CESARE DI GIROLAMO NETO

“DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS DE ALERTA PARA A FERRUGEM DO CAFEEIRO”

CAMPINAS
JUNHO DE 2013

UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE ENGENHARIA AGRÍCOLA

CESARE DI GIROLAMO NETO

"DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS DE ALERTA PARA A FERRUGEM DO CAFEEIRO"

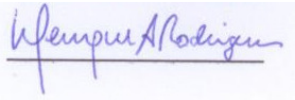
Orientador: Prof. Dr. Luiz Henrique Antunes Rodrigues

Coorientador: Prof. Dr. Carlos Alberto Alves Meira

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Agrícola da Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas para obtenção do título de Mestre em Engenharia Agrícola na área de Planejamento e Desenvolvimento Rural Sustentável.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO CESARE DI GIROLAMO NETO E ORIENTADA PELO PROF. DR. LUIZ HENRIQUE ANTUNES RODRIGUES

Assinatura do Orientador



CAMPINAS
JUNHO DE 2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

D569d

Di Girolamo Neto, Cesare, 1985-
Desenvolvimento e avaliação de modelos de alerta para a ferrugem do cafeeiro / Cesare Di Girolamo Neto. – Campinas, SP : [s.n.], 2013.

Orientador: Luiz Henrique Antunes Rodrigues.
Coorientador: Carlos Alberto Alves Meira
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Mineração de dados (Computação). 2. Hemileia vastatrix. 3. Café - Doenças e pragas. 4. Modelagem. I. Rodrigues, Luiz Henrique Antunes,1959-. II. Meira, Carlos Alberto Alves. III. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. IV. Título.

Informações para Biblioteca Digital

Título em inglês: Development and evaluation of warning models for coffee rust
Palavras-chave em inglês:
Data mining (Computing)
Hemileia vastatrix
Coffee - Diseases and pests
Modeling
Área de concentração: Planejamento e Desenvolvimento Rural Sustentável
Titulação: Mestre em Engenharia Agrícola
Banca examinadora:
Luiz Henrique Antunes Rodrigues [Orientador]
Stanley Robson de Medeiros Oliveira
Flávia Rodrigues Alves Patrício
Data de defesa: 19-06-2013
Programa de Pós-Graduação: Engenharia Agrícola

Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **Cesare Di Girolamo Neto**, aprovada pela Comissão Julgadora em 19 de junho de 2013 na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

Prof. Dr. Luiz Henrique Antunes Rodrigues – Presidente e Orientador
Feagri/Unicamp

Dra. Flávia Rodrigues Alves Patrício - Membro Titular
Instituto Biológico

Dr. Stanley Robson de Medeiros Oliveira – Membro Titular
Embrapa/CNPTIA

Dedico
à minha mãe FERNANDA,
aos meus irmãos, pais e avós,
com muito carinho.

AGRADECIMENTOS

Ao Prof. Dr. Carlos Alberto Alves Meira e ao Prof. Dr. Luiz Henrique Antunes Rodrigues, pela orientação, acompanhamento, amizade e oportunidade de crescimento profissional e pessoal.

À Empresa Brasileira de Pesquisa Agropecuária, pela oportunidade de capacitação profissional e pela oportunidade de utilizar as dependências físicas e a infraestrutura computacional durante o curso.

À Fundação PROCAFÉ, em especial ao Engenheiro Agrônomo Rodrigo Naves Paiva, por ceder os dados utilizados no desenvolvimento do trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo apoio financeiro.

Ao Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café pelo apoio aos projetos relacionados ao tema.

Ao professor Stanley Robson de Medeiros Oliveira pelos conselhos na área de mineração de dados.

Às professoras Raquel Ghini e Flávia Rodrigues Alves Patrício pelos conselhos na àrea de fitopatologia

Às professoras Mariangela Amendola e Maria Angela Fagnani pela amizade e conselhos para o desenvolvimento do trabalho.

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

Aos meus amigos e amigas que ajudaram durante o período de realização deste trabalho.

À mais alguém, que eu provavelmente esqueci.

À todos os meus familiares pelo apoio e compreensão nos momentos mais difíceis.

E a Deus.

"Na ciência, o crédito vai para o homem que convence o mundo de uma ideia, não para aquele que a teve primeiro"

William Osler

RESUMO

A ferrugem (causada pelo fungo *Hemileia vastatrix* Berk. e Br.) é a principal doença do cafeeiro. As perdas da produção causadas por esta doença podem chegar a 50%, caso nenhuma medida de controle seja adotada. O controle da ferrugem pode ser feito com fungicidas, entretanto métodos tradicionais de controle podem levar a aplicações desnecessárias, as quais são responsáveis por gerar gastos excessivos por parte do produtor, na compra e mão de obra para sua aplicação, além de causar impactos ambientais.

Ferramentas como modelos de predição, ou alerta, podem ser utilizadas para antecipar quando uma doença de planta pode ocorrer, sendo que uma predição correta evita aplicações desnecessárias de fungicidas. Neste sentido, modelos de alerta para a ferrugem do cafeeiro foram construídos por outros autores, entretanto, após o seu desenvolvimento, estes modelos não foram avaliados com dados externos ao conjunto de treinamento. Estes modelos passaram por um processo de validação neste trabalho e o resultado mostrou um desempenho abaixo do esperado, evidenciando a necessidade de se criarem novos modelos de alerta, com poder de predição maior do que os existentes.

O processo de descoberta de conhecimento em bases de dados foi realizado com o objetivo de gerar estes novos modelos de alerta, utilizando técnicas de mineração de dados como: redes neurais artificiais, máquinas de vetores suporte, florestas aleatórias e árvores de decisão. Dados meteorológicos e de espaçamento da lavoura foram as variáveis independentes do conjunto de dados. Os modelos de alerta foram desenvolvidos considerando taxa de progresso da ferrugem como atributo dependente, ou atributo meta, a qual consiste no aumento da incidência entre dois meses subsequentes. Este atributo foi de origem binária, seguindo os limites de 5 e 10 pontos percentuais - p.p. (classe ‘1’ para taxas maiores ou iguais ao limite; classe ‘0’, caso contrário). Foram desenvolvidos modelos de alerta para a cidade de Varginha e para a região Sul de Minas (com adição das cidades de Boa Esperança e Carmo de Minas), para dados entre 1998 e 2011. Os modelos são específicos para lavouras com alta carga pendente ou para lavouras com baixa carga, dado ao café ser uma cultura bianual.

Os modelos para a cidade de Varginha obtiveram, no geral, melhor desempenho do que aqueles contendo dados das 3 cidades juntas. Para alta carga pendente de frutos, a taxa de

acerto por validação cruzada, foi superior a 85%, considerando o alerta a partir de 5 p.p. Considerando o alerta a partir de 10 p.p., a taxa de acerto se aproximou dos 90%. Já para lavouras com baixa carga pendente, os modelos considerando o alerta a partir de 5 p.p. também chegaram a taxas de acerto próximas a 90%. Houve ainda equilíbrio entre outras medidas de desempenho importantes, como sensitividade, especificidade e confiabilidade positiva ou negativa em todos os modelos.

Os modelos mais bem avaliados mostraram ter desempenho superior aos modelos desenvolvidos por outros autores e têm potencial para servir como apoio na tomada de decisão referente à adoção de medidas de controle da ferrugem do cafeeiro.

PALAVRAS-CHAVE: mineração de dados; árvore de decisão; redes neurais artificiais, florestas aleatórias, máquinas de vetores suporte, previsão de doenças de plantas; *Hemileia vastatrix*.

ABSTRACT

Coffee rust (infection by the fungus *Hemileia vastatrix* Berk. e Br.) can cause up to 50% of yield losses, in the case no protective measures are taken. This disease can be controlled through fungicide applications, however, traditional control methods can lead to unnecessary use of these products, which cause, not only economic losses for the producer, on buying and applying the fungicides, but also major environmental impacts.

Tools like warning models can be used to predict when a plant disease may occur and a correct prediction might avoid unnecessary fungicide applications. According to this, some authors developed warning models for coffee rust, nevertheless, after their development, these models were not evaluated by a test set, besides the one used to create it. A Validation procedure was performed over these models, showing that their performance was way low than expected, highlighting the need for new warning models, with better performance than those previously developed.

The Knowledge Discovery in Database process was performed intending to develop new warning models by using four data mining techniques: Neural Networks, Support Vector Machines, Random Forests and Decision Trees. Meteorological and crop spacing data were designed as the independent variables. The dependent variable was labeled as the monthly progress rate of coffee rust, which consists on the increase of the incidence levels between two months in a row. It was mapped as a binary attribute, following the limits of five and ten percentage points (p.p.), considering the increase of the infection rate (class '1' for progress rate over or equal the limit, or class '0' otherwise). Models were developed considering 13 years (1998 – 2011) of incoming data for the city of Varginha – Minas Gerais – Brazil and for the South Minas Gerais region (by adding data from two more cities, Boa Esperança and Carmo de Minas). The models developed are specific for high or low fruit loads.

Warning models for Varginha obtained, usually, better performance than those developed with data from the three cities. For high fruit load, the accuracy by cross validation was higher than 85%, considering the warning over 5 p.p. Considering the warning over 10 p.p., the accuracy was close to 90%. For low fruit load, the models considering warning over 5

p.p. also obtained accuracy close to 90%. Other important performance measures, such as sensitivity (recall) and specificity, also obtained good values for all of these models.

The warning models developed on this study obtained better performance than others previously developed, and have a great potential to be used in decision-making systems, providing further information regarding the correct use of fungicides on controlling the coffee rust.

KEYWORDS: data mining; decision trees; neural networks, random forests, support vector machines, plant disease forecast; *Hemileia vastatrix*.

Lista de Figuras:

Lista de Tabelas:

Sumário:

1 Introdução

O Brasil é atualmente o maior produtor de café do mundo, sendo que em 2012 foi responsável por cerca de 37% da produção mundial, o equivalente a 55,9 milhões de sacas de 60 Kg. Com cerca de 60% da produção destinada ao mercado externo, os ganhos anuais do país com a exportação deste grão chegaram próximos a US$ 6 bilhões (USDA, 2013).

A ferrugem (*Hemileia vastatrix* Berk. e Br.) é a principal doença do cafeeiro, tanto para o café arábica (*Coffea arabica* L.) quanto para o café robusta (*Coffea canephora* P.) (WALLER et al., 2007). As perdas da produção causadas por esta doença podem chegar a 50%, caso nenhuma medida de controle seja adotada (ZAMBOLIM et al., 2002).

O controle da ferrugem pode ser feito com fungicidas, sendo sua aplicação recomendada para taxas de incidência da doença de 5% e 12%, para aplicação dos fungicidas protetores ou sistêmicos (curativos), respectivamente (ZAMBOLIM et al., 1997). Métodos tradicionais de controle, como a aplicação de fungicidas baseada em calendário fixo, em datas pré-estabelecidas, têm contribuído para um aumento da doença no final de seu ciclo de produção (CHALFOUN e CARVALHO, 1999). Estes métodos podem levar a aplicações desnecessárias de fungicidas, quando, por exemplo, uma aplicação for realizada sem a necessidade de se controlar a doença. Aplicações indiscriminadas geram gastos excessivos por parte do produtor, na compra dos fungicidas e na utilização de mão de obra para sua aplicação, além de causar impactos ambientais, como a contaminação do solo e de lençóis de água, podendo ainda ser responsável pela contaminação do próprio produto.

Estas observações evidenciam a importância de pesquisas capazes de gerar informações que possam ser utilizadas para fornecer suporte ao controle dessa doença, o que pode ser realizado por meio de modelos de predição, ou alerta. Um modelo desta natureza procura antecipar quando uma doença de planta pode atingir um nível crítico (HARDWICK, 2006). Uma predição correta evita aplicações desnecessárias, reduzindo os impactos ambientais e gastos por parte do produtor.

Neste sentido, Meira (2008) construiu modelos de alerta para a ferrugem do cafeeiro utilizando árvores de decisão. O desenvolvimento destes modelos contou com dados fornecidos pela Fundação PROCAFÉ para o município de Varginha/MG no período de 1998 a

2006. Estes modelos, no entanto, precisavam ser efetivamente validados. Não houve um processo de validação externa, ou seja, avaliar o desempenho destes modelos com dados externos ao conjunto utilizado para treiná-los, como dados a partir de 2007.

Atualmente, a fundação PROCAFÉ ampliou sua área de monitoramento de doenças cafeeiras e mantém três estações automáticas coletando dados meteorológicos nos municípios de Varginha, Carmo de Minas e Boa Esperança, todos no estado de Minas Gerais. Tais dados seguem atualizados e contêm mais de dez anos de registros, o que permite realizar o processo de validação externa dos modelos de Meira (2008), verificando se seu potencial de predição está mantido para condições mais recentes.

Os modelos desenvolvidos por Meira (2008) se enquadram na classe de modelos de simulação, nos quais os desenvolvedores e os usuários utilizam informações obtidas a partir dos resultados para dar suporte na tomada de decisões, por exemplo, determinar se é necessária a aplicação de fungicidas dado uma predição de aumento da taxa de progresso da ferrugem do cafeeiro. Entretanto, os indivíduos que são afetados pelas decisões baseadas em tais modelos precisam estar seguros de que os resultados estão corretos, por exemplo, se a previsão de que haverá um aumento na taxa de progresso da ferrugem está coerente. Esta preocupação é abordada por meio de um procedimento denominado “Verificação e Validação” de modelos (V&V) (SARGENT, 2013), a qual é normalmente definida como sendo a comprovação de que um modelo computadorizado, dentro de seu domínio de aplicação, possui uma taxa de acerto (acurácia) satisfatória, compatível com a aplicação pretendida (SCHLESINGER et al., 1979).

Processos de validação de modelos de simulação podem gerar informações que permitem incorporar mudanças estruturais nos mesmos, alterando suas características construtivas, sendo um processo iterativo na construção de um modelo de simulação válido (SARGENT, 2013). Normalmente, um modelo de simulação primário é desenvolvido e posteriormente passa por processos de ajuste e/ou adequações, visando o seu funcionamento “correto”. Cada alteração introduzida desta forma, pode ser capaz de trazer benefícios, como o aumento da taxa de acerto ou a ampliação da cobertura.

Também é possível alterar um modelo induzindo-o novamente, seja com a mesma técnica de modelagem utilizada na primeira indução, ou com técnicas diferentes. Na área de

mineração de dados existem diversas técnicas de modelagem além de árvores de decisão, como redes neurais artificiais, máquinas de vetores suporte e florestas aleatórias, as quais também podem ser utilizadas para uma nova indução de modelos de alerta para a ferrugem do cafeeiro. A escolha pelo uso de uma ou outra técnica de modelagem requer a análise do problema em questão, sendo que cada uma destas técnicas tem suas respectivas vantagens e desvantagens. No caso de modelos de alerta, a melhor opção poderia ser aplicar diversas técnicas de modelagem e no final escolher a combinação dos modelos que apresentassem os melhores resultados.

Sendo assim, a hipótese deste trabalho é que a geração de novos modelos de alerta para a ferrugem do cafeeiro por meio de diferentes técnicas de modelagem, poderia gerar modelos com poder de predição maior do que os gerados anteriormente, fornecendo novos subsídios para o monitoramento da ferrugem do cafeeiro no campo.

2 Objetivos

2.1 Objetivo geral

Este trabalho teve como objetivo desenvolver, comparar e selecionar modelos de alerta, baseados em técnicas de mineração de dados, para determinar o aumento da taxa de progresso da ferrugem do cafeeiro, buscando obter melhor desempenho do que modelos já existentes.

2.2 Objetivos específicos

Os objetivos específicos foram:

- Desenvolver novos modelos de alerta para determinar a taxa de progresso da ferrugem do cafeeiro utilizando técnicas de mineração de dados como árvores de decisão, redes neurais artificiais, máquinas de vetores suporte e florestas aleatórias.

- Comparar os modelos novos de alerta gerados por meio de suas medidas de avaliação e desempenho e validar os modelos de alerta já existentes.

- Selecionar os modelos que se adaptam melhor para representar o aumento da taxa de progresso da ferrugem do cafeeiro.

3 Revisão bibliográfica

3.1 Processo de descoberta de conhecimento

3.1.1 Visão geral, fases e etapas

Com a popularização da internet nos anos 90 e os avanços tecnológicos nas áreas de coleta, armazenamento e transmissão de grandes volumes de dados, o mundo se encontrou em situação "rica em dados, mas pobre em informação" (HAN et al., 2011). À medida que se buscava trabalhar com esses dados, percebeu-se que havia uma desproporção entre a quantidade de dados gerados e capacidade de analisar os dados, tornando-se inviável uma análise manual. Uma forma automatizada de tratar os dados seria um avanço valioso e traria vantagens competitivas consideráveis (FRAWLEY et al., 1992).

A demanda por geração de uma técnica computacional e ferramentas para análise de dados originou o campo da descoberta de conhecimento em bases de dados (KDD - *Knowledge Discovery in Databases*). Inicialmente, o KDD foi definido como a extração de informação implícita, previamente desconhecida e potencialmente útil a partir de dados (FRAWLEY et al., 1992). Posteriormente, Fayyad et al. (1996) revisaram o conceito de KDD, sendo este redefinido como sendo o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em repositórios de dados.

O processo KDD é composto de várias fases, sendo que para Fayyad et al. (1996), a mineração de dados (*data mining*) representa uma parte deste processo. Ela é a responsável pela extração de padrões embutidos em grandes volumes de dados, por meio da aplicação de algoritmos específicos. Uma aplicação imprudente de métodos de mineração de dados pode ser uma atividade perigosa e pode conduzir a descoberta de padrões incorretos ou sem sentido (AGRAWAL et al., 1996). A Figura 1 representa uma visão geral do processo KDD, o qual é constituído das seguintes fases (FAYYAD et al., 1996):

- Fase de seleção: Dados são selecionados de acordo com critérios pré-definidos.
- Fase de pré-processamento: Neste estágio ocorre a limpeza dos dados, há a exclusão de informações desnecessárias, tratamento de dados ausentes e outras atividades.
- Fase de transformação: Os dados são configurados de forma a atender às exigências de uma dada técnica de mineração de dados, ocorre a conversão de dados e a derivação de novos atributos.
- Fase de mineração de dados: Extraem-se padrões de comportamento dos dados, a partir de uma técnica pré-determinada.
- Fase de interpretação: Os padrões são interpretados gerando conhecimento, os quais darão suporte à tomada de decisões humanas.

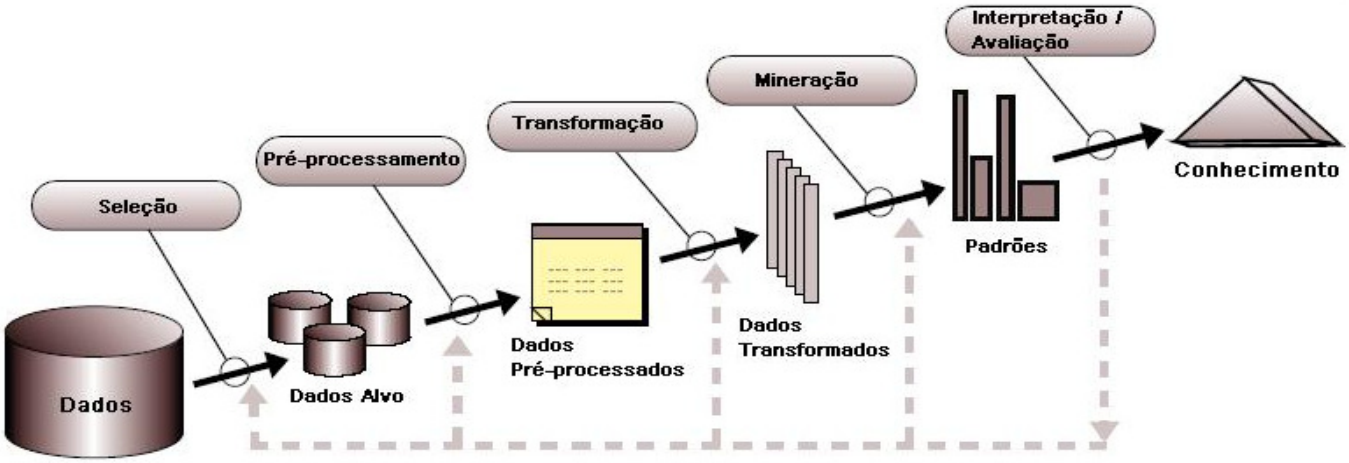


Figura 1: Fases do processo KDD (Adaptado de FAYYAD et al., 1996).

Esse processo pode envolver várias iterações e quase sempre é necessário o retorno para fases anteriores.

3.1.2 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 predizer 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 tarefas e a escolha de uma ou mais destas depende do problema em questão (HAN et al., 2011; FAYYAD et al., 1996). As tarefas tradicionais de mineração de dados estão representadas na Figura 2.

Cada tarefa de mineração de dados possui técnicas diferentes associadas, podendo haver abordagens híbridas, as quais 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 questão e a decisão de qual representação e estratégia de descoberta é a mais adequada.

Figura 2: Principais tarefas de mineração de dados (Adaptado de REZENDE, 2002).

TAREFAS DESCRITIVAS:

As tarefas descritivas procuram a identificação de padrões inerentes a determinado conjunto de dados, sendo que este conjunto não possui um atributo classe especificado. Entre essas tarefas, destacam-se as regras de associação, clusterização (agrupamento) e sumarização.

ASSOCIAÇÃO: As regras de associação foram introduzidas por Agrawal et al. (1993), elas descrevem a relação entre itens ou produtos de uma base de dados. Elas seguem um padrão da forma $X \rightarrow Y$ e $X \cap Y \rightarrow \phi$, onde X e Y são os conjuntos de valores (tais como artigos comprados por um cliente ou sintomas apresentados por um paciente). Por exemplo, o caso de um supermercado. O padrão "Clientes que compram pão também compram leite" representa uma regra de associação que reflete um padrão de comportamento dos clientes do supermercado.

AGRUPAMENTO: Este método consiste em agrupar os dados em grupos ou clusters, tal que os elementos de um determinado grupo tenham alta similaridade entre si e sejam diferentes dos elementos dos outros grupos. Seguindo o exemplo do supermercado, pode-se separar grupos de clientes pela frequência que estes vão às compras.

SUMARIZAÇÃO: Envolve meios de encontrar uma descrição compacta para um subconjunto de dados, como, por exemplo, derivação de regras resumidas ou visualização multivariada. No caso de um cliente de supermercado, suas compras poderiam ser sumarizadas aos principais itens adquiridos.

TAREFAS PREDITIVAS:

As tarefas preditivas têm como objetivo a construção de modelos, a partir de um determinado conjunto de dados, para posterior predição do comportamento de novos dados. As principais tarefas de predição são classificação e regressão.

CLASSIFICAÇÃO: Consiste na predição de um valor categórico ou discreto (atributo meta), e busca a construção de modelos a partir de um conjunto de exemplos pré-classificados corretamente, para posterior classificação de exemplos novos e desconhecidos (REZENDE,

2002). Seria como determinar se um cliente de um supermercado irá gastar muito ou pouco na sua próxima compra.

REGRESSÃO: Visa descobrir uma função que mapeie um item de dados para uma variável de predição de valor numérico contínuo. Análises de regressão ainda podem ser utilizadas quando o usuário está mais interessado em estimar alguns valores ausentes em seus dados, em vez de descobrir classes de objetos. O exemplo para este caso seria determinar quanto (em valor) o cliente do supermercado irá gastar em sua próxima compra.

Neste trabalho a tarefa utilizada é a classificação, com ênfase em quatro técnicas de aprendizado: árvores de decisão, redes neurais artificiais, florestas aleatórias e máquinas de vetores suporte.

3.1.2.1 Árvores de decisão

A indução de árvores de decisão é uma técnica de mineração de dados utilizada para descobrir regras de classificação para um atributo a partir da subdivisão dos dados em um conjunto que está sendo analisado. Árvores de decisão são simples representações de conhecimento e classificam exemplos em um número finito de classes (APTE e WEISS, 1997). Elas podem ser representadas graficamente por nós e ramos, parecido com uma árvore, mas no sentido invertido (WITTEN et al., 2011). Sua representação visual torna muito mais fácil para o usuário analisar e compreender os resultados (FAYYAD et al., 1996).

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 destes contém um teste sobre um ou mais atributos (variáveis independentes) e seus resultados formam os ramos da árvore. Cada regra tem início no nó raiz da árvore e caminha até uma de suas folhas (REZENDE, 2002).

Os algoritmos que constroem árvores de decisão buscam encontrar aqueles atributos e valores que provêm máxima segregação dos registros de dados, com respeito ao atributo que se quer classificar, a cada nível da árvore.

Normalmente a construção de uma árvore de decisão segue os seguintes passos:

1. Apresenta-se um conjunto de dados e a partir dele é criado o nó inicial (ou nó raiz), com um teste lógico que dividirá a árvore em dois ou mais ramos.
2. A partir da divisão do nó raiz, são gerados outros nós (ou nós internos), sendo que cada nó contém um novo teste lógico, que ramificará novamente a árvore.
3. A divisão dos nós internos continua até que se atinja um nó folha, o qual não irá ramificar mais a árvore.

A repetição deste procedimento caracteriza a recursividade da árvore de decisão (BREIMAN et al., 1984).

As árvores de decisão também podem ser chamadas de árvores de classificação ou de regressão, caso o atributo meta seja categórico ou numérico, respectivamente. Como exemplo simples, a Figura 3 apresenta a árvore de decisão sobre quem compra um computador. O atributo final é binário, sendo classificado como sim ou não (comprar).

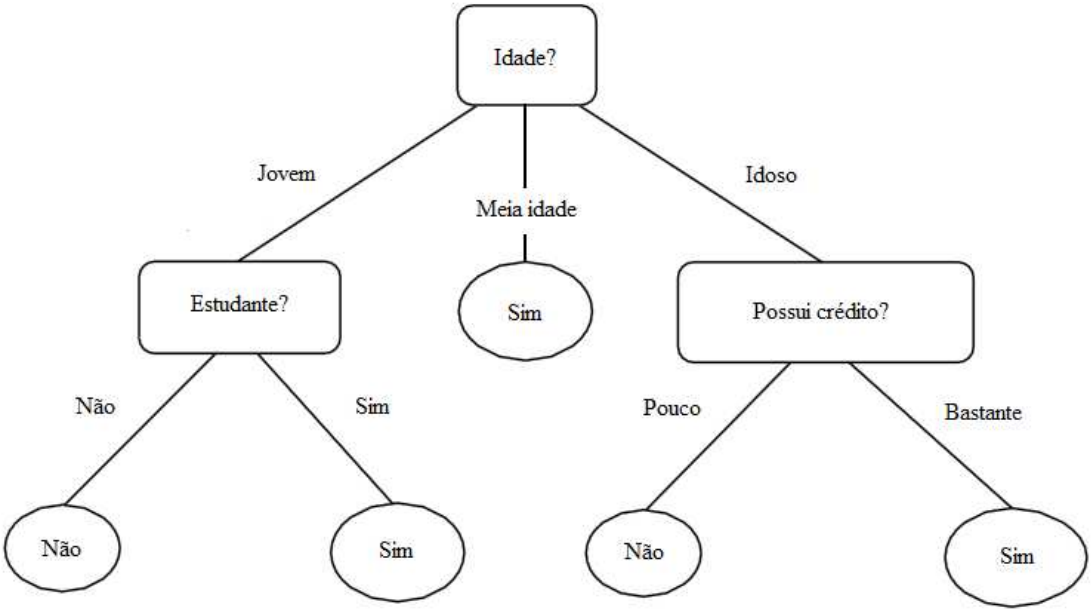


Figura 3: Representação de uma árvore de decisão (HAN et al., 2011)

3.1.2.2 Florestas aleatórias

As florestas aleatórias (*Random Forests*) são uma técnica de classificação desenvolvida por Breiman (2001). No algoritmo de árvore de decisão padrão, todo o conjunto de dados é utilizado para formular a árvore, já no algoritmo de florestas aleatórias, o conjunto de dados é

dividido aleatoriamente em diversos subconjuntos de tamanho menor. Cada um destes conjuntos é criado por um tipo de amostragem chamado de *bootstrap* (HAN et al., 2011), a qual é do tipo com reposição, ou seja, cada novo conjunto poderá ter alguns registros incluídos mais de uma vez e outros não inclusos nenhuma vez. A amostragem *bootstrap* garante que 1/3 dos exemplos são usados para testar as árvores após sua construção.

