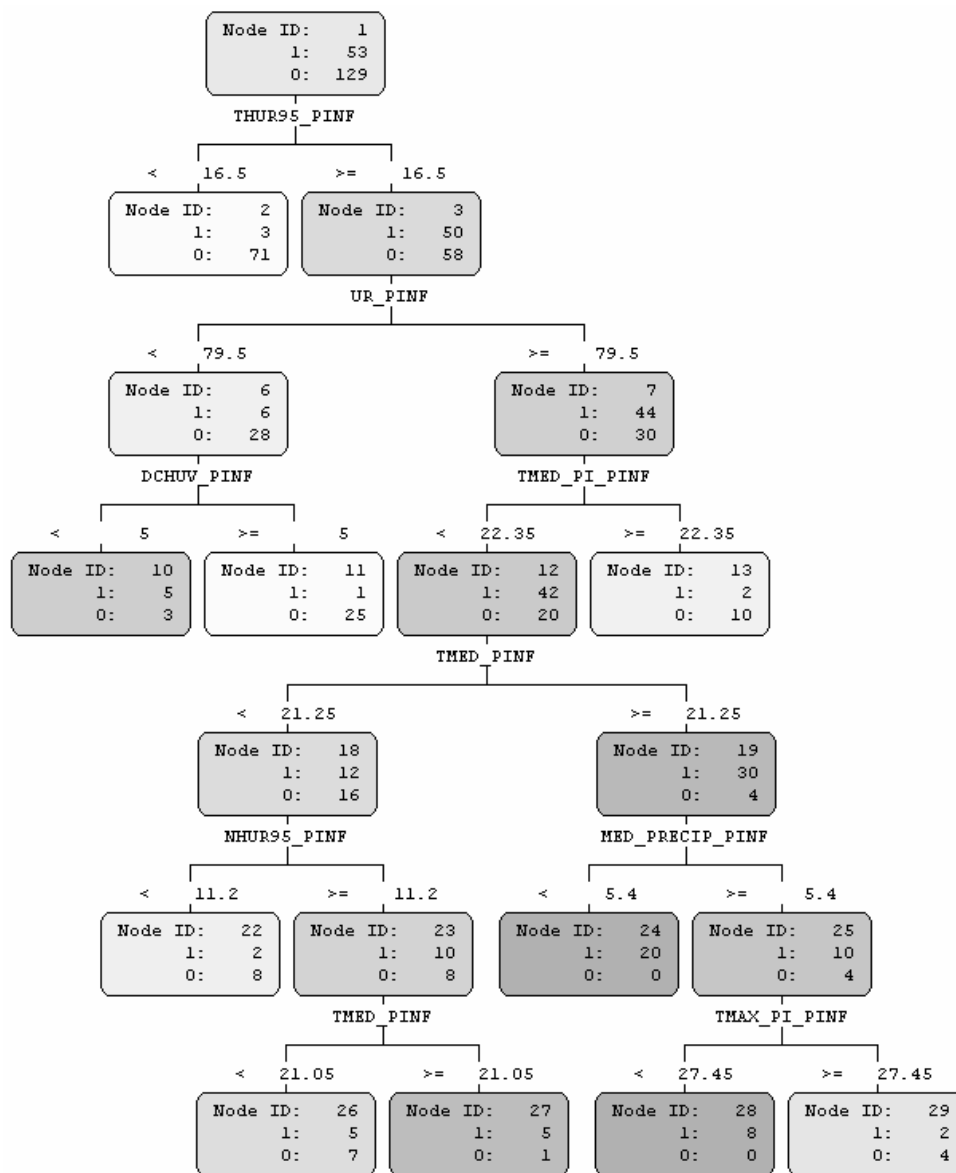


Figura 31: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 2; Geração: Enterprise Miner™.

A árvore de decisão correspondente gerada no Weka se resumiu a uma subárvore da gerada no EM, formada com os seguintes nós correspondentes da Figura 31: nó 3 (nó raiz da árvore do Weka), nós 7 e 12 (nós internos) e nós 6, 13, 18 e 19 (nós folhas).

Por não contemplar atributo de teste específico da Modelagem 2, ficou igual à árvore de decisão gerada com a opção de seleção de atributos Modelagem 3 (Figura 33, pg. 123). A razão disso foi a opção de poda *subtree raising* do Weka, que eliminou da árvore o nó raiz e o teste sobre o atributo THUR95_PINF, após o processo de indução, passando a ser o novo nó raiz o nó à direita do nó raiz eliminado.

A regra equivalente (*surrogate rule*) à regra do nó raiz da Figura 31 foi baseada no atributo TMIN_PINF: $TMIN_PINF < 14,2$ equivalendo a $THUR95_PINF < 16,5$ e $TMIN_PINF \geq 14,2$ equivalendo a $THUR95_PINF \geq 16,5$. Esta regra foi utilizada para classificar os dois exemplos de agosto de 2000 em que não houve períodos de molhamento foliar, com um mínimo de seis horas, no período de infecção correspondente ($NHUR95_PINF = 0$) e, portanto, os valores do atributo THUR95_PINF foram nulos. Os dois exemplos foram classificados no nó 2 ($TMIN_PINF = 10,4$ °C) como da classe ‘0’.

A regra de divisão do nó 6, baseada no atributo DCHUV_PINF, teve regras concorrentes com mesmo ganho de informação. Os atributos de teste dessas regras foram TMIN_PI_PINF (divisão em 14,2 °C) e TMED_PI_PINF (divisão em 19,75 °C).

As matrizes de confusão da árvore de decisão da Figura 31, obtidas pelos métodos de resubstituição e de validação cruzada, são apresentadas na Tabela 28. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 29.

Tabela 28: Matrizes de confusão da árvore de decisão da Figura 31.

Resubstituição				Validação cruzada			
TAXA_INF_M10	Pedita			TAXA_INF_M10	Pedita		
	1	0			1	0	
Verdadeira	1	38	15	Verdadeira	1	37	16
	0	4	125		0	22	107

Tabela 29: Avaliação da árvore de decisão da Figura 31.

Medida de avaliação	Método de estimativa	
	Resubstituição	Validação cruzada
Acurácia	89,6%	79,2% (3,0%)*
Taxa de erro	10,4%	20,8%
Sensitividade	71,7%	70,3% (4,6%)
Especificidade	96,9%	82,8% (3,3%)
Confiabilidade positiva	90,5%	65,3% (4,1%)
Confiabilidade negativa	89,3%	86,9% (2,4%)

* Desvio padrão da média.

As regras de classificação extraídas da árvore de decisão da Figura 31 e a avaliação individual de cada uma dessas regras estão apresentadas na Tabela 30. Junto com cada regra,

também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

Tabela 30: Regras extraídas da árvore de decisão da Figura 31 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1 - Nó 2	
SE THUR95_PINF < 16,5 ENTÃO TAXA_INF_M10 = 0	Precisão: 94,7% Novidade: 0,10 Sensitividade: 55,0% Cobertura: 40,7% Especificidade: 94,3% Suporte: 39,0%
Distribuição*: JUN(4VN;2FN); JUL(16VN); AGO(15VN;1FN); SET(16VN); OUT(14VN); NOV(4VN); DEZ(2VN).	
Regra 2 - Nó 10	
SE THUR95_PINF ≥ 16,5 E UR_PINF < 79,5 E DCHUV_PINF < 5 ENTÃO TAXA_INF_M10 = 1	Precisão: 60,0% Novidade: 0,01 Sensitividade: 9,4% Cobertura: 4,4% Especificidade: 97,7% Suporte: 2,7%
Distribuição: MAI(1VP;1FP); JUN (4VP;2FP).	
Regra 3 - Nó 11	
SE THUR95_PINF ≥ 16,5 E UR_PINF < 79,5 E DCHUV_PINF ≥ 5 ENTÃO TAXA_INF_M10 = 0	Precisão: 92,9% Novidade: 0,04 Sensitividade: 19,4% Cobertura: 14,3% Especificidade: 98,1% Suporte: 13,7%
Distribuição: JAN(4VN); ABR(2VN); MAI(2VN); NOV(10VN); DEZ(7VN;1FN).	
Regra 4 - Nó 13	
SE THUR95_PINF ≥ 16,5 E UR_PINF ≥ 79,5 E TMED_PI_PINF ≥ 22,35 ENTÃO TAXA_INF_M10 = 0	Precisão: 78,6% Novidade: 0,01 Sensitividade: 7,8% Cobertura: 6,6% Especificidade: 96,2% Suporte: 5,5%
Distribuição: JAN(2VN); FEV(4VN;2FN); MAR(4VN).	
Regra 5 - Nó 22	
SE THUR95_PINF ≥ 16,5 E UR_PINF ≥ 79,5 E TMED_PI_PINF < 22,35 E TMED_PINF < 21,25 E NHUR95_PINF < 11,2 ENTÃO TAXA_INF_M10 = 0	Precisão: 75,0% Novidade: 0,01 Sensitividade: 6,2% Cobertura: 5,5% Especificidade: 96,2% Suporte: 4,4%
Distribuição: ABR(2VN); MAI(3VN;1FN); DEZ(3VN;1FN).	
Regra 6 - Nó 24	
SE THUR95_PINF ≥ 16,5 E UR_PINF ≥ 79,5	Precisão: 95,5% Novidade: 0,08 Sensitividade: 37,7% Cobertura: 11,0%

E TMED_PI_PINF < 22,35
E TMED_PINF ≥ 21,25
E MED_PRECIP_PINF < 5,4
ENTÃO TAXA_INF_M10 = 1

Especificidade: 100% **Suporte:** 11,0%

Distribuição: JAN(2VP); FEV(2VP); MAR(4VP); ABR(8VP); MAI(4VP).

Regra 7 - Nó 26

SE THUR95_PINF ≥ 16,5
E UR_PINF ≥ 79,5
E TMED_PI_PINF < 22,35
E NHUR95_PINF ≥ 11,2
E TMED_PINF < 21,05

Precisão: 57,1% **Novidade:** -0,01
Sensitividade: 5,4% **Cobertura:** 6,6%
Especificidade: 90,6% **Suporte:** 3,8%

ENTÃO TAXA_INF_M10 = 0

Distribuição: JAN(3VN;1FN); MAI(2VN;2FN); JUN(2VN;2FN).

Regra 8 - Nó 27

SE THUR95_PINF ≥ 16,5
E UR_PINF ≥ 79,5
E TMED_PI_PINF < 22,35
E NHUR95_PINF ≥ 11,2
E 21,05 ≤ TMED_PINF < 21,25

Precisão: 75,0% **Novidade:** 0,02
Sensitividade: 9,4% **Cobertura:** 3,3%
Especificidade: 99,2% **Suporte:** 2,7%

ENTÃO TAXA_INF_M10 = 1

Distribuição: FEV(1VP;1FP); MAR(2VP); ABR(2VP).

Regra 9 - Nó 28

SE THUR95_PINF ≥ 16,5
E UR_PINF ≥ 79,5
E TMED_PI_PINF < 22,35
E TMED_PINF ≥ 21,25
E MED_PRECIP_PINF ≥ 5,4
E TMAX_PI_PINF < 27,45

Precisão: 90,0% **Novidade:** 0,03
Sensitividade: 15,1% **Cobertura:** 4,4%
Especificidade: 100% **Suporte:** 4,4%

ENTÃO TAXA_INF_M10 = 1

Distribuição: JAN(2VP); FEV(2VP); MAR(2VP); ABR(2VP).

Regra 10 - Nó 29

SE THUR95_PINF ≥ 16,5
E UR_PINF ≥ 79,5
E TMED_PI_PINF < 22,35
E TMED_PINF ≥ 21,25
E MED_PRECIP_PINF ≥ 5,4
E TMAX_PI_PINF ≥ 27,45

Precisão: 62,5% **Novidade:** 0,00
Sensitividade: 3,1% **Cobertura:** 3,3%
Especificidade: 96,2% **Suporte:** 2,2%

ENTÃO TAXA_INF_M10 = 0

Distribuição: JAN(1VN;1FN); MAR(3VN;1FN).

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

A Figura 32 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com alta carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M10 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 3 (seção 3.5.1).

A árvore de decisão correspondente gerada no Weka, apresentada na Figura 33, poderia ser obtida com uma poda na árvore da Figura 32. Da subárvore à esquerda do nó raiz, sobraria apenas o nó 2. Na subárvore à direita do nó raiz, seriam necessários dois tipos de poda: primeiro podar o nó 3 e erguer a subárvore (*subtree raising*) a partir do nó 7, redistribuindo os exemplos classificados no nó 6; depois podar a partir dos nós 18 e 19.

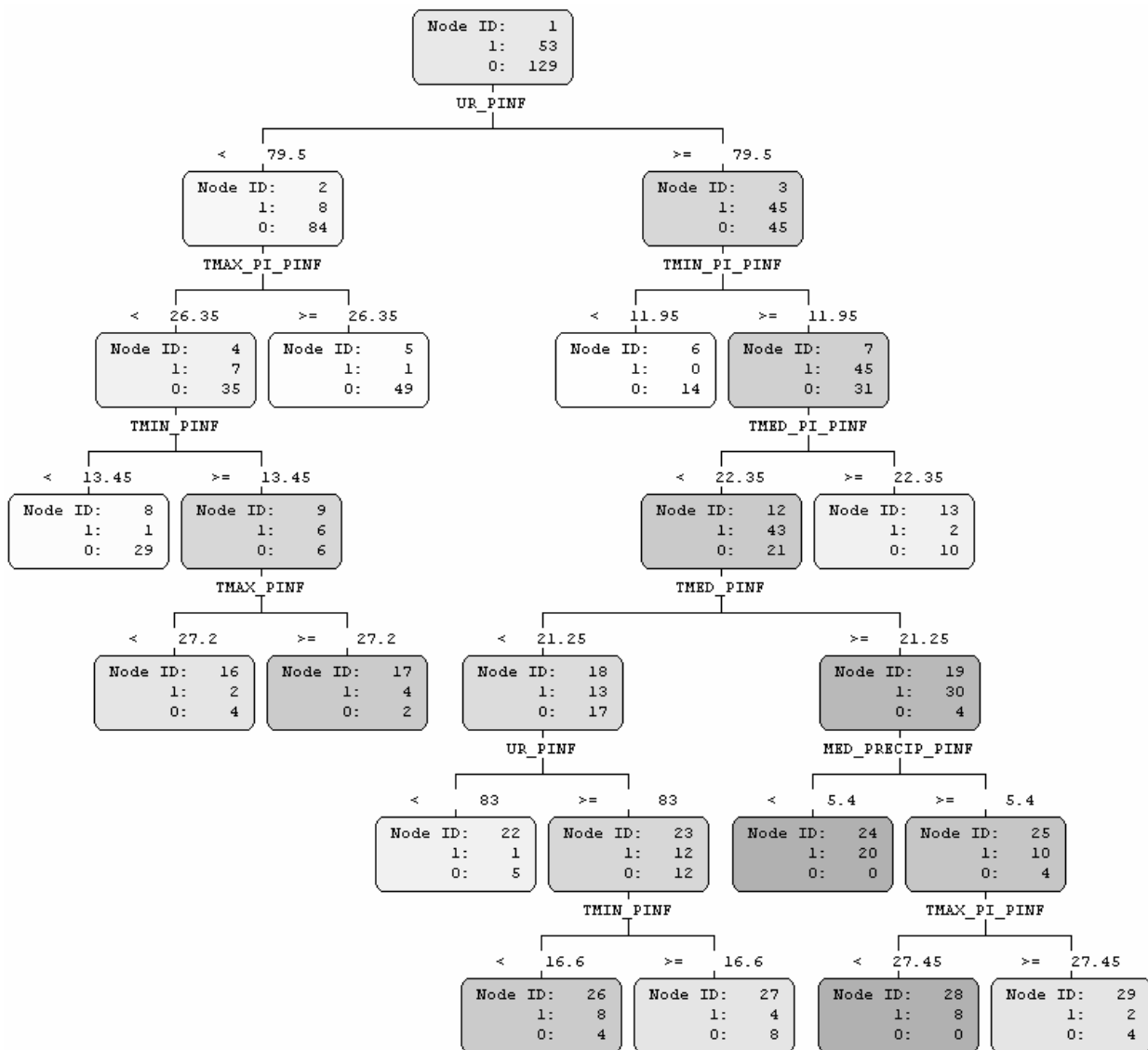


Figura 32: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 3; Geração: Enterprise Miner™.

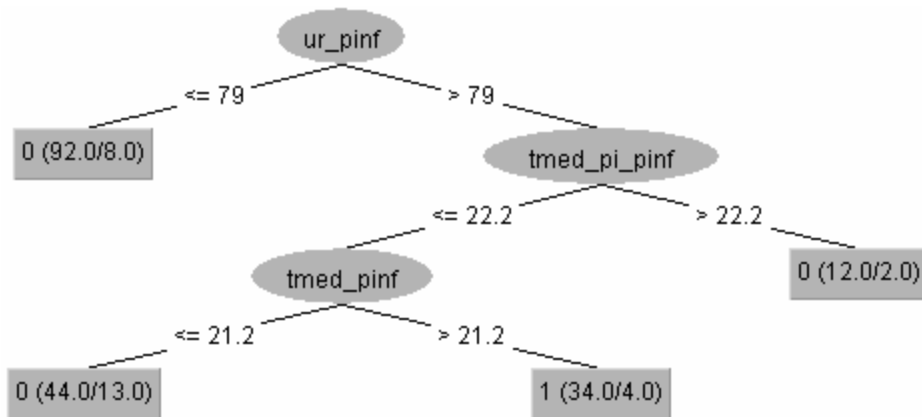


Figura 33: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 3; Geração: Weka.

O modelo do EM, com um maior número de regras, se ajustou melhor aos exemplos do conjunto de treinamento e, portanto, foi melhor avaliado na ressubstituição (acurácia = 89,6% contra 85,2%). O modelo do Weka, mais simples, melhorou seu desempenho na validação cruzada (acurácia = 79,2% contra 77% do modelo do EM).

O modelo do EM apresentou tendência de ser melhor avaliado pela sensibilidade, enquanto o modelo do Weka pela confiabilidade positiva, tanto na ressubstituição quanto na validação cruzada.

Nenhum modelo apresentou vantagem significativa sobre o outro. Então, pela simplicidade, optou-se pelo modelo gerado no Weka para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão.

As matrizes de confusão da árvore de decisão da Figura 33, obtidas pelos métodos de ressubstituição e de validação cruzada, são apresentadas na Tabela 31. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 32.

Tabela 31: Matrizes de confusão da árvore de decisão da Figura 33.

Ressubstituição				Validação cruzada			
TAXA_INF_M10	Pedita			TAXA_INF_M10	Pedita		
	1	0			1	0	
Verdadeira	1	30	23	Verdadeira	1	30	23
	0	4	125		0	15	114

Tabela 32: Avaliação da árvore de decisão da Figura 33.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	85,2%	79,2% (4,4%)*
Taxa de erro	14,8%	20,8%
Sensitividade	56,6%	57,3% (8,9%)
Especificidade	96,9%	88,5% (3,8%)
Confiabilidade positiva	88,2%	69,8% (8,9%)
Confiabilidade negativa	84,5%	83,7% (3,3%)

* Desvio padrão da média.

As regras de classificação extraídas da árvore de decisão da Figura 33 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 33. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

Tabela 33: Regras extraídas da árvore de decisão da Figura 33 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1	
SE UR_PINF \leq 79 ENTÃO TAXA_INF_M10 = 0	Precisão: 90,4% Novidade: 0,10 Sensitividade: 65,1% Cobertura: 50,5% Especificidade: 84,9% Suporte: 46,2%
Distribuição* : JAN(4VN); ABR(2VN); MAI(3VN;1FN); JUN(3VN;5FN); JUL(10VN); AGO(9VN;1FN); SET(16VN); OUT(14VN); NOV(14VN); DEZ(9VN;1FN).	
Regra 2	
SE UR_PINF > 79 E TMED_PI_PINF > 22,2 ENTÃO TAXA_INF_M10 = 0	Precisão: 78,6% Novidade: 0,01 Sensitividade: 7,8% Cobertura: 6,6% Especificidade: 96,2% Suporte: 5,5%
Distribuição: JAN(2VN); FEV(4VN;2FN); MAR(4VN).	
Regra 3	
SE UR_PINF > 79 E TMED_PI_PINF \leq 22,2 E TMED_PINF \leq 21,2 ENTÃO TAXA_INF_M10 = 0	Precisão: 69,6% Novidade: 0,00 Sensitividade: 24,0% Cobertura: 24,2% Especificidade: 75,5% Suporte: 17,0%
Distribuição: JAN(3VN;1FN); FEV(1VN;1FN); MAR(2FN); ABR(2VN;2FN); MAI(5VN;3FN); JUN(5VN;3FN); JUL(6VN); AGO(6VN); DEZ(3VN;1FN).	
Regra 4	
SE UR_PINF > 79	Precisão: 86,1% Novidade: 0,11 Sensitividade: 56,6% Cobertura: 18,7%

E $TMED_PI_PINF \leq 22,2$

E $TMED_PINF > 21,2$

ENTÃO $TAXA_INF_M10 = 1$

Especificidade: 96,9% **Suporte:** 16,5%

Distribuição: JAN(5VP;1FP); FEV(4VP); MAR(7VP;3FP); ABR(10VP); MAI(4VP).

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

As matrizes de confusão e a avaliação do modelo gerado no EM, bem como as regras de classificação extraídas da árvore de decisão da Figura 32 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

5.2.2.2 Discussão

No modelo gerado com a opção Modelagem 1 (Figura 30), o atributo de teste no nó raiz foi o mesmo do modelo para o atributo meta TAXA_INF_M5 (Figura 24), com o mesmo valor para a fronteira de decisão. A ocorrência de mais de 25 dias desfavoráveis à infecção ($DDI_PINF > 25$) correspondeu a taxas de infecção menores do que 10 p.p., na grande maioria dos casos.

A velocidade do vento exibiu efeito depressivo nas taxas de infecção maiores ou iguais a 10 p.p. quando a velocidade média diária ($VVENTO_PINF$) esteve menor ou igual a 2,2 km/h, diferente do ocorrido com as taxas de infecção maiores ou iguais a 5 p.p.

Mais de doze dias muito favoráveis à infecção ($DMFI_PINF > 12$) favoreceram as taxas de infecção maiores ou iguais a 10 p.p.

Vários dias chuvosos ($DCHUV_PINF > 13$) e chuvas com intensidade média ($MED_INDPLUVMAX_PINF$) acima de 0,3 mm/h causaram efeito depressivo nas taxas de infecção. Isto parece indicar o efeito negativo que as chuvas podem ocasionar, quando, ao invés de proporcionar a disseminação, lavam os esporos das pústulas (KUSHALAPPA, 1989a).

A influência dos demais atributos sobre as taxas de infecção da ferrugem do cafeeiro é parecida com o que já foi discutido no capítulo referente à análise da epidemia dessa doença com árvore de decisão (capítulo 4).

Com oito regras e baseado em sete atributos de teste, o modelo já não é tão simples e compacto quanto o modelo para TAXA_INF_M5. No geral, o modelo teve avaliação satisfatória (Tabela 26). A acurácia (78,7% - validação cruzada) foi maior do que a proporção

de exemplos da classe majoritária (71%), que seria a acurácia estimada de um classificador que atribuísse a todos os exemplos a classe '0'.

A especificidade (84,6%) e a confiabilidade negativa (85,6%) foram melhores do que a sensibilidade (64,7%) e a confiabilidade positiva (66,8%), provavelmente em decorrência da distribuição desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 15).

As predições para o período de dezembro a abril foram distribuídas entre quase todas as regras (Tabela 27). As regras de destaque foram:

- **Regra 1:** precisão (96,1%); sensibilidade (55,8%); especificidade (96,2%); novidade (0,11); cobertura (40,7%) e suporte (39,6%); cobriu corretamente exemplos de dezembro e janeiro, na maior parte dos casos: 7 VN e 1 FN.
- **Regras 2 e 3:** precisão (90% e 85,7%); especificidade (100% e 98,1%); as demais medidas foram baixas; cobriram corretamente o período de dezembro a março, na maior parte dos casos: 13 VN e apenas 1 FN.
- **Regras 4 e 7:** precisão (88,5% e 91,7%); sensibilidade (41,5% e 18,9%); especificidade (98,4% e 100%); novidade (0,08 e 0,04); cobertura (13,2% e 5,5%) e suporte (12,1% e 5,5%); cobriram corretamente vários exemplos no período de janeiro a abril: 30 VP e apenas 2 FP.
- **Regra 6:** precisão (72,4%); sensibilidade (24%); especificidade (79,2%); novidade (0,01); cobertura (23,1%) e suporte (17%); cobriu corretamente exemplos de dezembro a fevereiro e abril, na maior parte dos casos: 12 VN e 2 FN.

No modelo gerado com a opção Modelagem 2 (Figura 31), o atributo de teste no nó raiz também foi o mesmo do modelo para o atributo meta TAXA_INF_M5 (Figura 25), com o mesmo valor para a fronteira de decisão e, conseqüentemente, a mesma divisão dos exemplos. A diferença, é claro, foi que aumentou a proporção de casos com a classe '0' quando a temperatura média diária durante os períodos de molhamento foliar esteve mais baixa ($THUR95_PINF < 16,5\text{ °C}$).