A partir de cada subconjunto desenvolvido, uma árvore de decisão é criada. A construção destas árvores ocorre por meio de uma seleção aleatória de atributos dos subconjuntos, os quais são utilizados nos nós de cada uma das árvores desenvolvidas. Uma floresta aleatória é uma coleção dessas árvores de decisão.

Quando a floresta está formada, há um número grande de árvores de decisão a serem testadas e todas contribuem para a classificação do objeto em estudo, por meio de um voto sobre qual classe o atributo meta deve pertencer. Cada voto tem um certo "peso", o qual é afetado pela similaridade entre cada árvore, sendo que quanto menor a similaridade entre duas árvores melhor, e pela força que cada árvore tem individualmente, ou seja, quanto mais precisa uma árvore for, melhor será sua nota. O ideal é manter a precisão das árvores sem aumentar sua similaridade (HAN et al., 2011).

O algoritmo é escalar e pode lidar com conjuntos com um grande número de atributos. O uso de subconjuntos e amostragem *bootstrap* tornam o algoritmo mais poderoso do que uma simples árvore, apresentando boa taxa de acerto quando testado em diferentes conjuntos de dados (BELLE, 2008). Esta técnica também pode ser considerada uma das mais precisas quando comparada a outras, como redes neurais artificiais e máquinas de vetores suporte (CARUANA et al., 2008). As florestas aleatórias ainda são computacionalmente muito efetivas, além de evitarem sobreajuste (*overfitting*) e serem pouco sensíveis a ruídos (BREIMAN, 2001).

3.1.2.3 Redes neurais artificiais

As redes neurais artificiais (RNA) são sistemas de inteligência artificial baseados no cérebro humano (HAYKIN, 2009), solucionando problemas por meio do aprendizado, errando e fazendo descobertas. Uma grande rede neural artificial pode ter centenas ou milhares de

unidades de processamento, enquanto que o cérebro de mamíferos pode ter muitos bilhões de neurônios.

Rumelhart et al. (1986) demonstraram que é possível treinar redes neurais com múltiplas camadas, resultando no modelo de redes neurais artificiais mais utilizado atualmente, as redes Perceptron de Múltiplas Camadas (MLP – *Multi Layer Perceptron*).

A composição desta rede está baseada em neurônios artificiais ou elementos de processamento. As redes MLP apresentam uma camada de entrada, na qual os elementos de processamento são chamados de inputs. Esta camada se liga a uma ou mais camadas intermediárias, por meio de conexões, e chega a uma camada de saída, também chamada de output. A camada de saída fornece os resultados do aprendizado da rede. A estrutura de uma RNA está apresentada na Figura 4.

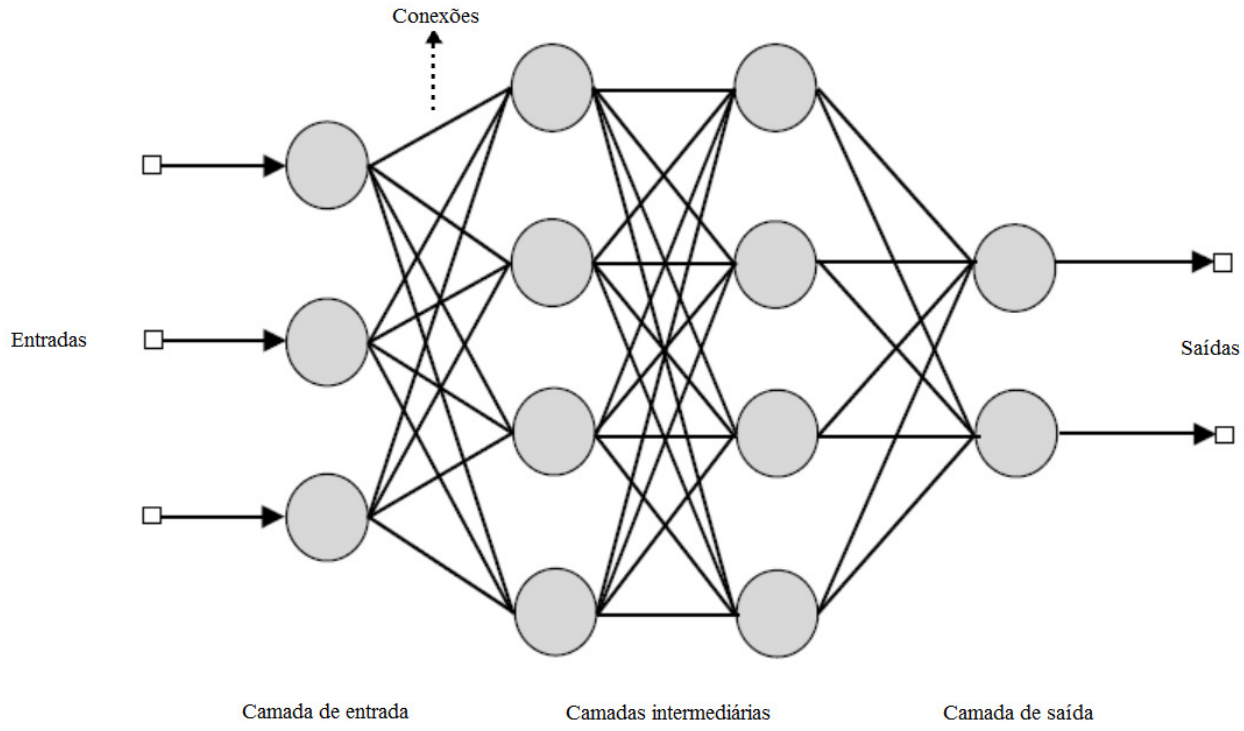


Figura 4: Estrutura de uma RNA (adaptada de JACOBS, 2011)

A rede apresentada como exemplo possui todas as conexões, o que significa que uma unidade (neurônio), em qualquer camada da rede, está conectada a todas as outras unidades na camada anterior.

O treinamento de redes MLP é do tipo supervisionado e, usualmente, utiliza-se o algoritmo chamado retro propagação do erro (*error backpropagation*). Este algoritmo é baseado numa regra de aprendizagem que corrige o erro durante o treinamento (HAYKIN, 2009). Os seguintes passos são utilizados no treinamento de uma rede:

1. Uma rede MLP recebe um sinal de entrada ou funcional (estímulo).
2. Este estímulo é propagado positivamente (neurônio a neurônio) através da rede, chegando à camada de saída como um sinal de saída.
3. O valor do sinal de saída é comparado ao valor desejado pelo usuário.
4. Caso este valor não atenda o valor desejado, é calculado um sinal que carrega um erro.
5. Este erro é retro propagado pela rede (neurônio a neurônio). As conexões sofrem ajustes de forma a minimizar este erro.

A grande vantagem de redes neurais é resolver problemas complexos que não encontram uma solução algorítmica nas tecnologias convencionais, ou que a solução seja muito difícil de ser encontrada (HAYKIN, 2009). Entretanto, as redes neurais funcionam como uma caixa preta, se tornando difíceis de serem compreendidas quando comparadas com algoritmos representativos, como as árvores de decisão (FAYYAD et al., 1996; HAYKIN, 2009).

3.1.2.4 Máquinas de vetores suporte

As máquinas de vetores suporte (SVM – *Support Vector Machines*) são classificadores que implementam modificações espaciais nos dados, levando-os a um plano onde a classificação se torna extremamente mais fácil.

Suponha que é necessário escolher um classificador para a Figura 5. Este classificador deve separar os círculos dos triângulos. Na parte (a) da Figura 5, tem-se um classificador que não erra, porém este pode ser muito específico a um determinado conjunto de dados. Isto pode causar uma situação de sobreajuste, quando o classificador reage muito bem apenas para o

conjunto em que foi treinado. Este classificador fica muito suscetível a erros quando confrontado a novos dados.

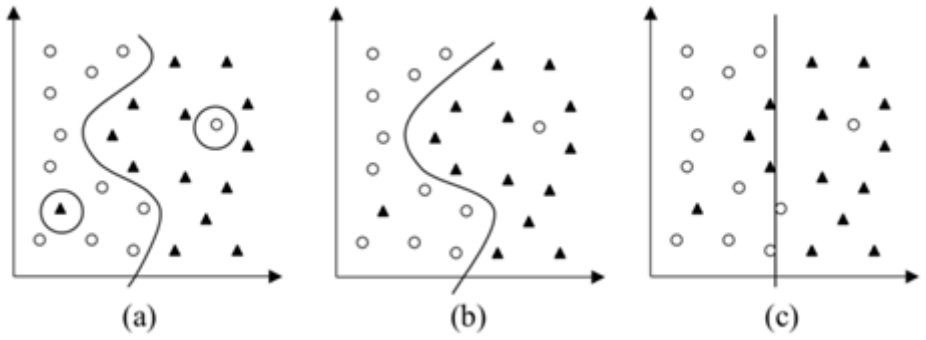


Figura 5: Exemplos de classificadores (SCHÖLKOPF e SAMOLA, 2002).

Na parte (c) há um classificador que comete muitos erros. Estes erros são cometidos pela baixa capacidade do classificador em distinguir pontos próximos pertencentes a diferentes classes. A taxa de acerto deste classificador seria muito inferior ao classificador (a), entretanto seria uma função muito mais simples. Um classificador mais complexo que (c) e não tão sobre ajustado quanto (a) seria o classificador (b). Ele tem uma complexidade intermediária e classifica bem grande parte dos exemplos.

A Teoria de Aprendizado Estatístico (VAPNIK, 2000) estabelece condições matemáticas que auxiliam na escolha de um classificador particular, a partir de um conjunto de dados de treinamento. Essas condições levam em conta o desempenho do classificador no conjunto de treinamento e a sua complexidade, com o objetivo de obter um bom desempenho também para novos dados do mesmo domínio (LORENA e CARVALHO, 2007). As Máquinas de Vetores Suporte foram desenvolvidas por meio dessa teoria.

As SVM implementam uma transcrição dos dados de entrada para um espaço característico de alta dimensão. Supondo-se que os dados da Figura 5 (b), ao serem levados a um espaço com dimensão maior, tivessem a distribuição da Figura 6. Nota-se que o processo de separação fica facilitado devido a essa transcrição.

Os pontos X1 e X2 são pertencentes a classes diferentes e podem ser separados por diversos hiperplanos compreendidos entre as linhas pontilhadas. A distância “d” representa os diversos hiperplanos que podem dividir este exemplo. As SVM trabalham em cima de um

processo de máximização da distância destes planos aos pontos de classes diferentes para determinar qual o melhor deles.

Para o exemplo da Figura 6, as transformações ocorreram em um hiperplano linear, entretanto as SVM são passíveis de representar planos de diversas formas. A elevação para um espaço hiperdimensional ocorre por meio de um produto interno chamado de *kernel*. Este produto pode ser do tipo linear ou não linear. Os não lineares são os polinomiais, os RBF (*Radial-Basis Function*) e os Sigmóides (LORENA e CARVALHO, 2007).

Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como as redes neurais artificiais (HAYKIN, 2009; BRAGA et al., 2007).

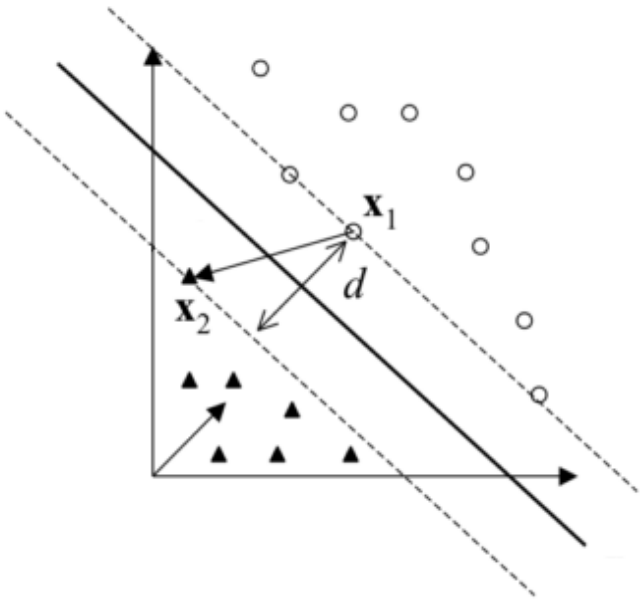


Figura 6: Hiperplanos gerados pelas SVMs (Adaptado de LORENA e CARVALHO, 2007).

3.1.3 Métodos de balanceamento de classes

Um problema que pode ocorrer na área de mineração de dados é quando um conjunto de dados tem uma classe representada por um grande número de exemplos enquanto que outra(s) classe(s) por um pequeno número, o que é comumente denominado de desbalanceamento de classes.

O procedimento para equilibrar o número de registros entre cada uma das classes, chamado de balanceamento de classes, deixa as classes do problema com um número próximo de registros, melhorando sua representatividade no conjunto de dados. Por exemplo: um conjunto de dados tem 100 registros, sendo 80 para a classe A e 20 para a classe B; ao se executar o balanceamento ele pode aumentar o número de registros da classe B para 60, totalizando 140 registros.

A presença de classes desbalanceadas pode influenciar no desempenho de um classificador, principalmente em casos em que há uma grande desproporção de exemplos entre duas classes. Isto pode causar uma dificuldade para o modelo reconhecer padrões ou regras que envolvam a classe com uma quantidade pequena de exemplos, chamada de classe minoritária.

Existem dois métodos para se realizar o balanceamento de classes, um deles é o método de *under-sampling* e o outro é o método de *over-sampling*. Estes métodos devem ser realizados antes da fase de modelagem.

O método de *under-sampling* consiste em balancear o conjunto de dados por meio da eliminação de exemplos da classe com mais exemplos, ou majoritária, enquanto que o método de *over-sampling* visa aumentar o número de exemplos da classe minoritária.

Existem diversas técnicas de *under-sampling* e *over-sampling*, sendo que Batista et al. (2004) fizeram uma revisão sobre quais são estas, além de estudarem a combinação entre elas. Um resultado que se destacou no trabalho dos autores foi a combinação de uma técnica de *over-sampling*, chamada de SMOTE (*Synthetic Minority Over-sampling Technique*), seguida de uma técnica de *under-sampling*, chamada de Pontos Tomek, ou *Tomek Links*. Esta combinação foi recomendada por Batista et al. (2004) pela sua eficiência em conjuntos de dados com poucos registros.

A técnica SMOTE (CHAWLA et al., 2002) consiste em aumentar exemplos da classe minoritária por meio da interpolação entre os pontos desta classe que se encontram próximos. As partes (a) e (b) da Figura 7 ilustram a ação do SMOTE em um conjunto de dados.

Já a técnica chamada de *Tomek Links* (TOMEK, 1976) consiste em identificar dois pontos de classes diferentes, de forma que não haja nenhum outro ponto mais próximo a estes dois, constituindo assim, um ponto Tomek. A partir da identificação deste ponto, o exemplo da

classe majoritária é removido. A Figura 7(c) mostra a identificação dos pontos tomek da figura Figura 7(b) e a Figura 7(d) mostra como fica o conjunto após a remoção da classe majoritária.

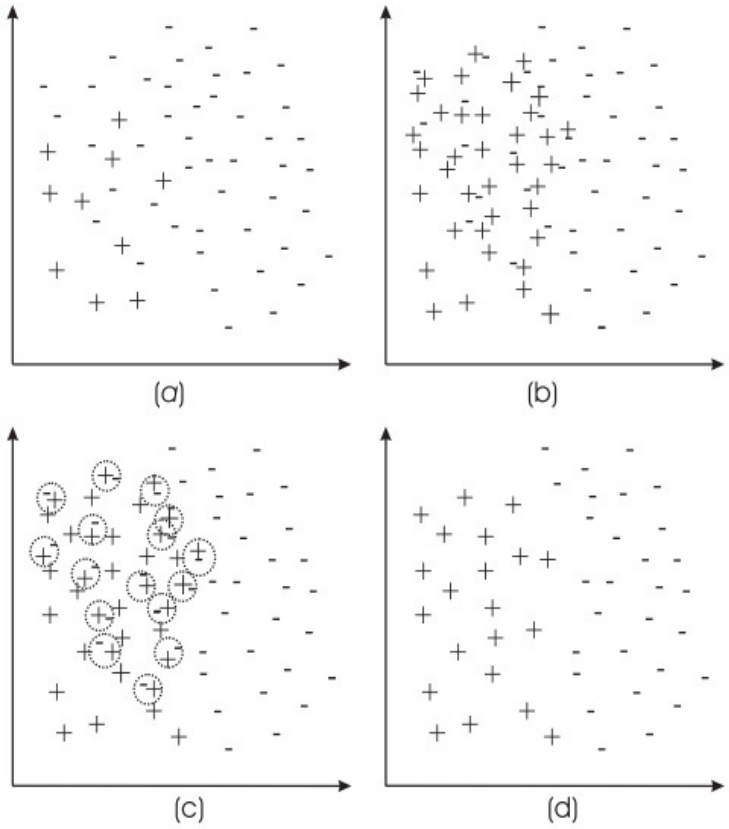


Figura 7: A união da técnicas de SMOTE + TOMEK LINKS (BATISTA et al., 2004)

Ainda se tratando de balanceamento de classes, resta saber qual a proporção de registros que se deve ter entre as classes de um conjunto de dados. Weiss e Provost (2001) realizaram um estudo sobre qual a quantidade de exemplos entre duas classes que deve ser utilizada na indução de um modelo. Um dos resultados do estudo foi que um conjunto de treinamento com 50% de exemplos da classe minoritária fornecem resultados frequentemente superiores aos resultados obtidos pela distribuição natural das classes. Às vezes estes resultados podem ser iguais, mas nunca são inferiores. Na prática, o procedimento seria deixar cada uma das duas classes com aproximadamente 50% dos registros.

3.1.4 Métodos de seleção de atributos

Seleção de atributos é o processo de refinamento do conjunto de treinamento pela escolha dos atributos mais importantes, sendo que a ideia por trás destes métodos é selecionar os atributos mais importantes do conjunto de dados e ignorar os demais.

A utilização de métodos de seleção de atributos tem duas finalidades. Primeiramente, faz o treinamento e a aplicação de um classificador mais eficiente, diminuindo a quantidade de atributos e tornando o processo de modelagem menos custoso computacionalmente. Em segundo lugar, pode aumentar a precisão de um classificador, eliminando atributos que podem confundi-lo. Tais atributos podem levar a uma generalização incorreta de uma característica acidental do conjunto de treinamento e, consequentemente, a erros de classificação por parte dos modelos (GUYON e ELISSEEFF, 2003).

Existem diversos métodos de seleção de atributos, cada um com uma dada característica. Um deles é o método do Qui quadrado (χ^2), o qual avalia os atributos individualmente com relação à classe. O teste estatístico do Qui quadrado é utilizado para determinar se há correlação entre o atributo e a classe. Após o teste, os atributos que contribuem efetivamente são ranqueados conforme o maior valor obtido no teste, sendo estes os que mais contribuem ao modelo (LIU e SETIONO, 1995).

Wrappers também são um método de seleção de atributos. Buscando um desempenho superior aos demais métodos, os wrappers levam em consideração o algoritmo que estará sendo usado sobre o conjunto de dados na fase de modelagem. Wrappers funcionam como uma caixa preta, calculando uma pontuação para um determinado subconjunto. Destes subconjuntos, o mais bem pontuado (que leva a maior taxa de acerto do classificador) é o selecionado. Wrappers tendem a levar a maior precisão, mas precisam de esforço computacional elevado quando comparados com outros métodos. Este método apresenta ganhos acentuados na taxa de acerto quando o algoritmo de árvores de decisão é utilizado (JOHN e KOHAVI, 1997).

Outro método de seleção de atributos é o CFS (do inglês, *Correlation Feature Selection*). Inicialmente, um conjunto aleatório é escolhido e classificado de acordo com uma medida de correlação com a classe, chamada de mérito. Quanto maior o mérito, mais o conjunto estará relacionado com a classe. O algoritmo busca novos conjuntos com méritos

superiores ao primeiro e após cinco testes com subconjuntos de mérito inferior, o conjunto atual é selecionado (HALL, 1999).

Além destes, existem métodos chamados de Infogain e Gainratio, os quais avaliam medidas como o ganho de informação e a taxa de ganho de informação entre atributos e classe (WITTEN et al., 2011).

A seleção de atributos em modelos de alerta do cafeeiro em árvores de decisão se mostrou eficaz para a melhora de medidas de avaliação e desempenho destes modelos. Girolamo Neto et al. (2012a) utilizaram as cinco técnicas mostradas anteriormente e a base de dados desenvolvida por Meira (2008) para avaliar os impactos da seleção de atributos em modelos de alerta da ferrugem do cafeeiro com alta carga pendente de frutos. Em 80% dos casos avaliados, a seleção de atributos promoveu ganhos na taxa de acerto.

3.1.5 Medidas de avaliação de desempenho

Diversas medidas de avaliação de desempenho podem ser usadas para analisar o comportamento de um modelo. Nesta seção elas se encontram divididas em três tópicos, o primeiro são medidas de desempenho derivadas de uma matriz que mostra os acertos e erros do modelo, chamada de matriz de confusão, seguido de uma análise gráfica do tipo ROC (*Receiver Operating Characteristics*) e o índice de concordância Kappa.

3.1.5.1 Matriz de confusão

A taxa de acerto e o erro são as medidas de avaliação mais comuns para modelos de classificação (WITTEN et al., 2011). São estimativas dos percentuais de acertos e erros do modelo na predição da classe de novos exemplos. Essas medidas podem ser calculadas a partir da matriz de confusão, que também oferece outros meios efetivos para a avaliação de um classificador (MONARD e BARANAUSKAS, 2002).

Para um problema com duas classes, denominadas classe positiva e classe negativa, a matriz de confusão (Figura 8) indica as quatro possibilidades de acertos e de erros do classificador:

- Verdadeiros positivos (**VP**): quando os exemplos de valor real “SIM” forem preditos como “SIM”.
- Falsos negativos (**FN**): quando os exemplos de valor real “SIM” forem preditos como “NÃO”.
- Verdadeiros negativos (**VN**): quando os exemplos de valor real “NÃO” forem preditos como “NÃO”.
- Falsos positivos (**FP**): quando os exemplos de valor real “NÃO” forem preditos como “SIM”.

	Predição: SIM	Predição: NÃO
Valor real: SIM	VP	FN
Valor real: NÃO	FP	VN

Figura 8: Exemplo de uma matriz de confusão.

Outras medidas de avaliação também podem ser derivadas da matriz de confusão (HAN et al., 2011): sensitividade, especificidade, confiabilidade positiva, e confiabilidade negativa.

A sensitividade *(Recall ou TPR – True Positive Rate)* é a proporção de exemplos positivos que foram classificados corretamente, já a especificidade *(TNR – True Negative Rate)* é a proporção de exemplos negativos que foram classificados corretamente.

A confiabilidade positiva *(Precision ou PPV – Positive Predicted Value)* é a proporção de exemplos positivos classificados corretamente dentre todos os exemplos classificados como positivos. A confiabilidade negativa *(NPV – Negative Predicted Value)* é a proporção de exemplos negativos classificados corretamente dentre todos os exemplos que foram classificados como negativos.

As equações de cálculo das medidas de avaliação são as seguintes:

$$Taxa.de.acerto = \frac{VP + VN}{n} \quad (1)$$

$$Erro = \frac{FP + FN}{n} \quad (2)$$

$$Sensitividade = \frac{VP}{VP + FN} \quad (3)$$

$$Especificidade = \frac{VN}{VN + FP} \quad (4)$$

$$Confiabilidade.Positiva = \frac{VP}{VP + FP} \quad (5)$$

$$Confiabilidade.Negativa = \frac{VN}{VN + FN} \quad (6)$$

Onde n é o número total de exemplos no conjunto.

As medidas de avaliação de um modelo podem ser geradas por diversas técnicas de amostragem, uma delas chama-se validação cruzada. Trata-se de um método que está baseado na divisão do conjunto de dados em N partições, mutuamente exclusivas, de tamanho igual. O modelo é gerado a partir de N-1 partes, e testado na parte que foi removida do conjunto de treinamento. Este procedimento é repetido N vezes, até que todas as partes tenham sido usadas para teste do modelo. A validação cruzada realizada com 10 partições permite obter as melhores medidas de avaliação sobre um modelo (WITTEN et al., 2011).

3.1.5.2 Gráfico ROC

Uma alternativa à avaliação utilizando medidas derivadas da matriz de confusão é a utilização de gráficos de dispersão e/ou diagramas. Gráficos permitem uma melhor visualização da multidimensionalidade do problema de avaliação (PRATI et al., 2008).

O gráfico ROC é baseado na probabilidade de detecção, ou taxa de verdadeiros positivos (*TPR – True Positive Rate*) - (VP/(VP+FN)), e na probabilidade de falsos positivos,

(FPR – *False Positive Rate*) - (FP/(FP+VN)). Para se construir o gráfico ROC, coloca-se FPR no eixo das ordenadas e FPR no eixo das abscissas (Figura 9).

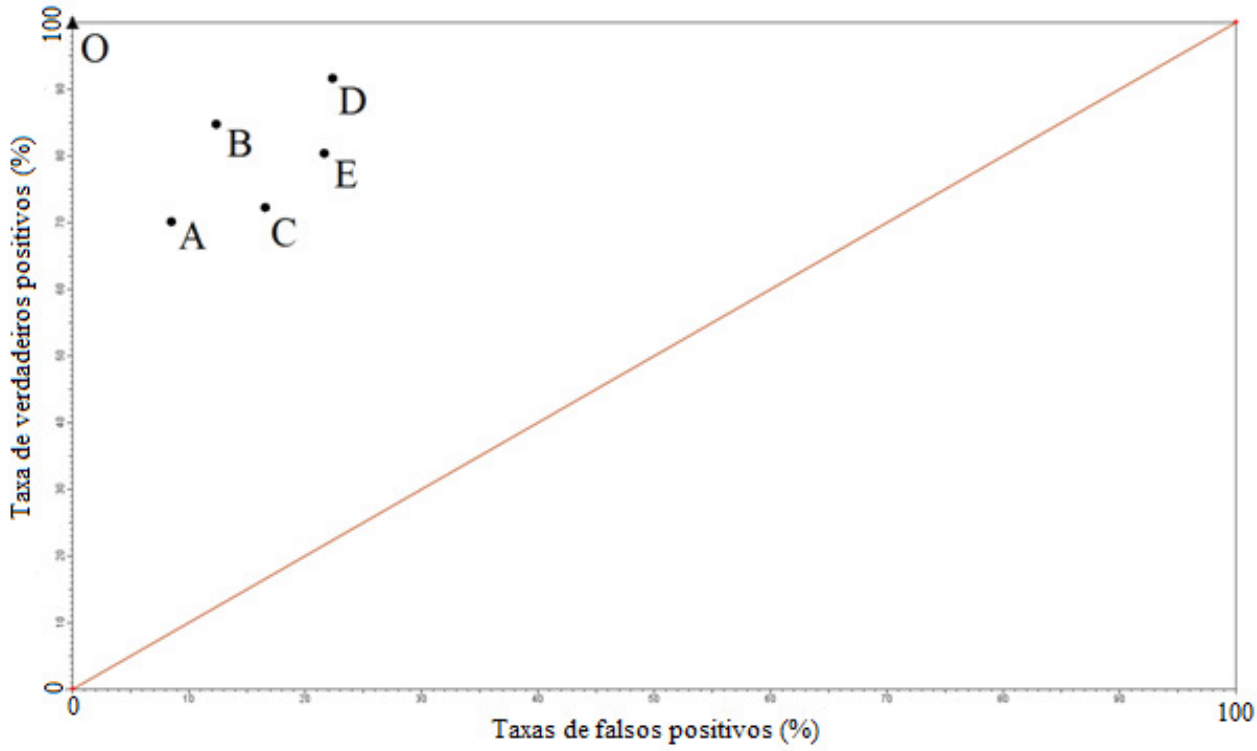


Figura 9: Exemplo de um Gráfico ROC.