O modelo está bastante relacionado com o modelo para TAXA_INF_M5: até o terceiro nível de profundidade das árvores, a estrutura é a mesma, com praticamente os mesmos atributos de teste e fronteiras de decisão. A principal diferença está na ramificação a

partir do nó 12 da árvore de decisão da Figura 31, que corresponde ao nó 14 da árvore da Figura 25.

Houve a inclusão no modelo da média diária do número de horas de molhamento foliar no período de infecção. Mais de onze horas de molhamento foliar ($NHUR95 \geq 11,2$ h) corresponderam a uma maior proporção de casos em que a taxa de infecção foi maior ou igual a 10 p.p. Segundo Kushalappa et al. (1983), a proporção de infecção aumentou exponencialmente conforme aumentou o número de horas de água líquida na superfície da folha para a germinação.

A troca de DCHUV_PINF como atributo de teste da regra de divisão do nó 6 por TMED_PI_PINF (regra concorrente com ganho de informação igual; divisão em 19,75 °C) manteria exatamente a mesma distribuição dos exemplos, da mesma forma como ocorreu com o modelo para TAXA_INF_M5. Sendo assim, o atributo TMED_PI_PINF poderia ser usado no lugar de ou em conjunto com DCHUV_PINF na condição da regra.

A influência dos demais atributos sobre as taxas de infecção da ferrugem do cafeeiro é parecida com o que já foi discutido no capítulo referente à análise da epidemia dessa doença com árvore de decisão (capítulo 4).

O modelo, formado por dez regras (o dobro do modelo para TAXA_INF_M5) e baseado em oito atributos de teste, teve avaliação satisfatória (Tabela 29). A acurácia (79,2% - validação cruzada) foi maior do que a proporção de exemplos da classe majoritária (71%).

A especificidade (82,8%) e a confiabilidade negativa (86,9%) foram melhores do que a sensibilidade (70,3%) e a confiabilidade positiva (65,3%). Conforme já comentado, isso deve ter ocorrido em virtude da distribuição desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 15).

As predições para o período de dezembro a abril foram distribuídas entre quase todas as regras (Tabela 30). As regras que se destacaram foram:

- **Regras 1 e 3:** precisão (94,7% e 92,9%); sensibilidade (55% e 19,4%); especificidade (94,3% e 98,1%); novidade (0,10 e 0,04); cobertura (40,7% e 14,3%) e suporte (39% e 13,7%); cobriram corretamente exemplos de dezembro, janeiro e abril, na maior parte dos casos: 15 VN e apenas 1 FN.

- **Regras 4 e 5:** precisão (78,6% e 75%); especificidade (96,2% para ambas); as demais medidas foram baixas; cobriram corretamente o período de dezembro a abril, na maior parte dos casos: 15 VN e 3 FN.
- **Regras 6 e 9:** precisão (95,5% e 90%); sensibilidade (37,7% e 15,1%); especificidade máxima (100% para ambas); novidade (0,08 e 0,03); cobriram corretamente todos os exemplos do período de janeiro a abril: 24 VP.
- **Regra 8:** precisão (75%); especificidade (99,2%); as demais medidas foram baixas; cobriu corretamente o período de fevereiro a abril, na maior parte dos casos: 5 VP e 1 FP.

No modelo gerado com a opção Modelagem 3 (Figura 33), o atributo de teste no nó raiz foi diferente do correspondente no modelo para o atributo meta TAXA_INF_M5 (Figura 28). A umidade relativa do ar média diária no período de infecção (UR_PINF) substituiu a média das temperaturas mínimas no mesmo período (TMIN_PINF).

Isso parece indicar que a umidade relativa, além de ter influenciado no processo de infecção, deve ter influenciado na esporulação de *H. vastatrix* – é sabido que a presença de umidade favorece o processo de esporulação do fungo (KUSHALAPPA, 1989a) –, prevalecendo em relação à influência da temperatura e aparecendo com maior importância nos períodos de evolução mais acelerada da ferrugem do cafeeiro.

Este comportamento sugerido é perfeitamente factível, pois, apesar de se ter procurado destacar, na preparação dos dados, a influência do ambiente no processo de infecção (seção 3.4.2), outros ciclos da doença também estão em curso (a ferrugem do cafeeiro é uma doença policíclica), em diferentes fases, como a disseminação ou a esporulação.

A influência dos demais atributos sobre as taxas de infecção da ferrugem do cafeeiro é parecida com o que já foi discutido no capítulo referente à análise da epidemia dessa doença com árvore de decisão (capítulo 4).

O modelo é bastante simples e compacto, com apenas quatro regras e três atributos de teste. No geral, teve avaliação satisfatória (Tabela 32). A acurácia (79,2% - validação cruzada) foi maior do que a proporção de exemplos da classe majoritária (71%).

A especificidade (88,5%) e a confiabilidade negativa (83,7%) foram melhores do que a sensibilidade (57,3%) e a confiabilidade positiva (69,8%). Como para os modelos gerados para as duas outras opções (Modelagem 1 e 2), isso deve ter ocorrido em virtude da

distribuição desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 15).

As predições para o período de dezembro a abril foram distribuídas entre as quatro regras do modelo (Tabela 33):

- **Regra 1:** precisão (90,4%); sensibilidade (65,1%); especificidade (84,9%); novidade (0,10); cobertura (50,5%) e suporte (46,2%); cobriu corretamente exemplos de dezembro, janeiro e abril, na maior parte dos casos: 15 VN e apenas 1 FN.
- **Regra 2:** precisão (78,6%); especificidade (96,2%); as demais medidas foram baixas; cobriu corretamente o período de janeiro a março, na maior parte dos casos: 10 VN e 2 FN.
- **Regra 3:** precisão (69,6%); sensibilidade (24%), especificidade (75,5%), novidade (0,00); cobertura (24,2%) e suporte (17%); cobriu corretamente exemplos de dezembro a abril, mas não para a maioria dos casos: 9 VN e 7 FN.
- **Regra 4:** precisão (86,1%); sensibilidade (56,6%); especificidade (96,9%); novidade (0,11); cobertura (18,7%) e suporte (16,5%); cobriu corretamente exemplos de janeiro a abril, na maior parte dos casos: 26 VP e apenas 4 FP.

Com relação à regra 3, caso se tivesse escolhido o modelo gerado no Enterprise Miner™, esta seria substituída por três regras, correspondentes aos nós folhas da ramificação a partir do nó 18 da Figura 32. Neste caso, 3 FN da regra 3 seriam substituídos por 3 VP e 1 VN por 1 FP, produzindo o seguinte resultado, no conjunto das três regras, para o período de dezembro a abril: 8 VN, 4 FN, 3 VP e 1 FP. Entretanto, o melhor ajuste aos dados de treinamento pelo modelo da Figura 32, como exemplificado, acarretou em desempenho inferior na validação cruzada, conforme foi mostrado na seção anterior.

Comparando-se os modelos obtidos com as três opções de seleção de atributos preditivos, observou-se o seguinte:

- Os modelos gerados com as opções Modelagem 1 e Modelagem 2 tiveram avaliação bastante semelhante, tanto na resubstituição quanto na validação cruzada.
- Os modelos gerados com as opções Modelagem 2 e Modelagem 3 tiveram a mesma acurácia na validação cruzada.
- O modelo gerado com a opção Modelagem 3 obteve maiores valores em relação à especificidade e à confiabilidade positiva, enquanto os outros dois modelos obtiveram maiores valores quanto à sensibilidade e à confiabilidade negativa.

Entre os modelos das opções Modelagem 1 e Modelagem 2, o segundo é preferível, apesar do primeiro ter menos regras e estar baseado em menor número de atributos de teste. Os atributos de teste da árvore de decisão da Figura 31 (Modelagem 2) são menos elaborados e se referem a dados meteorológicos de registro mais comum, se comparados com os atributos de teste da árvore de decisão da Figura 30 (Modelagem 1).

O modelo gerado com a opção Modelagem 2 obteve melhor relação de equilíbrio entre a acurácia, a sensibilidade, a especificidade e as confiabilidades positiva e negativa do que o modelo gerado com a opção Modelagem 3. Em compensação, o modelo da opção Modelagem 3 é bem mais simples e seus atributos de teste são bem menos custosos de preparar.

Portanto, as árvores de decisão da Figura 31 e da Figura 33 foram consideradas as melhores opções como modelos de alerta da ferrugem do cafeeiro para predizer quando a taxa de infecção da doença for atingir ou ultrapassar 10 p.p. em lavouras de café com alta carga pendente de frutos.

As melhores regras de cada modelo, considerando os verdadeiros positivos e os verdadeiros negativos para o período de dezembro a abril, além das outras medidas de avaliação de regras, foram:

- Modelagem 1 - VP: Regras 4 e 7; VN: Regras 1, 2, 3 e 6.
- Modelagem 2 - VP: Regras 6, 8 e 9; VN: Regras 1, 3, 4 e 5.
- Modelagem 3 - VP: Regra 4; VN: Regras 1 e 2.

5.3 Modelos para lavouras com baixa carga pendente de frutos

As distribuições percentual e absoluta dos exemplos do conjunto de treinamento entre as classes ‘1’ e ‘0’ dos atributos meta TAXA_INF_M5 e TAXA_INF_M10 são apresentadas na Tabela 34. Relembrando novamente, para TAXA_INF_M5, ‘1’ significa que a taxa de infecção da ferrugem do cafeeiro foi maior ou igual a 5 p.p. e ‘0’ significa o contrário. Para TAXA_INF_M10, a interpretação é semelhante, considerando o limite de 10 p.p.

Tabela 34: Distribuição dos exemplos de lavouras com baixa carga pendente entre as classes ‘1’ e ‘0’ dos atributos meta TAXA_INF_M5 e TAXA_INF_M10.

Atributo meta	Classe	
	1	0
TAXA_INF_M5	29% (53 exemplos)	71% (129 exemplos)
TAXA_INF_M10	12% (21 exemplos)	88% (161 exemplos)

5.3.1 Alerta quando a taxa de infecção for atingir ou ultrapassar 5 p.p.

5.3.1.1 Resultados

A Figura 34 apresenta a árvore de decisão, gerada no SAS[®] Enterprise Miner[™] (EM), para alertas da ferrugem do cafeeiro em lavouras com baixa carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M5 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 1 (seção 3.5.1).

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 35. Os dois modelos foram quase iguais. A diferença foi a divisão entre lavouras largas e adensadas no modelo do EM (nós 26 e 27 da Figura 34). Fora isso, os atributos de teste foram exatamente os mesmos e os valores para as fronteiras de decisão bem parecidos, proporcionando a mesma distribuição dos exemplos na classificação do conjunto de treinamento.

Como a divisão pelo espaçamento das lavouras não foi significativa (apenas cinco exemplos em cada nó folha) e indicou risco maior de taxas mais elevadas da ferrugem para lavouras largas, diferentemente do que seria esperado, optou-se pelo modelo gerado no Weka para a apresentação de sua avaliação e da avaliação individual das regras extraídas da árvore de decisão.

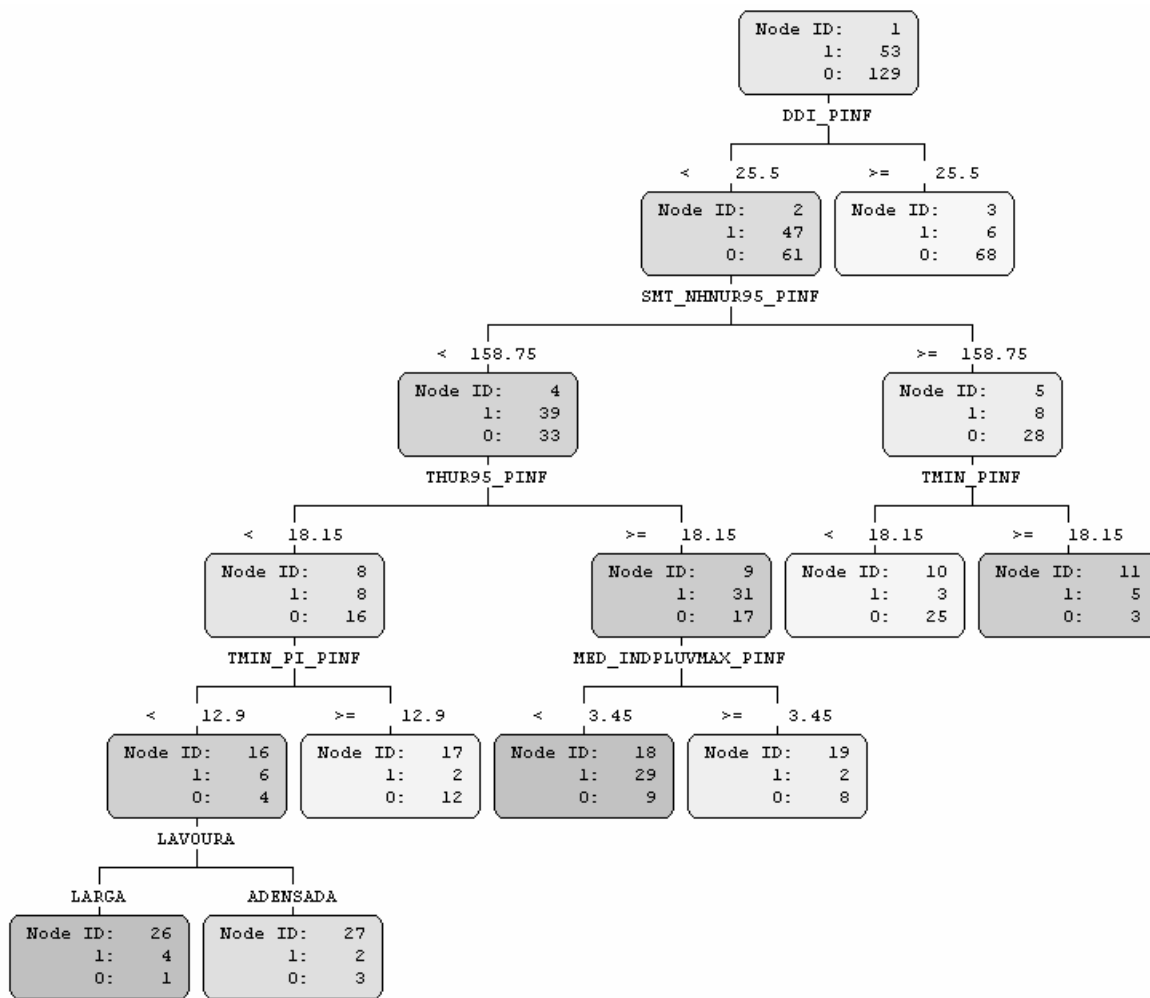


Figura 34: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 1; Geração: Enterprise Miner™.

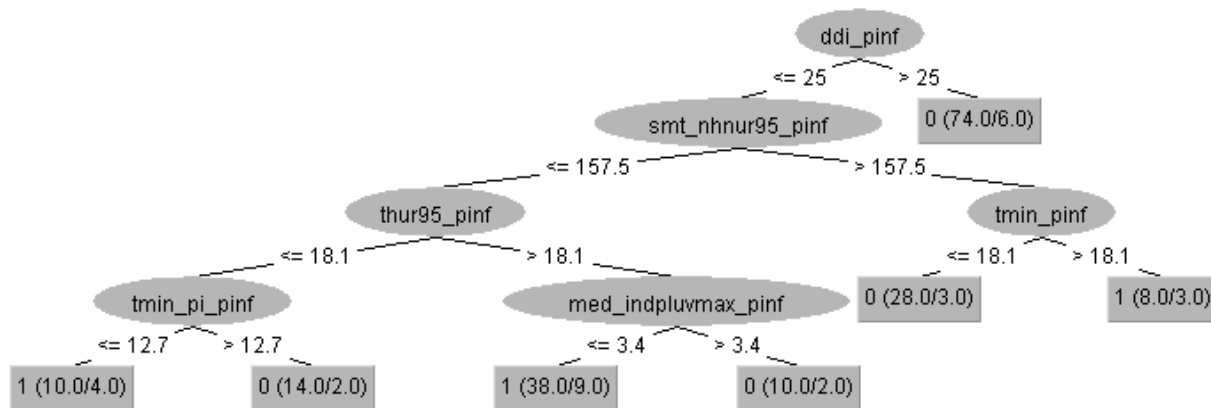


Figura 35: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 1; Geração: Weka.

As matrizes de confusão da árvore de decisão da Figura 35, obtidas pelos métodos de ressubstituição e de validação cruzada, são apresentadas na Tabela 35. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 36.

Tabela 35: Matrizes de confusão da árvore de decisão da Figura 35.

Ressubstituição				Validação cruzada			
TAXA_INF_M5	Predita			TAXA_INF_M5	Predita		
	1	0			1	0	
Verdadeira	1	40	13	Verdadeira	1	19	34
	0	16	113		0	20	109

Tabela 36: Avaliação da árvore de decisão da Figura 35.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	84,1%	70,3% (2,1%)*
Taxa de erro	15,9%	29,7%
Sensitividade	75,5%	34,7% (10,8%)
Especificidade	87,6%	84,5% (4,1%)
Confiabilidade positiva	71,4%	35,3% (8,5%)
Confiabilidade negativa	89,7%	77,7% (2,5%)

* Desvio padrão da média.

As regras de classificação extraídas da árvore de decisão da Figura 35 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 37. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

As matrizes de confusão e a avaliação do modelo gerado no EM, bem como as regras de classificação extraídas da árvore de decisão da Figura 34 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

Tabela 37: Regras extraídas da árvore de decisão da Figura 35 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação	
Regra 1		
SE DDI_PINF > 25	Precisão: 90,8%	Novidade: 0,09
ENTÃO TAXA_INF_M5 = 0	Sensitividade: 52,7%	Cobertura: 40,7%

Especificidade: 88,7% Suporte: 37,4%
Distribuição*: JAN(2VN); JUL(10VN;2FN); AGO(14VN;2FN); SET(15VN;1FN); OUT(14VN); NOV(8VN);
DEZ(5VN;1FN).

Regra 2

SE DDI_PINF \leq 25
E SMT_NHNUR95_PINF $>$ 157,5
E TMIN_PINF \leq 18,1
ENTÃO TAXA_INF_M5 = 0

Precisão: 86,7% Novidade: 0,03
Sensitividade: 19,4% Cobertura: 15,4%
Especificidade: 94,3% Suporte: 13,7%

Distribuição: JAN(5VN;1FN); FEV(2VN); MAR(4VN); ABR(6VN); MAI(2VN); JUN(2VN);
JUL(1VN;1FN); DEZ(3VN;1FN).

Regra 3

SE DDI_PINF \leq 25
E SMT_NHNUR95_PINF $>$ 157,5
E TMIN_PINF $>$ 18,1
ENTÃO TAXA_INF_M5 = 1

Precisão: 60,0% Novidade: 0,01
Sensitividade: 9,4% Cobertura: 4,4%
Especificidade: 97,7% Suporte: 2,7%

Distribuição: JAN(2VP;2FP); FEV(2VP); MAR(1VP;1FP).

Regra 4

SE DDI_PINF \leq 25
E SMT_NHNUR95_PINF \leq 157,5
E THUR95_PINF \leq 18,1
E TMIN_PI_PINF \leq 12,7
ENTÃO TAXA_INF_M5 = 1

Precisão: 58,3% Novidade: 0,02
Sensitividade: 11,3% Cobertura: 5,5%
Especificidade: 96,9% Suporte: 3,3%

Distribuição: JUN(4VP;4FP); JUL(2VP).

Regra 5

SE DDI_PINF \leq 25
E SMT_NHNUR95_PINF \leq 157,5
E THUR95_PINF \leq 18,1
E TMIN_PI_PINF $>$ 12,7
ENTÃO TAXA_INF_M5 = 0

Precisão: 81,3% Novidade: 0,01
Sensitividade: 9,3% Cobertura: 7,7%
Especificidade: 96,2% Suporte: 6,6%

Distribuição: FEV(2VN); ABR(1VN;1FN); JUN(4VN); NOV(5VN;1FN).

Regra 6

SE DDI_PINF \leq 25
E SMT_NHNUR95_PINF \leq 157,5
E THUR95_PINF $>$ 18,1
E MED_INDPLUVMAX_PINF \leq 3,4
ENTÃO TAXA_INF_M5 = 1

Precisão: 75,0% Novidade: 0,10
Sensitividade: 54,7% Cobertura: 20,9%
Especificidade: 93,0% Suporte: 15,9%

Distribuição: JAN(2VP;2FP); MAR(7VP;1FP); ABR(5VP;1FP); MAI(10VP;4FP); JUN(2VP);
DEZ(3VP;1FP).

Regra 7

SE DDI_PINF \leq 25
E SMT_NHNUR95_PINF \leq 157,5

Precisão: 75,0% Novidade: 0,01
Sensitividade: 6,2% Cobertura: 5,5%

E THUR95_PINF > 18,1
E MED_INDPLUVMAX_PINF > 3,4
ENTÃO TAXA_INF_M5 = 0

Distribuição: FEV(4VN;2FN); MAR(2VN); ABR (2VN).

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

A Figura 36 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com baixa carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M5 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 2 (seção 3.5.1).

A regra equivalente (*surrogate rule*) à regra do nó raiz foi baseada no atributo TMIN_PINF: TMIN_PINF < 16,35 equivalendo a THUR95_PINF < 17,55 e TMIN_PINF ≥ 16,35 equivalendo a THUR95_PINF ≥ 17,55. Esta regra foi utilizada para classificar os dois exemplos de agosto de 2000, em que não houve períodos de molhamento foliar, com um mínimo de seis horas, no período de infecção correspondente (NHUR95_PINF = 0) e, portanto, os valores do atributo THUR95_PINF foram nulos. Os dois exemplos foram atribuídos ao nó 2 (TMIN_PINF = 10,4 °C) e classificados como da classe '0' no nó 17.

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 37. O atributo de teste no nó raiz e a divisão dos exemplos neste nível foram os mesmos do modelo do EM. A subárvore à esquerda do nó raiz foi a do modelo do EM podada até a raiz. As subárvores à direita do nó raiz foram diferentes.

A avaliação dos dois modelos foi parecida, na ressubstituição (acurácia = 85,7% do modelo do EM contra 85,2%) e na validação cruzada (71,8% contra 72,1%). O modelo do EM apresentou tendência de ser melhor avaliado pela sensibilidade, enquanto o modelo do Weka pela confiabilidade positiva, tanto na ressubstituição quanto na validação cruzada. Houve diferença acima de 8 p.p. na sensibilidade, em favor do modelo do EM na validação cruzada (46,3% contra 38%), mas há de se considerar o desvio padrão da média dessa medida (6,9%). O modelo do EM apresentou melhor equilíbrio entre as medidas na validação cruzada: 71,8%, 46,3%, 82,1%, 49,9% e 79,4% em comparação com 72,1%, 38%, 86%, 52,5% e 77,5%, na seqüência das medidas apresentadas na Tabela 39, desconsiderando a taxa de erro.

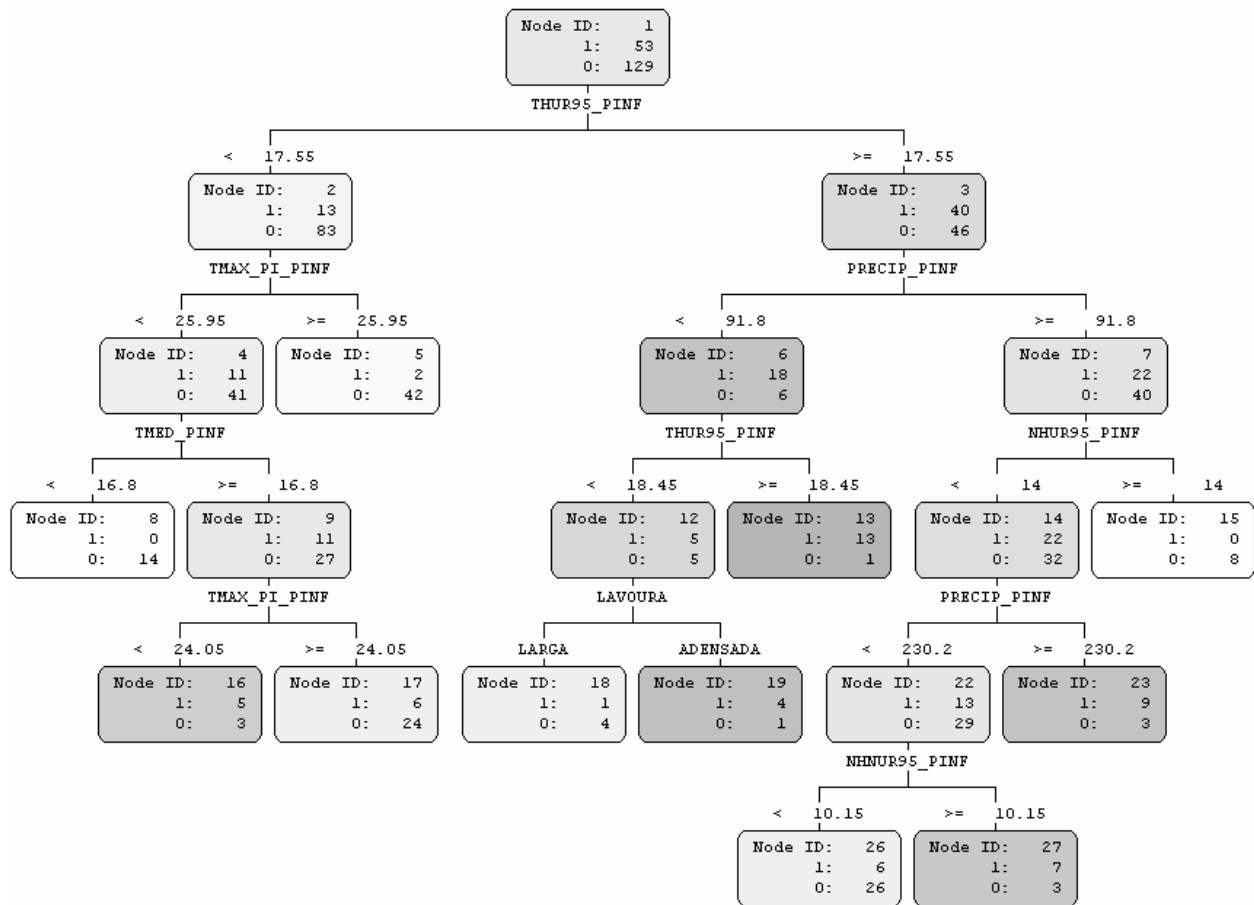


Figura 36: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 2; Geração: Enterprise Miner™.