Um modelo de classificação é representado por um ponto no espaço ROC, dado sua TPR e FPR. Para a Figura 9, os pontos A,B,C,D,E e O seriam exemplos de modelos dispostos no gráfico. Um classificador perfeito corresponderia ao ponto O no topo do gráfico, entretanto dificilmente esta perfeição é alcançada. A linha diagonal indica uma classificação aleatória, ou seja, um modelo que aleatoriamente seleciona saídas, obtendo 50% de acerto para cada classe em um problema com apenas duas classes.

Os modelos normalmente se situam entre a linha de desempenho aleatório e o ponto O, onde quanto maior a distância da linha diagonal, melhor o sistema. No entanto, um modelo que esteja localizada abaixo da diagonal ainda pode ser convertido num bom sistema – basta inverter suas saídas e então seu ponto também será invertido.

Ao se analisar um grupo de modelos no espaço ROC, pode-se notar a presença de um "envelope externo convexo" (*convex hull*), mostrado na Figura 10. Os modelos que se encontram nos "vértices" deste envelope são modelos considerados ótimos para uma

determinada situação. Esta situação está relacionada com a proporção das classes dos modelos induzidos e caso haja uma penalidade por um erro, ou um benefício por um acerto. Os modelos que não fazem parte do envelope têm desempenho inferior e podem ser descartados (PROVOST et al., 1998; PROVOST e FAWCETT, 2001; PRATI et al., 2008).

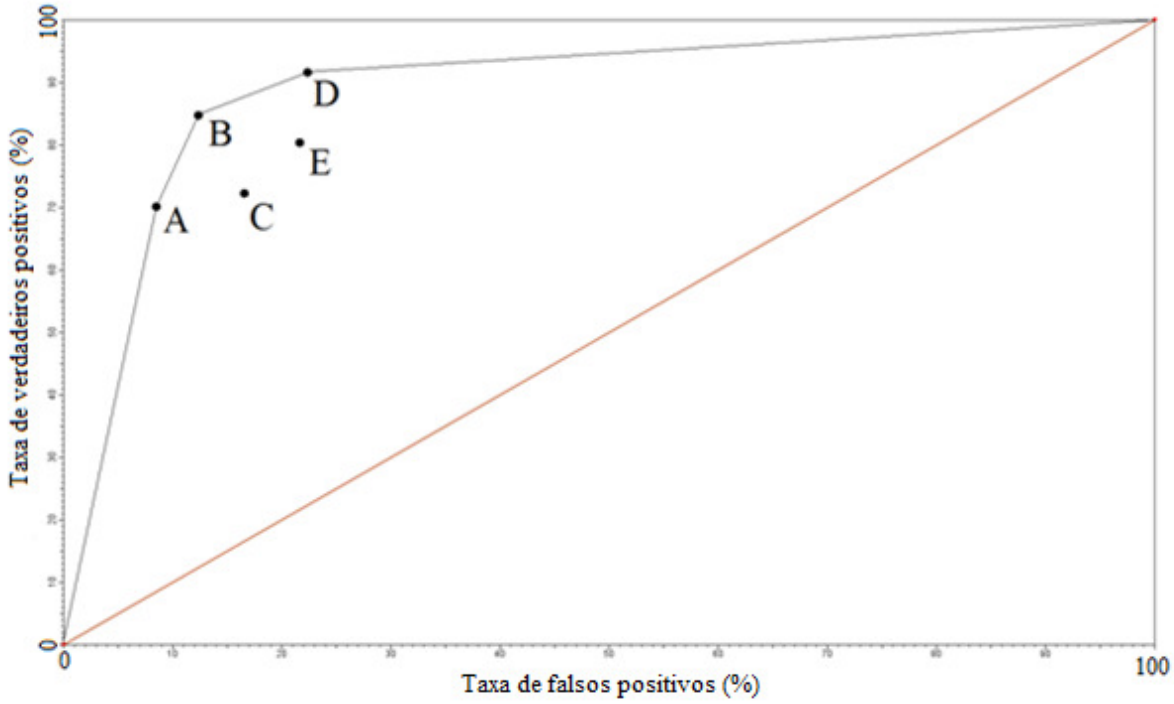


Figura 10: Envelope convexo (convex hull) em um gráfico ROC.

O gráfico ROC é uma dispersão bidimensional da performance do classificador. Para comparar classificadores, pode-se reduzir a análise ROC para um único valor escalar, o qual representa a performance esperada do classificador (FAWCETT, 2006).

O método mais comum é o cálculo da área sob a curva ROC, mais conhecido como AUC (*Area Under Curve*). A área é calculada por meio de métodos de integração numérica, sendo que é uma medida que varia de zero a um. Ela representa o desempenho médio de um classificador, ou seja, quanto maior a área, maior a probabilidade de um classificador classificar corretamente um exemplo (BRADLEY, 1997).

3.1.5.3 O índice Kappa

O índice Kappa foi introduzido por Cohen (1960). Este índice estatístico é uma medida de concordância usada em escalas nominais. Ele fornece subsídios sobre quanto as classificações são diferentes daquelas esperadas (ao acaso). No caso mais específico de mineração de dados, ele mede a correlação entre os valores preditos e os observados, corrigindo uma correlação que ocorre ao acaso (WITTEN et al., 2011).

Para uma matriz de confusão (Figura 8), o índice Kappa pode ser calculado por meio da determinação da probabilidade esperada, chance de ocorrer correlação ao acaso, e da probabilidade observada, correlação que realmente ocorreu. Isto pode ser feito seguindo as equações (8), (9) e (10):

Cálculo da Probabilidade Esperada (Pe):

$$Pe = \frac{(VP + FN) * (VP + FP) + (VN + FP) * (VN + FN)}{n^2} \quad (8)$$

Cálculo da Probabilidade Observada (Po):

$$Po = \frac{(VP + VN)}{n} \quad (9)$$

Cálculo do índice Kappa (IK):

$$IK = \frac{(Po - Pe)}{(1 - Pe)} \quad (10)$$

Os valores do índice Kappa podem variar de 0 até 1, onde valores intermediários representam diversos tipos de classificação quanto à correlação, ou ao comportamento de um modelo, demonstrados na Tabela 1.

Tabela 1: Índices de avaliação Kappa. (Adaptado de LANDIS e KOCK, 1977)

Valor de IK	Classificação do modelo
Menor de 0,40	Ruim
0,41 – 0,60	Regular
0,61 – 0,80	Bom
Maior de 0,81	Excelente

3.2 Verificação e Validação

Com o crescente uso de modelos de simulação para resolver os mais diversos problemas, os desenvolvedores e os usuários destes modelos utilizam informações obtidas a partir dos resultados expressos por estes para ajudar na tomada de decisões. Entretanto, os indivíduos afetados por decisões baseadas em tais modelos estão justamente preocupados com o fato de um modelo e seus resultados estarem "corretos". Esta preocupação é abordada pela verificação e validação (V&V) de modelos.

Existe uma grande diferença entre verificar e validar um modelo. Balci (1997) distingue verificação de validação atribuindo que um processo de verificação do modelo é analisar o modo em que o modelo se ajusta aos dados, como se pretendia, com precisão suficiente. Verificação está relacionada com a construção correta do modelo. Validação consiste em comprovar que o modelo, dentro do seu domínio de aplicabilidade, se comporta com taxa de acerto satisfatória e compatível com os objetivos do estudo. Validação é a construção do modelo certo.

Balci (1997) ainda aponta que o processo de teste de modelos é a detecção de falhas e erros em modelos já gerados, utilizando-se de conjunto de dados para avaliar se este está funcionando corretamente. Os processos de V&V fazem parte da etapa de teste.

Sargent (2013) aponta abordagens para decidir se um modelo de simulação é válido. Em uma delas, cabe aos desenvolvedores do modelo tomar uma decisão se o modelo de simulação é válido. A decisão é subjetiva s feita com base nos resultados dos vários testes e avaliações realizados como parte do processo de desenvolvimento do modelo.

Outra abordagem é envolver o usuário final na parte de validação do modelo, juntamente com os desenvolvedores. Nesta abordagem, o foco de determinar a validade do modelo de simulação muda dos desenvolvedores do modelo para os usuários finais.

Uma terceira abordagem, geralmente chamada de "verificação e validação independente" (IV & V – *Independent Validation and Verification*), usa um terceiro para decidir se o modelo de simulação é válido ou não. O terceiro é independente de ambas as equipes de desenvolvimento ou usuários finais. O terceiro precisa ter um entendimento completo da finalidade do modelo de simulação a fim de conduzir este processo. Existem duas maneiras comuns que o terceiro realiza o processo de IV & V: participando simultaneamente do desenvolvimento do modelo ou podendo realizar o processo após o modelo de simulação estar desenvolvido.

Banks (1998) aponta que os processos de validação encontram-se divididos em quatro grandes áreas, de acordo com o tipo de cada um dos testes: análises informais, estatísticas, dinâmicas e formais. Banks (1998) ainda afirma que as técnicas informais são as mais usadas, entretanto dependem fortemente do bom senso e da subjetividade do especialista, sem haver uma constatação matemática formal. Outra técnica muito usada é a análise estatística, a qual se preocupa com a taxa de acerto do modelo em questão, sendo que não é necessária a construção do modelo novamente, e sim uma análise dos seus dados de saída. Técnicas dinâmicas analisam o comportamento do modelo baseado em padrões de execução e as análises formais são constatadas na comprovação matemática dos modelos.

Os processos de V&V podem incorporar mudanças estruturais no modelo, mudando características construtivas, o que é apontado por Sargent (2007) como um processo iterativo para a construção de um modelo de simulação válido. Verificação e validação devem ser realizadas novamente quando qualquer mudança do modelo é feita. Normalmente, um modelo é desenvolvido e passa por processos de ajuste e adequação para seu funcionamento em condições de uso. Cada uma dessas alterações pode trazer benefícios, como o aumento da taxa de acerto ou a ampliação do suporte, por exemplo. Entretanto, também pode trazer alterações indesejadas, como promover o erro em uma parte em que o modelo anterior (sem alterações) estava acertando.

São inúmeras as formas para a validação de um modelo, sendo que Kiralj e Ferreira (2009) apontaram diversos métodos de validação de modelos. Um deles é o método de validação externa, o qual visa medir o poder de predição dos modelos para dados externos ao seu conjunto de treinamento. Os autores ainda atentam para a necessidade de avaliação estatística das amostras ou conjuntos de dados a serem testados.

3.2.1 Validação dos dados

Kleijnen (1999) afirma que comparando os dados reais com dados utilizados na simulação, ambos devem ser observados em cenários semelhantes. Por exemplo, um dia agitado no supermercado real deve ser comparado com um dia agitado no supermercado simulado.

A validade dos dados, mesmo que muitas vezes não seja considerada como parte da validação do modelo, pode trazer problemas. Muitas vezes essa pode ser a razão que leva um modelo a falhar. Dados são necessários para três propósitos: para a construção do modelo conceitual, para validar o modelo e para a realização de experimentos com o modelo validado (SARGENT, 2013).

As preocupações com os dados são de que estes devem estar apropriados, serem suficientes e disponíveis. Também é necessário atenção para que todas as transformações de dados, tais como desagregação ou junção, sejam feitas corretamente. Para garantir que os dados estejam corretos, deve-se desenvolver procedimentos para minimizar erros na fase de coleta e na manutenção dos dados, além do rastreamento de *outliers*.

A fim de verificar se dois conjuntos de dados são equivalentes, pode-se usar o teste estatístico chamado de teste-t. Diversas considerações sobre o teste-t podem ser encontradas em Pimentel-Gomes (2009).

3.3 A cultura do café e a ferrugem do cafeeiro

3.3.1 A cultura do café

O cafeeiro é uma planta da Família *Rubiaceae* e pode ser classificado em dois gêneros, *Coffea* e *Psilanthus*. Todas as espécies do gênero *Coffea* são nativas de regiões tropicais da África e ilhas do Oceano Índico, já as espécies do gênero *Psilanthus* são originárias da África, Ásia e Oceania (CROS et al., 1998). O gênero *Coffea* compreende duas espécies: *Coffea arabica* Linnaeus (café arábica) e *Coffea canephora* Pierre (café robusta). Estas espécies são responsáveis por praticamente todo o café comercializado mundialmente (FAZUOLI et al., 2002; MATIELLO et al., 2002).

A espécie *Coffea arabica* é nativa de uma região da África que compreende o sudoeste da Etiópia, sudeste do Sudão e norte do Quênia, regiões com altitude variando entre 1000 e 2000 metros. Já o *Coffea canephora* possui uma distribuição geográfica muito mais ampla, ocorrendo em grande parte do continente africano.

O café robusta apresenta um menor custo de produção, uma alta produtividade e um maior rendimento quando comparado ao café arábica, entretanto, o café arábica é mais aceito no mercado por possuir melhor sabor e aroma (MENDES et al., 2001).

O café foi introduzido no Brasil em 1727, vindo de plantações da Guiana Francesa. A frutificação do café ocorre, em média, cerca de dois anos após o plantio, dependendo dos tratos culturais utilizados. A brota do café depende muito do clima e da altitude do cultivo. Sua flor dá origem a um fruto de cor vermelha ou amarela, com aproximadamente 10 a 15 milímetros de diâmetro.

O Brasil tem cerca de 2,2 milhões de hectares plantados com café e produziu, em 2012, 55,9 milhões de sacas. O café é cultivado em 12 estados brasileiros, sendo que os três maiores produtores são Minas Gerais, com cerca de 51% da produção nacional, seguido do Espírito Santo, com 27%, e São Paulo, com 8% (MINISTÉRIO DA AGRICULTURA, 2013).

No âmbito mundial, o Brasil é o maior produtor e exportador de café em grão, tendo exportado cerca de 33,6 milhões de sacas (60% de sua produção) para seus principais importadores, Estados Unidos e União Européia, gerando ganhos de aproximadamente US$ 6 bilhões (USDA, 2013). A importância econômica do café é ainda mais exaltada por este ser, junto com o açúcar, a principal commoditie agrícola do Brasil, além de ser a 3ª commoditie mais exportada pelo país, ficando atrás apenas de petróleo bruto (2ª) e minério de ferro (1ª). (MINISTÉRIO DO DESENVOLVIMENTO, 2013)

3.3.2 Conceitos sobre epidemiologia

Um dos primeiros conceitos sobre o que é uma doença de planta foi: "As doenças de plantas devem ser atribuídas a mudanças anormais nos seus processos fisiológicos, decorrentes de distúrbios na atividade normal de seus órgãos" (KÜHN, 1858). Entretanto, essa definição pioneira referia-se apenas às condições específicas da planta em si, sem relacionar outros fatores importantes. Com este princípio, Ernst Albert Gaümann, em 1946, deu outro conceito para doenças de plantas: "Doença de planta é um processo dinâmico, no qual hospedeiro e patógeno, em íntima relação com o ambiente, se influenciam mutuamente, do que resultam modificações morfológicas e fisiológicas" (MICHEREFF, 2001).

Esta definição foi aprofundada e evoluiu para uma das definições mais utilizadas atualmente: "Doença é o mau funcionamento de células e tecidos do hospedeiro que resulta da sua contínua irritação por um agente patogênico ou fator ambiental e que conduz ao desenvolvimento de sintomas. Doença é uma condição envolvendo mudanças anormais na forma, fisiologia, integridade ou comportamento da planta. Tais mudanças podem resultar em dano parcial ou morte da planta ou de suas partes" (AGRIOS, 2004).

Logo, doenças de plantas são provenientes de alterações fisiológicas causadas por agentes infecciosos. Estes agentes têm a capacidade de ser transmitidos de uma planta infectada para uma planta sadia. Daí surgem as doenças infecciosas, onde seus agentes causadores podem ser vírus, fungos, bactérias e outros. Esses agentes podem debilitar e enfraquecer a planta, prejudicando a absorção de nutrientes, alterando o seu metabolismo, bloqueando funções de transporte e até consumir o conteúdo de uma célula hospedeira.

Além dos agentes infecciosos, existem fatores de natureza não infecciosa que podem facilitar ou dificultar o desenvolvimento de uma doença. Por exemplo, condições desfavoráveis do ambiente (temperatura muito baixa ou muito alta, deficiência ou excesso de umidade, luz e oxigênio) ou deficiências relacionadas à planta (desequilíbrio nutricional e susceptibilidade à doença).

As interações dos três componentes de uma doença podem ser representadas por um triângulo (Figura 11). As condições para ocorrência de uma doença são a presença de um patógeno virulento, condições ambientais favoráveis e hospedeiro susceptível. No caso de plantas resistentes ao agente infeccioso, o lado do triângulo do hospedeiro é pequeno ou inexistente, diminuindo a ocorrência da doença. 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 para o desenvolvimento do agente infeccioso, maior é o lado do triângulo referente ao ambiente.

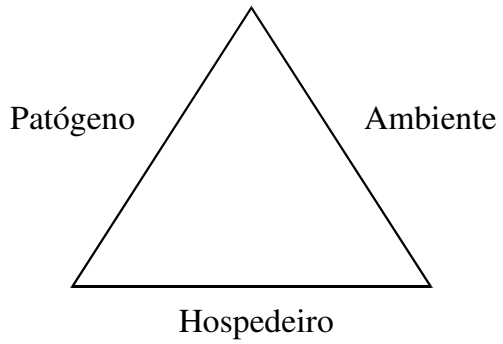


Figura 11: Triângulo de doença de planta (Adaptado de AGRIOS, 2004).

Quando uma doença se propaga em um curto período de tempo e afeta grande parte das plantas de um local ou aumenta a área infectada nessas plantas, ocorre uma epidemia. A epidemiologia é o estudo das epidemias e dos fatores que as influenciam.

Toda doença infecciosa tem um ciclo. Este ciclo da doença, também chamado de ciclo das relações patógeno-hospedeiro, é uma série de eventos ocorridos sequencialmente, que levam ao desenvolvimento de uma doença. Baseado no número de ciclos que uma determinada doença apresentar durante uma mesma estação de cultivo, ela pode ser classificada como doença monocíclica ou doença policíclica (MICHEREFF, 2001).

Eventos ou fases importantes de um ciclo de doença partem de uma fonte de inóculo, depois decorrem os subsequentes: disseminação, inoculação, germinação, penetração, colonização, aparecimento dos sintomas e lesões, reprodução do patógeno e geração de uma nova fonte de inóculo (Figura 12).

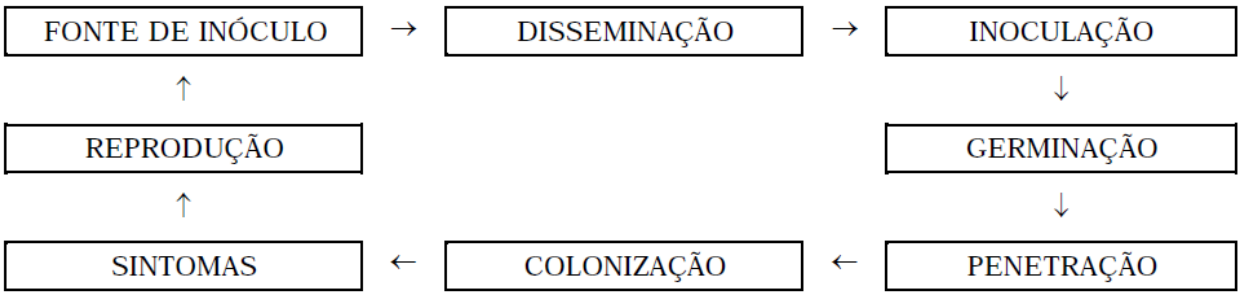


Figura 12: Ciclo de uma doença (MICHEREFF, 2001).

- Fonte de inóculo:

Fase inicial do ciclo, onde um inóculo é qualquer propágulo do patógeno capaz de causar infecção. Para fungos, essas estruturas podem ser esporos ou o micélio. O local onde o inóculo é produzido é chamado de fonte de inóculo.

- Disseminação:

É a dispersão ou transferência do patógeno de uma fonte de inóculo para diversos locais, a qual pode ocorrer por meio do próprio patógeno, como as larvas de nematóides, ou pelos agentes de disseminação, como vento e água.

- Inoculação:

O patógeno se transfere da fonte de inóculo para a planta. O processo só é completo quando o inóculo do patógeno consegue atingir o local de infecção.

- Germinação:

A fase de germinação ocorre com o inóculo junto à superfície do hospedeiro. O inóculo desenvolve a capacidade de penetrar no hospedeiro. No caso de fungos, esta fase é marcada pela emissão do tubo germinativo.

- Penetração:

Processo onde ocorre a introdução do patógeno no local da planta onde se iniciará o processo de colonização. No caso de fungos, o mais comum é que a penetração ocorra diretamente pela superfície do hospedeiro, por meio dos tubos germinativos. A penetração pode também ocorrer por ferimentos e por aberturas naturais do hospedeiro.

O intervalo de tempo entre a inoculação e o surgimento dos sintomas é chamado de período de incubação, já o período latente é o tempo decorrido desde a penetração do patógeno até a sua esporulação em pústulas ou lesões. A duração destes 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 a fase que ocorre quando o patógeno passa a se desenvolver e nutrir dentro do hospedeiro. Este processo só se caracteriza quando os mecanismos de ação de um determinado patógeno se sobrepõem aos mecanismos de defesa do hospedeiro. Ao final desta fase, os sintomas começam a aparecer e se tornar visíveis aos seres humanos.

A presença de plantas suscetíveis e de patógenos em um mesmo local não garante que ocorra infecção, ou mesmo, uma epidemia. Como mostrado na Figura 11, são necessárias condições ambientais favoráveis ao desenvolvimento da doença.

O molhamento foliar, que consiste em um acúmulo de água na superfície da folha do hospedeiro, é um desses fatores. Tal acúmulo pode decorrer de chuva, orvalho ou alta umidade relativa do ar. Em doenças causadas por fungos, este fator deve ser prolongado e repetitivo.

A temperatura ambiente deve se manter dentro de uma faixa favorável em cada uma das etapas do ciclo da doença, favorecendo o desenvolvimento do patógeno.

Juntando estes fatores ambientais e os relacionados ao patógeno-hospedeiro, a doença se desenvolve. É importante quantificar essa doença corretamente para fins de tratamento, acompanhamento e até mesmo para o processo de modelagem de uma epidemia. A medida genérica para quantificar uma doença é a intensidade e, mais especificamente, ela se divide entre a incidência e a severidade (CAMPBELL e MADDEN, 1990).

Incidência é a porcentagem de plantas doentes ou partes de plantas doentes em uma amostra ou população, enquanto que a severidade consiste na porcentagem da área ou do volume de tecido coberto por sintomas.

Existem dois métodos para a determinação da intensidade da doença. Os métodos diretos estimam a intensidade da doença diretamente pela presença ou não dos sintomas, enquanto que os métodos indiretos estimam a intensidade da doença pela população existente de patógenos.

Medir a incidência é um processo relativamente rápido e fácil, entretanto apresenta pouca relação com a severidade e 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, 2004).

O processo de quantificação de doenças é de extrema importância para o estudo de medidas de controle, na determinação de eficiência de fungicidas e na caracterização de espécies resistentes à doença.

3.3.3 A ferrugem do cafeeiro

A ferrugem é a mais importante doença do cafeeiro, podendo causar perdas na produção de 35 a 50%, dependendo das condições climáticas favoráveis à doença (ZAMBOLIN et al., 2002). O agente etiológico dessa doença é o fungo *Hemileia vastatrix* Berk. e Br., ele pertence à ordem Uredinales, Família Pucciniaceae e classe Basidiomycetes.

O ciclo da ferrugem (Figura 13) inicia-se com a produção de esporos, sendo que o fungo produz dois tipos de esporos morfologicamente diferentes e com função distinta. O primeiro tipo são os chamados de uredósporos ou urediniósporos, os quais permanecem na face inferior das folhas, unidos por uma mucilagem, e ao entrarem em contato com água libertam-se facilmente. Após se libertarem, os esporos contaminam outras folhas e posteriormente germinam, penetram e infectam a folha com a presença de água livre.

Após a contaminação das folhas, os uredósporos causam lesões e a partir dessas um segundo tipo de esporo pode ser produzido, chamado de teliósporo. Os teliósporos não têm função conhecida no ciclo da ferrugem (ZAMBOLIN et al., 1997), sabe-se apenas que estes produzem basidiósporos, os quais não foram relacionados à infecção do cafeeiro (FERNANDES et al., 2009). As lesões foliares são a fonte de inóculo, através delas também são produzidos os uredósporos, seguindo para um novo ciclo de infecção.

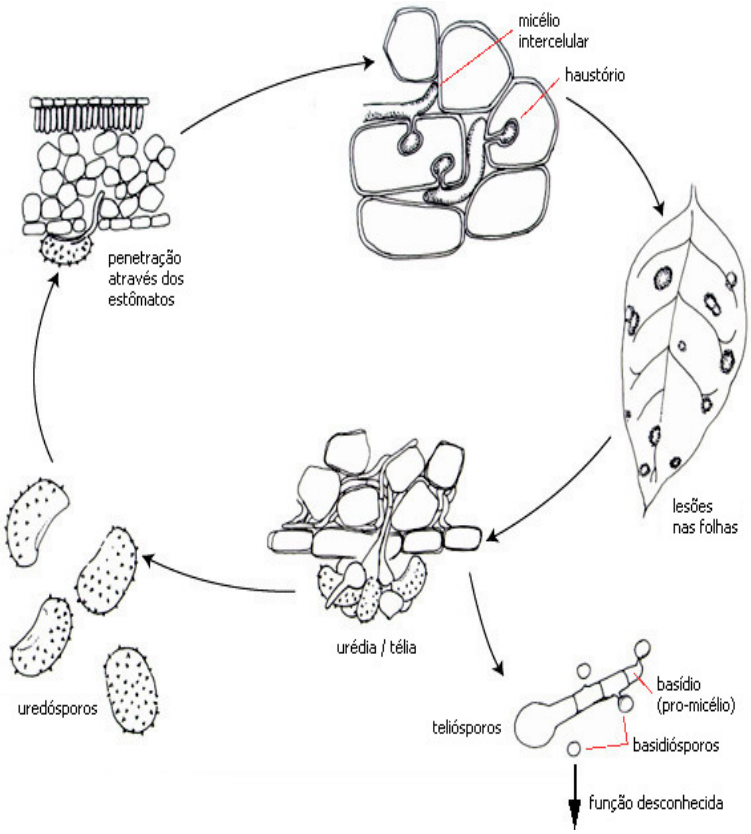


Figura 13: Ciclo da doença - ferrugem do cafeeiro (Adaptado de APSNET, 2013).

A ferrugem do cafeeiro é uma doença que ataca as folhas da planta causando, inicialmente, manchas cloróticas translúcidas com diâmetro de até 3 milímetros. Essas manchas evoluem chegando a atingir até 2 centímetros de diâmetro em cerca de poucos dias.

Em uma plantação, o sintoma mais evidente da ferrugem é a desfolha das plantas. Quando a desfolha ocorre antes do florescimento, o desenvolvimento dos botões florais e o processo de frutificação ficam prejudicados. A perda das folhas durante o desenvolvimento dos frutos leva à formação de grãos anormais, defeituosos e frutos sem sementes, afetando a produção (ZAMBOLIN et al., 2002).

Os prejuízos causados pela ferrugem podem ser atribuídos majoritariamente aos fatores climáticos, sendo que uma combinação ótima de fatores relacionados à temperatura e umidade é o seu principal componente (MORAES, 1983).