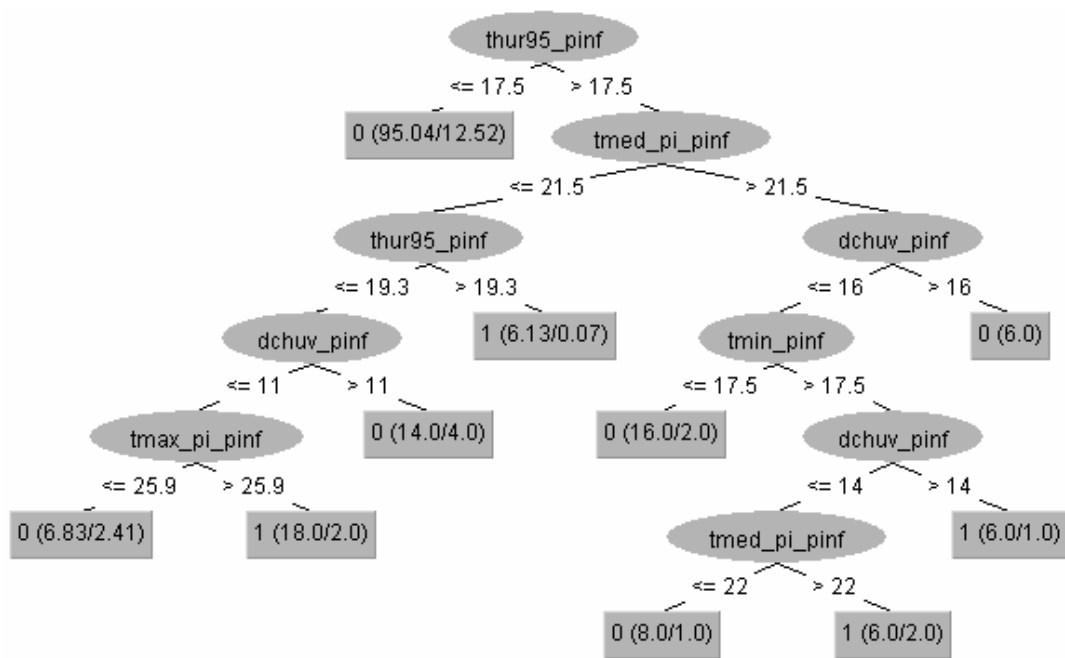


Figura 37: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 2; Geração: Weka.

No geral, conforme exposto, o modelo gerado no EM teve melhor desempenho. Optou-se, então, por ele para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão.

As matrizes de confusão da árvore de decisão da Figura 36, obtidas pelos métodos de resubstituição e de validação cruzada, são apresentadas na Tabela 38. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 39.

Tabela 38: Matrizes de confusão da árvore de decisão da Figura 36.

Resubstituição				Validação cruzada			
TAXA_INF_M5	Predita			TAXA_INF_M5	Predita		
	1	0			1	0	
Verdadeira	1	38	15	Verdadeira	1	25	28
	0	11	118		0	23	106

Tabela 39: Avaliação da árvore de decisão da Figura 36.

Medida de avaliação	Método de estimativa	
	Resubstituição	Validação cruzada
Acurácia	85,7%	71,8% (2,5%)*
Taxa de erro	14,3%	28,2%
Sensitividade	71,7%	46,3% (6,9%)
Especificidade	91,5%	82,1% (2,1%)
Confiabilidade positiva	77,6%	49,9% (5,9%)
Confiabilidade negativa	88,7%	79,4% (2,1%)

* Desvio padrão da média.

As regras de classificação extraídas da árvore de decisão da Figura 36 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 40. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

As matrizes de confusão e a avaliação do modelo gerado no Weka, bem como as regras de classificação extraídas da árvore de decisão da Figura 37 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

Tabela 40: Regras extraídas da árvore de decisão da Figura 36 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1 - Nó 5	
SE THUR95_PINF < 17,55 E TMAX_PI_PINF ≥ 25,95 ENTÃO TAXA_INF_M5 = 0	Precisão: 93,5% Novidade: 0,06 Sensitividade: 32,6% Cobertura: 24,2% Especificidade: 96,2% Suporte: 23,1%
Distribuição*: JUN(2VN); SET(8VN); OUT(14VN); NOV(13VN;1FN); DEZ(5VN;1FN).	
Regra 2 - Nó 8	
SE THUR95_PINF < 17,55 E TMAX_PI_PINF < 25,95 E TMED_PINF < 16,8 ENTÃO TAXA_INF_M5 = 0	Precisão: 93,8% Novidade: 0,02 Sensitividade: 10,9% Cobertura: 7,7% Especificidade: 100% Suporte: 7,7%
Distribuição: JUL(4VN); AGO(6VN); SET(4VN).	
Regra 3 - Nó 13	
SE THUR95_PINF ≥ 18,45 E PRECIP_PINF < 91,8 ENTÃO TAXA_INF_M5 = 1	Precisão: 87,5% Novidade: 0,05 Sensitividade: 24,5% Cobertura: 7,7% Especificidade: 99,2% Suporte: 7,1%
Distribuição: MAR(4VP); MAI(7VP;1FP); JUN(2VP).	
Regra 4 - Nó 15	
SE THUR95_PINF ≥ 17,55 E PRECIP_PINF ≥ 91,8 E NHUR95_PINF ≥ 14 ENTÃO TAXA_INF_M5 = 0	Precisão: 90,0% Novidade: 0,01 Sensitividade: 6,2% Cobertura: 4,4% Especificidade: 100% Suporte: 4,4%
Distribuição: JAN(8VN).	
Regra 5 - Nó 16	
SE THUR95_PINF < 17,55 E TMAX_PI_PINF < 24,05 E TMED_PINF ≥ 16,8 ENTÃO TAXA_INF_M5 = 1	Precisão: 60,0% Novidade: 0,01 Sensitividade: 9,4% Cobertura: 4,4% Especificidade: 97,7% Suporte: 2,7%
Distribuição: JUN (1VP;1FP); JUL(4VP;2FP).	
Regra 6 - Nó 17	
SE THUR95_PINF < 17,55 E 24,05 ≤ TMAX_PI_PINF < 25,95 E TMED_PINF ≥ 16,8 ENTÃO TAXA_INF_M5 = 0	Precisão: 78,1% Novidade: 0,02 Sensitividade: 18,6% Cobertura: 16,5% Especificidade: 88,7% Suporte: 13,2%
Distribuição: MAI(2VN); JUN(6VN;2FN); JUL(5VN;1FN); AGO(8VN;2FN); SET(3VN;1FN).	
Regra 7 - Nó 18	
SE 17,55 ≤ THUR95_PINF < 18,45 E PRECIP_PINF < 91,8	Precisão: 59,1% Novidade: -0,01 Sensitividade: 9,3% Cobertura: 11,0%

E LAVOURA = LARGA
ENTÃO TAXA_INF_M5 = 0

Distribuição: ABR(1VN); MAI(2VN;1FN); JUN(1VN).

Regra 8 - Nó 19

SE $17,55 \leq \text{THUR95_PINF} < 18,45$
E $\text{PRECIP_PINF} < 91,8$
E LAVOURA = ADENSADA

ENTÃO TAXA_INF_M5 = 1

Distribuição: ABR(1VP); MAI(2VP;1FP); JUN(1VP).

Regra 9 - Nó 23

SE $\text{THUR95_PINF} \geq 17,55$
E $\text{PRECIP_PINF} \geq 230,2$
E $\text{NHUR95_PINF} < 14$

ENTÃO TAXA_INF_M5 = 1

Distribuição: JAN(4VP); FEV(3VP;1FP); MAR(2VP;2FP).

Regra 10 - Nó 26

SE $\text{THUR95_PINF} \geq 17,55$
E $91,8 \leq \text{PRECIP_PINF} < 230,2$
E $\text{NHUR95_PINF} < 14$
E $\text{NHNUR95_PINF} < 10,15$

ENTÃO TAXA_INF_M5 = 0

Distribuição: JAN(3VN;1FN); FEV(7VN;1FN); MAR(4VN); ABR(9VN;3FN); DEZ(3VN;1FN).

Regra 11 - Nó 27

SE $\text{THUR95_PINF} \geq 17,55$
E $91,8 \leq \text{PRECIP_PINF} < 230,2$
E $\text{NHUR95_PINF} < 14$
E $\text{NHNUR95_PINF} \geq 10,15$

ENTÃO TAXA_INF_M5 = 1

Distribuição: MAR(2VP;2FP); ABR(2VP); DEZ(3VP;1FP).

Especificidade: 84,9% **Suporte:** 6,6%

Precisão: 71,4% **Novidade:** 0,01
Sensitividade: 7,5% **Cobertura:** 2,7%
Especificidade: 99,2% **Suporte:** 2,2%

Precisão: 71,4% **Novidade:** 0,03
Sensitividade: 17,0% **Cobertura:** 6,6%
Especificidade: 97,7% **Suporte:** 4,9%

Precisão: 79,4% **Novidade:** 0,02
Sensitividade: 20,2% **Cobertura:** 17,6%
Especificidade: 88,7% **Suporte:** 14,3%

Precisão: 66,7% **Novidade:** 0,02
Sensitividade: 13,2% **Cobertura:** 5,5%
Especificidade: 97,7% **Suporte:** 3,8%

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

A Figura 38 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com baixa carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M5 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 3 (seção 3.5.1).

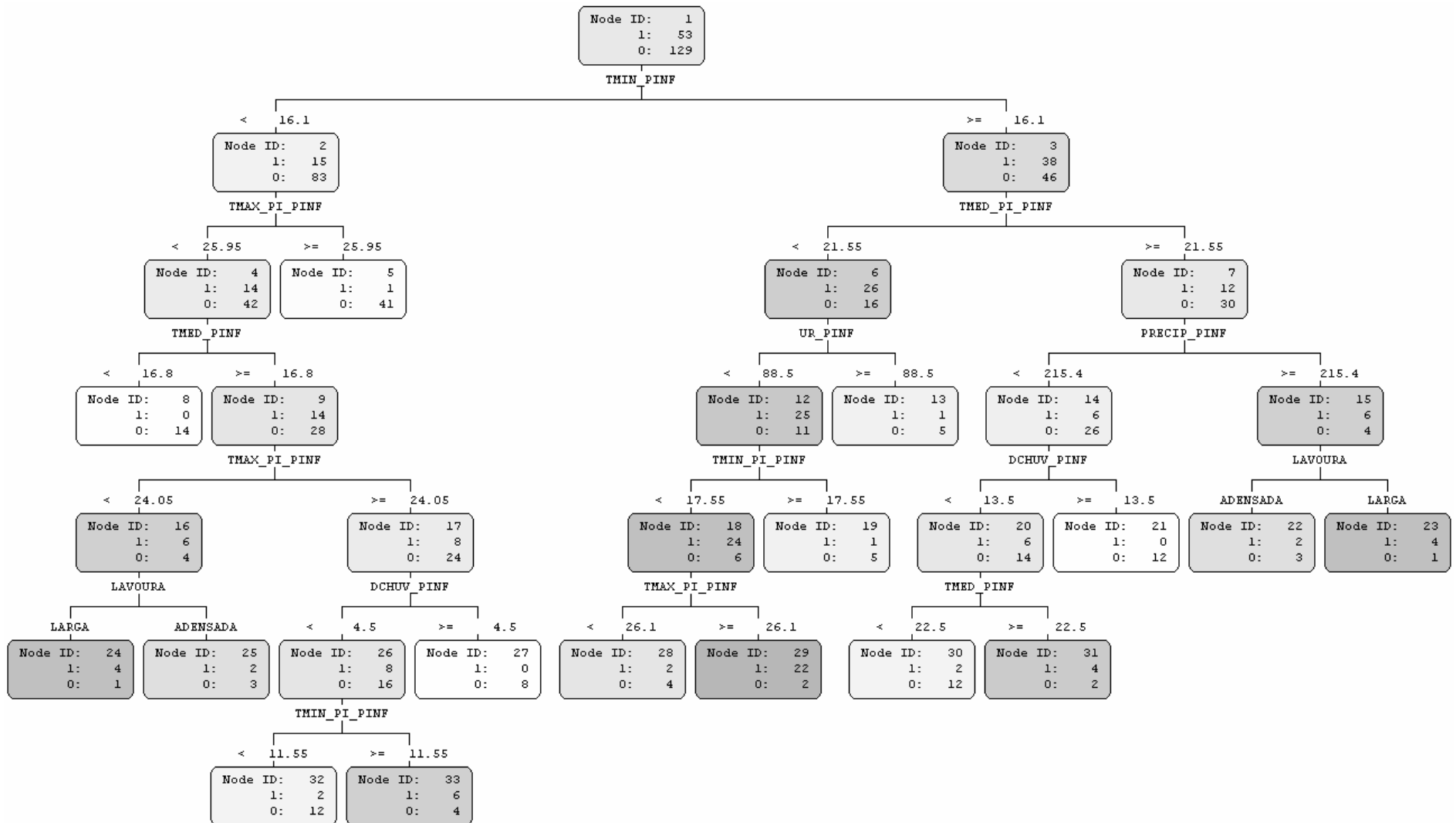


Figura 38: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 3; Geração: Enterprise Miner™.

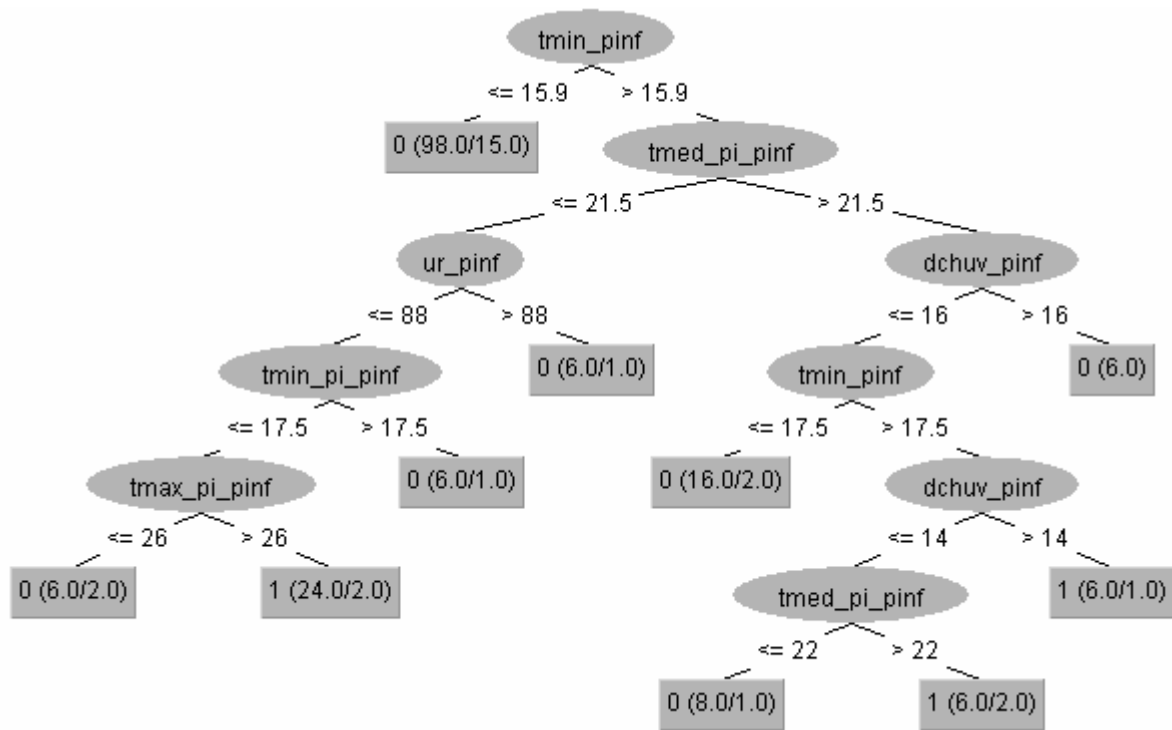


Figura 39: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M5; Seleção de atributos: Modelagem 3; Geração: Weka.

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 39. O atributo de teste no nó raiz e a divisão dos exemplos neste nível foram os mesmos do modelo do EM. A subárvore à esquerda do nó raiz foi a do modelo do EM podada até a raiz. As subárvores à esquerda das subárvores à direita do nó raiz dos dois modelos (p.ex. a partir do nó 6 na Figura 38) foram iguais, enquanto as subárvores à direita foram diferentes.

O modelo do EM, em comparação com o modelo do Weka, na validação cruzada, teve acurácia similar (70,3% contra 69,2%), mas sensibilidade (43,3% contra 26,3%) e confiabilidade positiva (50,6% contra 35,8%) melhores.

Esse desempenho melhor na sensibilidade e na confiabilidade positiva esteve relacionado com os verdadeiros positivos da subárvore à esquerda do nó raiz, que o modelo do Weka contabilizou como falsos negativos. Entretanto, esses casos se referem aos meses de junho, julho, agosto e setembro, de pouca importância no alerta da ferrugem do cafeeiro.

O modelo do EM também apresentou divisões com base no atributo LAVOURA (nós 15 e 16 da Figura 38) diferentemente do que seria esperado, com risco maior de taxas mais elevadas da ferrugem para lavouras largas do que para lavouras adensadas.

Sendo assim, conforme o exposto, e pelo modelo do Weka contar com um menor número de regras e de atributos de teste, optou-se por ele para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão.

As matrizes de confusão da árvore de decisão da Figura 39, obtidas pelos métodos de ressubstituição e de validação cruzada, são apresentadas na Tabela 41. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 42.

Tabela 41: Matrizes de confusão da árvore de decisão da Figura 39.

Ressubstituição			Validação cruzada			
TAXA_INF_M5	Preditada		TAXA_INF_M5	Preditada		
	1	0		1	0	
Verdadeira	1	31	22	1	14	39
	0	5	124	0	17	112

Tabela 42: Avaliação da árvore de decisão da Figura 39.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	85,2%	69,2% (2,0%)*
Taxa de erro	14,8%	30,8%
Sensitividade	58,5%	26,3% (7,3%)
Especificidade	96,1%	86,7% (3,7%)
Confiabilidade positiva	86,1%	35,8% (9,9%)
Confiabilidade negativa	84,9%	74,6% (1,8%)

* Desvio padrão da média.

As regras de classificação extraídas da árvore de decisão da Figura 39 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 43. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

As matrizes de confusão e a avaliação do modelo gerado no EM, bem como as regras de classificação extraídas da árvore de decisão da Figura 38 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

Tabela 43: Regras extraídas da árvore de decisão da Figura 39 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1	
SE TMIN_PINF \leq 15,9 ENTÃO TAXA_INF_M5 = 0	Precisão: 84,0% Novidade: 0,07 Sensitividade: 64,3% Cobertura: 53,8% Especificidade: 71,7% Suporte: 45,6%
Distribuição*: MAI(2VN); JUN(10VN;6FN); JUL(11VN;5FN); AGO(14VN;2FN); SET(15VN;1FN); OUT(14VN); NOV(13VN;1FN); DEZ(4VN).	
Regra 2	
SE TMIN_PINF > 15,9 E TMED_PI_PINF \leq 21,5 E UR_PINF > 88 ENTÃO TAXA_INF_M5 = 0	Precisão: 75,0% Novidade: 0,00 Sensitividade: 3,9% Cobertura: 3,3% Especificidade: 98,1% Suporte: 2,7%
Distribuição: JAN(1VN;1FN); ABR(4VN).	
Regra 3	
SE TMIN_PINF > 15,9 E TMED_PI_PINF > 21,5 E DCHUV_PINF > 16 ENTÃO TAXA_INF_M5 = 0	Precisão: 87,5% Novidade: 0,01 Sensitividade: 4,7% Cobertura: 3,3% Especificidade: 100% Suporte: 3,3%
Distribuição: JAN(2VN); FEV(2VN); MAR(2FN).	
Regra 4	
SE TMIN_PINF > 15,9 E TMED_PI_PINF \leq 21,5 E UR_PINF \leq 88 E TMIN_PI_PINF > 17,5 ENTÃO TAXA_INF_M5 = 0	Precisão: 75,0% Novidade: 0,00 Sensitividade: 3,9% Cobertura: 3,3% Especificidade: 98,1% Suporte: 2,7%
Distribuição: JAN(2VN); MAR(2VN); DEZ(1VN;1FN).	
Regra 5	
SE 15,9 < TMIN_PINF \leq 17,5 E TMED_PI_PINF > 21,5 E DCHUV_PINF \leq 16 ENTÃO TAXA_INF_M5 = 0	Precisão: 83,3% Novidade: 0,01 Sensitividade: 10,9% Cobertura: 8,8% Especificidade: 96,2% Suporte: 7,7%
Distribuição: JAN(4VN); FEV(4VN); MAR(2VN); ABR(1VN;1FN); DEZ(3VN;1FN).	
Regra 6	
SE TMIN_PINF > 15,9 E TMED_PI_PINF \leq 21,5 E UR_PINF \leq 88 E TMIN_PI_PINF \leq 17,5 E TMAX_PI_PINF \leq 26 ENTÃO TAXA_INF_M5 = 0	Precisão: 62,5% Novidade: 0,00 Sensitividade: 3,1% Cobertura: 3,3% Especificidade: 96,2% Suporte: 2,2%

Distribuição: MAI(3VN;1FN); DEZ(1VN;1FN).

Regra 7

SE TMIN_PINF > 15,9
E TMED_PI_PINF ≤ 21,5
E UR_PINF ≤ 88
E TMIN_PI_PINF ≤ 17,5
E TMAX_PI_PINF > 26

ENTÃO TAXA_INF_M5 = 1

Precisão: 88,5% **Novidade:** 0,08
Sensitividade: 41,5% **Cobertura:** 13,2%
Especificidade: 98,4% **Suporte:** 12,1%

Distribuição: JAN(2VP); FEV(2VP); MAR(2VP); ABR(5VP;1FP); MAI(9VP;1FP); DEZ(2VP).

Regra 8

SE TMIN_PINF > 17,5
E TMED_PI_PINF > 21,5
E 14 < DCHUV_PINF ≤ 16

ENTÃO TAXA_INF_M5 = 1

Precisão: 75,0% **Novidade:** 0,02
Sensitividade: 9,4% **Cobertura:** 3,3%
Especificidade: 99,2% **Suporte:** 2,7%

Distribuição: JAN(2VP); FEV(1VP;1FP); MAR(2VP).

Regra 9

SE TMIN_PINF > 17,5
E 21,5 < TMED_PI_PINF ≤ 22
E DCHUV_PINF ≤ 14

ENTÃO TAXA_INF_M5 = 0

Precisão: 80,0% **Novidade:** 0,01
Sensitividade: 5,4% **Cobertura:** 4,4%
Especificidade: 98,1% **Suporte:** 3,8%

Distribuição: JAN(2VN); FEV(1VN;1FN); ABR(4VN).

Regra 10

SE TMIN_PINF > 17,5
E TMED_PI_PINF > 22
E DCHUV_PINF ≤ 14

ENTÃO TAXA_INF_M5 = 1

Precisão: 62,5% **Novidade:** 0,01
Sensitividade: 7,5% **Cobertura:** 3,3%
Especificidade: 98,4% **Suporte:** 2,2%

Distribuição: MAR(4VP;2FP).

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

5.3.1.2 Discussão

No modelo gerado com a opção Modelagem 1 (Figura 35), o atributo de teste no nó raiz e a sua fronteira de decisão foram os mesmos dos modelos para as lavouras com alta carga pendente (Figura 24 e Figura 30). A ocorrência de mais de 25 dias desfavoráveis à infecção ($DDI_PINF > 25$) correspondeu a taxas de infecção menores do que 5 p.p., na maioria dos casos.

O somatório de horas noturnas de molhamento foliar no período de infecção ($SMT_NHNUR95_PINF$) foi incluído no modelo, sendo que valores acima de 157,5 h tiveram

efeito depressivo nas taxas de infecção. O somatório total de molhamento foliar no período de infecção (SMT_NHUR95_PINF) foi o atributo de teste da regra concorrente à regra principal do nó 2 da Figura 34, com praticamente o mesmo ganho de informação.

Com relação a isso, Silva-Acuña et al. (1998), utilizando análise de trilha na epidemiologia da ferrugem do cafeeiro, observaram efeito direto negativo do molhamento foliar diurno sobre a doença, coincidentemente no ano em que a carga pendente de frutos foi baixa.

A influência dos demais atributos sobre as taxas de infecção da ferrugem do cafeeiro é parecida com o que já foi discutido anteriormente, neste mesmo capítulo, ou no capítulo referente à análise da epidemia dessa doença com árvore de decisão (capítulo 4).

O modelo, com sete regras e baseado em seis atributos de teste, teve acurácia (70,3% - validação cruzada) equivalente à proporção de exemplos da classe majoritária (71%), que seria a acurácia estimada de um classificador que atribuísse a todos os exemplos a classe '0'. Entretanto, diferentemente deste último, que não resultaria em nenhum VP (verdadeiro positivo), o modelo da Figura 35 produziu 40 VP na ressubstituição e 19 VP na estimativa pela validação cruzada (Tabela 35).

A especificidade (84,5%) e a confiabilidade negativa (77,7%) foram melhores do que a sensibilidade (34,7%) e a confiabilidade positiva (35,3%), provavelmente em decorrência da distribuição desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 34). Vale mencionar que esses níveis de sensibilidade e de confiabilidade positiva ficaram bem abaixo dos obtidos pelos modelos para lavouras com alta carga pendente em que o atributo meta foi TAXA_INF_M10, cuja distribuição dos exemplos entre as classes no conjunto de treinamento foi exatamente a mesma (Tabela 15).

As predições para o período de dezembro a abril foram distribuídas entre quase todas as regras (Tabela 37). As regras de destaque foram:

- **Regra 1:** precisão (90,8%); sensibilidade (52,7%); especificidade (88,7%); novidade (0,09); cobertura (40,7%) e suporte (37,4%); cobriu corretamente exemplos de dezembro e janeiro, na maior parte dos casos: 7 VN e 1 FN.
- **Regra 2:** precisão (86,7%); sensibilidade (19,4%); especificidade (94,3%); novidade (0,03); cobertura (15,4%) e suporte (13,7%); cobriu corretamente vários exemplos no período de dezembro a abril: 20 VN e apenas 2 FN.

- **Regras 5 e 7:** precisão (81,3% e 75%); especificidade (96,2% para ambas); as demais medidas foram baixas; cobriram corretamente exemplos no período de fevereiro a abril, na maior parte dos casos: 11 VN e 3 FN.
- **Regra 6:** precisão (75%); sensibilidade (54,7%); especificidade (93%); novidade (0,10); cobertura (20,9%) e suporte (15,9%); cobriu corretamente exemplos de dezembro, janeiro, março e abril, na maior parte dos casos: 17 VP e 5 FP.

No modelo gerado com a opção Modelagem 2 (Figura 36), o atributo de teste no nó raiz foi o mesmo dos modelos para lavouras com alta carga pendente (Figura 25 e Figura 31), mas com a fronteira de decisão 1°C acima. Temperaturas médias diárias durante o molhamento foliar (THUR95_PINF) abaixo de 17,55 °C foram menos favoráveis e maiores ou iguais a esse limite foram mais favoráveis à infecção.

A média do número de horas de molhamento foliar no período de infecção (NHUR95_PINF) apareceu com influência no modelo, com valores a partir de 14 h causando efeito depressivo nas taxas de infecção (Figura 36, nó 15). Novamente, o modelo parece ter capturado efeito negativo do molhamento foliar sobre a doença semelhante ao observado por Silva-Acuña et al. (1998) em ano de baixa carga pendente de frutos.

A influência dos demais atributos sobre as taxas de infecção da ferrugem do cafeeiro já foi discutida anteriormente.

O modelo, com 11 regras e 7 atributos de teste, também teve acurácia (71,8% - validação cruzada) equivalente à proporção de exemplos da classe majoritária (71%). Mas, diferentemente de um classificador que atribuísse a todos os exemplos a classe '0', o que resultaria somente em VN e FN, o modelo produziu 38 VP na resubstituição e 25 VP na estimativa pela validação cruzada (Tabela 38).

A especificidade (82,1%) e a confiabilidade negativa (79,4%) foram melhores do que a sensibilidade (46,3%) e a confiabilidade positiva (49,9%), provavelmente em decorrência da distribuição desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 34). Novamente, esses níveis de sensibilidade e de confiabilidade positiva ficaram bem abaixo dos obtidos pelos modelos para lavouras com alta carga pendente em que o atributo meta foi TAXA_INF_M10, cuja distribuição dos exemplos entre as classes foi igual (Tabela 15).

As predições para o período de dezembro a abril foram distribuídas entre oito das onze regras (Tabela 40). As regras de destaque foram:

- **Regra 1:** precisão (93,5%); sensibilidade (32,6%); especificidade (96,2%); novidade (0,06); cobertura (24,2%) e suporte (23,1%); classificou corretamente casos do mês de dezembro, na maioria das vezes: 5 VN e 1 FN.
- **Regra 3:** precisão (87,5%); sensibilidade (24,5%); especificidade (99,2%); novidade (0,05); cobertura (7,7%) e suporte (7,1%); classificou corretamente casos do mês de março: 4 VP.
- **Regra 4:** precisão (90%); especificidade máxima (100%); as demais medidas foram baixas; classificou corretamente casos do mês de janeiro: 8 VN.
- **Regras 9 e 11:** precisão (71,4% e 66,7%); sensibilidade (17% e 13,2%); especificidade (97,7% para ambas); as demais medidas foram baixas; cobriram corretamente o período de dezembro a abril, na maior parte dos casos: 16 VP e 6 FP.
- **Regra 10:** precisão (79,4%); sensibilidade (20,2%); especificidade (88,7%); novidade (0,02); cobertura (17,6%) e suporte (14,3%); cobriu corretamente o período de dezembro a abril, na maior parte dos casos: 26 VN e 6 FN.

No modelo gerado com a opção Modelagem 3 (Figura 39), o atributo de teste no nó raiz e a sua fronteira de decisão (TMIN_PINF; divisão em 15,9 °C) foram os mesmos do modelo para lavouras com alta carga pendente em que o atributo meta foi também TAXA_INF_M5 (Figura 28).

Ainda comparando com a árvore de decisão da Figura 28, a subárvore à esquerda do nó raiz foi podada até a raiz. Na subárvore à direita, houve ramificação a partir do nó em que a decisão se baseou na temperatura média durante o período de incubação (TMED_PI_PINF).

Conforme indicou um dos testes incluídos nessa ramificação, altos valores de umidade relativa média diária ($UR_PINF > 88\%$) tiveram efeito negativo sobre as taxas de infecção. Analisando-se as regras concorrentes ao nó correspondente da Figura 38 (nó 6), verificou-se que o atributo de teste PRECIP_PINF (divisão em 160,6 mm) foi escolhido como opcional, proporcionando praticamente o mesmo ganho de informação que a regra principal.

Sendo assim, aqueles períodos de alta umidade relativa estiveram relacionados com períodos bastante chuvosos, o que parece indicar que a chuva foi o principal fator associado a esse efeito negativo nas taxas de infecção da ferrugem do cafeeiro, conduzindo os esporos ao chão em vez de disseminá-los para novos sítios de infecção (KUSHALAPPA, 1989a).

A influência dos demais atributos sobre as taxas de infecção da ferrugem do cafeeiro já foi discutida anteriormente.

O modelo, com 10 regras e 6 atributos de teste, também teve acurácia (69,2% - validação cruzada) equivalente à proporção de exemplos da classe majoritária (71%). Mas, em vez de atribuir a todos os exemplos a classe '0', o modelo produziu 31 VP na ressubstituição e 14 VP na estimativa pela validação cruzada (Tabela 41).

A especificidade (86,7%) e a confiabilidade negativa (74,6%) foram melhores do que a sensibilidade (26,3%) e a confiabilidade positiva (35,8%), provavelmente em decorrência da distribuição desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 34). Mais uma vez, esses níveis de sensibilidade e de confiabilidade positiva ficaram bem abaixo dos obtidos pelos modelos para lavouras com alta carga pendente em que o atributo meta foi TAXA_INF_M10, cuja distribuição dos exemplos entre as classes foi igual (Tabela 15).

As predições para o período de dezembro a abril foram distribuídas entre todas as regras (Tabela 43). As regras de destaque foram:

- **Regra 1:** precisão (84%); sensibilidade (64,3%); especificidade (71,7%); novidade (0,07); cobertura (53,8%) e suporte (45,6%); classificou corretamente casos do mês de dezembro: 4 VN.
- **Regras 2 e 4:** precisão (75% para ambas); especificidade (98,1% para ambas); as demais medidas foram baixas; classificaram corretamente casos de dezembro, janeiro, março e abril, na maior parte dos casos: 10 VN e 2 FN.
- **Regras 3, 5 e 9:** precisão (87,5%, 83,3% e 80%); especificidade (100%, 96,2% e 98,1%); as demais medidas foram baixas; classificaram corretamente casos de dezembro a abril, na maior parte dos casos: 25 VN e 5 FN.
- **Regra 7:** precisão (88,5%); sensibilidade (41,5%); especificidade (98,4%); novidade (0,08); cobertura (13,2%) e suporte (12,1%); cobriu corretamente casos no período de dezembro a abril: 13 VP e apenas 1 FP.
- **Regra 8:** precisão (75%); especificidade (99,2%); as demais medidas foram baixas; classificou corretamente casos de janeiro a março, na maior parte dos casos: 5 VP e 1 FP.

Comparando-se os modelos obtidos com as três opções de seleção de atributos preditivos, a árvore de decisão gerada com a opção Modelagem 2 (Figura 36) se destacou

pelos melhores valores de sensibilidade e de confiabilidade positiva, com os menores desvios padrões das médias para essas duas medidas. As demais medidas foram similares.

Contudo, analisando-se o desempenho de cada modelo no período de dezembro a abril, verificou-se que o modelo da opção Modelagem 3 (Figura 39) superou o modelo da opção Modelagem 2 em dois meses, equiparando-se nos demais. Em dezembro, o desempenho foi diferente, mas equivalente: 3 VP, 1 FP, 8 VN e 2 FN do modelo da opção Modelagem 2 contra 2 VP, 9 VN e 3 FN do modelo da opção Modelagem 3. Em janeiro e fevereiro, os dois modelos tiveram o mesmo desempenho: 4 VP, 11 VN e 1 FN em janeiro; e 3 VP, 1 FP, 7 VN e 1 FN em fevereiro. Em março, o modelo da opção Modelagem 3 foi melhor: 8 VP, 2 FP e 6 VN contra 8 VP, 4 FP e 4 VN. Em abril, também: 5 VP, 1 FP, 9 VN e 1 FN contra 3 VP, 10 VN e 3 FN.

O modelo da opção Modelagem 1 não obteve desempenho superior aos dos outros dois modelos, sem justificar, portanto, o trabalho mais elaborado na preparação dos atributos preditivos especiais específicos desta opção de seleção de atributos.

Portanto, as árvores de decisão da Figura 36 e da Figura 39 foram consideradas as melhores opções como modelos de alerta da ferrugem do cafeeiro para predizer quando a taxa de infecção da doença for atingir ou ultrapassar 5 p.p. em lavouras de café com baixa carga pendente de frutos.

As melhores regras de cada modelo, considerando os verdadeiros positivos e os verdadeiros negativos para o período de dezembro a abril, além das outras medidas de avaliação de regras, foram:

- Modelagem 1 - VP: Regra 6; VN: Regras 1, 2, 5 e 7.
- Modelagem 2 - VP: Regras 3, 9 e 11; VN: Regras 1, 4 e 10.
- Modelagem 3 - VP: Regras 7 e 8; VN: Regras 1 a 5 e 9.

5.3.2 Alerta quando a taxa de infecção for atingir ou ultrapassar 10 p.p.

5.3.2.1 Resultados

A Figura 40 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com baixa carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M10 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 1 (seção 3.5.1).

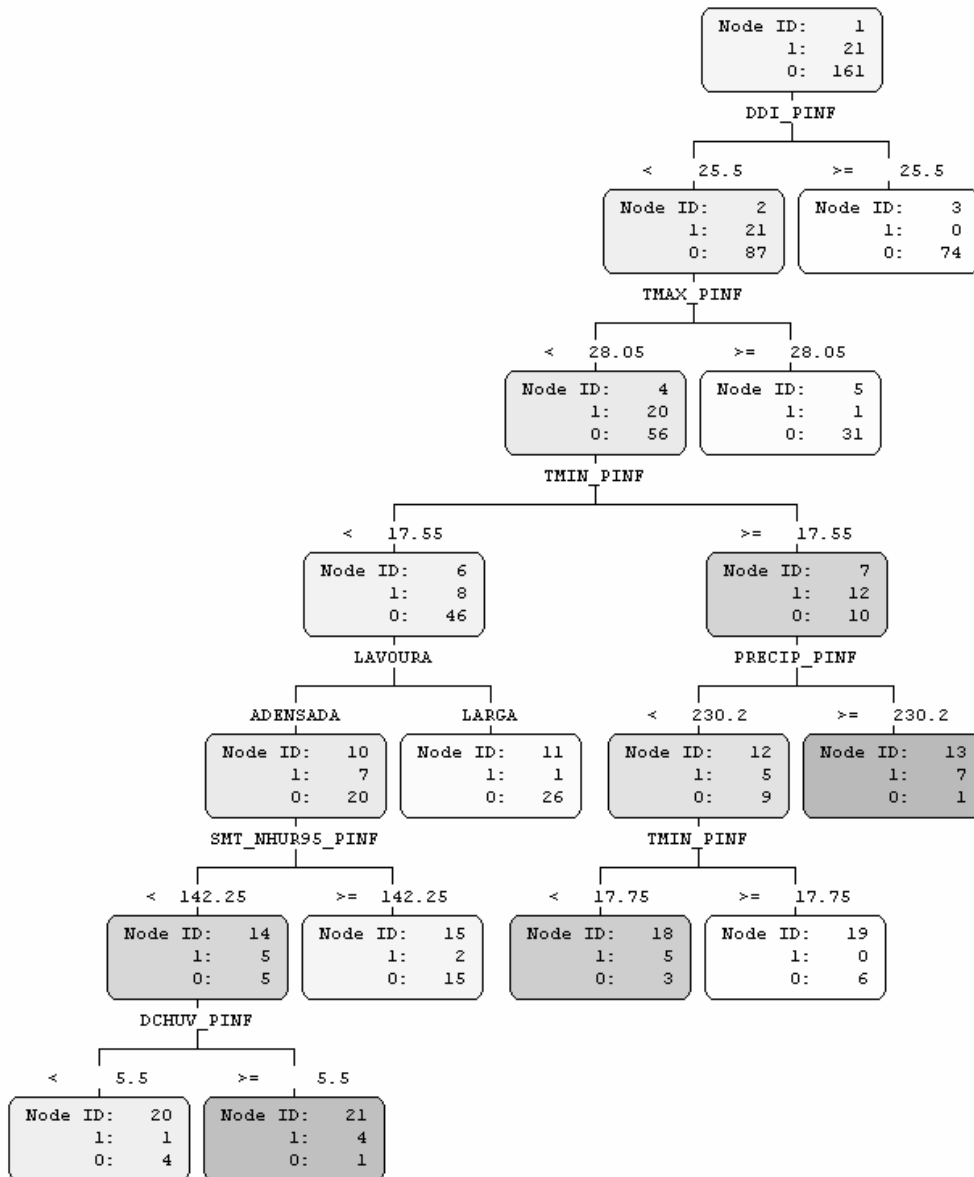


Figura 40: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 1; Geração: Enterprise Miner™.

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 41. Ela foi bem diferente da apresentada na Figura 40 – até mesmo o atributo de teste no nó raiz foi diferente em cada modelo. O modelo da Figura 41 não incluiu nenhum atributo preditivo especial (seção 3.4.3) específico da Modelagem 1. Entretanto, não foi igual à árvore gerada com a opção de seleção de atributos Modelagem 2 (Figura 43, pg. 156). A razão disso foi a opção de poda *subtree raising* do Weka, que eliminou da árvore o nó raiz e o teste sobre o atributo DDI_PINF após o processo de indução.

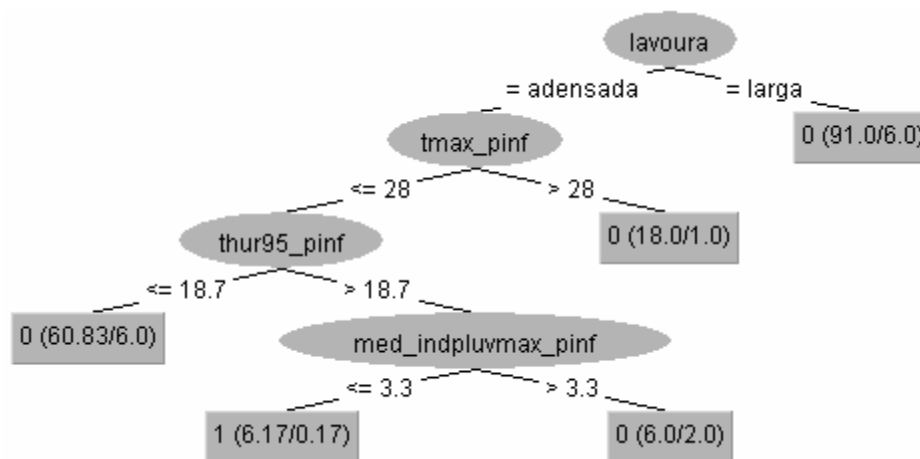


Figura 41: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 1; Geração: Weka.

Na avaliação pela ressubstituição, a acurácia do modelo do EM foi melhor (94,5% contra 91,8%). A confiabilidade positiva do modelo do Weka foi total (100%), mas sua sensibilidade foi baixa (28,6%).

O modelo do EM apresentou melhor equilíbrio entre as medidas: 94,5%, 76,2%, 96,9%, 76,2% e 96,9% em comparação com 91,8%, 28,6%, 100%, 100% e 91,5%, na seqüência das medidas apresentadas na Tabela 45, desconsiderando a taxa de erro.

O modelo do Weka apresentou melhor acurácia na validação cruzada (86,8% contra 82,4%). A sensibilidade (15% contra 10%) e a confiabilidade positiva (15,8% contra 7,5%) do modelo do EM foram melhores, mas deve-se considerar os altos valores dos desvios padrões das médias dessas duas medidas (7,6% e 10,1%, respectivamente).

Considerando o exposto em relação à avaliação dos dois modelos e pelo modelo do Weka não incluir atributo especial, optou-se pelo modelo gerado no EM para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão.

Tabela 44: Matrizes de confusão da árvore de decisão da Figura 40.

		Ressubstituição		Validação cruzada		
TAXA_INF_M10	Verdadeira	Preditada		TAXA_INF_M10	Preditada	
		1	0		1	0
1	1	16	5	1	3	18
0	0	5	156	0	14	147

Tabela 45: Avaliação da árvore de decisão da Figura 40.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	94,5%	82,4% (1,8%)*
Taxa de erro	5,5%	17,6%
Sensitividade	76,2%	15,0% (7,6%)
Especificidade	96,9%	91,3% (1,9%)
Confiabilidade positiva	76,2%	15,8% (10,1%)
Confiabilidade negativa	96,9%	89,2% (1,0%)

* Desvio padrão da média.

As matrizes de confusão da árvore de decisão da Figura 40, obtidas pelos métodos de ressubstituição e de validação cruzada, são apresentadas na Tabela 44. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 45.

As regras de classificação extraídas da árvore de decisão da Figura 40 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 46. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

As matrizes de confusão e a avaliação do modelo gerado no Weka, bem como as regras de classificação extraídas da árvore de decisão da Figura 41 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

Tabela 46: Regras extraídas da árvore de decisão da Figura 40 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1 - Nó 3	
SE DDI_PINF \geq 25,5 ENTÃO TAXA_INF_M10 = 0	Precisão: 98,7% Novidade: 0,05 Sensitividade: 46,0% Cobertura: 40,7% Especificidade: 100% Suporte: 40,7%
Distribuição*: JAN(2VN); JUL(12VN); AGO(16VN); SET(16VN); OUT(14VN); NOV(8VN); DEZ(6VN).	
Regra 2 - Nó 5	
SE DDI_PINF < 25,5 E TMAX_PINF \geq 28,05 ENTÃO TAXA_INF_M10 = 0	Precisão: 94,1% Novidade: 0,01 Sensitividade: 19,3% Cobertura: 17,6% Especificidade: 95,2% Suporte: 17,0%
Distribuição: JAN(4VN); FEV(2VN); MAR(8VN); ABR(8VN); MAI(3VN;1FN); JUN(2VN); NOV(2VN); DEZ(2VN).	
Regra 3 - Nó 11	
	Precisão: 93,1% Novidade: 0,01

SE DDI_PINF < 25,5
 E TMAX_PINF < 28,05
 E TMIN_PINF < 17,55
 E LAVOURA = LARGA
ENTÃO TAXA_INF_M10 = 0

Distribuição: JAN(3VN); FEV(2VN); MAR(1VN); ABR(2VN); MAI(5VN); JUN(6VN;1FN); JUL(2VN); NOV(2VN); DEZ(3VN).

Sensitividade: 16,1% **Cobertura:** 14,8%
Especificidade: 95,2% **Suporte:** 14,3%

Regra 4 - Nó 13

SE DDI_PINF < 25,5
 E TMAX_PINF < 28,05
 E TMIN_PINF ≥ 17,55
 E PRECIP_PINF ≥ 230,2
ENTÃO TAXA_INF_M10 = 1

Distribuição: JAN(3VP;1FP); FEV(2VP); MAR(2VP).

Precisão: 80,0% **Novidade:** 0,03
Sensitividade: 33,3% **Cobertura:** 4,4%
Especificidade: 99,4% **Suporte:** 3,8%

Regra 5 - Nó 15

SE DDI_PINF < 25,5
 E TMAX_PINF < 28,05
 E TMIN_PINF < 17,55
 E LAVOURA = ADENSADA
 E SMT_NHUR95_PINF ≥ 142,25
ENTÃO TAXA_INF_M10 = 0

Distribuição: JAN(2VN;1FN); FEV(2VN); MAR(1VN); ABR(1VN); MAI(2VN); JUN(3VN); JUL(1VN;1FN); NOV(1VN); DEZ(2VN).

Precisão: 84,2% **Novidade:** 0,00
Sensitividade: 9,3% **Cobertura:** 9,3%
Especificidade: 90,5% **Suporte:** 8,2%

Regra 6 - Nó 18

SE DDI_PINF < 25,5
 E TMAX_PINF < 28,05
 E 17,55 ≤ TMIN_PINF < 17,75
 E PRECIP_PINF < 230,2
ENTÃO TAXA_INF_M10 = 1

Distribuição: FEV(1VP;1FP); MAR(2VP); ABR(1VP;1FP); MAI(1VP;1FP).

Precisão: 60,0% **Novidade:** 0,02
Sensitividade: 23,8% **Cobertura:** 4,4%
Especificidade: 98,1% **Suporte:** 2,7%

Regra 7 - Nó 19

SE DDI_PINF < 25,5
 E TMAX_PINF < 28,05
 E TMIN_PINF ≥ 17,75
 E PRECIP_PINF < 230,2
ENTÃO TAXA_INF_M10 = 0

Distribuição: FEV(2VN); MAR(2VN); ABR(2VN).

Precisão: 87,5% **Novidade:** 0,00
Sensitividade: 3,7% **Cobertura:** 3,3%
Especificidade: 100% **Suporte:** 3,3%

Regra 8 - Nó 20

SE DDI_PINF < 25,5
 E TMAX_PINF < 28,05
 E TMIN_PINF < 17,55

Precisão: 71,4% **Novidade:** 0,00
Sensitividade: 2,5% **Cobertura:** 2,7%
Especificidade: 95,2% **Suporte:** 2,2%

E LAVOURA = ADENSADA
 E SMT_NHUR95_PINF < 142,25
 E DCHUV_PINF < 5,5
ENTÃO TAXA_INF_M10 = 0

Distribuição: MAI(1VN); JUN(3VN;1FN).

Regra 9 - Nó 21

SE DDI_PINF < 25,5
 E TMAX_PINF < 28,05
 E TMIN_PINF < 17,55
 E LAVOURA = ADENSADA
 E SMT_NHUR95_PINF < 142,25
 E DCHUV_PINF ≥ 5,5
ENTÃO TAXA_INF_M10 = 1

Precisão: 71,4% **Novidade:** 0,02
Sensitividade: 19,0% **Cobertura:** 2,7%
Especificidade: 99,4% **Suporte:** 2,2%

Distribuição: ABR(1VP); MAI(2VP); NOV(1VP); DEZ(1FP).

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

A Figura 42 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com baixa carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M10 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 2 (seção 3.5.1).

A regra equivalente (*surrogate rule*) à regra do nó raiz foi baseada no atributo TMIN_PINF: TMIN_PINF < 16,35 equivalendo a THUR95_PINF < 18,15 e TMIN_PINF ≥ 16,35 equivalendo a THUR95_PINF ≥ 18,15. Esta regra foi utilizada para classificar os dois exemplos de agosto de 2000, em que não houve períodos de molhamento foliar, com um mínimo de seis horas, no período de infecção correspondente (NHUR95_PINF = 0) e, portanto, os valores do atributo THUR95_PINF foram nulos. Os dois exemplos foram classificados no nó 2 (TMIN_PINF = 10,4 °C) como da classe ‘0’.