A germinação dos uredósporos, após caírem nas folhas, ocorre em um período de 6 a 8 horas em condições de alta umidade (MARTINS, 1988); caso a água seque antes da penetração, o processo é inibido (KUSHALAPPA e ESKES, 1989). Já a temperatura ótima de desenvolvimento do fungo ocorre entre 22 a 24º C (ZAMBOLIN et al., 2002), sendo que temperaturas superiores a 30ºC e inferiores a 14ºC foram consideradas limitantes para a infecção (KUSHALAPPA et al., 1983).

Moraes et al. (1976) formularam a expressão (7) para calcular o período de incubação. Entretanto, em seu trabalho, o período latente (tempo necessário entre a contaminação até a formação de 50% das pústulas) foi referido como período de incubação. Em meses mais quentes este período pode chegar a 28 dias e em meses mais frios 65 dias.

$$Y = 103,01 - 0,98 * T\ \max - 2,1 * T\ \min \qquad (7)$$

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

A disseminação dos esporos ocorre pela ação do vento, pelas gotas de chuva, pelo homem e por insetos e animais (MORAES, 1983). A disseminação pelo ser humano e animais se dá pelo contato dos mesmos com uma planta infectada, espalhando os esporos pelo resto da lavoura, já dentro da própria planta, o respingo de chuva é o principal meio de dispersão (ZAMBOLIN et al., 2002).

Fatores como a densidade de plantio e a presença de cultivares geneticamente resistentes influenciam no desenvolvimento da doença. A densidade de plantio pode afetar o microclima dentro da lavoura e os cultivares resistentes estão menos propícios a serem contaminados (VALE et al., 2000).

A carga pendente de frutos também é um fator que influencia sobre a severidade da ferrugem. O cafeeiro é uma planta tipicamente bianual, ou seja, alterna alta e baixa produtividade a cada ano. Nos anos de alta produtividade, a ferrugem atinge alta severidade, iniciando-se no final do ano, em dezembro, e atingindo seu pico em junho. Depois passa a decrescer dado às baixas temperaturas, a queda das folhas provocada pela colheita, senescência natural e a grande severidade da doença. Nos anos seguintes, de baixa produtividade, a doença não é severa, mesmo sob condições favoráveis do clima (ZAMBOLIN et al., 1997).

A fim de tentar evitar os danos causados pela doença, a utilização do controle químico tem se mostrado eficaz por meio de fungicidas protetores cúpricos ou fungicidas sistêmicos (MATIELLO et al., 2002). Também há a opção de se desenvolver cultivares resistentes geneticamente à doença, o que pode levar a uma eliminação total ou parcial do controle químico (FAZUOLI et al., 2002).

3.4 Modelos de previsão da ferrugem do cafeeiro

Modelos são ferramentas utilizadas para representar a realidade de forma simbólica, podendo ser muito complexos, utilizando muitas variáveis de difícil obtenção, por exemplo. Entretanto, também podem ser extremamente simples, mas não representar fielmente as condições reais. Um modelo ideal é aquele que contempla os aspectos essenciais de um sistema real, balanceando a complexidade e a simplicidade, as quais devem estar em acordo com o seu propósito, enquanto a simplicidade facilita o entendimento do modelo, a complexidade pode permitir maior taxa de acerto do sistema (CAMPBELL et al., 1988).

Na construção de modelos relativos à previsão de ocorrência de doenças de plantas, um período mínimo de oito a doze anos de registro de dados em campo é o recomendado para identificar, com mais segurança, quais podem ser os fatores climáticos que influenciam no

desenvolvimento de uma doença. Caso menos de oito anos de registro sejam utilizados, dados de diferentes regiões geográficas podem ser utilizados (COAKLEY, 1988).

Como pode ser encontrado em Madden e Ellis (1988) e Zambolim et al. (2002), os métodos e técnicas usuais para construção de modelos relacionados à predição de doenças de plantas são de cunho matemático, sendo predominante a análise de regressão. Como exemplo, pode-se citar a regressão linear múltipla no desenvolvimento de modelos para predizer a severidade de epidemias da ferrugem asiática da soja (DEL PONTE et al., 2006). Já para estudos epidemiológicos da ferrugem do cafeeiro, os trabalhos de Montoya e Chaves, (1974), Moraes et al. (1976), Kushalappa et al. (1983) e Zambolim et al. (2002) são alguns exemplos de trabalhos realizados a partir da técnica de regressão

Apesar das técnicas de regressão ainda dominarem esta área, tem-se notado que modelos têm sido gerados através de técnicas de mineração de dados, como redes neurais, SVMs e árvores de decisão, cada um seguindo um determinado detalhe construtivo.

Pinto et al. (2002) avaliaram o potencial das redes neurais para descrever epidemias da ferrugem do cafeeiro. As redes neurais foram utilizadas para tentar estabelecer as relações entre variáveis climáticas e de produção com a incidência da ferrugem do cafeeiro. Variáveis como precipitação, número de dias com e sem precipitação, umidade relativa do ar, horas de insolação, temperaturas média, máxima e mínima foram calculadas como médias ou somatórios para períodos de 15, 30, 45 e 60 dias anteriores à avaliação da incidência da ferrugem. Além destas variáveis, séries temporais da incidência isolada da doença também foram utilizadas na elaboração das redes neurais.

Diversas redes foram criadas e a escolha da melhor ocorreu em função dos menores valores do erro médio de previsão e quadrado médio do desvio. O melhor resultado foi de uma rede que envolvia variáveis de temperatura mínima, umidade relativa do ar, precipitação e insolação, referentes a 30 dias antes da data de avaliação da incidência da ferrugem. A partir de séries temporais, a melhor rede incluiu as observações da incidência da doença das quatro quinzenas anteriores à data de avaliação.

Redes neurais também foram utilizadas para analisar a severidade da ferrugem asiática da soja (BATCHELOR et al., 1997), enquanto que Alves et al. (2010) utilizaram métodos de regressão linear, regressão não-linear, sistemas fuzzy e neuro-fuzzy para avaliar a severidade no processo monocíclico da ferrugem do cafeeiro e da ferrugem asiática da soja.

Para a ferrugem asiática da soja, Alves et al. (2010) avaliaram medidas de temperatura média nos valores de 15ºC, 20ºC, 25ºC e 30ºC além da duração do período de molhamento foliar de 0, 6, 12, 18 ou 24 horas. Já para a ferrugem do cafeeiro avaliaram-se as mesmas medidas, entretanto com valores diferentes. Para a temperatura média, os valores foram de 15ºC, 20ºC e 30ºC; já para o molhamento foliar, foram de 6, 12, 18, 24 e 48 horas. A intensidade da doença foi calculada a partir da área sob a curva de progresso da doença.

Para avaliar seus modelos, Alves et al. (2010) utilizaram os valores do coeficiente de determinação (R^2), o coeficiente de determinação ajustado (adj.R^2) e do erro médio quadrático (*Root mean squared error – RMSE*). Os resultados mostraram que os sistemas neuro fuzzy foram superiores aos demais, seguidos dos sistemas fuzzy, regressão não linear e regressão linear. O melhor sistema neuro-fuzzy foi capaz de explicar a variação na severidade em 99% dos casos, para a ferrugem do cafeeiro, e em 85% dos casos, para a ferrugem asiática da soja.

Luaces et al. (2011) estudaram a incidência da ferrugem do cafeeiro por meio de uma regressão utilizando máquinas de vetores suporte. A incidência da ferrugem foi medida pela porcentagem de plantas infectadas pelo fungo causador da ferrugem. Atributos meteorológicos foram incluídos em um conjunto de dados utilizado para determinação da taxa de incidência da doença. A abordagem inicial utilizou valores pontuais na regressão feita por meio do método de máquinas de vetores suporte. Essa primeira tentativa obteve uma taxa de acerto de 94% em resultados de validação cruzada, entretanto a quantidade de falsos negativos era muito grande quando um dado limite de infecção era ultrapassado. A partir deste problema, os autores transformaram as medidas de incidência em faixas, ao invés de números pontuais. Essa transformação gerou classificadores de regressão chamados de não determinísticos, os quais foram capazes de reduzir o número de falsos negativos.

Meira (2008) utilizou árvores de decisão para descrever a epidemia da ferrugem do cafeeiro e como modelos de alerta da doença. A evolução da taxa de progresso da doença (diferença da incidência entre dois meses subsequentes) foi o objeto avaliado nesse estudo, podendo ser um aumento de 5 ou 10 p.p. (pontos percentuais). Foram relacionados 8 anos de dados meteorológicos e dados relativos à lavoura, como carga de frutos e espaçamento. Dentre os principais dados meteorológicos, havia a temperatura do ar, umidade relativa, precipitação, molhamento foliar (em horas) e temperatura durante o período de incubação do fungo causador da ferrugem.

Os modelos foram avaliados para medidas de taxa de acerto, erro, sensitividade, especificidade e confiabilidades positiva e negativa, por meio de validação cruzada 10 partes. Os modelos para baixa carga pendente de frutos, apesar de apresentarem alto valor de taxa de acerto, não obtiveram um equilíbrio entre as demais medidas de avaliação. Já os modelos de alta carga pendente de frutos (MEIRA et al., 2009) apresentaram medidas mais interessantes, sendo que sua taxa de acerto superou 80% e 78% para casos de aumento de 5 p.p. e 10 p.p., respectivamente.

A grande vantagem destes modelos com relação aos demais foi a alta taxa de acerto e a facilidade de compreensão das regras geradas por árvores de decisão. Meira (2008) também caracterizou a metodologia que utilizou, permitindo uma fácil reprodução e adaptação para outros problemas similares. Esta reprodução foi tomada como base da metodologia para o presente trabalho.

Cintra et al. (2011) basearam-se nos modelos de previsão gerados por Meira (2008) para desenvolver sistemas unindo as técnicas de árvore de decisão e sistemas Fuzzy, desenvolvendo árvores de decisão fuzzy para prever a taxa de progresso da ferrugem do cafeeiro. As árvores fuzzy foram criadas a partir de três conjuntos de dados e duas opções de atributo meta, totalizando a construção de seis árvores. A taxa de erro e a taxa de acerto foram comparadas às das árvores de Meira (2008). As taxas de acerto obtidas pelas árvores fuzzy foram superiores às árvores simples, sendo que a taxa de erro foi, em média, 3,7 pontos percentuais menores nas árvores fuzzy. Para o caso mais extremo, houve uma diferença de aproximadamente 6,5 pontos percentuais. Os autores ainda utilizaram técnicas de modelagem como redes neurais e florestas aleatórias para induzir novos modelos e compará-los às árvores fuzzy. Em nenhum dos casos a taxa de erro foi inferior ao sistema fuzzy, ou seja, o sistema de árvores de decisão fuzzy obtinha melhores medidas de avaliação.

4 Material e Métodos

A metodologia usada para desenvolver o processo de descoberta de conhecimento (*KDD – Knowledge Discovery in Databases*) foi a CRISP-DM (*CRoss Industry Standard Process for Data Mining*) (CHAPMAN et al., 2000). Esta divide o ciclo de vida de um projeto em 6 fases: compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição (Figura 14). A metodologia é cíclica, sendo que a sequência lógica entre as fases não é rígida, sendo comum, e quase sempre necessário, voltar e avançar entre diferentes fases. O resultado de uma fase realizada anteriormente é fundamental para determinar qual a próxima fase que deverá ser executada na sequência.

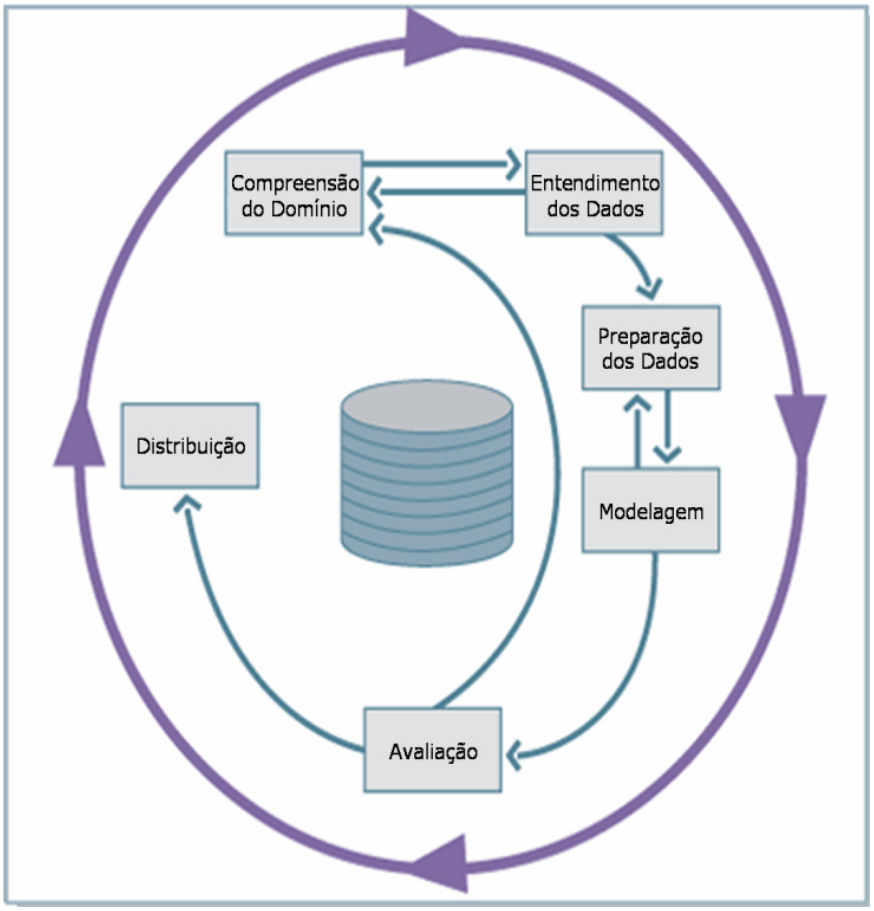


Figura 14: Fases do modelo de processo CRISP-DM. (Adaptado de CHAPMAN et al., 2000).

Fases do processo KDD:

Compreensão do domínio: Esta fase inicial visa entender os objetivos e requisitos do projeto, pela perspectiva do domínio de aplicação, e depois converter esse conhecimento na definição do problema e em um plano projetado para atingir os objetivos.

Neste projeto esta fase pôde ser descrita como sendo o exame de qualificação de mestrado, o qual conteve as justificativas para o desenvolvimento do trabalho, bem como a hipótese de pesquisa, além de uma revisão bibliográfica sobre a doença da ferrugem do cafeeiro e os modelos de predição desenvolvidos para tal doença.

Entendimento dos dados: A fase de entendimento dos dados começa com a coleta inicial de dados e continua com atividades para se familiarizar com os dados, identificar problemas de qualidade nesses dados e buscar as primeiras compreensões (*insights*) a partir deles. Esta fase está descrita na seção 4.1.

Preparação dos dados: Fase que abrange todas as atividades necessárias para construir, a partir dos dados iniciais brutos, o conjunto de dados final para a modelagem. Atividades como transformação de dados e geração dos atributos do conjunto de dados são realizadas nesta fase. A seção 4.2 detalhará estes procedimentos.

Modelagem: Várias técnicas de mineração de dados são selecionadas e aplicadas nesta fase, além de seus parâmetros serem calibrados para valores ótimos. Existem várias técnicas que podem ser utilizadas em um mesmo tipo de problema, algumas têm requisitos específicos para o conjunto de dados, então voltar para a fase de preparação de dados é normalmente necessário. As etapas e procedimentos desta fase estão descritos na seção 4.3.

Avaliação: Esta fase inicia-se com um ou mais modelos construídos. Estes modelos apresentam, aparentemente, boa qualidade na perspectiva da análise de dados. Antes de proceder com 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 Verificação e Validação (V&V) foi tarefa desta fase, a qual tem os resultados descritos no capítulo 5. Para a avaliação, comparação e seleção de novos modelos de alerta, os resultados encontram-se no capítulo 6.

Distribuição: Esta fase finaliza o processo colocando os modelos gerados disponíveis para uso, mesmo que seja com a intenção de aumentar o conhecimento obtido pela extração de padrões do conjunto de dados. A distribuição dos modelos desenvolvidos faz parte do projeto "Análise do risco de epidemias da ferrugem do cafeeiro a partir de estações de avisos fitossanitários com o auxílio de modelos de alerta da doença", em fase final de execução no Programa Pesquisa Café, coordenado pela Embrapa Informática Agropecuária em parceria com a Fundação PROCAFÉ e a Faculdade de Engenharia Agrícola da Unicamp.

Durante a execução das fases deste projeto, os resultados parciais foram divulgados em eventos acadêmicos e científicos (GIROLAMO NETO et al., 2012a e GIROLAMO NETO et al., 2012b).

4.1 Entendimento dos dados

4.1.1 O conjunto de dados

Os dados utilizados foram coletados seguindo Japiassú et al. (2007) e referem-se ao acompanhamento mensal da incidência da ferrugem do cafeeiro em fazendas experimentais da Fundação PROCAFÉ. Estas fazendas estão em 3 municípios de Minas Gerais, em Varginha, latitude sul de 21º 34’ 00’’, longitude oeste de 45º 24’ 22’’ e altitude de 940 m, em Carmo de Minas, latitude sul de 22º 10' 31", longitude oeste de 45º 09' 03" e altitude de 1180 m e Boa Esperança, latitude sul de 21º 03' 59", longitude oeste de 45º 34' 37" e altitude de 830 m.

VARGINHA:

Para a cidade de Varginha foram realizados registros entre outubro de 1998 até outubro de 2011. As lavouras selecionadas tinham 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, havia 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 (Vermelho e Amarelo) 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 foi realizado ao final de cada mês, conforme recomendação de 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) até outubro de 2006. A partir deste período houve uma troca da estação meteorológica, passando a ser do modelo WeatherLink; esta estação coletou os dados até outubro de 2011.

CARMO DE MINAS:

Para a cidade de Carmo de Minas foram realizados registros entre dezembro de 2006 até outubro de 2011. As lavouras selecionadas tinham idade entre 6 e 20 anos e espaçamento adensado, sendo que havia duas com alta carga pendente de frutos e duas com baixa carga pendente de frutos. Em cada par de lavouras, uma foi da cultivar Catuaí (Vermelho e Amarelo) 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 foi o mesmo realizado para a cidade de Varginha (CHALFOUN, 1997). Os dados meteorológicos foram registrados a cada 30 min. por uma estação meteorológica automática (marca Davis, modelo WeatherLink).

BOA ESPERANÇA:

Para a cidade de Boa Esperança foram realizados registros entre de junho de 2010 até outubro de 2011. As lavouras selecionadas tinham idade entre 6 e 20 anos e espaçamento largo, sendo que havia duas com alta carga pendente de frutos e duas com baixa carga pendente de frutos. Em cada par de lavouras, uma foi da cultivar Catuaí (Vermelho e Amarelo) 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 foi o mesmo realizado para as cidades anteriores (CHALFOUN, 1997). Os dados meteorológicos foram registrados a cada 30 min. por uma estação meteorológica automática (marca Davis, modelo WeatherLink).

Arquivos obtidos:

O conjunto de dados recebido foi composto por três tipos de arquivos, para cada mês de cada ano analisado:

- 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/.docx/.pdf) 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 combinações de espaçamento e produção das lavouras, de cada doença ou praga.

A obtenção destes dados ocorreu de forma gradual, sendo que os dados colhidos entre outubro de 1998 a outubro de 2006 foram obtidos de Meira (2008), e os demais foram obtidos em duas remessas, provenientes de visitas realizadas à Fundação PROCAFÉ nos meses de novembro de 2010 e novembro de 2011.

4.1.2 Descrição dos dados

Os dados recebidos de Meira (2008) foram derivados de medidas extraídas da estação de marca Davis, modelo Groweather Industrial. Já os dados recebidos após outubro de 2006 foram obtidos da estação meteorológica de marca Davis, modelo WeatherLink.

Para facilitar o entendimento, a estação que coletou os dados até outubro de 2006 (Groweather Industrial) foi chamada de estação antiga e a estação que coletou os dados após outubro de 2006 foi tratada por estação nova. Esta situação ocorreu apenas para a cidade de Varginha.

A estação antiga de Varginha coletou 24 atributos e as demais estações coletaram 35 atributos, entretanto nem todos estes atributos foram usados para gerar o conjunto de dados final. Os principais atributos que foram usados para gerar o conjunto final estão descritos na Tabela 2.

Tabela 2: Descrição dos atributos relevantes das estações meteorológicas que foram utilizados para gerar o conjunto de dados.

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:		

<table>
<tr><td colspan="3">Hora em que os dados foram obtidos no dia correspondente.
As estações realizavam leituras a cada 30 min., de 0:00 hora até 23:30.</td></tr>
<tr><td>TMED</td><td>Numérico</td><td>°C</td></tr>
<tr><td colspan="3">Significado:
Temperatura do ar (média dos últimos 30 min.) medida através de um sensor de temperatura, dada em graus Celsius (ºC).
Precisão: +/- 0,5°C.
Nome original do atributo no arquivo da estação antiga: ‘Ar Temp’.
Nome original do atributo no arquivo da estação nova: ‘Temp Out’.</td></tr>
<tr><td>TMAX</td><td>Numérico</td><td>°C</td></tr>
<tr><td colspan="3">Significado:
Temperatura máxima do ar, dada em ºC, obtida a cada intervalo de tempo (30 min.).
Precisão: +/- 0,5°C.
Nome original do atributo no arquivo da estação antiga: ‘Temp Max’.
Nome original do atributo no arquivo da estação nova: ‘Hi Temp’.</td></tr>
<tr><td>TMIN</td><td>Numérico</td><td>°C</td></tr>
<tr><td colspan="3">Significado:
Temperatura mínima do ar, dada em ºC, obtida a cada intervalo de tempo (30 min.).
Precisão: +/- 0,5°C.
Nome original do atributo no arquivo da estação antiga: ‘Ar Min’.
Nome original do atributo no arquivo da estação nova: ‘Low Temp’.</td></tr>
<tr><td>VVENTO</td><td>Numérico (≥0)</td><td>km/h</td></tr>
<tr><td colspan="3">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 antiga: ‘Vento Vel.’.
Nome original do atributo no arquivo da estação nova: ‘Wind Speed’.</td></tr>
<tr><td>PRECIP</td><td>Numérico (≥0, múltiplo de 0,2)</td><td>mm</td></tr>
</table>

Significado: Medida por um coletor, registra os dados de precipitação pluviométrica durante o último período (30 min.). É dada em milímetros (mm). Nome original do atributo no arquivo da estação antiga: ‘Prec’. Nome original do atributo no arquivo da estação nova: ‘Rain’.		
INDPLUVMAX	Numérico (≥0)	mm/h
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 antiga: ‘Max Índ.’. Nome original do atributo no arquivo da estação nova: ‘Rain rate’.		
UR	Numérico inteiro (≥0)	%
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 antiga: ‘Umid’. Nome original do atributo no arquivo da estação nova: ‘Out Hum’.		

Além dos atributos derivados das estações meteorológicas, houve atributos derivados dos boletins de avisos. Estes atributos estão descritos na Tabela 3.

Tabela 3: Descrição dos atributos relevantes dos boletins de avisos que foram utilizados para gerar o conjunto de dados.

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 Atacados - Ferrugem'.		

4.1.3 Verificação da qualidade dos dados

Antes de preparar os dados, foi necessário verificar as inconsistências nos dados recebidos. Por meio de inconsistências detectadas, foi possível melhorar a qualidade dos dados a serem preparados.

A primeira avaliação foi quanto aos boletins de avisos fitossanitários. Todos os boletins obtidos foram abertos e conferidos para detecção de possíveis problemas, como a presença de arquivos errados, entretanto, nenhuma inconsistência ou problema foi detectado.

Em seguida, foram verificados os arquivos das estações meteorológicas, buscando-se falhas de medição, como por exemplo, dias em que a estação não funcionou. Dependendo da

quantidade de dias com falhas, o mês poderia ficar comprometido e não ser utilizado para a composição do conjunto de dados para a modelagem. Estes arquivos também foram inspecionados buscando-se inconsistências nos atributos medidos pela estação meteorológica, como, por exemplo, temperatura mínima maior que temperatura máxima. Estas verificações ocorreram por meio de um software de manipulação de dados chamado Google Refine 2.0 (GOOGLE, 2011), sendo que as regras de verificação de atributos estão na Tabela 4.

Houve pouquíssimos registros com hora diferente do padrão, os quais foram excluídos ao se construir o conjunto de dados final. Encontraram-se registros com UR < 20%, entretanto, verificando que estes ocorreram à tarde e aliados a altas temperaturas, considerou-se que não se tratavam de inconsistências. Quanto aos dados ausentes, alguns meses ficaram comprometidos devido ao grande número de falhas na obtenção dos dados (Tabela 5).

Tabela 4: Regras de verificação de inconsistências em atributos das estações meteorológicas.

Atributo DATA	Teste: DATA = DD/MM/AAAA
Atributo HORA	Teste: HORA = HH:[00|30]
Atributo TEMP	Teste: TEMP > 0
Atributo TMAX	Testes: TMAX > 0 TMAX ≥ TEMP
Atributo TMIN	Testes: TMIN > 0 TMIN ≤ TEMP
Atributo VVENTO	Testes: VVENTO ≥ 0 VVENTO < 150
Atributo PRECIP	Testes: PRECIP ≥ 0 PRECIP múltiplo de 0,2 (a estação marca apenas nestes intervalos)
Atributo INDPLUVMAX	Teste: INDPLUVMAX ≥ 0
Atributo UR	Testes: 1. UR ≥ 0 2. UR ≥ 20

Tabela 5: Meses comprometidos e descartados por falhas diárias nas estações meteorológicas.

Cidade	Ano	Mês
Varginha	1998	Outubro
Varginha	2000	Fevereiro
Varginha	2000	Novembro
Varginha	2000	Dezembro
Varginha	2003	Fevereiro
Varginha	2011	Julho
Varginha	2011	Setembro
Carmo de Minas	2011	Fevereiro

4.1.4 Verificação de compatibilidade dos dados entre estações de Varginha

Duas estações meteorológicas (Estação antiga – modelo Groweather e Estação nova – modelo WeatherLink) funcionaram concomitantemente para a cidade de Varginha no período de Setembro de 2006 até Janeiro de 2007. Foi feita uma junção dos arquivos ".txt" provenientes de cada uma delas em um grande arquivo, no qual foram excluídos os registros em que ao menos uma apresentou problemas de medição. O conjunto continha registros de 30 em 30 minutos e cinco atributos diferentes: Temperatura Máxima, Temperatura Mínima, Temperatura Média, Precipitação e Umidade relativa.

Primeiramente foi efetuada uma comparação gráfica, onde os resultados encontram-se nas Figura 15 a Figura 19. Para essa comparação, foi calculada a diferença de um atributo entre a estação antiga e a estação nova (Dif = Eantiga – Enova). No eixo "y" há a diferença entre o atributo analisado e no eixo "x" o número do registro que foi comparado.