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 43. Ela foi bem diferente da apresentada na Figura 42 – até mesmo o atributo de teste no nó raiz foi diferente em cada modelo. Os exemplos de agosto de 2000 também foram classificados como da classe ‘0’, um deles (lavoura adensada) de acordo com o peso estabelecido na árvore (peso 0,63 para classificar como ‘0’ e peso 0,38 para classificar como ‘1’).

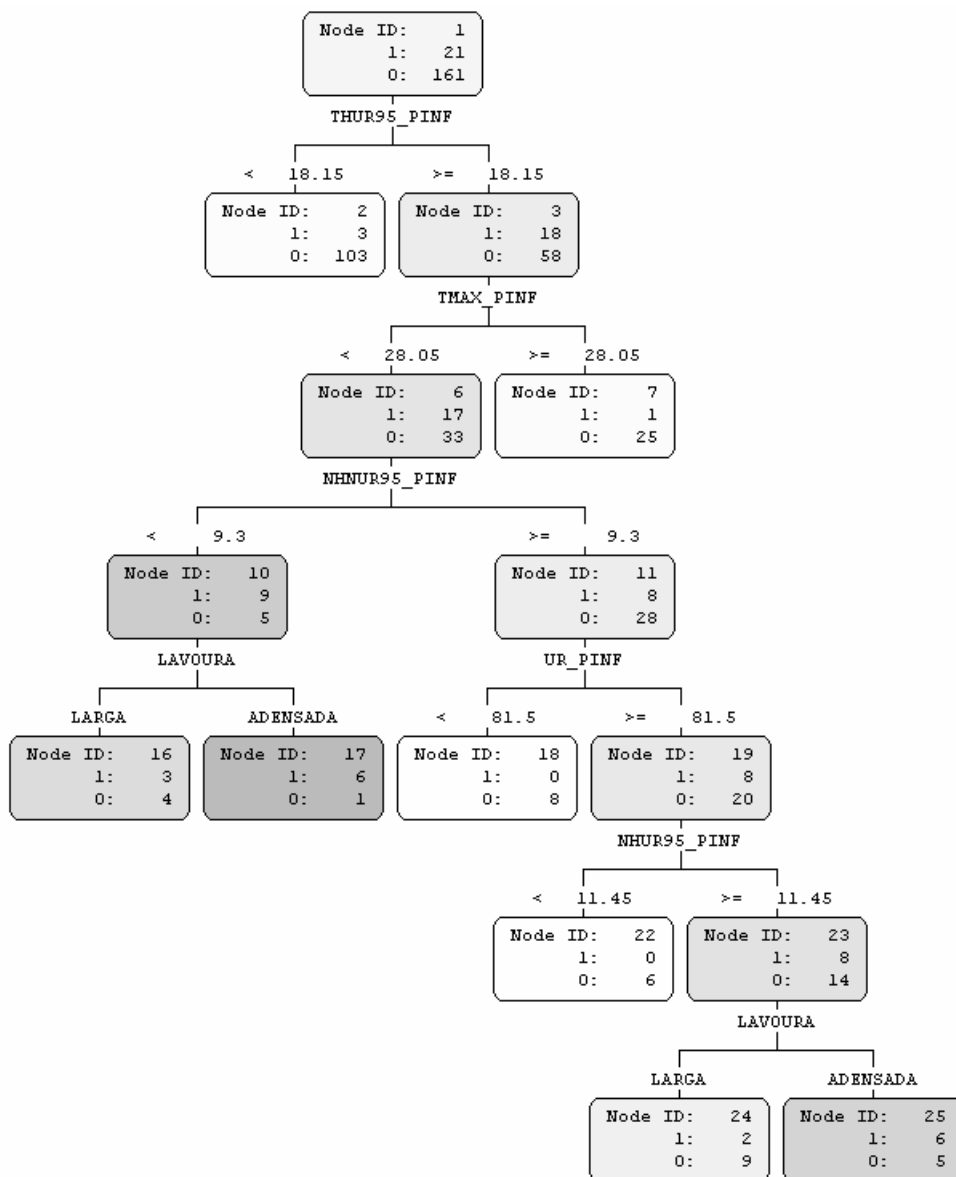


Figura 42: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 2; Geração: Enterprise Miner™.

O modelo do Weka teve melhor avaliação na ressubstituição (acurácia = 94,5% contra 91,8%), enquanto o modelo do EM foi melhor avaliado na validação cruzada (acurácia = 88,5% contra 86,3%). Houve diferença considerável a favor do modelo do EM, na validação cruzada, com relação à sensibilidade (30% contra 18,3%) e à confiabilidade positiva (50% contra 28,3%), mesmo considerando os altos desvios padrões das médias dessas duas medidas (8,2% e 14,9%, respectivamente). O modelo do EM apresentou melhor equilíbrio entre as medidas na validação cruzada: 88,5%, 30%, 96,3%, 50% e 91,3% em comparação com 86,3%,

18,3%, 95,1%, 28,3% e 90,1%, na seqüência das medidas apresentadas na Tabela 48, desconsiderando a taxa de erro.

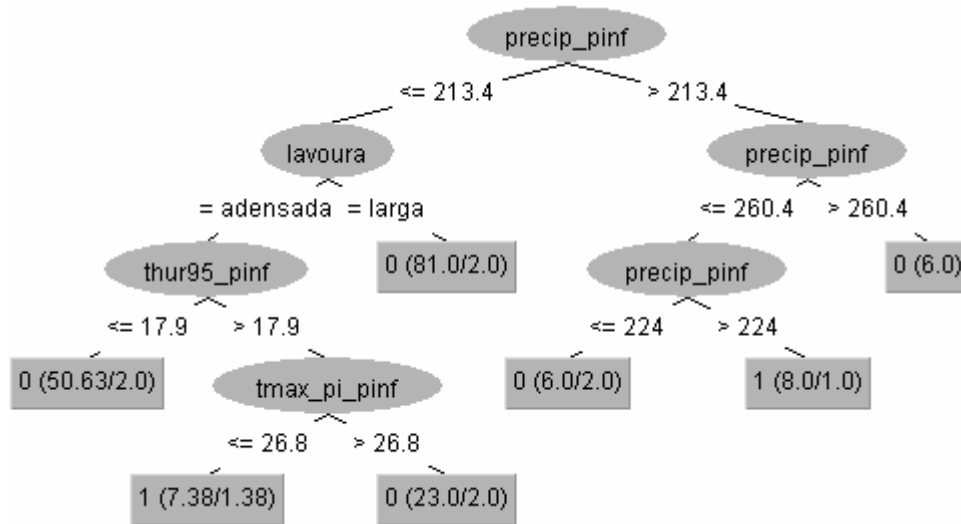


Figura 43: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 2; Geração: Weka.

Tabela 47: Matrizes de confusão da árvore de decisão da Figura 42.

Resubstituição			Validação cruzada				
TAXA_INF_M10	Predita		TAXA_INF_M10	Predita			
	1	0		1	0		
Verdadeira	1	12	9	Verdadeira	1	6	15
	0	6	155		0	6	155

Tabela 48: Avaliação da árvore de decisão da Figura 42.

Medida de avaliação	Método de estimativa	
	Resubstituição	Validação cruzada
Acurácia	91,8%	88,5% (1,8%)*
Taxa de erro	8,2%	11,5%
Sensitividade	57,1%	30,0% (8,2%)
Especificidade	96,3%	96,3% (1,0%)
Confiabilidade positiva	66,7%	50,0% (14,9%)
Confiabilidade negativa	94,5%	91,3% (1,2%)

* Desvio padrão da média.

Pela melhor avaliação na validação cruzada, optou-se pelo modelo gerado no EM para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão.

As matrizes de confusão da árvore de decisão da Figura 42, obtidas pelos métodos de resubstituição e de validação cruzada, são apresentadas na Tabela 47. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 48.

As regras de classificação extraídas da árvore de decisão da Figura 42 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 49. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

As matrizes de confusão e a avaliação do modelo gerado no Weka, bem como as regras de classificação extraídas da árvore de decisão da Figura 43 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

Tabela 49: Regras extraídas da árvore de decisão da Figura 42 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1 - Nó 2	
SE THUR95_PINF < 18,15 ENTÃO TAXA_INF_M10 = 0	Precisão: 96,3% Novidade: 0,05 Sensitividade: 64,0% Cobertura: 58,2% Especificidade: 85,7% Suporte: 56,6%
Distribuição*: JAN(2VN); FEV(2VN); ABR(2VN); MAI(2VN); JUN(13VN;1FN); JUL(15VN;1FN); AGO(16VN); SET(16VN); OUT(14VN); NOV(13VN;1FN); DEZ(8VN).	
Regra 2 - Nó 7	
SE THUR95_PINF ≥ 18,15 E TMAX_PINF ≥ 28,05 ENTÃO TAXA_INF_M10 = 0	Precisão: 92,9% Novidade: 0,01 Sensitividade: 15,5% Cobertura: 14,3% Especificidade: 95,2% Suporte: 13,7%
Distribuição: JAN(4VN); FEV(2VN); MAR(8VN); ABR(6VN); MAI(3VN;1FN); DEZ(2VN).	
Regra 3 - Nó 16	
SE THUR95_PINF ≥ 18,15 E TMAX_PINF < 28,05 E NHNUR95_PINF < 9,3 E LAVOURA = LARGA ENTÃO TAXA_INF_M10 = 0	Precisão: 55,6% Novidade: -0,01 Sensitividade: 2,5% Cobertura: 3,8% Especificidade: 85,7% Suporte: 2,2%
Distribuição: JAN(1FN); FEV(1VN); MAR(2FN); ABR(1VN); MAI(1VN); JUN(1VN).	
Regra 4 - Nó 17	
SE THUR95_PINF ≥ 18,15 E TMAX_PINF < 28,05	Precisão: 77,8% Novidade: 0,03 Sensitividade: 28,6% Cobertura: 3,8%

E NHNUR95_PINF < 9,3
E LAVOURA = ADENSADA
ENTÃO TAXA_INF_M10 = 1

Distribuição: JAN(1VP); FEV(1FP); MAR(2VP); ABR(1VP) ; MAI(1VP); JUN(1VP).

Regra 5 - Nó 18

SE THUR95_PINF ≥ 18,15
E TMAX_PINF < 28,05
E NHNUR95_PINF ≥ 9,3
E UR_PINF < 81,5
ENTÃO TAXA_INF_M10 = 0

Distribuição: JAN(2VN); MAI(2VN); DEZ(4VN).

Regra 6 - Nó 22

SE THUR95_PINF ≥ 18,15
E TMAX_PINF < 28,05
E NHNUR95_PINF ≥ 9,3
E UR_PINF ≥ 81,5
E NHUR95_PINF < 11,45
ENTÃO TAXA_INF_M10 = 0

Distribuição: FEV(2VN); MAR(2VN); MAI(2VN).

Regra 7 - Nó 24

SE THUR95_PINF ≥ 18,15
E TMAX_PINF < 28,05
E NHNUR95_PINF ≥ 9,3
E UR_PINF ≥ 81,5
E NHUR95_PINF ≥ 11,45
E LAVOURA = LARGA
ENTÃO TAXA_INF_M10 = 0

Distribuição: JAN(3VN); FEV(2FN); MAR(1VN); ABR(3VN) ; MAI(2VN).

Regra 8 - Nó 25

SE THUR95_PINF ≥ 18,15
E TMAX_PINF < 28,05
E NHNUR95_PINF ≥ 9,3
E UR_PINF ≥ 81,5
E NHUR95_PINF ≥ 11,45
E LAVOURA = ADENSADA
ENTÃO TAXA_INF_M10 = 1

Distribuição: JAN(2VP;1FP); FEV(1VP;1FP); MAR(1FP); ABR(1VP;2FP) ; MAI(2VP).

Especificidade: 99,4% **Suporte:** 3,3%

Precisão: 90,0% **Novidade:** 0,01
Sensitividade: 5,0% **Cobertura:** 4,4%
Especificidade: 100% **Suporte:** 4,4%

Precisão: 87,5% **Novidade:** 0,00
Sensitividade: 3,7% **Cobertura:** 3,3%
Especificidade: 100% **Suporte:** 3,3%

Precisão: 76,9% **Novidade:** 0,00
Sensitividade: 5,6% **Cobertura:** 6,0%
Especificidade: 90,5% **Suporte:** 4,9%

Precisão: 53,8% **Novidade:** 0,03
Sensitividade: 28,6% **Cobertura:** 6,0%
Especificidade: 96,9% **Suporte:** 3,3%

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

A Figura 44 apresenta a árvore de decisão, gerada no Enterprise Miner™ (EM), para alertas da ferrugem do cafeeiro em lavouras com baixa carga pendente, em que o atributo meta foi a taxa de infecção binária TAXA_INF_M10 (seção 3.4.1) e os atributos preditivos foram escolhidos segundo a opção de seleção de atributos Modelagem 3 (seção 3.5.1).

A árvore de decisão correspondente gerada no Weka é apresentada na Figura 45. Ela pode ser vista como a árvore da Figura 44 podada até a raiz na subárvore à esquerda do nó raiz. Os testes sobre o atributo PRECIP_PINF nos nós correspondentes foram parecidos e produziram a mesma distribuição dos exemplos na classificação do conjunto de treinamento.

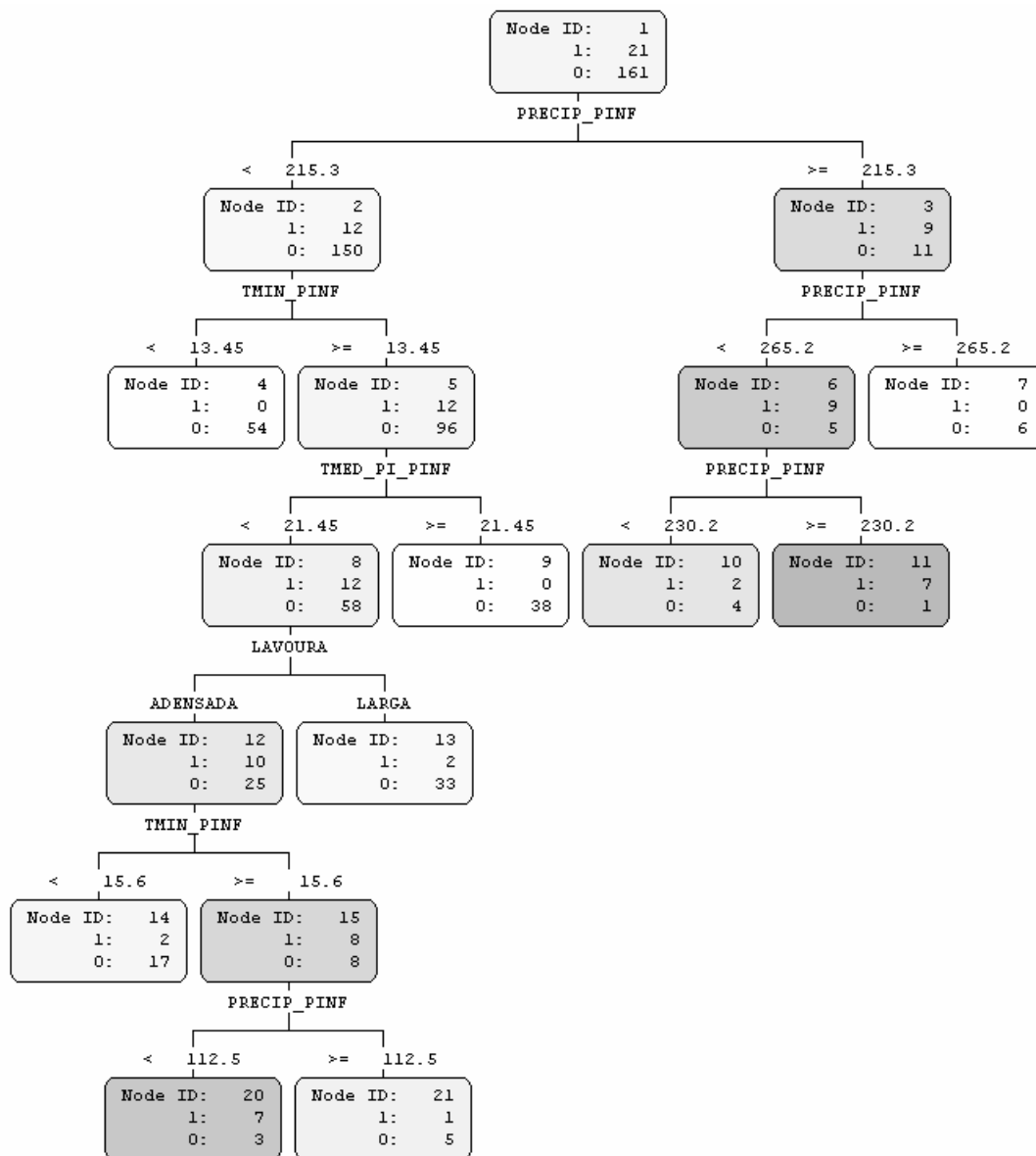


Figura 44: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 3; Geração: Enterprise Miner™.

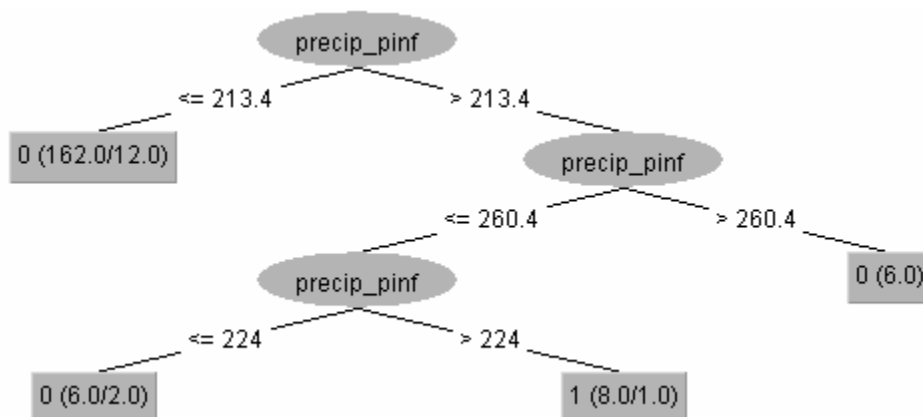


Figura 45: Árvore de decisão para alerta da ferrugem do cafeeiro em lavouras com baixa carga pendente. Atributo meta: TAXA_INF_M10; Seleção de atributos: Modelagem 3; Geração: Weka.

O modelo do EM teve desempenho melhor na ressubstituição (acurácia = 94% contra 91,8%), com o dobro da sensibilidade (66,7% contra 33,3%). Na validação cruzada, os desempenhos foram parecidos (acurácia = 86,9% do modelo do EM contra 86,2%).

A subárvore à esquerda do nó raiz do modelo do EM contabilizou verdadeiros positivos em um período importante de evolução da ferrugem do cafeeiro, para os meses de novembro, março e abril. No modelo do Weka, ao contrário, esses casos foram classificados de maneira incorreta e contabilizados como falsos negativos.

Por este motivo, e também considerando o exposto em relação à avaliação dos dois modelos, optou-se pelo modelo gerado no EM para a apresentação de sua avaliação completa e da avaliação individual de cada regra extraída da árvore de decisão.

As matrizes de confusão da árvore de decisão da Figura 44, obtidas pelos métodos de ressubstituição e de validação cruzada, são apresentadas na Tabela 50. A avaliação do modelo, pelos mesmos métodos de estimativa, é apresentada na Tabela 51.

Tabela 50: Matrizes de confusão da árvore de decisão da Figura 44.

Ressubstituição				Validação cruzada			
TAXA_INF_M10	Preditada			TAXA_INF_M10	Preditada		
	1	0			1	0	
Verdadeira	1	14	7	Verdadeira	1	4	17
	0	4	157		0	7	154

Tabela 51: Avaliação da árvore de decisão da Figura 44.

Medida de avaliação	Método de estimativa	
	Ressubstituição	Validação cruzada
Acurácia	94,0%	86,9% (2,1%)*
Taxa de erro	6,0%	13,1%
Sensitividade	66,7%	20,0% (8,2%)
Especificidade	97,5%	95,7% (1,3%)
Confiabilidade positiva	77,8%	35,0% (15,0%)
Confiabilidade negativa	95,7%	90,1% (1,2%)

* Desvio padrão da média.

As regras de classificação extraídas da árvore de decisão da Figura 44 e a avaliação individual de cada uma dessas regras, com base na classificação do conjunto de treinamento, estão apresentadas na Tabela 52. Junto com cada regra, também é apresentada a distribuição mensal das predições em relação aos exemplos do conjunto de treinamento.

As matrizes de confusão e a avaliação do modelo gerado no Weka, bem como as regras de classificação extraídas da árvore de decisão da Figura 45 e a avaliação individual de cada uma dessas regras, podem ser encontradas no CD-ROM anexo.

Tabela 52: Regras extraídas da árvore de decisão da Figura 44 e avaliação de cada regra individualmente.

Regras	Medidas de avaliação
Regra 1 - Nó 4	
SE PRECIP_PINF < 215,3 E TMIN_PINF < 13,45 ENTÃO TAXA_INF_M10 = 0	Precisão: 98,2% Novidade: 0,03 Sensitividade: 33,5% Cobertura: 29,7% Especificidade: 100% Suporte: 29,7%
Distribuição* : JUL(14VN); AGO(16VN); SET(16VN); OUT(8VN).	
Regra 2 - Nó 7	
SE PRECIP_PINF ≥ 265,2 ENTÃO TAXA_INF_M10 = 0	Precisão: 87,5% Novidade: 0,00 Sensitividade: 3,7% Cobertura: 3,3% Especificidade: 100% Suporte: 3,3%
Distribuição: JAN(2VN); FEV(2VN); MAR(2VN).	
Regra 3 - Nó 9	
SE PRECIP_PINF < 215,3 E TMIN_PINF ≥ 13,45 E TMED_PI_PINF ≥ 21,45 ENTÃO TAXA_INF_M10 = 0	Precisão: 97,5% Novidade: 0,02 Sensitividade: 23,6% Cobertura: 20,9% Especificidade: 100% Suporte: 20,9%
Distribuição: JAN(6VN); FEV(6VN); MAR(10VN); ABR(6VN); NOV(2VN); DEZ(8VN).	

Regra 4 - Nó 10

SE $215,3 \leq \text{PRECIP_PINF} < 230,2$
ENTÃO $\text{TAXA_INF_M10} = 0$

Precisão: 62,5% **Novidade:** -0,01
Sensitividade: 2,5% **Cobertura:** 3,3%
Especificidade: 90,5% **Suporte:** 2,2%

Distribuição: JAN(1VN;1FN); FEV(1VN;1FN); ABR(2VN).

Regra 5 - Nó 11

SE $230,2 \leq \text{PRECIP_PINF} < 265,2$
ENTÃO $\text{TAXA_INF_M10} = 1$

Precisão: 80,0% **Novidade:** 0,03
Sensitividade: 33,3% **Cobertura:** 4,4%
Especificidade: 99,4% **Suporte:** 3,8%

Distribuição: JAN(3VP;1FP); FEV(2VP); MAR(2VP).

Regra 6 - Nó 13

SE $\text{PRECIP_PINF} < 215,3$
E $\text{TMIN_PINF} \geq 13,45$
E $\text{TMED_PI_PINF} < 21,45$
E $\text{LAVOURA} = \text{LARGA}$
ENTÃO $\text{TAXA_INF_M10} = 0$

Precisão: 91,9% **Novidade:** 0,01
Sensitividade: 20,5% **Cobertura:** 19,2%
Especificidade: 90,5% **Suporte:** 18,1%

Distribuição: JAN(1VN); MAR(1FN); ABR(4VN); MAI(8VN); JUN(7VN;1FN); JUL(1VN); OUT(3VN); NOV(6VN); DEZ(3VN).

Regra 7 - Nó 14

SE $\text{PRECIP_PINF} < 215,3$
E $13,45 \leq \text{TMIN_PINF} < 15,6$
E $\text{TMED_PI_PINF} < 21,45$
E $\text{LAVOURA} = \text{ADENSADA}$
ENTÃO $\text{TAXA_INF_M10} = 0$

Precisão: 85,7% **Novidade:** 0,00
Sensitividade: 10,6% **Cobertura:** 10,4%
Especificidade: 90,5% **Suporte:** 9,3%

Distribuição: MAI(1VN); JUN(7VN;1FN); JUL(1FN); OUT(3VN); NOV(5VN); DEZ(1VN).

Regra 8 - Nó 20

SE $\text{PRECIP_PINF} < 112,5$
E $\text{TMIN_PINF} \geq 15,6$
E $\text{TMED_PI_PINF} < 21,45$
E $\text{LAVOURA} = \text{ADENSADA}$
ENTÃO $\text{TAXA_INF_M10} = 1$

Precisão: 66,7% **Novidade:** 0,03
Sensitividade: 33,3% **Cobertura:** 5,5%
Especificidade: 98,1% **Suporte:** 3,8%

Distribuição: MAR(1VP); ABR(1VP); MAI(4VP;3FP); NOV(1VP).