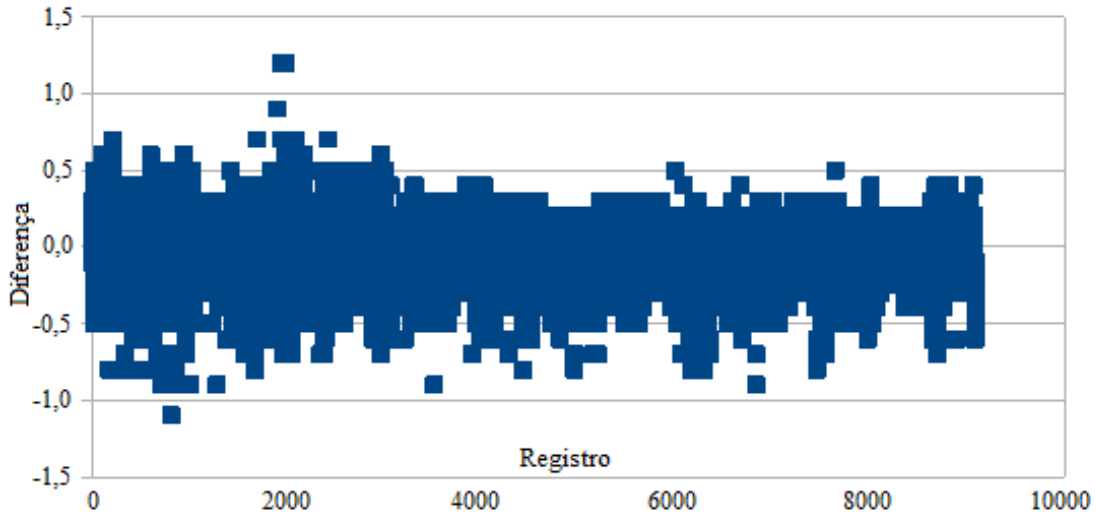


Figura 15: Diferença entre as temperaturas médias das duas estações meteorológicas.

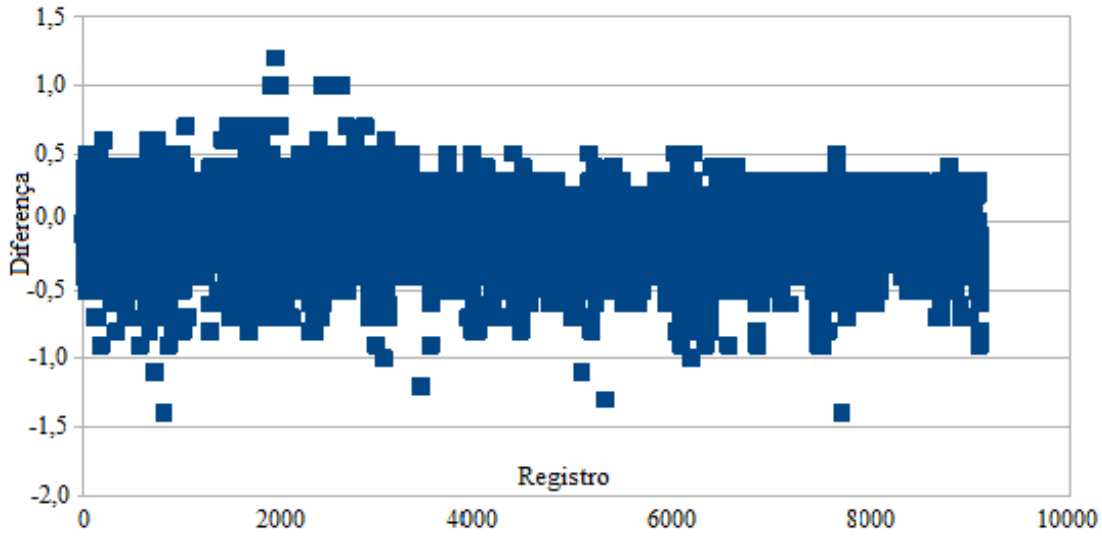


Figura 16: Diferença entre as temperaturas máximas das duas estações meteorológicas.

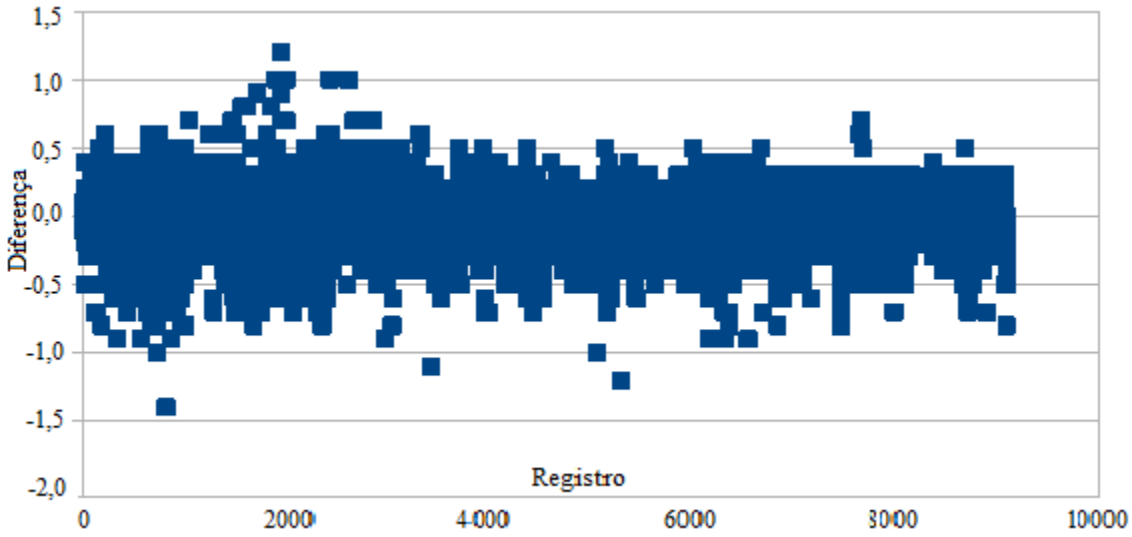


Figura 17: Diferença entre as temperaturas mínimas das duas estações meteorológicas.

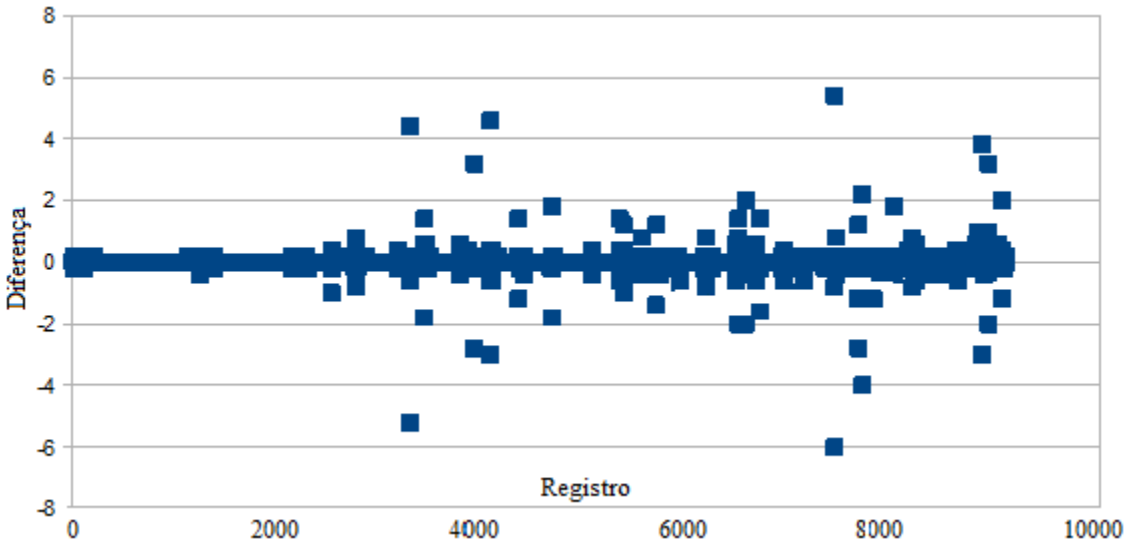


Figura 18: Diferença entre as precipitações das duas estações meteorológicas.

Figura 19: Diferença entre as umidades relativas das duas estações meteorológicas.

A comparação gráfica mostrou algumas diferenças nas medidas realizadas pelas estações e a fim de avaliar a equivalência dos conjuntos de dados provenientes de cada uma delas, foi utilizado o teste-t (PIMENTEL-GOMES, 2009). Os resultados deste teste mostraram que o atributo relacionado à Umidade Relativa (UR) não foi equivalente para os dois conjuntos, já os demais atributos foram.

Analisando a Figura 19, observou-se que a estação antiga media, normalmente, valores mais altos de UR do que a estação nova, fato que gerou a incompatibilidade desses dados. Assim, uma regressão linear simples foi realizada com o intuito de ajustar os dois conjuntos.

Foram feitas duas regressões distintas. Na primeira, os dados da estação nova foram ajustados ao padrão da estação antiga, possibilitando a utilização do conjunto de dados na etapa de validação dos modelos de Meira (2008), uma vez que estes foram construídos com dados da estação antiga. Na segunda, o ajuste foi feito de forma contrária, ou seja, adequando os dados da estação antiga para o padrão da estação nova, deixando estes dados prontos para serem utilizados na geração de novos modelos. Após realizadas as duas regressões, os conjuntos foram novamente testados pelo teste-t e obtiveram resultados de equivalência em ambos os casos.

4.2 Preparação dos dados

4.2.1 Atributos do conjunto de dados

Atributo Meta:

O atributo meta ou a variável dependente foi a taxa de progresso da doença, onde valores da incidência da ferrugem foram utilizados para gerar este atributo. A taxa de progresso consiste no aumento, diminuição ou manutenção da incidência entre dois meses subsequentes. Para efeito de cálculo, a taxa de progresso é igual a incidência em um mês subtraída da incidência do mês anterior.

Os valores da taxa de progresso foram, inicialmente, numéricos, entretanto passaram a ser mapeados em categorias ou classes. A primeira opção escolhida foi mapear um aumento de até 5 pontos percentuais (p.p.) e a segunda um aumento de até 10 p.p.. Além disso, os valores foram transformados em valores binários, sendo que a primeira opção criou o atributo TAXA_INF_M5, com valor '1' para taxas de infecção maiores ou iguais a 5 p.p. e valor '0', caso isso não ocorresse. A segunda opção criou o atributo TAXA_INF_M10, com valor '1' para taxas de infecção maiores ou iguais a 10 p.p. e valor '0', caso isso não ocorresse.

O valor de 5 p.p. foi baseado em Zambolim et al. (1997). Já o valor de 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 (ZAMBOLIN et al., 1997).

Atributos preditivos:

Os atributos preditivos, ou variáveis independentes, partiram de um nível de construção horário (forma em que foram coletados) e passaram por transformações, levando-os até um nível mensal. Essa transformação possibilitou uma análise em conjunto com o atributo meta.

A construção deste conjunto iniciou-se com o período de incubação (PI) e com o período de infecção (PINF). Cada dia foi tratado como um eventual dia de infecção. Considerando um período de incubação estimado, cada um destes dias foi associado ao mês correspondente de avaliação da incidência da ferrugem (Figura 20). O período de incubação para cada dia foi estimado pela expressão (1). Dessa forma, cada dia foi associado a uma taxa de infecção, para a qual possivelmente teve parcela de contribuição. A nomenclatura período de infecção (PINF) foi utilizada para representar quais os possíveis dias em que ocorreu uma infecção da planta pelo patógeno.

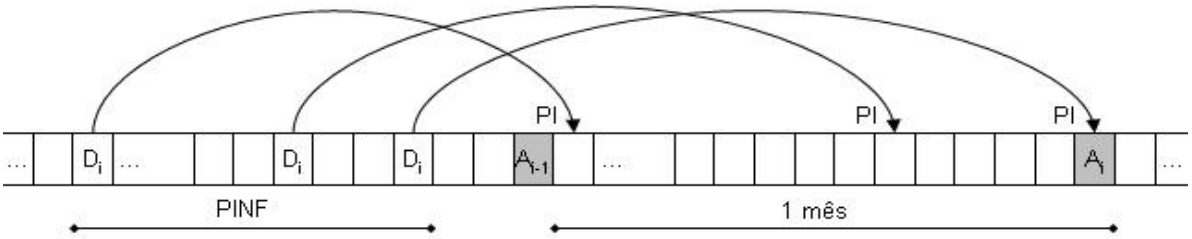


Figura 20: Representação dia-a-dia do esquema usado na preparação dos dados meteorológicos (MEIRA, 2008).

Di - dia de infecção; Ai - avaliação da incidência da ferrugem do cafeeiro; Ai-1 – avaliação da incidência no mês anterior; PI - período de incubação; PINF - período de infecção.

Diversos atributos foram calculados utilizando-se de médias e somatórios simples das variáveis meteorológicas, alguns com os prefixos MED, para a média, e SMT para somatórios.

O atributo DCHUV_PINF corresponde a dias com precipitação maior ou igual a 1 mm. Os atributos como NHNUR90_PINF e NHUR90_PINF, relacionaram-se a períodos de molhamento foliar prolongado (mínimo de 6 h). A germinação do fungo só ocorre se a folha estiver molhada por no mínimo 6 horas, ou seja, é necessária a presença de água livre na superfície da folha durante este período. O período de molhamento foliar e a quantidade de precipitação basearam-se em Kushalappa et al. (1983).

O número de horas com alta umidade relativa do ar (maior ou igual a 90%) foi utilizado como medida indireta de molhamento foliar contínuo e baseado nos trabalho de Sutton et al. (1984), Meira et al. (2008) e Finholdt (2012). Em dias com períodos de molhamento disjuntos, foi considerado o maior, com tolerância de até uma hora para uni-los em um único período.

Os períodos de molhamento foliar foram avaliados para o período contínuo e para sua porção noturna (período das 20:00 h às 8:00 h), uma vez que a infecção ocorre preferencialmente na ausência de ou com pouca luminosidade (MONTOYA e CHAVES, 1974).

A temperatura média durante o período total também foi calculada, tendo em vista que a temperatura é o principal fator que determina o percentual de germinação dos esporos e de penetração do fungo na folha durante o período em que a superfície da folha está molhada (KUSHALAPPA et al., 1983). Não foi calculada a temperatura média durante o período noturno, pois esta apresenta valores praticamente iguais aos do período contínuo (MEIRA, 2008).

Os atributos preditivos foram completados com o atributo LAVOURA, o qual dava a característica do espaçamento de cada lavoura utilizada. A lista dos atributos preditivos está na Tabela 6.

Alguns outros atributos foram criados com a intenção de reunir aspectos conhecidos da epidemiologia da ferrugem do cafeeiro encontrados na literatura, os quais foram os mesmos desenvolvidos por Meira (2008). Estes atributos reuniram informações sobre condições diárias de molhamento foliar, de luminosidade (M-L) e temperatura (T) durante o período de molhamento, com relação ao processo de infecção. Trabalhos como o de Montoya e Chaves (1974) e Kushalappa et al. (1983) serviram de base para a geração desses atributos.

A partir das correlações presentes na Tabela 7, um dia poderia ser classificado como desfavorável, favorável ou muito favorável para a ocorrência da ferrugem. Cada dia classificado obteve uma espécie de nota, variando de 0-4. Valores inferiores a 1, o dia seria desfavorável. No intervalo de 1 até valores inferiores a 3, o dia seria favorável. Para valores de 3 até 4, o dia seria muito favorável. As notas foram somadas em um outro atributo chamado de acúmulo da condição diária (ACDINF_PINF).

Tabela 6: Atributos meteorológicos e de espaçamento presentes no conjunto de dados.

	Descrição		
Nome	Tipo	Medida	Significado
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.
NHNUR90_PINF	numérico	h	Média diária do número de horas noturnas com umidade relativa do ar ≥ 90% no PINF.
NHUR90_PINF	numérico	h	Média diária do número de horas com umidade relativa do ar ≥ 90% no PINF.
PRECIP_PINF	numérico	mm	Precipitação pluvial acumulada no PINF.
SMT_NHNUR90_PINF	numérico	h	Somatório de NHNUR90 no PINF.
SMT_NHUR90_PINF	numérico	h	Somatório de NHUR90 no PINF.
SMT_VVENTO_PINF	numérico	km/h	Média do somatório da velocidade do vento de cada dia do PINF.
THUR90_PINF	numérico	ºC	Temperatura média diária durante as horas com umidade relativa ≥ 90% no PINF.
TMAX_PINF	numérico	ºC	Média das temperaturas máximas diárias no PINF.
TMAX_PI_PINF	numérico	ºC	Média das temperaturas máximas diárias no período de incubação para os dias do PINF.
TMED_PINF	numérico	ºC	Média das temperaturas médias diárias no PINF.
TMED_PI_PINF	numérico	ºC	Média das temperaturas médias diárias no período de incubação para os dias do PINF.
TMIN_PINF	numérico	ºC	Média das temperaturas mínimas diárias no PINF.
TMIN_PI_PINF	numérico	ºC	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.

Tabela 7: Matriz de condições diárias de infecção e seus respectivos índices numéricos.

T M-L	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 (NHNUR90 < 4 ou NHUR90 < 6)	Desfavorável 0	Desfavorável 0	Desfavorável 0	Desfavorável 0
Pouco favorável (NHNUR90 ≥ 4 e NHUR90 ≥ 6)	Desfavorável 0	Desfavorável 0*	Favorável 1*	Favorável 2*
Favorável (NHNUR90 ≥ 8 e NHUR90 ≥ 12)	Desfavorável 0	Favorável 1	Favorável 2	Muito favorável 3
Muito favorável (NHNUR90 ≥ 8 e NHUR90 ≥ 18)	Desfavorável 0	Favorável 2	Muito favorável 3	Muito favorável 4

* Caso NHNUR90 ≥ 8 ou NHUR90 ≥ 12, incrementa-se 0,5 no índice.

Os atributos derivados da Tabela 7 estão relacionados na

Tabela *8*.

Tabela 8: Atributos especiais derivados da matriz de condições diárias.

	Descrição		
Nome	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.

4.2.2 Transformação dos dados

A transformação dos dados foi um processo que visou modificar os dados coletados das estações meteorológicas de meia em meia hora para dados mensais, os quais puderam ser relacionados aos dados de incidência da ferrugem do café (coletados mensalmente de acordo com os boletins de aviso da fundação PROCAFÉ).

Este processo está ilustrado na Figura 21. Cada passo foi realizado por meio de scripts computacionais desenvolvidos por Meira (2008) na linguagem de programação Perl (versão 5.8.7). Estes scripts tiveram que ser modificados de forma a atender os dados provenientes do novo modelo de estação meteorológica.

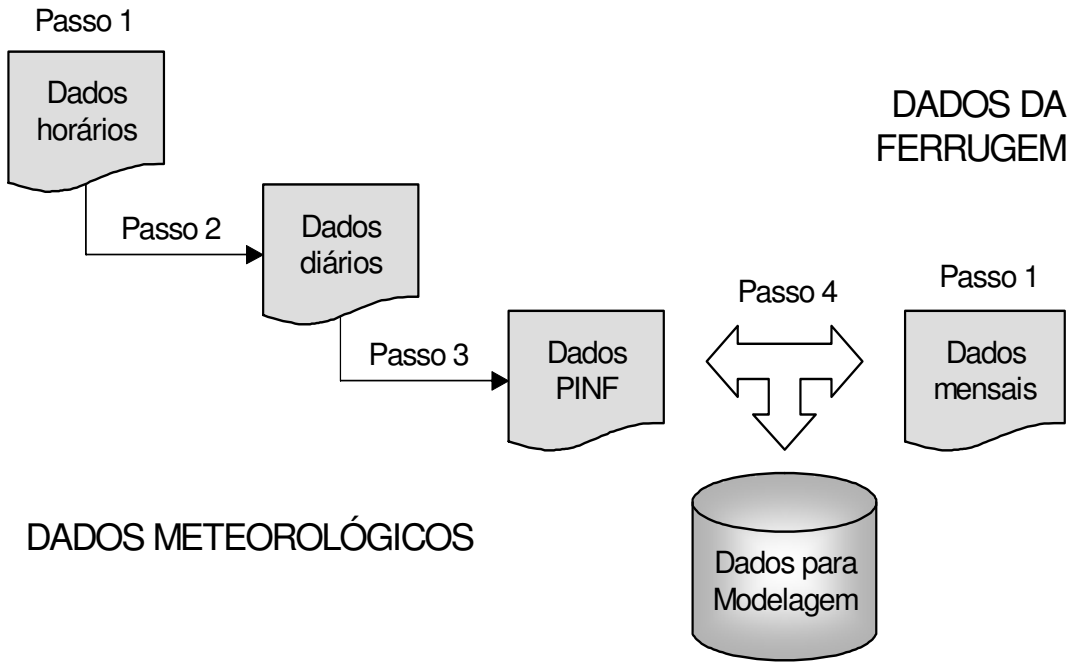


Figura 21: Esquema geral da preparação dos dados para a modelagem (MEIRA, 2008).

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.
- 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 teve inicio com a junção (concatenação) de todos os arquivos do formato ".txt" provenientes das estações meteorológicas, criando-se um longo arquivo de registros para cada uma das cidades. Foi criado um outro fator chamado de dia epidemiológico (atributos DATA_EPID e HORA_EPID). Este transformava um período de 24 horas entre dois dias em um dia meteorológico. Por exemplo, o dia meteorológico 20/01/2001 era equivalente ao período das 12:00 às 24:00 do dia 19/01/2001 acrescido do período das 00:00 às 12:00 do dia 20/01/2001. Este procedimento foi realizado a fim de evitar a quebra do período noturno, responsável por valores de molhamento foliar maiores.

Em seguida, os dados referentes à incidência da doença e outros atributos importantes como lavoura e carga, foram extraídos dos boletins de avisos e reunidos em um outro arquivo. O atributo meta correspondente foi o avaliado nas duas classes mencionadas na seção 4.2.1.

O arquivo com dados dos boletins foi guardado para ser utilizado no passo 4, já o arquivo com os dados meteorológicos foi utilizado no passo 2.

Passo 2 - Criação de atributos no nível diário

O segundo passo da preparação dos dados foi criar os atributos no nível diário. Inicialmente, foram criados os seguintes atributos: 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.

Depois foram criados os atributos relacionados ao período de molhamento foliar contínuo. Este período foi classificado como a duração de horas com umidade relativa maior

ou igual a 90%. Também foram calculados a temperatura média nesse período e o índice da condição diária de infecção.

Para finalizar este passo, foram criados atributos relacionados ao período de incubação. Foi estimado o eventual dia de infecção (Figura 20), além do mês e ano correspondentes à provável incidência da ferrugem. Estes atributos permitiram relacionar todos os dias (eventuais dias de infecção) com as taxas mensais de infecção da ferrugem do cafeeiro. Todos estes atributos estavam no arquivo final deste passo.

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.

Passo 4 - Integração dos dados

O quarto passo da preparação de dados foi a integração do arquivo gerado no passo 3 (dados meteorológicos relacionados ao PINF) com os primeiros dados retirados dos boletins de aviso (Passo 1). Esta integração foi realizada com valores baseados no 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.

Passo 5 – Integração dos dados de cada cidade

Todos os passos anteriores foram realizados isoladamente para cada uma das cidades onde havia estações meteorológicas. O passo 5 reuniu todos os arquivos de saída do passo 4 para cada uma das cidades em um arquivo final.

4.2.3 Conjunto preparado para a modelagem

O conjunto de dados preparado totalizou 738 registros referentes às respectivas cidades e às datas coletadas (Apêndice B).

4.3 Modelagem

A fase de modelagem foi dividida em duas grandes etapas: a primeira foi a de pré-indução e a segunda de indução. Na primeira etapa, o conjunto de dados foi separado em diversos cenários de indução, depois foi feito o balanceamento de classes e, em seguida, foram aplicadas as técnicas de seleção de atributos. Já na segunda fase, ocorreu a indução dos modelos utilizando as técnicas de mineração de dados. A modelagem foi realizada em um processo cíclico junto com a fase de avaliação dos modelos.

4.3.1 Etapa de pré-indução

4.3.1.1 Cenários de indução

Foram utilizados, inicialmente, quatro cenários para a indução dos modelos de predição da taxa de progresso da ferrugem do cafeeiro, que foram escolhidos de acordo com a distribuição espacial dos dados e as características da doença.

A ferrugem é mais acentuada em anos de alta carga pendente de frutos e a mistura de um ano com medidas mais severas e outro com medidas mais amenas pode "confundir" o modelo. Modelos mais simples também podem ser gerados pela separação entre as cargas, obtendo como resultado árvores de decisão menores, por exemplo. A característica bianual dos cafezais também é determinante para essa divisão, um ano há carga alta e outro há carga baixa, assim os modelos seriam mais específicos e úteis em cada uma dessas situações (MEIRA et al., 2009).

Já a distribuição geográfica fez com que as cidades fossem separadas em dois grupos, um conteve as três cidades e o outro apenas a cidade de Varginha. Essa divisão foi feita pelo fato de que Varginha possuia dados coletados desde 1998, enquanto que Carmo de Minas começou a coleta no final de 2006 e Boa Esperança em 2010.

Utilizando os dados apenas da cidade de Varginha, foram escolhidos dois cenários. Um deles conteve dados de 1998 até 2011 e foi chamado de "Varginha". No outro cenário, apenas dados de 2007 a 2011 foram utilizados, sendo descartados os dados utilizados por Meira

(2008), formando o cenário "Varginha-Novo". Os conjuntos de dados para estes cenários contaram, respectivamente, com 596 e 228 registros.

Para os outros dois cenários, foram utilizados os dados das 3 cidades. O cenário "Tudo" utilizou todos os dados do conjunto, enquanto que o cenário "Tudo-Novo" apenas dados de 2007 em diante. Os conjuntos de dados para estes cenários contaram, respectivamente, com 738 e 370 registros.

Após a construção destes quatro cenários, cada um deles recebeu outras três divisões, de acordo com a carga e atributo meta: carga alta e taxa 5 p.p; carga baixa e taxa 5 p.p.; carga alta e taxa 10 p.p. As informações sobre cada um destes cenários estão na Tabela 9.

Tabela 9: Detalhe dos diferentes cenários utilizados para a indução.

Nome do cenário	Carga	Atributo meta	Data dos registros	Cidade(s)	Número de registros
Varginha-alta-tx5	Alta	Taxa_inf_M5	1998 até 2011	Varginha	298
Varginha-baixa-tx5	Baixa	Taxa_inf_M5	1998 até 2011	Varginha	298
Varginha-alta-tx10	Alta	Taxa_inf_M10	1998 até 2011	Varginha	298
Varginha-Novo-alta-tx5	Alta	Taxa_inf_M5	2007 até 2011	Varginha	114
Varginha-Novo-baixa-tx5	Baixa	Taxa_inf_M5	2007 até 2011	Varginha	114
Varginha-Novo-alta-tx10	Alta	Taxa_inf_M10	2007 até 2011	Varginha	114
Tudo-alta-tx5	Alta	Taxa_inf_M5	1998 até 2011	Todas	369
Tudo-baixa-tx5	Baixa	Taxa_inf_M5	1998 até 2011	Todas	369
Tudo-alta-tx10	Alta	Taxa_inf_M10	1998 até 2011	Todas	369
Tudo-Novo-alta-tx5	Alta	Taxa_inf_M5	2007 até 2011	Todas	185
Tudo-Novo-baixa-tx5	Baixa	Taxa_inf_M5	2007 até 2011	Todas	185
Tudo-Novo-alta-tx10	Alta	Taxa_inf_M10	2007 até 2011	Todas	185

O cenário de carga baixa e taxa de 10 p.p. não foi gerado, pois em anos de baixa carga pendente de frutos a ferrugem tem uma evolução mais contida do que em anos de alta carga pendente de frutos, o que levou a pouquíssimos registros de ocorrências superiores a 10 p.p.. Meira (2008) não obteve bons resultados utilizando este cenário quando comparado aos

demais, o que se tornou mais uma justificativa para o mesmo não ser reproduzido neste trabalho.

4.3.1.2 Balanceamento de classes

O uso de balanceamento de classes ocorreu apenas para os cenários de carga alta - taxa 10 p.p. e carga baixa - taxa 5 p.p. A distribuição das classes para estes cenários estava com exemplos minoritários na faixa de 20% a 30%. As classes foram balanceadas conforme o método de SMOTE+TOMEK, explicado na seção 3.2.3.

Este procedimento foi realizado por uma implementação do software WEKA (versão 3.6.2) desenvolvida pelo Instituto de Ciências Matemáticas e de Computação da USP São Carlos (http://www.icmc.usp.br/Portal/). Maiores informações sobre o software WEKA estão na seção 4.3.2.2.

Após o balanceamento, o conjunto de dados ficou com cerca de 50% dos exemplos da classe minoritária. Para o cenário de carga alta e taxa de 5 p.p., não foi realizado o balanceamento, pois a classe minoritária se encontrava na faixa de 45 a 50% dos exemplos.

Sempre que o balanceamento foi realizado, o arquivo original foi mantido, pois os modelos foram induzidos no conjunto balanceado e testados no conjunto original.

4.3.1.3 Métodos de seleção de atributos

A partir do conjunto de dados preparado e balanceado para cada um dos cenários, foram utilizados os métodos de seleção de atributos. Estes métodos têm como finalidade reduzir o número de atributos do conjunto, melhorando o desempenho dos modelos gerados e proporcionando um ganho computacional. Eles foram divididos em dois grupos:

Métodos objetivos: métodos em que um algoritmo de seleção, já implementado, é utilizado para filtrar o conjunto de dados.

Métodos subjetivos: consistem na seleção de atributos baseada em características escolhidas pelo desenvolvedor do modelo. Neste caso, foram adotadas as mesmas seleções executadas por Meira (2008).

Métodos objetivos:

Os métodos objetivos de seleção de atributos utilizados foram:

- Wrapper (WRP)
- Chi-quadrado (Chi^2)
- Correlation Feature Selection (CFS)
- InfoGain (IG)
- GainRatio (GR)

O método do Wrapper necessitou de vários processamentos, pois a cada um deles estava associado um método de modelagem, por exemplo, o resultado do wrapper com árvores de decisão foi diferente do resultado com redes neurais artificiais.

Já o método do *Correlation Feature Selection* não é específico a um método, como o Wrapper, sendo que em apenas um processamento o conjunto selecionado serviu para todas as técnicas de modelagem.

O Wrapper e o CFS eliminam os atributos pouco relevantes do conjunto de dados, diferentemente dos demais métodos, os quais são responsáveis por ranquear os atributos de acordo com sua contribuição. A partir desse ranking, foram feitas diversas escolhas de atributos manualmente até se chegar ao melhor conjunto. O ranking dos atributos é o mesmo independentemente da técnica de modelagem, entretanto, a seleção manual mudou de técnica para técnica. O software utilizado para realizar o balanceamento de classes foi o WEKA 3.6.7.

Métodos subjetivos:

Os métodos chamados de subjetivos foram as seleções de atributos feitas por Meira (2008). Estas seleções foram feitas de acordo com a complexidade de obtenção dos atributos do conjunto de dados. A primeira opção foi chamada de modelagem 1 (M1) e contou com todos os atributos do conjunto de dados. A segunda opção foi a modelagem 2 (M2), a qual contou com atributos meteorológicos mais simples e de ampla disponibilidade, como temperatura, precipitação e umidade relativa, além de atributos relacionados ao molhamento foliar. Finalmente, a modelagem 3 (M3) incluiu os mesmos da modelagem 2, com exceção dos atributos relacionados ao molhamento foliar. Os atributos selecionados em cada uma das modelagens estão dispostos na Tabela 10.

Ao final do procedimento de seleção de atributos, cada cenário ficou com 8 conjuntos de dados, sendo 5 provenientes da seleção objetiva e 3 da seleção subjetiva. Nos casos em que houve balanceamento de classes, a seleção foi feita em cima do conjunto balanceado para gerar os modelos balanceados e depois feita sobre o conjunto original para gerar os modelos não balanceados, como mostra a próxima seção.

Tabela 10: Conjuntos de dados utilizados nas induções (M1, M2 e M3).

	Opção de seleção dos atributos		
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	*		
NHUR90_PINF	*	*	
SMT_NHUR90_PINF	*		
THUR90_PINF	*	*	
NHNUR90_PINF	*	*	
SMT_NHNUR90_PINF	*		
TMAX_PI_PINF	*	*	*
TMIN_PI_PINF	*	*	*
TMED_PI_PINF	*	*	*
VVENTO_PINF	*		
SMT_VVENTO_PINF	*		

* De acordo com a ordem que aparecem no conjunto de dados

4.3.2 Fase de indução

4.3.2.1 Geração de novos modelos

A fase de indução utilizou quatro técnicas de modelagem para gerar os modelos:

- Redes Neurais
- SVM
- Árvores de decisão
- Florestas aleatórias

Os modelos foram desenvolvidos para os cenários mostrados na Tabela 9. Lembrando que, para cada um dos cenários, havia 8 conjuntos de dados provenientes dos métodos de seleção de atributos.

Foram utilizadas 4 técnicas de modelagem, com os doze cenários da Tabela 9 (combinações de carga, atributo meta e distribuição espacial dos dados) para 8 conjuntos de dados provenientes da seleção de atributos, gerando inicialmente 384 induções. Porém, este número foi ainda maior quando a indução também foi realizada para os conjuntos desbalanceados dos cenários de carga alta – taxa 10 p.p. e carga baixa – 5 p.p.

Como mencionado anteriormente, estes cenários utilizaram o balanceamento de classes, e o conjunto de dados balanceado foi utilizado para gerar os modelos, os quais foram posteriormente testados no conjunto original. O conjunto original também foi utilizado para a indução dos modelos, justamente para verificar qual seria o efeito do balanceamento de classes nesses modelos.

Assim, aos 8 cenários que continham combinações de carga alta – taxa 10 p.p. e carga baixa – taxa 5 p.p., foi aplicada uma nova rodada de induções, desta vez com 4 técnicas e 8 conjuntos provenientes da seleção de atributos, acrescentando 256 induções, gerando um total de 640 induções.

Os modelos gerados foram avaliados de acordo com as medidas de desempenho propostas na seção 3.1.5, já a avaliação e seleção do melhor modelo está no capítulo 6.

4.3.2.2 Configurações do software

O software utilizado para a indução dos modelos foi o WEKA, versão 3.7.9. (HALL et al., 2009). O WEKA é um conjunto de algoritmos de aprendizado de máquina e de ferramentas relacionadas, que também oferece suporte ao processo completo de mineração de dados (WITTEN et al., 2011). Ele é um *software* livre, gratuito (http://www.cs.waikato.ac.nz/ml/weka/), distribuído sob a licença de uso GNU (*General Public License*).

Em pesquisas recentes (KDNUGGETS, 2012), este software foi uma das ferramentas mais usadas em trabalhos com mineração de dados. Esta pesquisa incluía tanto softwares livres quanto softwares com licença paga. As configurações do WEKA são mostradas mais detalhadamente para cada técnica de modelagem.

- Arvores de decisão:

As árvores de decisão foram geradas por meio do classificador nomeado "J48" no WEKA. Uma opção utilizada foi o número mínimo de objetos por folha, a qual gera uma árvore onde cada folha teve uma quantidade mínima de registros. Para a escolha desta opção, foram gerados diversos gráficos avaliando o número mínimo de objetos por folha pela taxa de acerto do modelo gerado (Figura 22).

Quanto menos objetos por folha, maior ficava a árvore e melhor era sua taxa de acerto. Porém, um número muito pequeno de registros por folha poderia gerar um sobreajuste no modelo, pois ele acertaria casos específicos ao conjunto de treinamento.

Após análises de diversos gráficos similares ao da Figura 22, decidiu-se utilizar a opção de 5 objetos por folha. Esta opção evita um sobreajuste muito grande, como o caso de 1 ou 2 objetos por folha, e não deixa que o modelo perca um percentual grande de sua taxa de acerto (notou-se uma perda acentuada após 10 registros por folha). Além disso, esta foi a

mesma opção de geração dos modelos de Meira (2008), o que permitiu uma comparação mais precisa. As demais opções foram as mesmas utilizadas por Meira (2008), as recomendadas como “configurações padrão” do software WEKA.

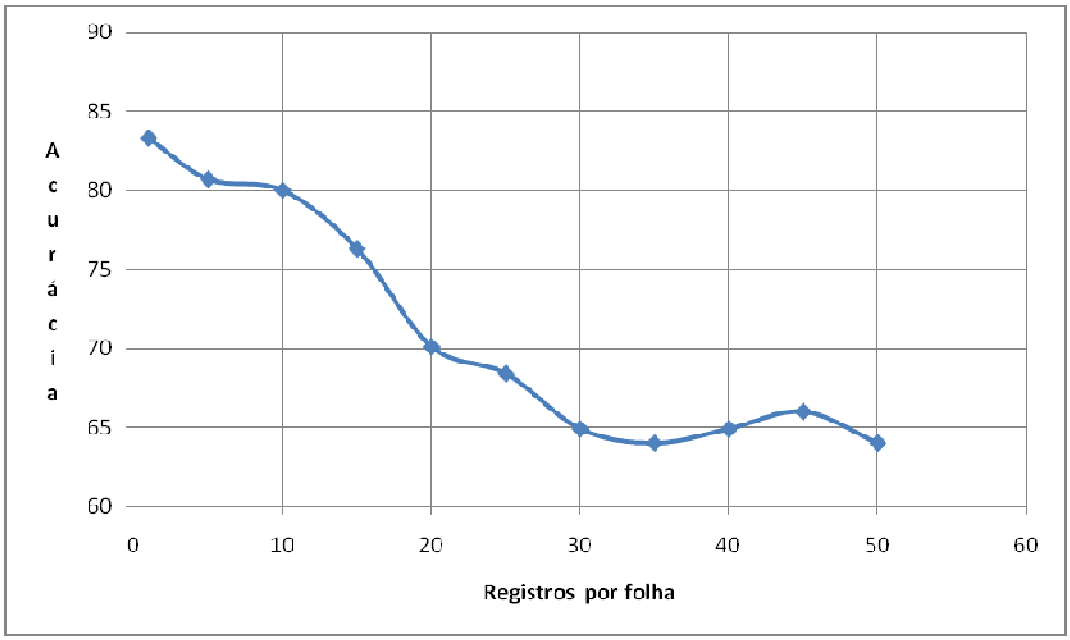


Figura 22: Escolha do número de registros por folha para a indução.

- Redes neurais artificiais:

A topologia de uma rede neural é o seu formato. Ela é composta pelo número de camadas intermediárias e a quantidade de neurônios que cada camada deve ter. O software WEKA recomenda a utilização de algumas medidas padrões para a quantidade de neurônios por camada, entretanto, não há recomendações quanto ao número de camadas intermediárias de uma rede.

Realizando um procedimento similar ao realizado para árvores de decisão, foi avaliada a topologia das redes com a taxa de acerto dos modelos induzidos por essa estrutura. Foram geradas redes neurais de uma, duas e três camadas intermediárias e cada camada teve seu número de neurônios avaliado de 1 até 10, além das opções do software WEKA.

Após a geração de todas as redes neurais, duas delas obtiveram desempenho muito próximo, uma com duas camadas intermediárias de dois neurônios (2,2) e a outra com duas camadas intermediárias utilizando o número de atributos do conjunto de dados como número

de neurônios das camadas (está é uma das opções do weka). Para a seleção dentre estas duas opções, foi comparado o tempo de geração de cada um dos modelos e notou-se que a rede 2,2 foi construída, em média, cerca de 80% mais rápida do que a rede com a opção do WEKA, portanto ela foi a escolhida (Figura 23). As demais opções de indução foram mantidas as padrões do software.

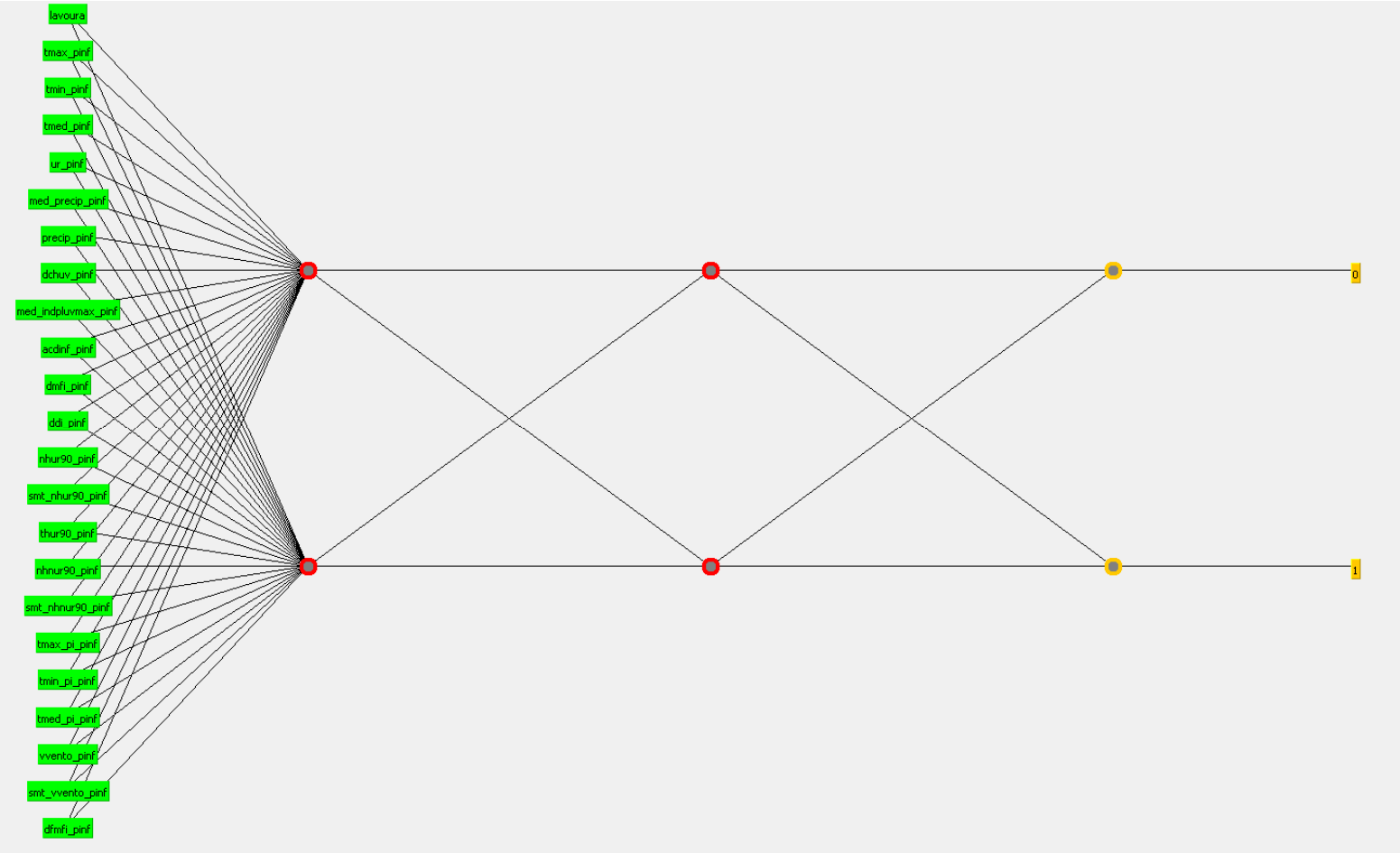


Figura 23: Exemplo de uma rede neural com duas camadas intermediárias.

- Máquinas de vetores suporte

Dentre as técnicas utilizadas para a modelagem, as SVM foram as que mais necessitaram de ajustes. Foi necessário calibrar quatro parâmetros para realizar a indução, os quais foram calibrados seguindo procedimento análogo às redes neurais e às árvores de decisão.

O primeiro parâmetro calibrado foi o seu "kernel". O kernel é o produto interno que elevou os dados a uma dimensão maior, para posteriormente classificá-los. São 4 opções de Kernel: linear, polinomial, RBF e Sigmóide

Dentro do Kernel, uma variável que necessitou de calibração foi o parâmetro gamma (γ). Foram testados valores de gamma de $10^{(-4)}$ até $10^{(4)}$, variando-se em potências de 10, além da opção padrão do weka (1 / número de atributos). O melhor valor de gamma escolhido foi de $10^{(-1)}$.

Outros dois parâmetros precisaram de calibração: o coeficiente de custo (c) e o parâmetro epsilon (ε). Foram avaliados os mesmos valores da variável gamma e as opções padrões do software se mostraram as melhores. O valor do coeficiente de custo ficou em 1, enquanto que épsilon ficou em $10^{(-3)}$.

Para a indução por SVM, foi utilizada uma biblioteca chamada de LIBSVM (CHANG e LIN, 2011), a qual pode ser instalada pelo próprio aplicativo WEKA.

- Florestas aleatórias

As florestas aleatórias necessitaram de dois parâmetros para calibração. Estes dois parâmetros foram a profundidade das árvores e o número de atributos aleatórios utilizados nas árvores. Realizando o mesmo procedimento para as outras técnicas, foram testados os valores de 2 a 12 para os dois parâmetros. A opção padrão do WEKA para a profundidade da árvore também foi testada, ela gera uma árvore com a máxima profundidade possível.

O valor que apresentou melhor resultado para ambos os casos foi o valor 8. Assim, tem-se até 8 atributos aleatórios nas árvores geradas, cada uma com uma profundidade de até 8 níveis. Um nível pode ser considerado como uma "linha horizontal" da árvore, onde o nó raiz simboliza o primeiro nível; os nós derivados deste o segundo nível e assim por diante. Por exemplo, a árvore da Figura 3 tem 2 níveis. Houve casos em que o conjunto original de modelagem continha menos do que 8 atributos (pelos métodos de seleção de atributos objetivos) e também casos em que as árvores geraram menos de 8 níveis de profundidade. A quantidade de árvores geradas em cada floresta foi definida em 100, seguindo recomendação de Breiman (2001).

4.4 Avaliação dos modelos

Nesta dissertação, foi realizado o processo de Validação dos modelos de Meira (2008) e também a avaliação do desempenho dos diversos modelos gerados, bem como a seleção dos melhores modelos. É importante ressaltar que a validação é um processo que foi realizado separadamente da avaliação dos modelos.

4.4.1 Validação dos modelos em árvore de decisão

Para realizar o processo de validação dos modelos de Meira (2008), o conjunto de dados foi separado em duas partes, uma idêntica à utilizada pelo autor em sua pesquisa, que variou de outubro de 1998 até outubro de 2006, e a outra de novembro de 2006 em diante, a qual foi usada como conjunto de teste. Apenas dados de Varginha foram utilizados nestes conjuntos.

O conjunto de 1998 a 2006 foi carregado no WEKA, e foram gerados modelos idênticos aos de Meira (2008). Em seguida, os arquivos do conjunto de teste foram carregados através da opção "*supplied test set*", e foram avaliados para os modelos gerados anteriormente.

A regra para determinar se um modelo foi considerado validado é a obtenção de uma taxa de acerto igual ou superior à obtida na sua construção. Caso a taxa de acerto fosse inferior, o modelo não seria aceito. Outras medidas de avaliação também foram utilizadas para detalhar o comportamento destes modelos para dados não utilizados em sua construção.

Neste procedimento, houve apenas a seleção de atributos subjetiva. Não foram avaliados aspectos construtivos dos modelos, pois este se trata de um estudo de validação e não de verificação. Os resultados obtidos por Meira (2008) estão dispostos na Tabela 11.

Tabela 11: Medidas de avaliação* dos modelos de Meira (2008).

Nome do Modelo	Atributo Meta	Carga	Taxa de acerto (%)	Erro (%)	Sensitividade (%)	Especificidade (%)
tx5altaM1	5 p.p.	Alta	80,8	19,2	78,5	82,6
tx5altaM2	5 p.p.	Alta	81,9	18,1	78,5	84,7
tx5altaM3	5 p.p.	Alta	81,8	18,2	75,3	87,8
tx5baixaM1	5 p.p.	Baixa	70,3	29,7	34,7	84,5
tx5baixaM2	5 p.p.	Baixa	72,1	27,9	38,0	86,0
tx5baixaM3	5 p.p.	Baixa	69,2	30,8	26,3	86,7
tx10altaM1	10 p.p.	Alta	78,7	21,3	64,7	84,6
tx10altaM2	10 p.p.	Alta	76,4	23,6	57,3	84,6
tx10altaM3	10 p.p.	Alta	79,2	20,8	57,3	88,5
tx10baixaM1	10 p.p.	Baixa	86,8	13,2	10,0	96,9
tx10baixaM2	10 p.p.	Baixa	86,3	13,7	18,3	95,1
tx10baixaM3	10 p.p.	Baixa	86,2	13,8	23,3	94,4

**Os valores da tabela são para validação cruzada.*

Para cada um desses 12 modelos, ocorreu um processo de validação, sendo que os resultados encontram-se discutidos no capítulo 5.

4.4.2 Desempenho e escolha dos melhores modelos

A avaliação de desempenho e seleção dos melhores modelos ocorreu para cada um dos cenários mencionados na seção 4.3.1.1. Os modelos tiveram seus atributos selecionados e medidas de avaliação analisadas, e a escolha do(s) melhor(es) ocorreu, principalmente, pela curva ROC.

5 Validação de modelos de alerta da ferrugem do cafeeiro

Este capítulo visou validar os modelos de Meira (2008) para dados que não foram utilizados em sua construção, ou seja, de novembro de 2006 em diante. O processo de validação ocorreu como um ciclo do processo KDD (Figura 14), onde a fase de avaliação foi a responsável por averiguar se ocorreu a validação. Vale lembrar que, em seu trabalho, Meira (2008) gerou modelos com dois softwares, sendo um deles o WEKA. Os modelos gerados com este software são considerados neste capítulo.

Os dados provenientes da estação nova, para a cidade de Varginha, foram coletados e tratados, de forma a se gerar o conjunto utilizado na validação. Estes dados estavam com o padrão de umidade relativa da respectiva estação e foram corrigidos para o padrão da estação antiga, como explicado na seção 4.1.4. O conjunto de dados foi separado nos quatro cenários utilizados por Meira (2008): carga alta e atributo meta 10 p.p.; carga alta e atributo meta 5 p.p.; carga baixa e atributo meta 10 p.p.; carga baixa e atributo meta 5 p.p. (Tabela 11). Cada um destes conjuntos conteve 122 registros.

Os modelos de Meira (2008) foram carregados no software WEKA e os conjuntos de dados citados anteriormente foram utilizados como conjunto de teste, por meio da opção "*supplied test set*" no software. Cada conjunto foi avaliado ao seu respectivo modelo, sendo que este procedimento foi feito por 12 vezes, uma vez para cada uma das quatro opções carga e atributo meta, e para 3 opções de modelagem (M1, M2 e M3). Os resultados estão expressos a seguir (Tabela 12 a Tabela 15), e foram baseados na matriz de confusão e nas medidas de avaliação citadas na seção 3.1.5. A regra para determinar se um modelo foi aceito é que este obtivesse uma taxa de acerto igual ou superior à obtida na sua construção (seção 4.4.1).

Os modelos de Meira (2008) para o cenário de carga alta e atributo meta 5 p.p. não foram aceitos (Tabela 12), afinal a taxa de acerto foi menor na avaliação do que na construção em todos os conjuntos (M1, M2 e M3). De maneira geral, ela caiu da faixa de 80% para a faixa de 60%, já as outras medidas de avaliação como sensitividade e especificidade também foram inferiores no conjunto de teste, mostrando, ainda mais, um desempenho inferior na fase de avaliação.

Tabela 12: Resultado da validação para os modelos de carga alta e taxa 5 p.p.

	Medidas de avaliação					
	M1		M2		M3	
	Avaliação	Meira	Avaliação	Meira	Avaliação	Meira
Taxa de acerto	62,3	80,8	65,6	81,9	60,7	81,8
Erro	37,7	19,2	34,4	18,1	39,3	18,2
Sensitividade	58,6	78,5	67,2	78,5	51,7	75,3
Especificidade	65,6	82,6	64,1	84,7	68,7	87,8

Os modelos de Meira (2008) para o cenário de carga alta e atributo meta 10 p.p. também não foram aceitos (Tabela 13), uma vez que a taxa de acerto foi menor na avaliação do que na construção para todos os conjuntos de dados. A diferença na taxa de acerto entre a fase de construção e avaliação foi maior para os conjuntos M1 e M3 do que para o conjunto M2, chegando à casa dos 5 p.p.. Este fato indicou uma perda de desempenho menor do que os demais conjuntos, fazendo com que este fosse o modelo mais próximo a ser aceito.

Os valores de especificidade na avaliação foram levemente superiores aos valores da construção em dois casos, M1 e M2, e inferior para o caso de M3. Nos modelos de Meira (2008), verificou-se valores baixos de sensitividade quando comparados à especificidade, e tal diferença foi mais acentuada quando submetidos à avaliação. Todos os valores de sensitividade foram inferiores em sua avaliação.

Tabela 13: Resultado da validação para os modelos de carga alta e taxa 10 p.p.

	Medidas de avaliação					
	M1		M2		M3	
	Avaliação	Meira	Avaliação	Meira	Avaliação	Meira
Taxa de acerto	69,7	78,7	71,3	76,4	71,3	79,2
Erro	30,3	21,3	28,7	23,6	28,7	20,8
Sensitividade	32,4	64,7	37,8	57,3	37,8	57,3
Especificidade	85,9	84,6	85,9	84,6	85,9	88,5

Analisando o cenário de carga baixa e atributo meta 5 p.p., houve um modelo que foi validado e dois que não foram aceitos (Tabela 14). O modelo que foi validado foi o gerado pelo conjunto M1, onde a taxa de acerto foi superior na fase de avaliação do que na fase de construção, mostrando que esse modelo manteve sua taxa de acerto para dados externos ao conjunto de treinamento. Já as taxas de acerto de M2 e M3 foram levemente inferiores para o caso de avaliação e, portanto, os modelos não foram considerados aceitos.

A especificidade foi inferior na fase de avaliação para todos os modelos, mesmo para o gerado por M1, que foi um modelo validado. A sensitividade deste modelo também foi inferior na fase de avaliação, já para o modelo gerado por M3, esta medida foi superior na avaliação do que na construção, mas não foi suficiente para validá-lo.

O valor de 25% para sensitividade na fase de avaliação do modelo gerado por M1 mostra que este modelo está classificando corretamente apenas um quarto dos exemplos positivos. Apesar de validado, este modelo teria um desempenho melhor caso sua sensitividade obtivesse valores maiores. Mesmo considerando o valor de sensitividade de sua construção (34,7%), quando este modelo é comparado aos modelos do capítulo 6, ele tem um desempenho muito inferior e não seria recomendado predizer o aumento da taxa de progresso da ferrugem do cafeeiro.

Tabela 14: Resultado da validação para os modelos de carga baixa e taxa 5 p.p.

	Medidas de avaliação					
	M1		M2		M3	
	Avaliação	Meira	Avaliação	Meira	Avaliação	Meira
Taxa de acerto	70,5	70,3	68,8	72,1	68,8	69,2
Erro	29,5	29,7	31,1	27,9	31,1	30,8
Sensitividade	25,0	34,7	28,6	38,0	35,7	26,3
Especificidade	84,0	84,5	80,8	86,0	78,7	86,7

Analisando o cenário de carga baixa e atributo meta 10 p.p., houve apenas um modelo que não foi aceito, que foi o modelo gerado por M2 (Tabela 15). Os modelos gerados por M1 e M3 foram considerados validados, pois suas taxas de acerto foram superiores na fase de avaliação do que na fase de construção.

O modelo gerado por M1 obteve valor zero de sensitividade, o que mostrou que o modelo não classificou corretamente nenhum caso positivo de aumento da taxa de progresso. Neste caso, o que ocorreu foi que o modelo gerado por M1 classificou todos os casos do conjunto de teste como "não aumento da taxa de progresso", não detectando nenhum caso de aumento da taxa de progresso. Neste caso, mesmo validado, este modelo não é recomendado para determinar a taxa de progresso da ferrugem do cafeeiro. Já o modelo gerado por M3 apresentou o mesmo problema do modelo M1 do cenário anterior, que foi o baixo valor de sensitividade, apenas 13,8%. Este modelo está classificando uma baixa porcentagem de exemplos positivos corretamente e quando este modelo é comparado aos modelos do capítulo 6, e pelo seu desempenho ele também não seria recomendado predizer o aumento da taxa de progresso da ferrugem do cafeeiro.