Regra 9 - Nó 21

SE $112,5 \leq \text{PRECIP_PINF} < 215,3$
E $\text{TMIN_PINF} \geq 15,6$
E $\text{TMED_PI_PINF} < 21,45$
E $\text{LAVOURA} = \text{ADENSADA}$
ENTÃO $\text{TAXA_INF_M10} = 0$

Precisão: 75,0% **Novidade:** 0,00
Sensitividade: 3,1% **Cobertura:** 3,3%
Especificidade: 95,2% **Suporte:** 2,7%

Distribuição: JAN(1VN); ABR(2VN;1FN); DEZ(2VN).

* Distribuição mensal das predições entre VP (verdadeiros positivos), FP (falsos positivos), VN (verdadeiros negativos) e FN (falsos negativos).

5.3.2.2 Discussão

No modelo gerado com a opção Modelagem 1 (Figura 40), o atributo de teste no nó raiz e a sua fronteira de decisão foram os mesmos do modelo para TAXA_INF_M5 (Figura 34) e dos modelos para as lavouras com alta carga pendente (Figura 23 e Figura 29). A ocorrência de mais de 25 dias desfavoráveis à infecção ($DDI_PINF \geq 25,5$) correspondeu a taxas de infecção menores do que 10 p.p., na maioria dos casos.

De forma semelhante que no modelo para TAXA_INF_M5, o somatório total de molhamento foliar no período de infecção (SMT_NHUR95_PINF) foi incluído no modelo, com valores mais altos (mais de 142 h) ocasionando efeito depressivo nas taxas de infecção.

A influência dos demais atributos sobre as taxas de infecção da ferrugem do cafeeiro é parecida com o que já foi discutido anteriormente, neste mesmo capítulo, ou no capítulo referente à análise da epidemia dessa doença com árvore de decisão (capítulo 4).

O modelo, com nove regras e baseado em sete atributos de teste, teve desempenho fraco na avaliação (Tabela 44 e Tabela 45). A acurácia (82,4% - validação cruzada) foi inferior à proporção de exemplos da classe majoritária (88%) e estimou-se, na validação cruzada, o acerto de apenas 3 alertas (verdadeiros positivos) em 21.

A sensibilidade (15%) e a confiabilidade positiva (15,8%) foram bem baixas e os desvios padrões das médias dessas medidas foram altos (7,6% e 10,1%, respectivamente), em decorrência da distribuição bastante desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 34).

As predições para o período de dezembro a abril foram distribuídas entre quase todas as regras (Tabela 46). As regras de destaque foram:

- **Regra 1:** precisão (98,7%); sensibilidade (46%); especificidade máxima (100%); novidade (0,05); cobertura (40,7%) e suporte (40,7%); cobriu corretamente exemplos de dezembro e janeiro: 8 VN.
- **Regras 2 e 3:** precisão (94,1% e 93,1%); sensibilidade (19,3% e 16,1%); especificidade (95,2% para ambas); novidade (0,01 para ambas); cobertura (17,6% e 14,8%) e suporte (17% e 14,3%); cobriram corretamente exemplos do período de dezembro a abril: 35 VN.
- **Regra 4:** precisão (80%); sensibilidade (33,3%); especificidade (99,4%); novidade (0,03); cobertura (4,4%) e suporte (3,8%); cobriu corretamente exemplos de janeiro a março, na maior parte dos casos: 7 VP e 1 FP.

- **Regras 5 e 7:** precisão (84,2% e 87,5%); especificidade (90,5% e 100%); as demais medidas foram baixas; cobriram corretamente exemplos do período de dezembro a abril: 14 VN e apenas 1 FN.
- **Regra 6:** precisão (60%); sensibilidade (23,8%); especificidade (98,1%); as demais medidas foram baixas; cobriu corretamente exemplos de fevereiro a abril, na maior parte dos casos: 4 VP e 2 FP.

No modelo gerado com a opção Modelagem 2 (Figura 42), o atributo de teste no nó raiz foi o mesmo do modelo para TAXA_INF_M5 (Figura 36) e dos modelos para as lavouras com alta carga pendente (Figura 25 e Figura 31). A fronteira de decisão, como já havia aumentado no modelo para TAXA_INF_M5, em relação à dos modelos para as lavouras com alta carga pendente, aumentou mais um pouco. Temperaturas médias diárias durante o molhamento foliar (THUR95_PINF) abaixo de 18,15 °C foram menos favoráveis e maiores ou iguais a esse limite foram mais favoráveis à infecção.

A influência dos demais atributos de teste da árvore de decisão sobre as taxas de infecção da ferrugem do cafeeiro já foi discutida anteriormente.

O modelo, com oito regras e baseado em seis atributos de teste, também não teve desempenho bom na avaliação (Tabela 47 e Tabela 48). A acurácia (88,5% - validação cruzada) foi equivalente à proporção de exemplos da classe majoritária (88%) e estimou-se, na validação cruzada, o acerto de 6 alertas (verdadeiros positivos) em 21.

A sensibilidade (30%) e a confiabilidade positiva (50%) foram baixas e os desvios padrões das médias dessas medidas foram altos (8,2% e 14,9%, respectivamente), em decorrência da distribuição bastante desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 34).

As predições para o período de dezembro a abril foram distribuídas entre todas as regras (Tabela 49). As regras de destaque foram:

- **Regras 1 e 2:** precisão (96,3% e 92,9%); sensibilidade (64% e 15,5%); especificidade (85,7% e 95,2%); novidade (0,05 e 0,01); cobertura (58,2% e 14,3%) e suporte (56,6% e 13,7%); cobriram corretamente exemplos no período de dezembro a abril: 36 VN.
- **Regra 4:** precisão (77,8%); sensibilidade (28,6%); especificidade (99,4%); novidade (0,03); cobertura (3,8%) e suporte (3,3%); cobriu corretamente exemplos de janeiro a abril: 4 VP e 1 FP.

- **Regras 5 e 6:** precisão (90% e 87,5%); especificidade (100% para ambas); as demais medidas foram baixas; cobriram corretamente exemplos do período de dezembro a março: 10 VN.
- **Regra 7:** precisão (76,9%); especificidade (90,5%); as demais medidas foram baixas; cobriu corretamente exemplos de janeiro a abril, na maior parte dos casos: 7 VN e 2 FN.

No modelo gerado com a opção Modelagem 3 (Figura 44), o atributo de teste no nó raiz foi diferente do de todos os outros modelos gerados com esta opção de seleção de atributos preditivos. O atributo PRECIP_PINF, que representa a precipitação pluvial acumulada no período de infecção, não só foi escolhido como o atributo de teste no nó raiz, mas teve influência determinante na construção da árvore de decisão.

No caso do modelo gerado no Weka (Figura 45), PRECIP_PINF foi o único atributo utilizado nas regras de decisão. O acúmulo de precipitação entre 224 mm e 260 mm foi escolhido como a condição favorável para as taxas de infecção da ferrugem do cafeeiro maiores ou iguais a 10 p.p. em lavouras com baixa carga pendente. Para além desses extremos, com precipitações acumuladas menores ou iguais a 224 mm e maiores do que 260 mm, as condições passam a ser desfavoráveis para as taxas de infecção àquele nível.

A faixa intermediária de precipitação deve estar relacionada com o efeito positivo das chuvas nos processos de germinação e de disseminação (KUSHALAPPA, 1989a; ZAMBOLIM et al., 1997). Precipitações menores ou iguais a 224 mm não foram favoráveis a esses processos ao ponto de proporcionarem taxas de infecção de 10 p.p. ou mais. Por outro lado, as precipitações acima de 260 mm parecem indicar o efeito negativo das chuvas quando transportam para o chão a maior parte dos esporos (KUSHALAPPA, 1989a).

A influência dos demais atributos de teste da árvore de decisão da Figura 44 sobre as taxas de infecção da ferrugem do cafeeiro já foi discutida anteriormente.

O modelo, com nove regras e baseado em quatro atributos de teste, também não teve desempenho bom na avaliação (Tabela 50 e Tabela 51). A acurácia (86,9% - validação cruzada) foi equivalente à proporção de exemplos da classe majoritária (88%) e estimou-se, na validação cruzada, o acerto de 4 alertas (verdadeiros positivos) em 21.

A sensibilidade (20%) e a confiabilidade positiva (35%) foram baixas e os desvios padrões das médias dessas medidas foram altos (8,2% e 15%, respectivamente), em

decorrência da distribuição bastante desbalanceada dos exemplos entre as classes no conjunto de treinamento (Tabela 34).

As previsões para o período de dezembro a abril foram distribuídas entre quase todas as regras (Tabela 52). As regras de destaque foram:

- **Regras 2 e 7:** precisão (87,5% e 85,7%); especificidade (100% e 90,5%); as demais medidas foram baixas; cobriram corretamente exemplos do período de dezembro a março: 7 VN.
- **Regras 3 e 6:** precisão (97,5% e 91,9%); sensibilidade (23,6% e 20,5%); especificidade (100% e 90,5%); novidade (0,02 e 0,01); cobertura (20,9% e 19,2%) e suporte (20,9% e 18,1%); cobriram corretamente exemplos no período de dezembro a abril, na maior parte dos casos: 44 VN e apenas 1 FN.
- **Regras 5 e 8:** precisão (80% e 66,7%); sensibilidade (33,3% para ambas); especificidade (99,4% e 98,1%); novidade (0,03 para ambas); cobertura (4,4% e 5,5%) e suporte (3,8% para ambas); cobriram corretamente exemplos no período de janeiro a abril, na maior parte dos casos: 9 VP e apenas 1 FP.
- **Regra 9:** precisão (75%); especificidade (95,2%); as demais medidas foram baixas; cobriu corretamente exemplos de dezembro, janeiro e abril, na maior parte dos casos: 5 VN e 1 FN.

Nenhum dos modelos obtidos com as três opções de seleção de atributos preditivos teve bom desempenho na avaliação. A árvore de decisão gerada com a opção Modelagem 2 (Figura 42) foi a que apresentou os melhores valores para as medidas de avaliação.

O fraco desempenho dos modelos certamente está relacionado com a distribuição bastante desbalanceada dos exemplos entre as classes '0' e '1' no conjunto de treinamento (88% e 22%, respectivamente). Para as lavouras com baixa carga pendente, foram apenas 21 os casos em que a taxa de infecção da ferrugem do cafeeiro alcançou ou ultrapassou os 10 p.p. (classe '1'), dos 182 exemplos do conjunto de treinamento.

Quando a distribuição dos exemplos está desbalanceada a esse ponto, o processo de indução de árvores de decisão, sem a adoção de mecanismos que procurem contornar o problema, privilegia a classe majoritária, resultando em baixos valores para a sensibilidade e a confiabilidade positiva.

As melhores regras de cada modelo, considerando os verdadeiros positivos e os verdadeiros negativos para o período de dezembro a abril, além das outras medidas de avaliação de regras, foram:

- Modelagem 1 - VP: Regras 4 e 6; VN: Regras 1 a 3, 5 e 7.
- Modelagem 2 - VP: Regra 4; VN: Regras 1, 2 e 5 a 7.
- Modelagem 3 - VP: Regras 5 e 8; VN: Regras 2, 3, 6, 7 e 9.

5.4 Considerações finais

Segundo Zambolim et al. (2002), nos anos de alta carga pendente na lavoura, não são recomendadas as atomizações tardias, após a constatação de nível de incidência da ferrugem do cafeeiro maior do que 5%, mesmo se tratando dos fungicidas sistêmicos, para que o nível de controle da doença alcance 90% a 95% na colheita. Dessa maneira, evitar-se-á que o percentual de desfolha ultrapasse os 10% até a colheita e que inóculo residual passe para a estação seguinte.

Então, os modelos em que o atributo meta foi TAXA_INF_M5, para alertar quando a taxa de infecção da ferrugem for esperada atingir ou ultrapassar 5 p.p. no prazo de um mês, são os mais indicados para a utilização no suporte à decisão dos momentos oportunos para a adoção de medidas de controle da doença.

Os modelos para quando a taxa de infecção for esperada atingir ou ultrapassar 10 p.p. (TAXA_INF_M10) podem servir como instrumento adicional, alertando que, além de medidas de controle serem necessárias, estas deveriam ser urgentes e/ou mais eficazes, pois as condições estariam sendo propícias para um desenvolvimento ainda mais acelerado da doença.

Nos ciclos de baixa produção do cafeeiro, uma única aplicação de fungicida, no momento certo, é suficiente para se alcançar controle eficiente e racional da doença (MATIELLO e MANSK, 1984; GARÇON et al., 2004). Nesses anos de baixa carga pendente, a incidência da ferrugem é esperada diminuir ou retardar sua evolução (MORAES, 1983).

Segundo os dados utilizados neste trabalho, em condições de clima parecidas, é pouco provável que a taxa de infecção da ferrugem do cafeeiro atinja ou supere os 10 p.p. em lavouras com baixa carga pendente – apenas 22% dos exemplos analisados.

Sendo assim, quando as lavouras estiverem no ciclo de baixa produção, os modelos de alerta em que o atributo meta foi TAXA_INF_M5 podem ser suficientes no suporte à

decisão. Os modelos em que o atributo meta foi TAXA_INF_M10 não teriam tanta utilidade como no caso das lavouras com alta carga pendente.

Os modelos obtidos com a opção de seleção de atributos Modelagem 3, no geral, tiveram tão bom desempenho quanto os modelos obtidos com as outras duas opções de seleção de atributos preditivos. A vantagem desses modelos é o menor trabalho na preparação dos dados necessários para emitir os alertas e, principalmente, a não necessidade de registros horários dos dados meteorológicos.

Quando se tiver a disponibilidade de registros horários dos dados meteorológicos, os modelos obtidos com a opção de seleção de atributos Modelagem 2 são recomendados. Pode-se até levar em conta a adoção desses modelos em conjunto com os modelos da opção Modelagem 3. Uma decisão pode ficar melhor assegurada caso se considere concomitantemente as saídas de modelos das duas opções de seleção de atributos. Nesse caso, possíveis conflitos podem ser resolvidos com base na avaliação individual das regras disparadas de cada modelo (detalhes sobre isso mais adiante).

Os atributos preditivos especiais da opção de seleção de atributos Modelagem 1 não foram capazes de melhorar o desempenho dos modelos obtidos. Apesar da contabilização dos dias desfavoráveis à infecção ter se mostrado, de forma isolada, com alto poder preditivo – DDI_PINF foi escolhido como o atributo de teste no nó raiz de todas as árvores de decisão induzidas a partir do conjunto de treinamento da opção Modelagem 1 –, o desempenho final dos modelos não apresentou nenhum ganho em relação aos modelos obtidos com as outras duas opções de seleção de atributos, segundo as medidas de avaliação adotadas.

Os modelos de alerta para lavouras com alta carga pendente foram os que obtiveram os melhores resultados na avaliação. A acurácia dos melhores modelos em que o atributo meta foi TAXA_INF_M5 foi estimada entre 81% e 83%, podendo alcançar em torno de 90%, segundo a estimativa mais otimista. No caso dos modelos em que o atributo meta foi TAXA_INF_M10, a acurácia foi estimada em 79%, podendo alcançar entre 85% e 90%, pela estimativa mais otimista.

Os modelos de alerta para lavouras com baixa carga pendente obtiveram resultados inferiores. A acurácia dos melhores modelos em que o atributo meta foi TAXA_INF_M5 foi estimada entre 69% e 72%, podendo alcançar até 86%, de acordo com a estimativa mais otimista. Os modelos em que o atributo meta foi TAXA_INF_M10 não foram bem avaliados.

Não há como se comparar os resultados obtidos neste trabalho com os resultados relatados na literatura referentes a outros modelos de previsão da ferrugem do cafeeiro (seção 2.3.3), devido às diferenças entre as técnicas utilizadas e às diferentes formas de avaliação adotadas em cada trabalho.

Apenas para apontar a dificuldade de se obter modelos com melhores níveis de acurácia do que os obtidos, resolveu-se considerar o trabalho desenvolvido por Kushalappa et al. (1984), que está entre os mais conceituados e referenciados na literatura. Os autores desenvolveram equações de regressão para prever a taxa de infecção da ferrugem do cafeeiro com base no que denominaram de razão de sobrevivência líquida para o processo monocíclico de *H. vastatrix*. A melhor equação explicou 76% da variação na taxa de infecção da doença (coeficiente de determinação $R^2 = 0,76$).

As estimativas de acurácia dos modelos para lavouras com alta carga pendente foram parecidas com as de outras aplicações das árvores de decisão como modelos de alerta de doenças de plantas, cujos maiores valores estimados não ultrapassaram 80% (seção 2.3.4).

Referente à fase de distribuição dos modelos obtidos no processo de KDD realizado, ou seja, com relação à adoção de qualquer dos modelos desenvolvidos neste trabalho, algumas considerações a respeito da preparação dos dados para o uso dos modelos devem ser feitas.

Seja a data do alerta (D_{AL}) um dia em que o modelo venha a ser usado para predizer se a taxa de infecção da ferrugem pode atingir ou ultrapassar, no prazo de um mês, o limite de 5 p.p. ou de 10 p.p., dependendo do modelo.

A partir de D_{AL} , deve-se determinar o período de infecção (PINF) correspondente à taxa de infecção esperada um mês após D_{AL} , de maneira similar quando da preparação dos dados para o treinamento dos modelos (Figura 13, pg. 62). D_{AL} corresponde, na Figura 13, a um mês antes da data de avaliação da incidência da ferrugem (A_i), ou seja, a data de avaliação da incidência no mês anterior (A_{i-1}).

Conforme visto na seção 3.4.2, a determinação do PINF foi dependente da temperatura durante o período de incubação (PI) dos eventuais dias de infecção, que compreendeu o primeiro dia do PINF até o dia da avaliação A_i (Figura 13). Portanto, a determinação do PINF foi dependente da temperatura nos dias posteriores à data de avaliação A_{i-1} . Na preparação dos dados para o conjunto de treinamento, esses valores de temperatura eram conhecidos.

No caso da preparação dos dados para o uso dos modelos, não se vai ter a temperatura dos dias além da data do alerta D_{AL} . Nesses casos, ter-se-á que usar uma estimativa da temperatura para esses dias, seja com base em temperaturas médias históricas do local ou, se houver disponibilidade, com base em previsões de temperatura.

Ajustada a preparação dos dados, conforme mencionado, os modelos podem ser usados diariamente, desde que os dados meteorológicos também sejam atualizados diariamente.

Emitida a saída do modelo, é sempre bom consultar o resultado da avaliação do mesmo, principalmente na validação cruzada, mas também na resubstituição. O que pode também auxiliar na tomada de decisão é verificar qual foi a regra do modelo disparada e consultar a sua avaliação individual, salientando-se que a avaliação das regras se baseou nos resultados da resubstituição e, portanto, incorpora um viés otimista.

A avaliação individual das regras pode ser útil também na resolução de conflitos entre modelos. Caso venham a ser usados dois modelos para o mesmo tipo de alerta (p.ex. os modelos das opções Modelagem 2 e Modelagem 3 em que o atributo meta foi TAXA_INF_M5), é possível que um modelo emita um alerta e o outro não. Nessas circunstâncias, a avaliação individual de cada regra disparada pode ajudar a resolver o conflito.

Antes, porém, da adoção de qualquer dos modelos desenvolvidos neste trabalho, é fundamental que seja realizada uma validação do modelo, ou dos modelos, de interesse. Seria interessante que a validação, primeiro, fosse realizada com dados mais recentes produzidos na própria Fundação Procafé. Essa validação deveria ser realizada pelo tempo necessário para que se tenha a segurança de que os alertas são confiáveis.

A Fundação Procafé, mais recentemente, vem coletando dados de outras duas localidades no estado de Minas Gerais, nos municípios de Carmo de Minas (Latitude 22° 10' 31'' S; Longitude 45° 09' 03'' W; Altitude 1080 m) e de Boa Esperança (Latitude 21° 03' 59'' S; Longitude 45° 34' 37'' W; Altitude 830 m). Os modelos poderiam ser validados com os dados destes locais também.

Uma vez validados, a Fundação Procafé poderia fazer uso dos alertas emitidos pelos modelos para incluir informação adicional nos seus boletins de avisos mensais divulgados aos produtores, técnicos e órgãos ligados ao setor de produção de café.

6 CARACTERIZAÇÃO DO PROCESSO

6.1 Considerações iniciais

Este capítulo apresenta a caracterização do processo realizado de descoberta de conhecimento em bases de dados, com o intuito de que as experiências adquiridas neste projeto possam servir em projetos futuros no mesmo domínio de aplicação e em contextos semelhantes.

O conteúdo e a forma de apresentação foram elaborados de acordo com as orientações para o mapeamento entre os níveis genérico e especializado da metodologia CRISP-DM (seção 3.6).

A instância do processo, que é a denominação do projeto segundo os termos da metodologia, está inserida em um contexto de mineração de dados específico (seção 6.2).

Com base nesse contexto e na experiência adquirida, foram elaboradas as tarefas especializadas (seção 6.3), que particularizam algumas das tarefas genéricas das fases do processo.

6.2 Contexto de mineração de dados

O contexto de mineração de dados usado na especialização do modelo do processo é apresentado na Tabela 53. Este contexto conduziu o mapeamento entre os níveis genérico e especializado da metodologia CRISP-DM.

Tabela 53: Contexto de mineração de dados usado na especialização do modelo do processo.

Domínio de aplicação	Tipo de problema	Aspecto técnico	Ferramenta e técnica
Epidemiologia e alertas de doenças de plantas.	Classificação	Definição de classes	SAS [®] Enterprise Miner [™]
		Indução interativa	Weka (J48)
		Critérios de poda	Árvore de decisão

6.3 Tarefas especializadas

As tarefas especializadas estão organizadas conforme as fases do processo de descoberta de conhecimento em bases de dados e as tarefas genéricas correspondentes.

6.3.1 Compreensão do domínio

Tarefa

Determinar os objetivos

Tarefa especializada

Verificar pré-requisitos da doença

Desenvolver um modelo de alerta requer que a doença atenda quatro requisitos (COAKLEY, 1988): (1) a doença ocasiona perdas economicamente significativas na qualidade ou na quantidade da produção; (2) a doença varia entre cada estação de cultivo (p.ex. severidade na colheita e taxa de aumento); (3) medidas de controle da doença estão disponíveis e são economicamente viáveis; e (4) informação sobre a natureza da dependência da doença em relação às condições meteorológicas é suficientemente conhecida.

Relato da experiência na instância do processo

A ferrugem do cafeeiro causa decréscimos significativos na produção e é considerada a principal doença da cultura. Além da importância econômica, os demais requisitos da doença também foram atendidos: ela varia em intensidade a cada ano agrícola; existem várias opções de medidas de controle economicamente viáveis; e diversos estudos foram encontrados na literatura sobre a influência das condições meteorológicas no seu desenvolvimento.

Tarefa Especializada

Identificar o esquema de análise de interesse

Esquemas de análise são adotados de acordo com o interesse em uma ou mais dentre cinco características de uma epidemia (BUTT e ROYLE, 1990): (1) o interesse pode estar focado no *progresso da epidemia* no decorrer do tempo; (2) pode estar concentrado na *taxa de aumento da doença*; (3) pode estar relacionado com os fatores que determinam o nível de *severidade da doença* em um determinado momento (p.ex. na época da colheita); (4) pode estar centrado em *eventos do ciclo da doença*; (5) e pode estar vinculado com *perda de produção*, que é uma consequência e não parte de uma epidemia.

Relato da experiência na instância do processo

No caso da ferrugem do cafeeiro, como a doença é policíclica, o interesse

estava em relacionar mudanças nas condições meteorológicas com aumentos na intensidade da doença.

Considerar os próprios níveis de incidência da doença não era adequado, pois o nível de incidência em um mês dependeria do nível no mês anterior, o que não era desejado.

Com a área sob a curva de progresso da doença, uma medida comum para comparação entre epidemias, ocorreria a mesma coisa: o valor de um mês seria dependente do valor no mês anterior.

O interesse estava, então, na taxa de aumento da doença, que foi calculada como a taxa de infecção, subtraindo-se a incidência no mês em questão com a incidência no mês anterior.

Chegou-se a cogitar o uso de uma taxa que indicasse, ao invés do aumento absoluto, o aumento percentual no nível de incidência da ferrugem. No entanto, percebeu-se que haveria dois tipos de problema: aumentos a partir do nível zero de incidência (não há como fazer o cálculo do aumento percentual); e aumentos significativos a partir de níveis baixos de incidência (p.ex. um aumento na incidência de 0,5% para 13% corresponderia a um aumento percentual muito alto de 2500%; nesse caso, a taxa de infecção seria de 12,5 p.p.). Esses problemas dificultariam a definição das classes de taxa de infecção. Além disso, se fosse se considerar aumentos percentuais, não se teria encontrado referencial na literatura para a definição das classes.

Tarefa

Avaliar a situação

Tarefa especializada

Identificar as fontes de dados e garantir acesso a elas

Mencionar a importância disso parece óbvio, mas é essencial que se identifique todas as fontes de dados e se garanta o acesso a elas desde o início da concepção do projeto. É enganoso acreditar que o importante está na proposta do projeto e que se garantir acesso a parte dos dados é suficiente. Sem os dados completos, depois, principalmente relativos à doença, não há o que fazer.

Relato da experiência na instância do processo

Este projeto foi concebido como parte de um projeto maior (EMBRAPA, 2001a), que, por sua vez, era um subprojeto de um projeto mais amplo, no âmbito da Embrapa, referente ao desenvolvimento e evolução de um sistema de monitoramento agroclimático para o estado de São Paulo (EMBRAPA, 2001b).

À época considerou-se que já se tinha uma grande quantidade de dados meteorológicos e “só faltavam” os dados das doenças e pragas consideradas do cafeeiro. Havia a percepção de que aqueles dados eram importantes, mas preocupou-se, naquele momento, apenas com a proposta do projeto.

Com o passar do tempo, percebeu-se que o mais difícil estava em conseguir as fontes de dados sobre as doenças e pragas do cafeeiro. No estado de São Paulo, para onde se planejava originalmente desenvolver os modelos, não foram encontradas tais fonte de dados.

Naquela época, a Fundação Procafé foi o único lugar em que se encontrou um monitoramento sistemático e duradouro de doenças e pragas da cultura do café, dentre elas a ferrugem do cafeeiro.

Tarefa Especializada

Conferir a disponibilidade de especialistas do domínio

A participação no projeto de especialistas do domínio é fundamental. Fitopatologistas, atuantes na epidemiologia de doenças de plantas e, de preferência, que tenham experiência e interesse na doença, são os mais indicados. A participação de agrometeorologistas, com experiência no desenvolvimento de modelos de previsão de doenças de plantas, também é bem-vinda.

Contudo, a participação de um especialista não desobriga o analista de dados de se inteirar sobre a epidemiologia da doença em questão. O nível de conhecimento do analista deve estar estreitamente relacionado com o nível de participação do especialista no projeto. De qualquer forma, quanto mais o analista de dados conhecer do domínio de aplicação, melhor.

Relato da experiência na instância do processo

O especialista do domínio, neste projeto, foi o Dr. Sérgio Almeida de Moraes, fitopatologista e pesquisador do Instituto Agronômico de Campinas (IAC). O Dr. Moraes desenvolveu estudos importantes sobre a ferrugem do cafeeiro (MORAES et al., 1976; MORAES, 1983), durante alguns anos. Posteriormente, esteve envolvido com modelos e sistemas de previsão da mancha preta do amendoim (PEDRO JÚNIOR et al., 1994; MORAES, 1999).

A sua participação foi como um consultor. Diversas reuniões foram realizadas, com a sua presença, desde a elaboração do plano de pesquisa até a fase final de execução do projeto.

6.3.2 Entendimento dos dados

Tarefa **Explorar os dados**

Tarefa especializada **Explorar os dados da doença**

Recomenda-se a elaboração de gráficos de evolução da doença, como os da Figura 8 e da Figura 9, para se conhecer e conferir a sua periodicidade estacional e para se identificar possíveis erros no processo de avaliação da intensidade da doença (p.ex. erros na amostragem).

Relato da experiência na instância do processo

Com esse tipo de gráfico foi possível identificar claramente um problema na amostragem que determinou um dos valores de incidência da ferrugem do cafeeiro (Figura 9).

Tarefa Especializada **Explorar os dados meteorológicos**

A exploração visual dos dados meteorológicos também é importante. Histogramas (Figura 46), gráficos do tipo *box plot* (Figura 10, Figura 11 e Figura 12) e diagramas de dispersão (Figura 15) permitem um melhor entendimento dos dados e podem fornecer subsídios para as fases posteriores do processo.

Estatísticas descritivas dos dados meteorológicos (valores máximos e

mínimos, quantidade de registros e de valores nulos para cada atributo etc.) também são úteis, tanto no entendimento dos dados como no auxílio à verificação da qualidade desses dados.

Relato da experiência na instância do processo

Apenas para exemplificar a capacidade representativa de um gráfico de exploração de dados, já que a pressão barométrica não foi atributo de interesse no projeto, considere o histograma a seguir.

Percebe-se, claramente, pela visualização da distribuição dos valores, que houve algum problema. Esse problema, esclarecido posteriormente, foi um período em que o sensor da estação meteorológica esteve descalibrado (faixa de distribuição à direita).

O pessoal da Fundação Procafé nunca havia detectado esse problema.

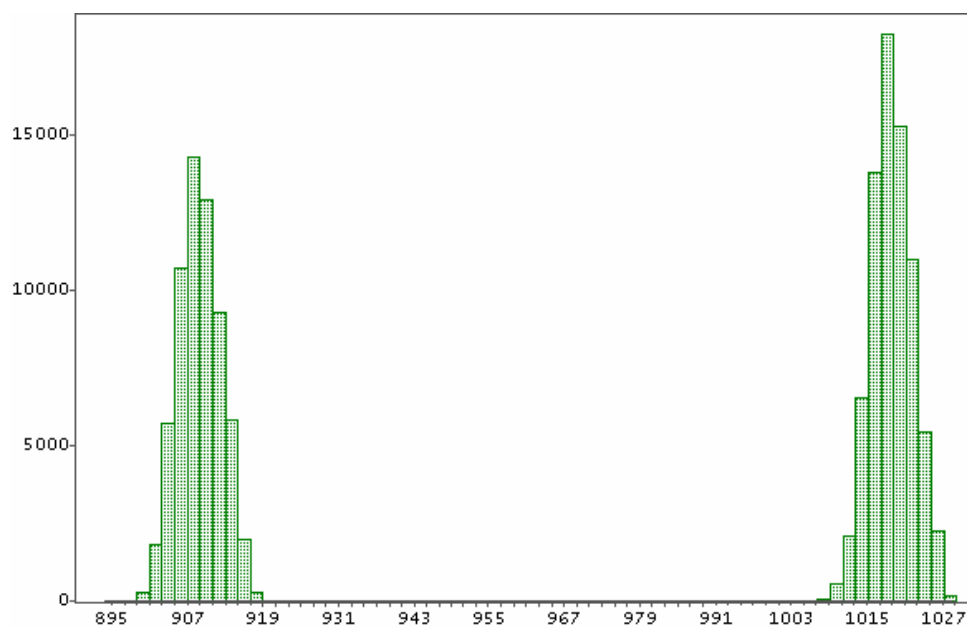


Figura 46: Histograma com distribuição de valores de pressão barométrica.

Tarefa

Verificar a qualidade dos dados

Tarefa especializada

Inspecionar os dados meteorológicos

Além de elaborar e conferir regras de consistência dos dados meteorológicos (Tabela 7), é importante inspecionar os arquivos gerados por uma estação meteorológica. Pode parecer uma tarefa desnecessária,

capaz até de causar certo receio no analista de dados, dado ao volume de dados registrado, mas uma simples inspeção visual pode revelar bastante sobre a qualidade dos dados.

Relato da experiência na instância do processo

A inspeção visual nos arquivos gerados pela estação meteorológica da Fundação Procafé permitiu identificar problemas do tipo: registros ausentes; registros repetidos; registros não pertencentes ao mês correspondente do arquivo da estação; registros para um dia que não existe (29 de fevereiro de um ano que não foi bissexto); e valores inconsistentes de determinados atributos.

6.3.3 Preparação dos dados

Tarefa

Construir dados

Tarefa especializada

Definir as classes do atributo meta

Em problemas de classificação, o atributo meta deve ser categórico. O ideal é quando já se tem esse atributo com as classes definidas. Em algumas situações, no entanto, é preciso derivar o atributo meta a partir de um atributo numérico contínuo. A divisão do intervalo contínuo em intervalos delimitados, que formam as classes, pode ser feita com o apoio de ferramentas computacionais ou com o auxílio de especialistas e/ou de literatura específica.

Relato da experiência na instância do processo

Embora o Enterprise Miner™ e o Weka possuam a funcionalidade de transformar atributos numéricos em categóricos, procurou-se na literatura resultados que permitissem a definição das classes da taxa de infecção da ferrugem do cafeeiro, por recomendação do especialista do domínio.

Tarefa Especializada

Derivar os atributos preditivos meteorológicos

É comum se ter a impressão de que atributos com derivação mais elaborada podem dotar os modelos de maior acurácia. Contudo, pode ser que não funcione dessa forma.

Uma sugestão é começar pela derivação de atributos mais diretos e, portanto, mais simples; em seguida, proceder a uma rodada de modelagem e de avaliação dos modelos; depois, conforme a necessidade, ou mesmo se já estiver planejado dessa forma, partir para a derivação de atributos mais elaborados, com vistas a melhorar o desempenho dos modelos.

Esta sugestão permite que os resultados, mesmo que parciais, sejam obtidos de maneira mais rápida do que realizar toda a fase de preparação dos dados pretendida antes de se encaminhar para a fase de modelagem.

Relato da experiência na instância do processo

A sugestão dada decorre da própria experiência com a preparação dos dados para a obtenção dos modelos de alerta da ferrugem do cafeeiro. Procurou-se pensar e realizar a preparação dos dados de maneira completa; os programas foram implementados para derivar todos os atributos imaginados, desde os mais simples até os mais elaborados.

Isso se justifica num projeto de pesquisa de doutorado, mas pode não ser adequado para projetos em que seja importante apresentar resultados num período de tempo mais curto.

Além disso, os programas de preparação dos dados foram implementados para proporcionar flexibilidade, permitindo a atribuição de alguns parâmetros (p.ex. limite de umidade para considerar molhamento foliar; período de incubação fixo ou dependente de equação, qual equação considerar etc.). Quase a totalidade desses parâmetros foi mantida fixa desde a determinação dos valores iniciais, pela adequação presumida dos mesmos, mas também pela impossibilidade de se testar todas as possíveis alternativas.

▪ **Considerar o período de incubação na derivação**

No caso de doenças fúngicas, é preciso considerar o período de incubação (PI) do fungo (referenciado por alguns autores como período latente) na derivação dos atributos preditivos meteorológicos.

O PI pode ser considerado fixo (KUSHALAPPA et al., 1983) ou variável,

segundo alguma equação já desenvolvida. No caso do PI fixo, a estimativa dos períodos de infecção é mais grosseira e podem ocorrer distorções/contaminações mais significativas nos dados preparados.

Relato da experiência na instância do processo

No projeto, o período de incubação foi considerado variável, dependente das médias das temperaturas máximas e mínimas durante o período, segundo a equação desenvolvida por Moraes et al. (1976).

▪ **Estimar o período de molhamento foliar**

Caso se queira considerar o molhamento foliar na análise dos dados e não se tenha, ou não se queira usar, registros de sensores de molhamento, é possível se fazer uma estimativa da duração do molhamento foliar diário.

A forma mais comum utilizada para estimar o período de molhamento é por meio da duração da umidade relativa do ar acima de um limite específico, geralmente 90% ou 95% (SUTTON et al., 1984; JENSEN e BOYLE, 1966). Existem várias outras maneiras de se estimar os períodos de molhamento (HUBER e GILLESPIE, 1992). Inclusive, a indução de árvores de decisão já foi utilizada para obtenção de um modelo de estimativa da duração de períodos de molhamento (GLEASON et al., 1994).

Relato da experiência na instância do processo

De início, o molhamento foliar foi considerado quando a umidade relativa do ar era maior ou igual a 90%. Chegou-se a gerar alguns modelos com a taxa de infecção categórica com três classes como o atributo meta.

Posteriormente, um modelo experimental mostrou que 95% seria um limite de decisão mais adequado para o conjunto de dados disponível. Esse modelo foi uma árvore de decisão com a umidade relativa sendo o único atributo preditivo e o atributo meta indicando a existência ou não de molhamento foliar ('1' = molhado e '0' = seco) – esse atributo meta binário foi construído a partir dos registros do sensor de molhamento foliar da estação meteorológica.

DICA: quando se tem que calcular o número de horas do dia em que ocorreu, ou se manteve, determinada situação, como é o caso do número de horas de molhamento foliar, é preciso estar atento para considerar um registro às 24:00 horas de cada dia. Por exemplo, em uma estação que registra dados de 0:00 h às 23:30 h de um dia, é preciso considerar o registro à 0:00 h do dia seguinte como o registro às 24:00 h daquele dia; caso contrário, o dia não vai ter 24 horas.

6.3.4 Modelagem

Tarefa **Gerar os modelos**

Tarefa **Gerar modelo aplicado na epidemiologia de uma doença**
especializada

A técnica de indução de árvores de decisão se mostrou bastante interessante quanto à sua aplicação na epidemiologia de doenças de plantas.

No caso dessa aplicação, considera-se como características essenciais da ferramenta de modelagem: (1) a possibilidade de indução interativa da árvore de decisão e (2) a visualização da distribuição de probabilidade entre as classes nos nós internos da árvore, além dos nós folhas.

A indução interativa proporciona flexibilidade e liberdade ao analista de dados, em conjunto com o especialista do domínio, para verificarem e analisarem diversas possibilidades de configuração, ou estruturação, da árvore de decisão.

É interessante que o processo de indução interativo permita se construir a árvore de decisão desde o nó raiz. Em cada nó que se queira ramificar, deve aparecer uma relação dos atributos mais importantes, de acordo com o critério de seleção de atributos. A partir daí, deve-se permitir escolher o mesmo atributo que seria escolhido no processo automático de indução ou escolher outro atributo de interesse.

A indução interativa pode também partir da árvore de decisão gerada pelo processo automático. Neste caso, deve-se permitir a troca de atributos de teste nos nós internos e, também, a poda ou ramificação dos nós já criados.

A visualização da distribuição de probabilidade entre as classes nos nós

internos da árvore de decisão permite melhorar a percepção e a compreensão dos relacionamentos entre os atributos preditivos/explicativos e a doença.

Essa visualização fica ainda melhor se a ferramenta de modelagem permitir a coloração dos nós da árvore de decisão conforme a proporção de distribuição de uma das classes, que pode ser a classe correspondente ao nível mais intenso de ataque da doença.

O Enterprise Miner™ possui todas essas características e o Weka nenhuma delas.

Relato da experiência na instância do processo

A visualização da distribuição de probabilidade entre as classes, com coloração proporcional à classe de maior interesse, em todos os nós dos modelos obtidos para a taxa de infecção categórica com três classes (TAXA_INF3N), foi o que permitiu descobrir o potencial de aplicação da técnica de indução de árvores de decisão na epidemiologia da ferrugem do cafeeiro.

A indução interativa, apesar de não ter sido usada, diretamente, na indução da árvore de decisão para analisar a epidemia da ferrugem do cafeeiro, permitiu a verificação de todas as alternativas de escolha dos atributos de teste, em cada nó da árvore, auxiliando na discussão dos resultados.

Tarefa especializada

Gerar modelo aplicado em alertas de uma doença

Quando o interesse está em modelos de alerta de uma doença, parece ser mais adequado definir o atributo meta como binário. Primeiro, porque o alerta fica melhor caracterizado com apenas duas classes: ‘1’ e ‘0’ ou ‘positivo’ e ‘negativo’. Segundo, porque os modelos tendem a ter um número menor de regras (menor complexidade) e a acertar mais (melhor acurácia).

Relato da experiência na instância do processo

Os modelos desenvolvidos para a ferrugem do cafeeiro com as taxas de infecção binárias são realmente mais indicados para a emissão de alertas, do

que o modelo considerando a taxa de infecção categórica com três classes. A separação entre modelos para lavouras com alta carga pendente e para lavouras com baixa carga pendente também contribuiu para um melhor desempenho dos modelos com as taxas de infecção binárias, especialmente aqueles para lavouras com alta carga pendente.

Tarefa especializada

Escolher os critérios de poda

Os mecanismos e critérios de poda do Enterprise Miner™ e do Weka são os seguintes.

Enterprise Miner™:

- Pré-poda
 - Profundidade máxima da árvore (valor padrão = 6).
 - Número de exemplos requeridos para a divisão de um nó (valor sugerido calculado com base na quantidade de exemplos).
 - Número mínimo de exemplos em cada nó folha (valor sugerido calculado com base na quantidade de exemplos – máximo entre 5 e $N/1000$, onde N é o número de exemplos).
- Pós-poda
 - Escolha de uma subárvore (p.ex. menor subárvore com a melhor avaliação).

Weka:

- Pré-poda
 - Número mínimo de exemplos em cada nó folha (valor padrão = 2).
- Pós-poda
 - Fator de confiança (valor padrão = 0,25 – valores menores incorrem em poda mais acentuada).
 - *Subtree raising* (habilitada por padrão).
 - Outras opções (desabilitadas por padrão).

O único critério igual é o que define o número mínimo de exemplos em cada nó folha, mas os valores padrões são diferentes.

Relato da experiência na instância do processo

Os valores padrões dos critérios de poda foram mantidos no Enterprise Miner™, por se considerar que estavam adequados.

No Weka, o número mínimo de exemplos em cada nó folha foi alterado para 5, para ficar igual ao determinado pelo Enterprise Miner™. Não houve alteração nos demais critérios.

Com essas configurações, no geral, a poda nos modelos gerados pelo Weka foi maior. Em nenhuma situação houve poda maior do Enterprise Miner™.

Tarefa

Avaliar os modelos

Tarefa especializada

Avaliar os modelos por meio de validação cruzada

Epidemiologia e alertas de doenças de plantas é o tipo de aplicação com poucos dados para a modelagem. Então, para a avaliação dos modelos, a validação cruzada é uma das alternativas recomendáveis.

O Weka (versão 3.4.11) permite avaliar os modelos obtidos por meio de validação cruzada, sem nenhum problema.

Já o Enterprise Miner™ (versão 4.3) não possui essa funcionalidade pronta, seja como uma opção da ferramenta *Tree* ou seja como uma ferramenta própria (*node*). É preciso implementar a validação cruzada com programação específica na linguagem do SAS®.

Caso se pretenda usar outra ferramenta de modelagem, é interessante que se faça uma avaliação inicial dela no momento do planejamento do projeto. Nessa avaliação, todos os aspectos do processo de KDD devem ser considerados, de forma geral, mas ressalta-se que os aspectos da avaliação não podem ser esquecidos.

Relato da experiência na instância do processo

Na época da adoção do Enterprise Miner™ como ferramenta de modelagem, não se tinha conhecimento da sua limitação quanto ao procedimento de validação cruzada.

A implementação da validação cruzada com programas escritos na

linguagem de programação do SAS[®] demandou esforço e tempo consideráveis.

6.4 Considerações finais

O intuito, na elaboração deste capítulo, não foi o de produzir, por completo, um modelo do processo especializado, segundo o contexto de mineração de dados específico, que contemplasse todas as tarefas especializadas possíveis, de todas as fases e tarefas genéricas do modelo do processo CRISP-DM.

O que se procurou fazer foi contemplar o que aconteceu de importante (as “lições aprendidas”) no transcorrer da instância do processo para a obtenção dos modelos da ferrugem do cafeeiro, que pudesse vir a ser útil em iniciativas futuras.

Espera-se que o registro desses acontecimentos, organizado de acordo com a estrutura genérica do modelo hierárquico da metodologia CRISP-DM, possa servir de orientação para novos projetos no mesmo domínio de aplicação, e que sejam semelhantes, também, quanto às demais dimensões do contexto de mineração de dados.

7 CONCLUSÕES

A análise realizada nos dados de incidência da ferrugem do cafeeiro, junto com os dados meteorológicos obtidos da estação meteorológica padrão, de acordo com os preceitos do processo de descoberta de conhecimento em bases de dados, produziu modelos de alerta da doença que podem vir a ser usados como parte integrante de um sistema de monitoramento agrometeorológico de acesso público e gratuito.

O intuito, com esses modelos de alerta, é auxiliar na tomada de decisão das medidas apropriadas e dos momentos oportunos para o controle da ferrugem do cafeeiro, viabilizando recomendações aos produtores de café com base em informações precisas e confiáveis sobre a taxa de infecção da doença no campo.

Os modelos de alerta considerando as taxas de infecção binárias (capítulo 5), que predizem, para o prazo de um mês à frente, aumentos a partir de 5 p.p. e de 10 p.p. na taxa de infecção da ferrugem do cafeeiro, podem auxiliar na decisão de quando e de quais medidas, protetivas (fungicidas de contato) e/ou curativas (fungicidas sistêmicos), devem ser usadas no controle da doença.

A preocupação maior no controle da ferrugem do cafeeiro deve estar nos anos de alta carga pendente de frutos, quando o progresso da doença é mais rápido e o ataque é mais severo. Por felicidade, os modelos de alerta da ferrugem do cafeeiro para as lavouras com alta carga pendente de frutos tiveram melhores resultados na avaliação do que os modelos para as lavouras com baixa carga pendente.

No caso das lavouras com alta carga pendente, a estimativa de acurácia foi de até 83%, pela validação cruzada, para os modelos escolhidos para alertar quando a taxa de infecção da ferrugem for esperada atingir ou superar 5 p.p. em um mês, o que é um resultado bastante significativo. As estimativas, para esses modelos, de outras medidas importantes, como a sensibilidade, a especificidade, a confiabilidade positiva e a confiabilidade negativa, também atingiram níveis altos, bem próximos do patamar de acurácia.

Os modelos desenvolvidos para alertar quando a taxa de infecção da ferrugem for esperada atingir ou superar 10 p.p., em um mês, tiveram a acurácia estimada em 79%, o que também é um bom resultado. Entretanto, o mesmo equilíbrio entre a acurácia e as demais

medidas de avaliação não foi obtido, com estimativas superiores para a especificidade e a confiabilidade negativa e inferiores para a sensibilidade e a confiabilidade positiva.

A estimativa de acurácia dos modelos para as lavouras com baixa carga pendente de frutos foi de até 72%, no caso de alertar quando a taxa de infecção da ferrugem do cafeeiro for esperada atingir ou superar 5 p.p em um mês. Também não houve equilíbrio entre a acurácia e as demais medidas de avaliação, com estimativas superiores para a especificidade e a confiabilidade negativa e bastante inferiores para a sensibilidade e a confiabilidade positiva.

Os modelos para alertar quando a taxa de infecção da ferrugem for esperada atingir ou superar 10 p.p., em um mês, tiveram desempenho fraco na avaliação. No entanto, esses modelos teriam pouca utilidade, pois é pouco provável que as taxas de infecção da doença alcancem ou ultrapassem 10 p.p. em lavouras com baixa carga pendente de frutos, pelo menos em condições parecidas com as analisadas neste trabalho. A baixa quantidade de exemplos com taxas de infecção maiores ou iguais a 10 p.p., no conjunto de treinamento, deve ter contribuído para o fraco desempenho desses modelos. A distribuição bastante desbalanceada entre as classes provavelmente influenciou o processo de indução, fazendo com que a classe majoritária fosse privilegiada.

Independentemente do modelo a ser utilizado, é interessante que seja consultada a avaliação individual da regra que venha a ser disparada. A avaliação individual das regras pode fornecer subsídios adicionais, além da avaliação geral do modelo, para a tomada de decisão, mesmo para o caso dos modelos com os piores desempenhos.

Os modelos obtidos considerando a taxa de infecção categórica com três classes não foram considerados adequados como modelos de alerta. Entretanto, a árvore de decisão apresentada no capítulo 4, devido à sua representação simbólica e interpretável, foi apropriada para uma análise da epidemia da ferrugem do cafeeiro.

Os atributos de teste dos nós internos, as fronteiras de decisão para esses atributos e as distribuições de probabilidade das classes de taxa de infecção nos nós da árvore de decisão indicaram os principais fatores que interferiram no progresso da ferrugem do cafeeiro no campo e as faixas de valores de destaque para esses fatores, os quais se mostraram de acordo com diversos relatos encontrados na literatura referentes à epidemiologia da doença.

Espera-se que a caracterização do processo de descoberta de conhecimento em bases de dados (capítulo 6) possa vir a ser útil em iniciativas futuras no mesmo domínio de

aplicação, parecidas também quanto às demais dimensões do contexto específico de mineração de dados. Ou seja, que o processo especializado possa ser reproduzido e adaptado para problemas semelhantes com doenças/pragas de outras culturas agrícolas ou com outras doenças/pragas da própria cultura do café.

O uso de classificação e de indução de árvores de decisão no desenvolvimento de modelos de alerta de doenças de plantas é novidade na área de fitopatologia brasileira. Ademais, o uso dessa tarefa e dessa técnica de mineração de dados na análise de epidemias de doenças de plantas parece ser inovador em âmbito internacional, uma vez que não se encontrou nenhum relato na literatura referente a esse tipo de aplicação.

A abordagem metodológica, no geral, conduzindo a análise dos dados como uma instância do processo de descoberta de conhecimento em bases de dados, também é uma novidade. A preparação dos dados segundo elementos epidemiológicos da ferrugem do cafeeiro, especificamente considerando variável o período de incubação, segundo uma equação de estimativa conhecida, em vez de considerá-lo como um valor médio fixo, também pode ser considerada inovadora.

A principal limitação dos modelos de alerta desenvolvidos está relacionada com a sua abrangência. O uso desses modelos deve ficar restrito à região onde os dados foram coletados ou a regiões com condições de clima parecidas. Regiões com clima diferente podem apresentar condições meteorológicas que não foram representadas nos dados analisados e que, portanto, podem condicionar o progresso da ferrugem do cafeeiro de maneira diferente do comportamento capturado pelos modelos de alerta.

Em relação a isso, seria interessante que, além da Fundação Procafé, outras instituições realizassem um acompanhamento sistemático das doenças e pragas do cafeeiro, em paralelo com os registros meteorológicos. O Instituto Agrônomo de Campinas - IAC, por exemplo, recentemente iniciou um monitoramento agrometeorológico da cafeicultura na região Mogiana do estado de São Paulo, nos municípios de Campinas e de Mococa, de maneira parecida com o realizado pela Fundação Procafé (PEZZOPANE et al., 2007).