Como comentado anteriormente, neste trabalho e em Meira (2008), os modelos de carga baixa e atributo meta 10 p.p. não obtiveram um desempenho interessante para ser utilizado na predição da taxa de progresso da ferrugem do cafeeiro, dado seus baixos valores de sensitividade, e assim, mesmo validados, não devem ser utilizados.

Tabela 15: Resultado da validação para os modelos de carga baixa e taxa 10 p.p.

	Medidas de avaliação					
	M1		M2		M3	
	Avaliação	Meira	Avaliação	Meira	Avaliação	Meira
Taxa de acerto	89,6	86,8	85,2	86,3	87,7	86,2
Erro	10,7	13,2	14,7	13,7	12,3	13,8
Sensitividade	0,0	10,0	15,4	18,3	15,4	23,3
Especificidade	100,0	96,9	93,6	95,1	96,3	94,4

Considerações finais:

Verificou-se que dos 12 modelos desenvolvidos por Meira (2008), apenas 3 foram aceitos, todavia, a sua utilização não é recomendada, dado aos baixos valores de sensitividade tanto na sua construção quanto em sua avaliação. Essa situação evidenciou, ainda mais, a necessidade de uma nova indução de modelos de alerta para determinar a taxa de progresso do cafeeiro para dados mais recentes.

6 Desenvolvimento de modelos de alerta

Este capítulo trata da construção dos novos modelos de alerta. Foram desenvolvidos um total de 640 modelos, seguindo o proposto na seção 4.3, e estes modelos foram selecionados de acordo com o estabelecido na seção 4.4.2. A seção 6.1 é utilizada, como exemplo, para exibir alguns dos modelos de alerta desenvolvidos. Já os demais modelos são apresentados no apêndice A.

6.1 Modelos para o cenário Varginha

Esta seção apresenta os modelos desenvolvidos para o cenário de Varginha. Cada subseção contém uma figura referente ao gráfico ROC do respectivo cenário de indução, com os modelos selecionados destacados no envelope convexo. Para estes modelos, há uma tabela contendo suas medidas de avaliação e informações sobre os atributos utilizados no seu conjunto de treinamento.

A Figura 24 contém um gráfico ROC para os modelos com combinação de alta carga pendente de frutos e atributo meta 5 p.p. A numeração desses modelos ocorreu da seguinte maneira: modelos de 1-8 representam árvores de decisão, modelos de 9 a 16 florestas aleatórias, 17 a 24 redes neurais artificiais e 25 a 32 máquinas de vetores suporte. Esta numeração foi a mesma utilizada nos Apêndices A.1.1, A.2.1 e A.3.1.

Já a Figura 25 e a Figura 26 mostram os gráficos ROC para modelos com combinação de alta carga pendente de frutos e atributo meta 10 p.p. e baixa carga pendente de frutos e atributo meta 5 p.p., cenários em que o balanceamento de classes foi executado. Assim, os modelos foram numerados da seguinte maneira: os modelos de 1 a 16 representam árvores de decisão, os modelos de 17 a 32 florestas aleatórias, 33 a 48 redes neurais artificiais e 49 a 64 máquinas de vetores suporte, onde a primeira metade dos modelos de cada uma das técnicas representa os modelos gerados por arquivos não balanceados e a segunda metade modelos gerados por arquivos balanceados. Esta numeração se repetiu para os Apêndices A.1.2, A.1.3, A.2.2, A.2.3, A.3.2 e A.3.3.

Por exemplo, os modelos 1 a 8 são árvores de decisão geradas com arquivos não balanceados, já os modelos 9 a 16 são árvores de decisão geradas com arquivos balanceados. Os modelos 17 a 24 são florestas aleatórias geradas com arquivos não balanceados, já os modelos 25 a 32 são florestas aleatórias geradas por arquivos balanceados; e assim por diante.

6.1.1 Cenário Varginha-alta-tx5

O gráfico ROC da Figura 24 representa o desempenho dos modelos desenvolvidos para o cenário Varginha-alta-tx5. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas na Tabela 16 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 17.

Neste cenário, foram selecionados 2 modelos no envelope convexo: 22 e 28. Para os modelos pertencentes ao envelope convexo foi atribuído um símbolo diferente, para ajudar na sua identificação, conforma a legenda da Figura 24.

A discussão sobre os atributos presentes no conjunto de dados que gerou cada um destes modelos (Tabela 17) é realizada na seção 6.2.2, enquanto que as medidas de avaliação (Tabela 16) são discutidas na seção 6.2.1. A partir dessa discussão foi possível selecionar um dentre os 2 modelos para representar este cenário (seção 6.2.3). Após a seleção dos melhores modelos, eles são comparados com outros modelos de alerta na seção 6.2.4.

Na Figura 24 ainda há o realce de um grupo de modelos, dado a comportamentos específicos. Este grupo contém os modelos 20, 21, 22 e 24, os quais se destacaram por alto valor de sensitividade e foram todos gerados por redes neurais artificiais. A discussão sobre características peculiares de cada uma das técnicas de mineração de dados utilizadas neste trabalho está descrita na seção 6.2.6.

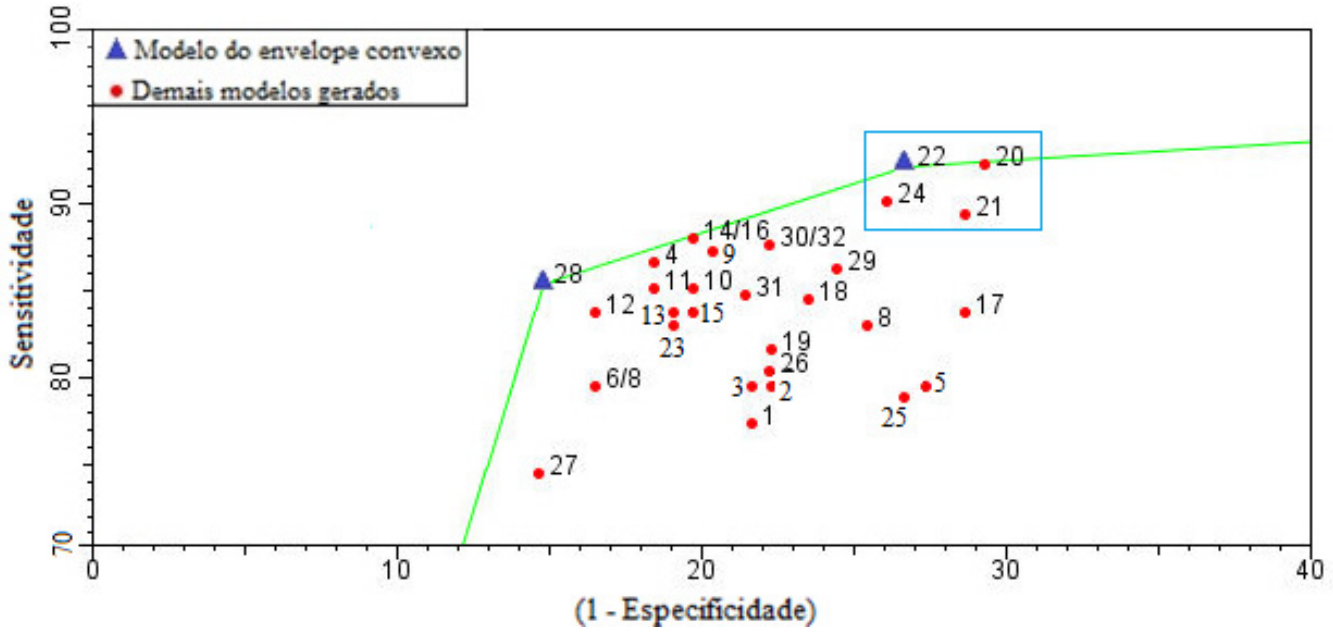


Figura 24: Gráfico ROC para o cenário Varginha-alta-tx5.

Tabela 16: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Varginha-alta-tx5.

Modelos	22	28
Técnica de modelagem	RNA	SVM
Seleção de atributos	Chi^2	WRP
Taxa de acerto	82,2	85,3
Erro	17,8	14,7
Sensitividade	92,2	85,4
Especificidade	73,3	85,2
Confiabilidade Positiva	75,6	85,4
Confiabilidade Negativa	91,3	85,2
TP Rate	92,2	85,4
FP Rate	26,8	14,8
AUC	0,841	0,853
Kappa	0,65	0,71

Tabela 17: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Varginha-alta-tx5.

	Modelos	
Atributos	22	28
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	*	
NHUR90_PINF		
SMT_NHUR90_PINF		
THUR90_PINF		*
NHNUR90_PINF		
SMT_NHNUR90_PINF		
TMAX_PI_PINF		
TMIN_PI_PINF		
TMED_PI_PINF		
VVENTO_PINF		
SMT_VVENTO_PINF		*

6.1.2 Cenário Varginha-alta-tx10

O gráfico ROC da Figura 25 representa o desempenho dos modelos desenvolvidos para o cenário Varginha-alta-tx10. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas na Tabela 18 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 19.

Houve casos em que dois ou mais modelos obtiveram as mesmas medidas de desempenho, assim eles passaram a ser tratados como apenas um e renomeados. Neste cenário os modelos 62 e 64 obtiveram as mesmas medidas de desempenho, resultando no modelo chamado de 62/64.

Tabela 18: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Varginha-alta-tx10.

Modelos	15	26	52	59	61	62/64
Técnica de modelagem	AD	RF	SVM	SVM	SVM	SVM
Seleção de atributos	GR	M2	WRP	M3	CFS	Chi^2/IG
Taxa de acerto	76,5	89,9	85,3	88,6	86,0	83,5
Erro	23,5	10,1	14,7	11,4	74,0	16,5
Sensitividade	96,7	90,0	70,1	93,3	94,3	95,4
Especificidade	67,8	89,9	92,4	86,5	82,2	77,8
Confiabilidade Positiva	56,5	79,4	81,3	75,0	71,3	66,9
Confiabilidade Negativa	97,9	95,4	86,8	96,8	96,8	97,3
TP Rate	96,7	90,0	70,1	93,3	94,3	95,4
FP Rate	32,2	10,1	7,6	13,5	17,8	22,2
AUC	0,857	0,930	0,813	0,936	0,882	0,866
Kappa	0,54	0,77	0,65	0,75	0,70	0,73

Neste cenário, foram selecionados 6 modelos no envelope convexo: 15, 26, 52, 59, 61 e 62/64. O modelo 52 foi gerado por meio de arquivos sem o balanceamento de classes, já os demais foram todos gerados com arquivos balanceados. As linhas transversais separam os modelos gerados por arquivos não balanceados (abaixo da linha de 77% de sensitividade) dos

modelos gerados por arquivos balanceados (acima da linha de 82% de sensitividade), seguindo o explicado na seção 6.2.5. Na Figura 25 ainda há o realce de um grupo de modelos, o qual contém 10 modelos, que se destacaram por apresentarem os melhores desempenhos para este cenário, além do fato que todos foram gerados por SVM ou florestas aleatórias.

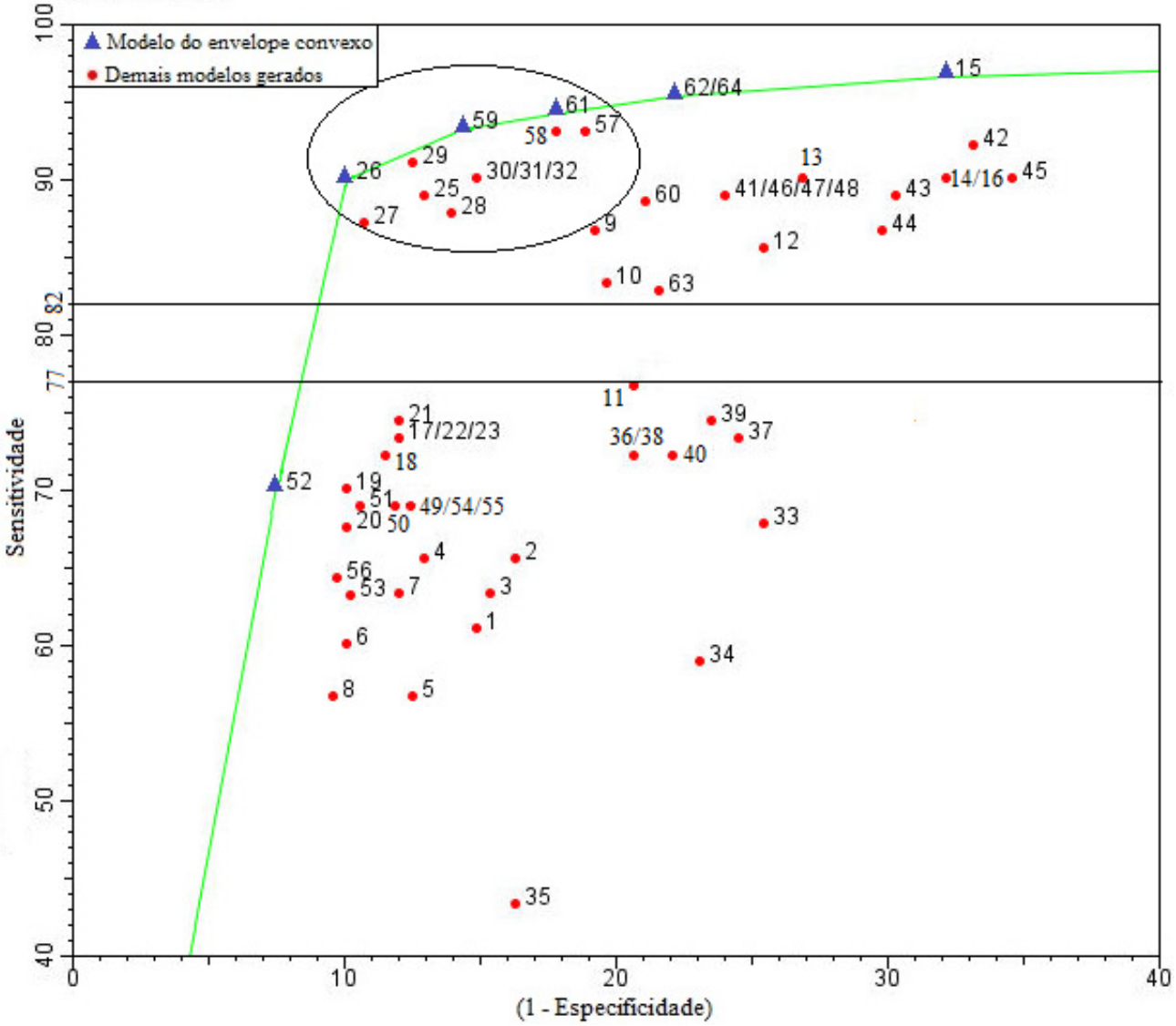


Figura 25: Gráfico ROC para o cenário Varginha-alta-tx10.

Tabela 19: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Varginha-alta-tx10.

Atributos	Modelos					
	15	26	52	59	61	62/64
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	*				*	*
NHUR90_PINF		*				
SMT_NHUR90_PINF						
THUR90_PINF	*	*			*	
NHNUR90_PINF		*				
SMT_NHNUR90_PINF						
TMAX_PI_PINF		*		*		
TMIN_PI_PINF		*		*		
TMED_PI_PINF		*		*		
VVENTO_PINF						
SMT_VVENTO_PINF					*	

6.1.3 Cenário Varginha-baixa-tx5

O gráfico ROC da Figura 26 representa o desempenho dos modelos desenvolvidos para o cenário Varginha-baixa-tx5. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas na Tabela 20 e o conjunto de atributos para o modelo selecionado foi o M1.

Neste cenário, foi selecionado apenas um modelo no envelope convexo (25/32). As linhas transversais separam os modelos gerados por arquivos não balanceados dos gerados por arquivos balanceados (seção 6.2.5). Na Figura 26 ainda há o realce de um grupo de modelos, o qual contém 14 modelos, que se destacaram por apresentarem os melhores desempenhos para este cenário, além do fato que todos foram gerados por SVM ou florestas aleatórias.

Tabela 20: Resultado da avaliação para o modelo selecionado no envelope convexo para o cenário Varginha-baixa-tx5.

Modelo	25/32
Técnica de modelagem	RF
Seleção de atributos	M1
Taxa de acerto	88,9
Erro	11,1
Sensitividade	86,3
Especificidade	89,9
Confiabilidade Positiva	75,8
Confiabilidade Negativa	94,7
TP Rate	86,3
FP Rate	10,1
AUC	0,917
Kappa	0,73

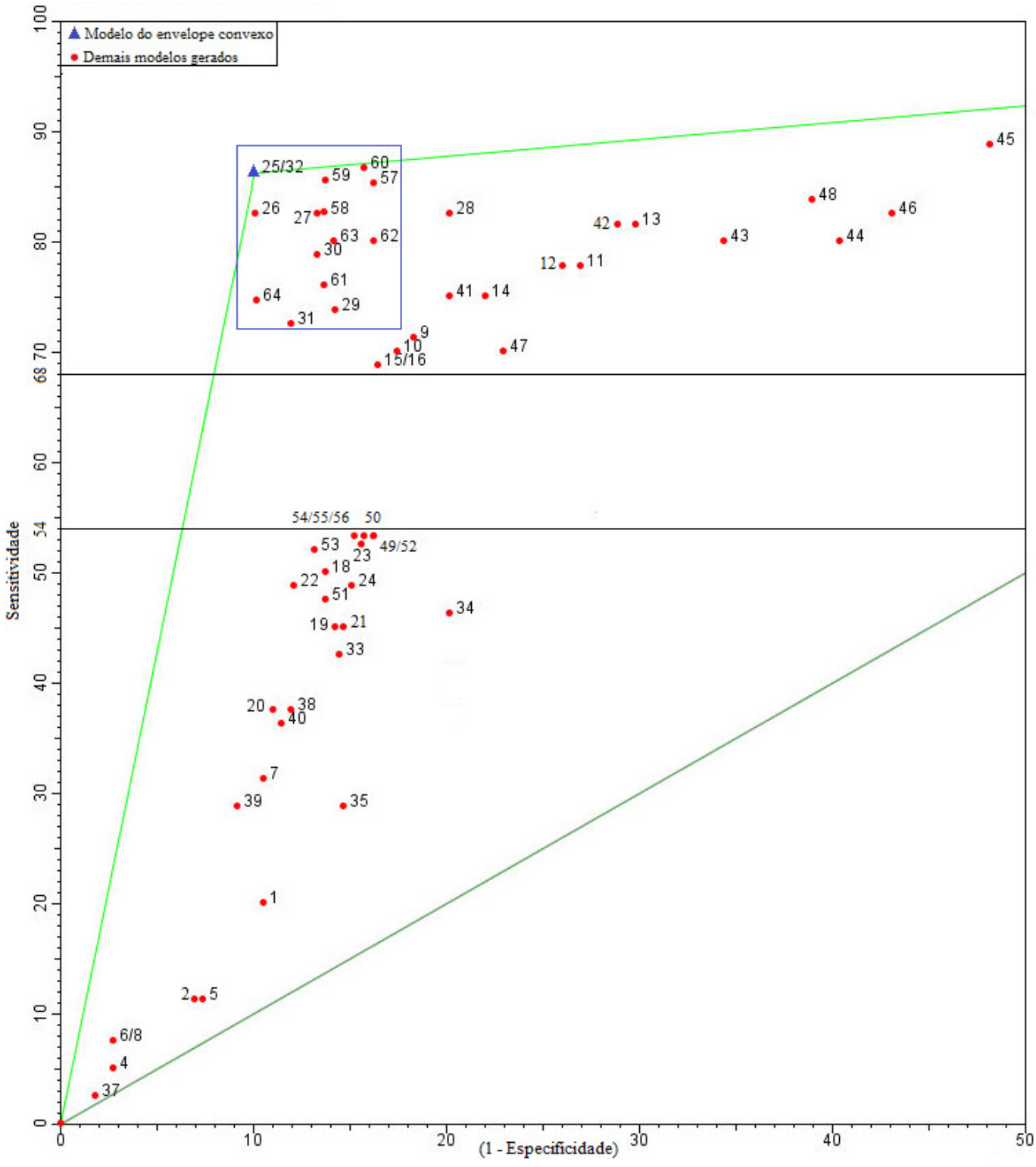


Figura 26: Gráfico ROC para o cenário Varginha-baixa-tx5.

6.2 Discussão sobre os modelos desenvolvidos

A discussão sobre os modelos de alerta ocorre em diversas etapas, abrangendo pontos importantes da mineração de dados, aspectos epidemiológicos e chegando à indicação de um modelo para cada um dos cenários.

6.2.1 Medidas de avaliação

As medidas de avaliação dos modelos selecionados nos diversos envelopes convexos foram apresentadas na seção 6.1 e no Apêncide A. A primeira medida de avaliação, a taxa de acerto, foi a principal medida utilizada para avaliar um classificador, afinal por meio dela é possível saber quantas predições relativas à taxa de progresso da ferrugem do cafeeiro foram feitas corretamente. Esta medida também é fator determinante para que um modelo seja descartado quando comparado a outro, que tenha uma taxa de acerto significativamente superior. Entretanto, esta medida aliada a valores de sensitividade e especificidade foi o que determinou o melhor desempenho de um modelo em relação a outro.

A sensitividade indica como o modelo trabalha com exemplos positivos (exemplos de aumento da taxa de progresso da ferrugem), quanto maior seu valor, uma maior porcentagem destes exemplos é classificada corretamente. Para a ferrugem do cafeeiro, uma predição correta de aumento da taxa de progresso faz com que o produtor fique atento e se prepare para a aplicação de fungicidas, a fim de conter o avanço predito da doença. Esta aplicação pode ser considerada uma aplicação "correta" do fungicida, afinal ela está sendo utilizada para combater o avanço da doença quando esta realmente ocorre, evitando perdas econômicas por parte do produtor. Caso ocorra uma predição incorreta do aumento da taxa de progresso, dizendo que não haveria o aumento, mas na realidade ele ocorrer, o produtor pode não acompanhar o desenvolvimento da doença na lavoura, já que o modelo não indicou que haveria risco da doença progredir. Considerando que a ferrugem pode causar danos de até 50% na produção, além de apresentar uma evolução rápida nos meses mais quentes, devido ao baixo valor do período de incubação do *H vastatrix.*, uma epidemia da ferrugem já pode ter se

alastrado pela lavoura devido a um descuido por parte do produtor, que não conseguirá efetuar mais o controle da doença, mesmo realizando a aplicação de fungicidas.

Já a especificidade mostra como o modelo trabalha com exemplos negativos (exemplos de não aumento da taxa de progresso), quanto maior seu valor, uma maior porcentagem destes exemplos é classificada corretamente. Para a ferrugem do cafeeiro, uma predição correta de não aumento da taxa de progresso evita que o produtor aplique o fungicida de forma desnecessária, reduzindo as perdas financeiras do produtor com a compra e mão de obra para aplicação do produto e evitando qualquer tipo de impacto ambiental, uma vez que o produto não será aplicado. Caso ocorra uma predição incorreta de não aumento da taxa de progresso, ou seja, o modelo predizer um aumento da taxa de progresso e na realidade este não ocorrer, o produtor ficaria tentado a aplicar o fungicida, com medo de um avanço da doença e estaria, consequentemente, desperdiçando o dinheiro gasto com a compra e a mão de obra para a aplicação do fungicida, isto sem contar o impacto ambiental causado por este tipo de agrotóxico.

Um modelo ideal seria o modelo que não cometesse erros, acertando todos os exemplos positivos e negativos. Assim, o que se busca em um modelo de alerta é um alto valor de taxa de acerto, aliado a valores altos, tanto de sensitividade, quanto de especificidade. Quanto mais próximo estes dois valores estiverem, mais equilibrado um modelo será, por classificar corretamente uma mesma porcentagem de exemplos positivos e negativos. Dentre diversos modelos com os mesmos valores, ou valores próximos, de taxa de acerto, o melhor é aquele que apresenta medidas de sensitividade e especificidade mais altas e próximas.

Com relação aos diversos modelos desenvolvidos e selecionados no envelope convexo, o maior valor de taxa de acerto obtido foi de 90,5% (Modelo 59 do cenário Tudo-alta-tx10) e a menor foi de 64,1% (Modelo 48 do cenário Varginha-Novo-baixa-tx5), mostrando a discrepância no comportamento entre os modelos gerados. A diferença de taxa de acerto chamou a atenção em três cenários: Tudo-Novo-alta-tx5, Varginha-alta-tx10 e Varginha-Novo-baixa-tx5.

Para o primeiro destes cenários (Tudo-Novo-alta-tx5 – seção A.2.1), a taxa de acerto apresentou uma diferença de 12,2 p.p. entre o modelo 12, com valor de 81,1%, e o modelo 25, com valor de 68,9%. Os outros dois modelos selecionados (11 e 28) para este cenário apresentaram valores de taxa de acerto de, respectivamente, 80,5% e 79,4%, mostrando um

comportamento muito mais próximo ao do modelo com maior taxa de acerto dentro do cenário (Tabela 34). Pelo valor da taxa de acerto, o modelo 25 se mostra impróprio para realizar qualquer tipo de predição relacionado à taxa de progresso da ferrugem quando comparado aos demais, o que é confirmado quando suas demais medidas são avaliadas. Por exemplo, notou-se que seu índice Kappa é inferior a 0,4, o que indicou que este é um modelo ruim. O fato deste modelo estar presente no envelope convexo para este cenário (Figura 30) se deve ao seu alto valor de especificidade, que foi de 90,3%, enquanto que para os demais modelos este valor foi, no máximo, de 81,6%.

No cenário Varginha-alta-tx10 (seção 6.1.2) ocorreu uma situação semelhante, sendo que a diferença na taxa de acerto entre o modelo 26, com maior taxa de acerto (89,9%), e o modelo 15, com menor taxa de acerto (76,5%), foi de 13,4 p.p.. Já os demais modelos gerados para este cenário obtiveram taxas de acerto com valores variando de 88,6% a 83,5% (Tabela 18), indicando que o modelo 15 apresentou um desempenho bem abaixo dos demais. O baixo desempenho deste modelo pode ser comprovado pelo seu índice Kappa, o qual foi de 0,54 e indicou que este é um modelo apenas razoável. Todos os outros modelos para este cenário obtiveram valores de índice Kappa superiores a 0,6, sendo classificados como bons. Este modelo esteve presente no envelope convexo (Figura 25) pelo seu altíssimo valor de sensitividade (96,7%), entretanto pelos baixos valores de taxa de acerto e especificidade (67,8%), quando comparado aos demais, ele se mostrou também um modelo inadequado para representar a taxa de progresso da ferrugem.

A maior discrepância que ocorreu nos valores de taxa de acerto entre modelos selecionados no envelope convexo chegou a 23,6 p.p. e ocorreu para o cenário Varginha-Novo-baixa-tx5 (seção A.3.3). Dois modelos deste cenário (25 e 26) obtiveram as melhores taxas de acerto (87,7%), enquanto que o modelo 48 obteve o pior valor (64,1%). Os demais modelos (27 e 63) obtiveram taxas de acerto de 86,8% e 83,3%, respectivamente (Tabela 40). O índice Kappa comprovou que o modelo 48 teve um desempenho baixo, seu valor de 0,32 o classificou como um modelo ruim, enquanto que os outros modelos desenvolvidos foram classificados como bons. Mais uma vez, o alto valor de sensitividade (92,6%) foi fator responsável pelo aparecimento deste modelo no envelope convexo (Figura 35). Para este modelo também chamou a atenção o alto valor de FP rate, que foi de 44,8%, ou seja, este modelo realizou vários alarmes falsos, ou seja, previsões de aumento da taxa de progresso da

doença enquanto esta não ocorreu na realidade. Este modelo foi mais um que foi descartado para representar a taxa de progresso da ferrugem do cafeeiro.