Algumas vantagens de se ter esse tipo de acompanhamento realizado em diferentes regiões são: maior volume e diversidade de dados para a modelagem; geração de modelos de alerta mais confiáveis; ampliação dos horizontes para a utilização dos modelos; e disponibilidade dos dados para outras pesquisas, o que pode reduzir a necessidade e os custos

com experimentos conduzidos no formato tradicional. É necessário que haja, entretanto, uma preocupação das diferentes instituições com um mínimo de padronização da metodologia utilizada nos levantamentos de dados.

7.1 Sugestões para a continuidade do trabalho

São diversas as sugestões para a continuidade deste trabalho. Essas sugestões passam pela validação dos modelos de alerta desenvolvidos, pela ampliação do conjunto de dados de análise, por incrementos e alterações na parte metodológica e pela aplicação do processo de descoberta de conhecimento em bases de dados para a obtenção de modelos de alerta de outras doenças e pragas do cafeeiro.

A validação dos modelos de alerta poderia ser iniciada na própria Fundação Procafé. Os dados coletados a partir de outubro de 2006 poderiam ser utilizados para o início da validação. Em seguida, a validação continuaria pelo tempo necessário para se avaliar, com segurança, se os alertas emitidos pelos modelos são mesmo confiáveis ou não.

Os dados coletados a partir de outubro de 2006 podem também ser adicionados aos dados já analisados. Pode-se ainda juntar os dados coletados nos municípios de Carmo de Minas e de Boa Esperança. A partir daí, uma nova rodada de modelagem pode ser realizada, para se verificar se alguma alteração nos modelos vai ser ocasionada com a inclusão desses novos dados. Projetos em parceria podem ser considerados também, caso haja o interesse de alguma outra instituição de participar de um trabalho conjunto.

Pode-se ainda pensar em incrementos e/ou alterações na parte metodológica de execução do processo de descoberta de conhecimento em bases de dados. Algumas delas são (WITTEN e FRANK, 2005):

- Utilizar métodos de seleção automática de atributos. Esses métodos vão permitir selecionar subconjuntos de atributos preditivos diferentes daqueles considerados pelas três opções de seleção de atributos utilizadas neste trabalho. Sendo assim, novos modelos vão poder ser gerados e avaliados.
- Levar em consideração o custo de decisões ou classificações incorretas. Uma das formas de contornar o problema do desbalanceamento entre as classes é atribuir um custo maior para os erros em relação à classe minoritária. Dessa forma, o algoritmo de indução vai dar maior importância aos acertos relativos a essa classe minoritária, permitindo um equilíbrio relativo entre os acertos e os erros para cada classe.

- Usar outros algoritmos de indução de árvores de decisão ou de indução de regras de classificação. O próprio Weka, além do classificador J48, possui outras opções de algoritmo para a indução de árvores de decisão e também opções para a indução de regras de classificação.

Por fim, modelos de alerta de outras doenças e pragas do cafeeiro poderiam ser desenvolvidos, de acordo com a mesma abordagem metodológica utilizada neste trabalho. A Fundação Procafé, além da ferrugem do cafeeiro, faz o monitoramento de outras doenças e pragas da cultura do café, como a cercosporiose ou mancha de olho pardo (*Cercospora coffeicola*), a mancha de phoma (*Phoma costaricensis*), o bicho-mineiro (*Leucoptera coffeella*) e a broca-do-café (*Hypothenemus hampei*).

REFERÊNCIAS BIBLIOGRÁFICAS

- AGRIOS, G. N. **Plant pathology**. 3rd ed. San Diego: Academic Press, 1988. 803 p.
- ALFONSI, R. R.; ORTOLANI, A. A.; PINTO, H. S.; PEDRO JUNIOR, M. J.; BRUNINI, O. Associação entre nível de infecção da ferrugem do cafeeiro, variáveis climáticas e área foliar, observadas em *Coffea arabica* L. In: CONGRESSO BRASILEIRO DE PESQUISAS CAFEEIRAS, 2., 1974, Poços de Caldas, MG. **Resumos...** 1974. p. 80-83.
- APSNET. **APSnet Education Center**: plant disease lessons - coffee rust - disease cycle and epidemiology. Disponível em: <www.apsnet.org/education/lessonsPlantPath/Coffeerust/discycle.htm>. Acesso em: 16 fev. 2008.
- APTE, C.; WEISS, S. Data mining with decision trees and decision rules. **Future Generation Computer Systems**, v. 13, p. 197-210, 1997.
- BAKER, F. A.; VERBYLA, D. L.; HODGES, C. S.; ROSS, E. W. Classification and regression tree analysis for assessing hazard of pine mortality caused by *Heterobasidion annosum*. **Plant Disease**, v. 77, n. 2, p. 136-139, 1993.
- BATCHELOR, W. D.; YANG, X. B.; TSCHANZ, A. T. Development of a neural network for soybean rust epidemics. **Transactions of the ASAE**, v. 40, n. 1, p. 247-252, 1997.
- BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. Boca Raton: CRC Press, 1984. 358 p.
- BUTT, D. J.; ROYLE, D. J. Multiple regression analysis in the epidemiology of plant diseases. In: KRANZ, J. (Ed.) **Epidemics of plant diseases: mathematical analysis and modeling**. 2nd ed. Berlin: Springer-Verlag, 1990. p. 143-180.
- CAMPBELL, C. L.; MADDEN, L. V. **Introduction to plant disease epidemiology**. New York: John Wiley & Sons, 1990. 532 p.
- CAMPBELL, C. L.; REYNOLDS, K. M.; MADDEN, L. V. Modeling epidemics of root diseases and development of simulators. In: KRANZ, J.; ROTEM, J. (Ed.) **Experimental techniques in plant disease epidemiology**. Berlin: Springer-Verlag, 1988. p. 253-265.
- CHALFOUN, S. M. **Doenças do cafeeiro**: importância, identificação e métodos de controle. Lavras, MG: UFLA/FAEPE. 1997.
- CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0**: step-by-step data mining guide. [Illinois]: SPSS, 2000. 78 p.
- CHAVES, G. M. et al. Ferrugem do cafeeiro: resultados preliminares de ensaios sobre avaliação de fungicidas, em Minas Gerais, e recomendação para controle químico da enfermidade, **Seiva**, 31, 1970 apud KUSHALAPPA, A. C. Rust management: an epidemiological approach and chemical control. In: KUSHALAPPA, A. C.; ESKES, A. B.

(Ed.) **Coffee rust: epidemiology, resistance, and management**. Boca Raton, Florida: CRC Press, 1989b. p. 81-139.

COAKLEY, S. M. Variation in climate and prediction of disease in plants. **Annual Review of Phytopathology**, v. 26, p. 163-181, 1988.

COSTA, R. V.; ZAMBOLIM, L.; VALE, F. X. R.; MIZUBUTI, E. S. G. Previsão da requeima da batateira. **Fitopatologia Brasileira**, Brasília, v. 27, n. 4, p. 349-354, 2002.

DE WOLF, E. D.; MADDEN, L. V.; LIPPS, P. E. Risk assessment models for wheat Fusarium head blight epidemics based on within-season weather data. **Phytopathology**, v. 93, n. 4, p. 428-435, 2003.

DEL PONTE, E. M.; GODOY, C. V.; LI, X.; YANG, X. B. Predicting severity of Asian soybean rust epidemics with empirical rainfall models. **Phytopathology**, v. 96, n. 7, p. 797-803, 2006.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **Mineração de dados para aplicação em alertas agrometeorológicos e fitossanitários para as culturas de café e cana-de-açúcar no Estado de São Paulo**. Campinas: Embrapa Informática Agropecuária, 2001a. 21 p.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **Desenvolvimento e evolução de um sistema de monitoramento agroclimático para o Estado de São Paulo**. Campinas: Embrapa Informática Agropecuária, 2001b. 15 p.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. In: FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). **Advances in knowledge discovery and data mining**. Menlo Park: AAAI Press, 1996a. p. 1-34.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996b.

FAYYAD, U., PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: towards a unifying framework. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2., 1996. **Proceedings...** Menlo Park: AAAI Press, p. 82-88, 1996c.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: an overview. **AI Magazine**, v. 14, n. 3, p. 57-70, 1992.

GARÇON, C. L. P., ZAMBOLIM, L., MIZUBUTI, E. S. G., VALE, F. X. R.; COSTA, H. Controle da ferrugem do cafeeiro com base no valor de severidade. **Fitopatologia Brasileira**, Brasília, v. 29, n. 5, p. 486-491, 2004.

- GENT, D. H.; SCHWARTZ, H. F. Validation of potato early blight disease forecast models for Colorado using various sources of meteorological data. **Plant Disease**, v. 87, n. 1, p. 78-84, 2003.
- GLEASON, M. L.; TAYLOR, S. E., LOUGHIN, T. M.; KOEHLER, K. J. Development and validation of an empirical model to estimate the duration of dew periods. **Plant Disease**, v. 78, n. 10, p. 1011-1016, 1994.
- HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. San Francisco: Morgan Kaufmann Publishers, 2001. 550 p.
- HARDWICK, N. V. Disease forecasting. In: JONES, D. G. (Ed.) **The epidemiology of plant diseases**. Boston: Kluwer Academic Publishers, 1998. p. 207-230.
- HAU, B.; KRANZ, J. Mathematics and statistics for analysis in epidemiology. In: KRANZ, J. (Ed.) **Epidemics of plant diseases**: mathematical analysis and modeling. 2nd ed. Berlin: Springer-Verlag, 1990. p. 12-52.
- HUBER, L.; GILLESPIE, T. J. Modeling leaf wetness in relation to plant disease epidemiology. **Annual Review of Phytopathology**, Palo Alto, v. 30, p. 553-577, 1992.
- JAPIASSÚ, L. B.; GARCIA, A. W. R.; MIGUEL, A. E.; CARVALHO, C. H. S.; FERREIRA, R. A.; PADILHA, L.; MATIELLO, J. B. Influência da carga pendente, do espaçamento e de fatores climáticos no desenvolvimento da ferrugem do cafeeiro. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 5., 2007, Águas de Lindóia, SP. **Anais...** Brasília: Embrapa, 2007. 1 CD-ROM.
- JENSEN, R. E.; BOYLE, L. W. A technique for forecasting leafspot on peanuts. **Plant Disease Reporter**, v. 50, n. 11, p. 810-814, 1966.
- KDNUGGETS. **KDnuggets**: data mining, web mining, text mining, and knowledge discovery. Disponível em: <www.kdnuggets.com>. Acesso em: 23 jan. 2008.
- KIM, K. S.; TAYLOR, S. E.; GLEASON, M. L.; KOEHLER, K. J. Model to enhance site-specific estimation of leaf wetness duration. **Plant Disease**, v. 86, n. 2, p. 179-185, 2002.
- KUSHALAPPA, A. C. Epidemiologia da ferrugem do cafeeiro sob alta densidade de plantio: um enfoque de sistema. In: SIMPÓSIO INTERNACIONAL SOBRE CAFÉ ADENSADO, 1994, Londrina, PR. **Anais...** Londrina: IAPAR, p. 131-147, 1994.
- KUSHALAPPA, A. C. Biology and epidemiology. In: KUSHALAPPA, A. C.; ESKES, A. B. (Ed.) **Coffee rust**: epidemiology, resistance, and management. Boca Raton, Florida: CRC Press, 1989a. p. 13-80.
- KUSHALAPPA, A. C. Rust management: an epidemiological approach and chemical control. In: KUSHALAPPA, A. C.; ESKES, A. B. (Ed.) **Coffee rust**: epidemiology, resistance, and management. Boca Raton, Florida: CRC Press, 1989b. p. 81-139.

KUSHALAPPA, A. C.; ESKES, A. B. Advances in coffee rust research. **Annual Review of Phytopatology**, v. 27, p. 503-531, sept. 1989.

KUSHALAPPA, A. C.; HERNANDEZ, T. A.; LEMOS, H. G. Evaluation of simple and complex coffee rust forecasts to time fungicide application. **Fitopatologia Brasileira**, Brasília, v. 11, p. 515-26, out. 1986.

KUSHALAPPA, A. C.; AKUTSU, M.; OSEGUERA, S. H.; CHAVES, G. M.; MELLES, C. Equations for predicting the rate of coffee rust development based on net survival ratio for monocyclic process of *Hemileia vastatrix*. **Fitopatologia Brasileira**, Brasília, v. 9, p. 255-271, jun. 1984.

KUSHALAPPA, A. C.; AKUTSU, M.; LUDWIG, A. Application of survival ratio for monocyclic process of *Hemileia vastatrix* in predicting coffee rust infection rates. **Phytopathology**, St. Paul, v. 73, n. 1, p. 96-103, 1983.

LAVRAC, N.; FLACH, P.; ZUPAN B. Rule evaluation measures: a unifying view. In: DZEROSKI, S.; FLACH, P. (Ed.) Proceedings of the 9th International Workshop on Inductive Logic Programming. **Lecture Notes in Artificial Intelligence**, v. 1634, p. 174-185, 1999.

LIMA, L. C.; SILVA, V. O. da. Bases sólidas para a cafeicultura brasileira. **Revista do Café**, Rio de Janeiro, v. 85, n. 820, p. 26-27, dez. 2006.

MADDEN, L. V.; LIPPS, P. E.; DE WOLF, E. Developing forecasting systems for fusarium head blight. In: INTERNATIONAL SYMPOSIUM ON FUSARIUM HEAD BLIGHT, 2., 2004, Orlando. **Proceedings...** Michigan: East Lansing, v. 2, p. 471, 2004.

MADDEN, L. V.; ELLIS, M. A. How to develop plant disease forecasters. In: KRANZ, J.; ROTEM, J. (Ed.) **Experimental techniques in plant disease epidemiology**. Berlin: Springer-Verlag, 1988. p. 191-208.

MADDEN, L.; PENNYPACKER, S. P.; MAC NAB, A. A. FAST, a forecast system for *Alternaria solani* on tomato. **Phytopathology**, v. 68, p. 1354-1358, 1978.

MAGAREY, R. D.; SEEM, R. C.; RUSSO, J. M.; ZACK, J. W.; WAIGHT, K. T.; TRAVIS, J. W.; OUDEMANS, P. V. Site-specific weather information without on-site sensors. **Plant Disease**, v. 85, n. 12, p. 1216-1226, 2001.

MATIELLO, J. B.; MANSK, Z. Estudo de esquemas de controle à ferrugem do cafeeiro em lavouras com alta, média e baixa produção, no estado do Espírito Santo. In: CONGRESSO BRASILEIRO DE PESQUISAS CAFEIEIRAS, 11., 1984, Londrina, PR. **Resumos...** 1984. p. 107-108.

MEIRA, C. A. A.; RODRIGUES, L. H. A.; MORAES, S. A. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. **Tropical Plant Pathology**, v. 33, n. 2, p. 114-124, Mar./Apr. 2008.

MEIRA, C. A. A.; RODRIGUES, L. H. A. Preparação de dados para obtenção de modelos de alerta da ferrugem do cafeeiro. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 6., 2007, São Pedro. **Anais...** Campinas: Embrapa Informática Agropecuária/SBI-AGRO, 2007a. p. 381-385. SBI-Agro 2007. 1 CD-ROM.

MEIRA, C. A. A.; RODRIGUES, L. H. A. **Processo de descoberta de conhecimento em bases de dados aplicado em alertas de doenças de culturas agrícolas.** Campinas: FEAGRI-UNICAMP, 2007. Trabalho apresentado no VI Workshop de Pós-graduação, Faculdade de Engenharia Agrícola, UNICAMP, Campinas, SP, jun. 2007b. Não paginado. 1 CD-ROM.

MEIRA, C. A. A.; RODRIGUES, L. H. A. Entendimento e preparação de dados no processo de descoberta de conhecimento aplicado a sistema de alerta da ferrugem do cafeeiro. In: CONGRESSO BRASILEIRO DE ENGENHARIA AGRÍCOLA, 35., 2006, João Pessoa. **Agroenergia e desenvolvimento tecnológico: anais.** João Pessoa: SBEA, 2006a. Não paginado. 1 CD-ROM.

MEIRA, C. A. A.; RODRIGUES, L. H. A. **Processo de descoberta de conhecimento em bases de dados aplicado em alertas de doenças de culturas agrícolas.** Campinas: FEAGRI-UNICAMP, 2006. Trabalho apresentado no I Workshop de Mineração de Dados Agrícolas, Faculdade de Engenharia Agrícola, UNICAMP, Campinas, SP, abr. 2006b. Não paginado. Resumo.

MEIRA, C. A. A.; RODRIGUES, L. H. A. Mineração de dados no desenvolvimento de sistemas de alerta contra doenças de culturas agrícolas. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 5., 2005, Londrina. **Agronegócio, tecnologia e inovação: anais.** Londrina: FAPEAGRO/SBI-AGRO, 2005. Não paginado. SBI-Agro 2005. 1 CD-ROM.

MICHALSKI, R. S.; BRATKO, I.; KUBAT, M. (Ed.). **Machine learning and data mining: methods and applications.** Baffins Lane (UK): John Wiley & Sons, 1998. 456 p.

MICHEL, C. A.; MENDES, C. S.; REIS, E. M. Validação de sistemas de previsão de epidemias causadas por *Phytophthora infestans* na cultura da batata. I – safra 1996/97. **Fitopatologia Brasileira**, Brasília, v. 22, Suplemento, p. 285, 1997. (Resumo).

MOLINEROS, J. E.; DE WOLF, E. D.; FRANCL, L.; MADDEN, L.; LIPPS, P. Modeling epidemics of fusarium head blight: trials and tribulations. **Phytopathology**, v. 95(Suppl.):S71. 2005.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Org.). **Sistemas inteligentes: fundamentos e aplicações.** Barueri: Editora Manole, 2002a. p. 89-114.

MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. In: REZENDE, S. O. (Org.). **Sistemas inteligentes: fundamentos e aplicações.** Barueri: Editora Manole, 2002b. p. 115-139.

- MONTOYA, R. H.; CHAVES, G. M. Influência da temperatura e da luz na germinação, infectividade e período de geração de *Hemileia vastatrix* Berk. & Br. **Experientiae**, v. 18, n. 11, p. 239-266, 1974.
- MORAES, S. A. Monitoramento das doenças foliares do amendoim e avisos climáticos para indicar as pulverizações com fungicidas. **O Agrônomo**, Campinas, v. 51(2,3): 86-89, 1999.
- MORAES, S. A. **A ferrugem do cafeeiro**: importância, condições predisponentes, evolução e situação no Brasil. Campinas: Instituto Agrônomo, 1983. 50 p. (IAC. Circular, 119).
- MORAES, S. A.; SUGIMORI, M. H.; RIBEIRO, I. J. A.; ORTOLANI, A. A.; PEDRO JR., M. J. Período de incubação de *Hemileia vastatrix* Berk. et Br. em três regiões do Estado de São Paulo. **Summa Phytopathologica**, Piracicaba, v. 2, n. 1, p. 32-38, 1976.
- PARVIN JR., D. W.; SMITH, D. H.; CROSBY, F. L. Development and evaluation of a computerized forecasting method for *Cercospora* leafspot of peanuts. **Phytopathology**, St. Paul, v. 64, p. 385-388, 1974.
- PAUL, P. A.; MUNKVOLD, G. P. Regression and artificial neural network modeling for the prediction of gray leaf spot of maize. **Phytopathology**, St. Paul, v. 95, n. 4, p. 388-396, 2005.
- PAUL, P. A.; MUNKVOLD, G. P. A model-based approach to preplanting risk assessment for gray leaf spot of maize. **Phytopathology**, St. Paul, v. 94, n. 12, p. 1350-1357, 2004.
- PEDRO JÚNIOR, M. J.; MORAES, S. A.; GODOY, I. J. Agrometeorological forecasting method for cercospora leafspot in peanuts. **Fitopatologia Brasileira**, Brasília, v. 19, n. 1, p. 69-73, 1994.
- PEZZOPANE, J. R. M.; SOUZA, P. S. de; PEREIRA, S. P.; GALLO, P. B.; THOMAZIELLO, R. A.; ROLIM, G. S.; CAMARGO, M. B. P. de; FAZUOLI, L. C. Monitoramento agrometeorológico da cafeicultura na região Mogiana do estado de São Paulo - Safra 2005-2006. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 5., 2007, Águas de Lindóia, SP. **Anais...** Brasília: Embrapa, 2007. 1 CD-ROM.
- PINTO, A. C. S.; POZZA, E. A.; SOUZA, P. E.; POZZA, A. A. A.; TALAMINI, V.; BOLDINI, J. M.; SANTOS, F. S. Descrição da epidemia da ferrugem do cafeeiro com redes neuronais. **Fitopatologia Brasileira**, Brasília, v. 27, n. 5, p. 517-524, 2002.
- PYLE, D. **Data preparation for data mining**. San Francisco: Morgan Kaufmann, 1999. 540 p.
- QUINLAN, J. R. **C4.5: programs for machine learning**. San Francisco: Morgan Kaufmann, 1993.
- REFAAT, M. **Data preparation for data mining using SAS**. San Francisco: Morgan Kaufmann, 2006. 424 p.
- REIS, E. M. (Ed.) **Previsão de doenças de plantas**. Passo Fundo: UPF, 2004. 316 p.

- REIS, E. M.; BRESOLIN, A. C. R. Sistemas de previsão de doenças de plantas. In: REIS, E. M. (Ed.) **Previsão de doenças de plantas**. Passo Fundo: UPF, 2004. p. 155-287.
- REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. de Mineração de dados. In: REZENDE, S. O. (Org.). **Sistemas inteligentes: fundamentos e aplicações**. Barueri: Editora Manole, 2002. p. 307-335.
- SAS INSTITUTE INC. **SAS/INSIGHT[®] 9.1 user's guide**. Cary, NC: SAS Institute Inc., 2004a. 816 p.
- SAS INSTITUTE INC. **Getting started with SAS[®] Enterprise Miner[™] 4.3**. Cary, NC: SAS Institute Inc., 2004b. 126 p.
- SAS INSTITUTE INC. **SAS[®] Enterprise Miner[™] 4.3 reference help**. Cary, NC: SAS Institute Inc., 2004c. Não paginado.
- SHTIENBERG, D.; ELAD, Y. Incorporation of weather forecasting in integrated, biological-chemical management of *Botrytis cinerea*. **Phytopathology**, St. Paul, v. 87, n. 3, p. 332-340, 1997.
- SILVA-ACUÑA, R.; ZAMBOLIM, L.; CRUZ, C. D.; VALE, F. X. R. Estudo epidemiológico da ferrugem do cafeeiro (*Hemileia vastatrix*) utilizando a análise de trilha. **Fitopatologia Brasileira**, Brasília, v. 23, n. 4, p. 425-430, 1998.
- SOUZA, R. T.; FORCELINI, C. A.; REIS, E. M.; CALVETE, E. O. Validação de dois sistemas de previsão para a queima das folhas da cenoura. **Fitopatologia Brasileira**, Brasília, v. 27, n. 1, p. 87-90, 2002.
- SUTTON, J. C.; GILLESPIE, T. J.; HILDEBRAND, P. D. Monitoring weather factors in relation to plant disease. **Plant Disease**, v. 68, n. 1, p. 78-84, 1984.
- VALE, F. X. R.; ZAMBOLIM, L.; JESUS JUNIOR, W. C. Efeito de fatores climáticos na ocorrência e no desenvolvimento da ferrugem do cafeeiro. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 1., 2000, Poços de Caldas, MG. **Resumos Expandidos...** p. 171-174, 2000.
- WALLIN, J. R. Summary of recent progress in predicting late blight epidemics in United States and Canada. **American Potato Journal**, v. 39, p. 306-312, 1962.
- WITTEN, I. H.; FRANK, E. **Data mining: practical machine learning tools and techniques**. 2nd ed. San Francisco: Morgan Kaufmann, 2005. 525 p.
- XU, X.; HARRIS, D. C.; BERRIE, A. M. Modeling infection of strawberry flowers by *Botrytis cinerea* using field data. **Phytopathology**, St. Paul, v. 90, n. 12, p. 1367-1374, 2000.
- ZADOKS, J. C. A quarter century of disease warning, 1958 – 1983. **Plant Disease**, v. 68, n. 4, p. 352-355, 1984.

ZADOKS, J. C.; SCHEIN, R. D. **Epidemiology and plant disease management**. Oxford: Oxford University Press, 1979. 427 p.

ZAMBOLIM, L.; VALE, F. X. R.; COSTA, H.; PEREIRA, A. A.; CHAVES, G. M. Epidemiologia e controle integrado da ferrugem-do-cafeeiro. In: ZAMBOLIM, L. (Ed.). **O estado da arte de tecnologias na produção de café**. Viçosa: Suprema Gráfica e Editora, 2002. p. 369-449.

ZAMBOLIM, L.; VALE, F. X. R. do; PEREIRA, A. A.; CHAVES, G. M. Café (*Coffea arabica* L.): controle de doenças – doenças causadas por fungos, bactérias e vírus. In: VALE, F. X. R. do; ZAMBOLIM, L. (Ed.). **Controle de doenças de plantas: grandes culturas**. Viçosa: UFV, v. 1, 1997. p. 83-139.