Para os demais modelos apresentados na seção 6.1 e no apêndice A, as diferenças de taxa de acerto não obtiveram discrepâncias grandes como as apresentadas anteriormente, onde essa diferença chegou, no máximo, na casa de 7 p.p., não promovendo descartes de mais modelos dado apenas a esta medida de desempenho. O cenário de Varginha-Novo-alta-tx5 (seção A.3.1) e foi o que apresentou o menor valor desta diferença, 2,1 p.p., já o cenário Varginha-baixa-tx5 (seção 6.1.3) apresentou apenas um modelo no envelope convexo, não possibilitando tal comparação.

Para os modelos de alerta que obtiveram taxas de acerto próximas, foi necessário analisar cuidadosamente os parâmetros de sensitividade e especificidade, além das demais medidas de desempenho. Houve situações que se repetiram para diversos cenários, sendo que a mais comum foi que o modelo com maior taxa de acerto também foi o que obteve as maiores e mais próximas medidas de sensitividade e especificidade, o que ocorreu para os três cenários "Tudo" (seção A.1), para os três cenários "Varginha" (seção 6.1) e para o cenário Varginha-Novo-alta-tx5 (seção A.3.1).

Para o último cenário do parágrafo anterior, dois modelos foram selecionados no envelope convexo (Figura 33), um deles foi o modelo 28 e o outro foi o modelo 12/14/16. Suas medidas de taxa de acerto foram próximas, sendo de 83,9% e 86,0%, respectivamente. O valor de sensitividade para o modelo 28 (82,5%) foi 5,2 p.p. menor do que o do modelo 12/14/16 (87,7%), já a sua especificidade foi de 85,5%, contra 84,2% do modelo 12/14/16, resultando em uma diferença de 1,3 p.p. O modelo 12/14/16 é relativamente superior ao modelo 28 em termos de sensitividade e tem uma levíssima desvantagem em termos de especificidade, o que leva o modelo 12/14/16 a ser mais adequado a representar a taxa de progresso da ferrugem, pois obteve uma taxa de acerto e sensitividade maiores do que o modelo 28 e os valores de especificidade foram muito próximos. Este fato pode ser ainda comprovado pelas melhores medidas do modelo 12/14/16 no índice Kappa e na área sob a curva do gráfico ROC (AUC - Tabela 35).

Para os cenários "Tudo", o cenário Tudo-alta-tx5 (seção A.1.1) foi utilizado de exemplo para descrever o comportamento dos modelos deste grupo. Neste cenário, seis modelos fizeram parte do envelope convexo (Figura 27) e todos obtiveram taxas de acertos

próximas, variando de 78,0% a 83,2%. Entretanto, dois destes seis modelos receberam destaque, eles foram os modelos 10 e 11, os quais obtiveram as melhores taxas de acerto aliadas às menores diferenças entre as medidas de sensitividade e especificidade, essa diferença foi de 3 p.p. para o modelo 10 e 4,1 p.p. para o modelo 11. Para os demais modelos, essa diferença foi de, no mínimo, 9,4 p.p., chegando a 21,1 p.p. no pior caso. Os modelos 10 e 11 também foram os que apresentaram melhores valores de índice Kappa e AUC dentre os modelos desta situação, sinalizando, ainda mais, que estes são mais adequados para representar o aumento da taxa de progresso neste cenário.

Para o cenário Tudo-alta-tx10 (seção A.1.2), o modelo mais indicado para representar a taxa de progresso foi o modelo 59, já o cenário Tudo-baixa-tx5 (seção A.1.3) mostrou que dois modelos (30 e 26) foram os mais adequados.

Dentre os cenários "Varginha", o cenário Varginha-alta-tx5 (seção 6.1.1) teve o modelo 28 como destaque, ele obteve a melhor taxa de acerto dentre todos o modelos deste cenário e as medidas mais próximas de sensitividade e especificidade, com uma diferença de apenas 0,2 p.p. Além disso, ele também obteve o melhor índice Kappa e valor de AUC. O mesmo aconteceu para o cenário Varginha-alta-tx10 (seção 6.1.2), onde o destaque foi o modelo 26, o qual também obteve a maior taxa de acerto e as medidas mais próximas de sensitividade e especificidade, com uma diferença de apenas 0,1 p.p. A exemplo do cenário anterior, este modelo também obteve o melhor índice Kappa e o segundo melhor valor de AUC (com uma diferença de 0,006 para o melhor). O cenário Varginha-baixa-tx5 (seção 6.1.3) selecionou apenas um modelo e não permitiu tal análise.

Todos os modelos mencionados anteriormente tiveram as melhores taxas de acerto aliadas às melhores medidas de sensitividade e especificidade. Todavia, houve casos em que nem sempre o modelo com melhor taxa de acerto é o que apresenta as melhores, e mais próximas medidas de sensitividade e especificidade, o que foi o caso dos cenários Tudo-Novo-alta-tx10, Tudo-Novo-baixa-tx5 e Varginha-Novo-baixa-tx5.

O exemplo a ser discutido para explicar estes casos é o cenário Tudo-Novo-baixa-tx5 (seção A.2.3). Neste cenário, três modelos foram selecionados no envelope convexo (Figura 32): 27, 57 e 59. O modelo 57 foi o que apresentou a pior medida de taxa de acerto (81,7%), uma grande diferença entre sensitividade e especificidade (14,9 p.p.) e um índice Kappa inferior aos demais, sendo considerado um modelo regular, enquanto que os outros dois foram

considerados bons. A maior discussão acontece entre as medidas de desempenho do modelo 59 e do modelo 27. O primeiro obteve melhores valores de taxa de acerto (88,7% contra 87,0%) e índice Kappa (0,70 contra 0,67), já o segundo obteve medidas mais equilibradas de sensitividade e especificidade (diferença de 1,2 p.p. contra 6,4 p.p) e maior valor de AUC (0,914 contra 0,869), tornando a seleção do melhor modelo complicada. Há uma leve tendência que indica o modelo 27 como o mais adequado, pois não perde tanto em taxa de acerto e aproximou muito as medidas de sensitividade e especificidade, indicando que este modelo trabalha bem em classificar tanto exemplos positivos quanto negativos corretamente. Além do mais, apesar da diferença no índice Kappa, ambos os modelos se enquadraram como bons modelos. Uma análise mais cautelosa deve ser feita em relação à complexidade dos atributos no conjunto de dados de cada um destes modelos (seção 6.2.2), mas as medidas de desempenho mostraram uma leve tendência que o modelo 27 seria o mais indicado neste cenário.

O mesmo procedimento de análise se aplica ao cenário Tudo-Novo-alta-tx10 (seção A.2.2), sendo que o modelo 56 já é descartado inicialmente, deixando os modelos 25, 30, 57 e 59 equiparados. Torna-se difícil uma escolha dentre os três, mas a principio o que demonstra ter uma leve tendência de ser o mais indicado para este cenário é o modelo 59, pela baixa diferença entre sensitividade e especificidade quando comparado aos demais. O cenário Varginha-Novo-baixa-tx5 (seção A.3.3) também teve modelos descartados inicialmente (modelos 48 e 63, com taxa de acerto inferior aos demais), deixando outros três modelos remanescentes, os modelos 25, 26 e 27. A tendência é que o mais indicado neste caso seja o 27, novamente pela baixa diferença de sensitividade e especificidade quando comparado aos demais.

Finalmente, chega-se a o caso mais difícil de análise quanto às medidas de desempenho. Por enquanto foram vistos casos em que a melhor taxa de acerto obteve melhores valores de sensitividade e especificidade, casos em que nem sempre a melhor taxa de acerto resultou nos melhores valores de sensitividade e especificidade e agora o caso em que taxas de acerto similares têm medidas de sensitividade e especificidade muito próximas, sem grandes divergências.

Um destes casos ocorreu para os modelos do cenário Tudo-Novo-alta-tx5 (seção A.2.1). Quatro modelos (11, 12, 25 e 28) foram selecionados no envelope convexo para este

caso, e um deles (25) já foi apontado como impróprio pelos baixos valores de taxa de acerto em relação aos demais. Outro modelo (28) também apresenta deficiências, como menor taxa de acerto, maior diferença entre sensitividade e especificidade e classificação inferior quanto ao índice Kappa quando comparado aos modelos 11 e 12. Estes dois modelos obtiveram medidas de desempenho muito próximas (Tabela 30), praticamente com valores iguais de taxa de acerto, diferença entre sensitividade e especificidade e índice Kappa. O que chama atenção é que o modelo 11 obteve valor de especificidade mais alto do que o modelo 12, que por sua vez obteve valor mais alto de sensitividade. Neste caso, é necessária uma análise do conjunto de dados utilizado na indução de cada um destes modelos (seção 6.2.2).

Outro cenário que apresentou uma situação similar foi o Varginha-Novo-alta-tx10 (seção A.3.2), onde, novamente, quatro modelos (4/6/8, 28, 54 e 59) foram selecionados pelo envelope convexo. Pela taxa de acerto inferior e, principalmente, a grande diferença entre valores de sensitividade e especificidade, os modelos 54 e 59 não são recomendados para predizer a taxa de progresso da ferrugem. Dentre os dois modelos restantes, 4/6/8 e 28, nota-se um comportamento inverso quanto à classificação de exemplos positivos e negativos. O modelo 28 apresentou valores de sensitividade e especificidade de 91,9% e 84,4%, respectivamente, mostrando ser um modelo que classifica corretamente uma maior porcentagem de exemplos positivos, ou seja, de aumento da taxa de progresso da ferrugem. Já o modelo 4/6/8 obteve valores de 81,1% e 94,8%, respectivamente, para sensitividade e especificidade, mostrando-se um modelo que classifica corretamente uma maior quantidade de exemplos negativos, ou seja, de não aumento da taxa de progresso. O modelo 28 ainda mostrou ser o mais equilibrado, com uma diferença de 7,5 p.p. entre sensitividade e especificidade, contra 13,7 p.p. do modelo 4/6/8, que por sua vez obteve uma taxa de acerto de 90,4% (a segunda maior dentre todos os modelos gerados) contra 86,8% do modelo 28. Novamente, a escolha de um ou de outro, para predizer a taxa de progresso da ferrugem, requer uma análise mais cautelosa do conjunto de dados utilizado para gerá-los (seção 6.2.2).

6.2.2 Atributos do conjunto de dados

Esta seção visa uma discussão cautelosa sobre o conjunto de dados utilizado nas induções. São discutidos os atributos presentes nos conjuntos de dados, a complexidade na sua obtenção e cálculo, além da sua relação com a ferrugem do cafeeiro. Também foi feita uma análise sobre os atributos selecionados por meio de métodos de seleção de atributos objetivos e como estes conjuntos ficaram.

Para dois modelos de alerta com medidas de avaliação similares, o conjunto de dados foi o responsável por determinar qual foi o indicado para predizer a taxa de progresso da ferrugem do cafeeiro. Modelos com uma maior quantidade de atributos no seu conjunto de dados, além de atributos difíceis de serem calculados, têm sua aplicabilidade reduzida, justamente pela dificuldade de obtenção dos dados para gerá-los. Em contrapartida, modelos com poucos atributos podem ser considerados impróprios por especialistas da área de doenças, pelo fato destes não apresentarem atributos relevantes às mais diversas condições de desenvolvimento da doença.

Para os 12 cenários da Tabela 9, um total de 43 modelos de alerta foram selecionados nos envelopes convexos. Destes, 21 foram gerados por meio de conjuntos de dados com seleção de atributos subjetiva, sendo que o conjunto mais simples (M3) foi o que mais apareceu (10 casos). Já os 22 modelos restantes foram gerados por métodos de seleção de atributos objetivos, onde o método que mais se repetiu foi o Wrapper (10 casos). Um ponto interessante foi que para 4 casos o mesmo conjunto foi selecionado pelo método do Chi2 e InfoGain, mostrando uma similaridade na escolha de atributos por estes dois métodos.

Em seguida, é feita uma discussão sobre os atributos utilizados nos conjuntos de dados para realizar a indução e como eles se relacionaram com as principais características da ferrugem do cafeeiro.

Atributos de temperatura do ar:

No conjunto de dados inicial havia três atributos de temperatura, sendo a média da temperatura média, mínima e máxima (Tabela 6), os quais foram amplamente utilizados pelos modelos selecionados no envelope convexo. Dos 43 modelos selecionados no envelope

convexo, 36 (84%) utilizaram, ao menos, um destes atributos de temperatura. Estes atributos são importantes para os modelos de alerta à medida que representam condições que inibem o desenvolvimento do fungo, como temperaturas muito altas ou muito baixas, por exemplo. Também são atributos fáceis e simples de serem calculados e obtidos.

O atributo de temperatura mínima (TMIN_PINF) foi o que mais apareceu nos modelos selecionados no envelope convexo, em 33 destes (77%). A temperatura mínima para a cidade de Varginha no período de 1998 até 2011 teve valor médio de 15,1^{o}C, mas cerca de 35% de seus registros foram inferiores a 14^{o} C, valor considerado limitante para a infecção de *H. vastatrix* (KUSHALAPPA et al., 1983). A importância deste atributo deve ser ressaltada pois ele apareceu em 12 (55%) dos 22 modelos gerados com métodos de seleção de atributos objetivos, o que forneceu subsídios matemáticos para sua presença no conjunto de dados.

O atributo de temperatura máxima (TMAX_PINF) foi o que menos apareceu nos modelos selecionados no envelope convexo, em 24 destes (54%). A temperatura máxima para a cidade de Varginha neste mesmo período de 22,1 ^{o}C a 30,5^{o}C, praticamente não superando temperaturas acima de 30ºC, as quais também são limitantes ao desenvolvimento do fungo (KUSHALAPPA et al., 1983). Este atributo se mostrou menos importante que o de temperatura mínima neste trabalho, o que pode ser comprovado pela sua baixa frequência nos modelos provenientes de métodos de seleção de atributos objetivos, apenas 3 (13%) de 23 modelos.

O atributo de temperatura média (TMED_PINF) também é importante, afinal ele determina se o ambiente irá proporcionar condições ótimas de desenvolvimento do fungo, com temperaturas entre 22 ^{o}C e 24^{o}C (ZAMBOLIM et al., 2002), entretanto a presença destas condições serão mais detalhadas no atributo de temperatura durante o molhamento foliar.

Atributos relacionados à precipitação:

O conjunto inicial de dados conteve 4 atributos relacionados à precipitação: PRECIP_PINF, MED_INDPLUVMAX_PINF, DCHUV_PINF e MED_PRECIP_PINF (Tabela 6). Estes atributos são responsáveis por dois fatores importantes relacionados à doença, sendo um deles relativo à disseminação dos esporos e o outro à ocorrência de períodos úmidos. Houve uma ampla presença destes atributos nos modelos gerados, sendo que em 33

(77%) dos 43 modelos, ao menos um destes atributos esteve presente. Para os 22 modelos provenientes de seleção de atributos objetiva, esta estatística foi de 50%.

Dentre estes quatro atributos, o MED_INDPLUVMAX_PINF foi poucas vezes selecionado pelos métodos de seleção de atributos objetivos, aparecendo apenas em 3 (13%) dos 22 modelos gerados por tais métodos. Este atributo mede o aumento de intensidade em uma chuva, sendo mais complexo de ser calculado do que apenas uma medida simples de pluviosidade.

A relação mais interessante dos atributos relacionados à precipitação envolve os três atributos restantes, com eles é possível saber quantos dias choveram no mês, a média de chuva diária no mês e a quantidade total mensal de chuva. Em um mês em que ocorram diversos períodos de chuva, eles podem ser de chuvas leves e constantes ou de chuvas intensas e rápidas.

No primeiro caso, as chuvas leves e constantes aumentam a umidade relativa do ar, deixando água livre na superfície das folhas por mais tempo e levando a períodos de molhamento foliar prolongados, favorecendo a ocorrência das condições mínimas de molhamento foliar (6 horas contínuas) para o desenvolvimento do fungo (KUSHALAPPA et al., 1983). Além disso, chuvas leves e seus respingos são apontadas como um dos principais fatores de disseminação do fungo na lavoura e dentro da própria planta (ZAMBOLIM et al., 2002). Para o segundo caso, chuvas intensas podem conduzir a maior parte dos esporos para o chão (KUSHALAPPA e ESKES, 1989), além de não promoverem longas horas de molhamento foliar, desfavorecendo o desenvolvimento do fungo.

Como um exemplo, pode-se analisar alguns meses do conjunto de dados para mostrar a importância da interação destes atributos. Iniciando pelo mês de fevereiro de 2001, o qual obteve uma precipitação acumulada de 285mm com 16 dias chuvosos, ou seja, cada dia chuvoso obteve uma média de 17,8mm de chuva, o que tende a chuvas de grande intensidade quando comparadas ao mês de março de 2003, que registrou uma precipitação acumulada de 161mm com 22 dias chuvosos, ou seja, em cada dia chuvoso a média foi de apenas 7,3 mm, o que se inclina a chuvas de baixa intensidade em um período chuvoso maior. Essa discussão será complementada com a análise do atributo de umidade relativa.

Atributo de Umidade Relativa (UR):

O atributo UR_PINF representa a umidade relativa do ar média durante o período de infecção, além de que a prórpia UR é a base para se gerar os atributos de medição indireta de duração de molhamento foliar (número de horas com umidade relativa acima de 90%). Este atributo é importante para o desenvolvimento da doença, pois quanto maior a umidade relativa média, maior a tendência de períodos de molhamento foliar, o que favorece o desenvolvimento do fungo.

Este atributo apareceu consideravelmente nos modelos gerados, presente em 27 (63%) dos 43 modelos selecionados no envelope convexo. Entretanto, quando considerados modelos gerados por seleção de atributos objetiva, este atributo apareceu em apenas 6 modelos (27%), o que indicou que ele foi filtrado poucas vezes pelos métodos de seleção de atributos. Este baixo valor não é preocupante, uma vez que os atributos de precipitação mencionados anteriormente e os atributos de duração de molhamento foliar também expressam condições relacionadas à umidade relativa. É importante relembrar que o cálculo e obtenção do atributo UR_PINF é mais simples do que os que relacionam o número de horas com umidade relativa acima de 90%.

Seguindo o exemplo dos atributos de precipitação, o mês de fevereiro de 2001 obteve uma umidade relativa média de 72%, enquanto que o mês de março de 2003 obteve 83%. Tais valores evidenciam ainda mais que o mês de fevereiro de 2001 foi propício a menos períodos favoráveis ao desenvolvimento do fungo do que o mês de março de 2003. Para as lavouras largas e carga alta esta diferença ficou evidente, sendo que o mês de fevereiro de 2001 apresentou uma taxa de progresso inferior a 5 p.p., enquanto que o mês de março de 2003 apresentou uma taxa de progresso superior a 5 p.p.

Atributos de duração de molhamento foliar:

No conjunto de dados completo, existiam quatro atributos relacionados com a duração do molhamento foliar, dois deles com relação à média de horas com molhamento foliar em todo um dia epidemiológico ou só no período noturno (NHUR90_PINF e NHNUR90_PINF,

respectivamente) e outros dois relacionados à somatória destes períodos durante o PINF (SMT_NHUR90_PINF e SMT_NHNUR90_PINF, respectivamente).

Estes atributos são de extrema importância para o desenvolvimento da ferrugem do cafeeiro, pois por meio deles é possível saber se houve ou não molhamento foliar mínimo necessário para que o fungo se desenvolva. Um período mínimo de 6 horas de molhamento foliar é necessário para que ocorra a germinação do fungo (KUSHALAPPA et al., 1983). Um exemplo de como este atributo está relacionado a períodos ótimos será mostrado em uma relação com o atributo de temperatura durante o período de molhamento foliar, na próxima seção.

Os atributos relativos aos somatórios da horas com umidade relativa acima de 90% foram pouco utilizados nos modelos, apenas 9 (21%) dos 43 modelos utilizaram, ao menos, um destes atributos. Com relação aos modelos gerados por seleção de atributos objetiva, sua seleção apareceu em apenas 2 modelos (9%) dos 22 selecionados. Estes atributos são complexos der ser calculados e não se mostraram importantes nos modelos de predição da taxa de progresso da ferrugem.

Os outros dois atributos de medição indireta de molhamento foliar (NHUR90_PINF e NHNUR90_PINF) foram mais representativos do que os anteriores, eles apareceram em 19 dos 43 modelos, cerca de 44%. Também foram selecionados 8 vezes pelos 22 modelos provenientes de seleção de atributos objetiva (36%).

Atributos de temperatura durante o molhamento foliar:

O atributo de medida da temperatura durante o período de molhamento foliar (THUR90_PINF) esteve presente em 20 (46%) dos 43 modelos selecionados nos envelopes convexos, e em 9 (41%) dos 22 modelos gerados por seleção de atributos objetiva, mostrando que este foi representativo para os modelos de predição da ferrugem do cafeeiro. Outro ponto interessante é que, dos 20 modelos que apresentaram este atributo, 16 também apresentaram, ao menos, um atributo de medição indireta de molhamento foliar, ressaltando ainda mais a importância destes dois tipos de atributos estarem presentes em modelos de alerta para predição da taxa de progresso da ferrugem. A complexidade deste atributo se equipara aos atributos de medida indireta de molhamento foliar.

O atributo de medida da temperatura durante o período de molhamento foliar reúne, juntamente com os atributos de duração de molhamento foliar, condições que podem ser as mais propícias para o desenvolvimento do *H. vastatrix*. No caso de existir um período com, no mínimo, 6 horas de molhamento foliar contínuo e a temperatura média neste período estiver entre 22 ºC a 24ºC (ZAMBOLIM et al., 2002), a condição ótima é atingida e o fungo tem as condições propícias para germinar e infectar a planta. Todavia, essa temperatura também pode ser responsável por inibir o desenvolvimento do fungo, mesmo com a presença de molhamento foliar, caso as temperaturas sejam muito altas ou muito baixas, como explicado para os atributos de temperatura.

Um exemplo de como a combinação dos atributos de medida indireta de molhamento foliar e o atributo de temperatura durante esse período é importante para o desenvolvimento da doença é o mês de janeiro de 2003, o qual obteve um período médio diário de quase 15 horas de molhamento foliar contínuo com temperatura média próxima aos 20 ºC nestas horas se aproximando bastante das condições muito favoráveis ao desenvolvimento da doença. Já o mês de agosto de 2011 também obteve um período médio diário de molhamento foliar elevado, quase 11 horas, entretanto a temperatura média deste período foi inferior a 11 ºC, fator que inibiu o desenvolvimento do *H. vastatrix*. Para lavouras adensadas e carga alta, o mês de janeiro de 2003 obteve uma taxa de progresso superior a 5 p.p., enquanto que para agosto de 2011 ela foi inferior a 5 p.p.

Atributos especiais:

Os atributos especiais (Tabela 8) são muito representativos com relação à doença, por reunirem condições propícias ao seu desenvolvimento (luminosidade, temperatura e duração do período de molhamento foliar - Tabela 7). Apesar de representativos, estes atributos são difíceis de serem calculados, sendo os mais complexos do conjunto de dados, o que pode dificultar a aplicabilidade de modelos que contenham estes atributos, em casos que não haja uma disponibilidade suficiente de dados coletados para gerá-los.

Estes atributos foram utilizados por diversos modelos selecionados no envelope convexo, ao menos um destes atributos esteve presente em 23 (53%) dos 43 modelos. Já para os modelos provenientes de seleção de atributos objetiva essa porcentagem foi muito superior,

dos 22 modelos provenientes destes métodos, 18 (82%) contiveram, ao menos, um atributo especial. Estes atributos foram os mais indicados para fazer parte dos modelos de predição da taxa de progresso da ferrugem pelos métodos de seleção de atributos objetivos. Caso haja disponibilidade dos dados necessários para gerar estes atributos, eles devem ser utilizados nos modelos de alerta

Atributo de espaçamento:

O atributo de espaçamento (LAVOURA) foi considerado em vários trabalhos como fator importante para o desenvolvimento da doença. Como mencionado na seção 3.3.3, ele seria responsável por afetar o microclima dentro da lavoura (VALE et al., 2000), além de que a ferrugem tende a ser mais agressiva em lavouras adensadas do que largas (JAPIASSÚ et al., 2007).

Este atributo esteve presente em todos os modelos gerados com métodos de seleção de atributos subjetivos. Entretanto, apesar da influência que este atributo pode ter no desenvolvimento da ferrugem do cafeeiro, ele não foi muito utilizado nos modelos presentes no envelope convexo que realizaram a seleção de atributos objetiva, apenas 3 (13%) dos 22 modelos selecionaram este atributo.

Este atributo é muito simples de ser determinado, basta saber a densidade de plantio, entretanto, foi pouco representativo nos modelos de alerta.

Atributos relacionados à temperatura durante o período de incubação:

Três atributos com relação à temperatura no período de incubação (TMAX_PI_PINF, TMIN_PI_PINF e TMED_PI_PINF e) estavam no conjunto inicial de dados e, a exemplo dos atributos de temperatura, eles retrataram os valores médios para as temperaturas máximas, mínimas e médias deste período. Estes valores influenciam o desenvolvimento da ferrugem na medida em que são os responsáveis por determinar a duração do período de incubação, lembrando que em temperaturas mais amenas o PI é maior, enquanto que para temperaturas altas o PI é menor. Isto representa um aumento mais rápido da ferrugem em meses mais quentes do que em meses mais frios.

A exemplo do atributo de espaçamento, estes atributos estavam presentes em todos os modelos gerados com seleção de atributos subjetiva. Entretanto, sua importância não foi ressaltada pelos métodos de seleção de atributos subjetiva, afinal eles só foram utilizados em 3 (13%) dos 22 modelos gerados por estes métodos. Estes atributos são parcialmente complexos para serem calculados, pois necessitam de uma estimativa do PI, o que torna, junto com sua baixa utilização, mais um fator para que estes atributos não sejam recomendados nos modelos de alerta.

Atributos relacionados ao vento:

O conjunto de dados contou com dois atributos relacionados à velocidade do vento (VVENTO_PINF e SMT_VVENTO_PINF), que é um fator, junto com a chuva, considerado importante para a disseminação dos esporos dentro da lavoura (ZAMBOLIM et al., 2002). A presença de atributos relacionados à precipitação e velocidade do vento no mesmo modelo poderia indicar que este estaria levando em conta fatores relacionados à disseminação dos esporos do *H. vastatrix.*

A presença dos atributos relacionados ao vento nos modelos de alerta provenientes de conjuntos oriundos de métodos de seleção objetivos foi muito pequena, apenas 4 (18%) dos 22 modelos contaram com estes atributos. Em contrapartida, estes quatro modelos contaram também com atributos relacionados à precipitação, cobrindo a disseminação dos esporos do fungo pelos modelos de alerta.

Os atributos de velocidade do vento requerem um anemômetro para serem medidos, com este aparelho eles são facilmente calculados. Mesmo com a baixa representatividade nos modelos, seria recomendada a inclusão destes atributos para que os efeitos da disseminação do fungo possam ser completamente cobertos.

Considerações da seção:

Todos os atributos tiveram aspectos relacionados à ferrugem do cafeeiro, entretanto deve-se pensar na complexidade de obtenção destes atributos e qual foi sua representatividade nos modelos selecionados. Quanto mais complexo, mais difícil torna-se a utilização de um modelo que contenha tais atributos, e quanto maior foi sua representatividade (mais usados nos modelos de alerta), maior foi a importância para métodos de seleção de atributos objetivos e, consequentemente, para os modelos de alerta.

A representatividade de um atributo foi baseada na quantidade de vezes em que este foi selecionado, considerando apenas os modelos induzidos por meio de métodos de seleção de atributos objetivos e selecionados nos envelopes convexos (22 modelos), não sobre todos os modelos selecionados nos envelopes (incluindo os métodos subjetivos – 43 modelos). Isto foi feito pelo fato dos métodos de seleção objetivos apresentarem uma forte comprovação matemática do porque um atributo foi selecionado, seja por sua correlação com a classe, mérito, ganho de informação ou taxa de ganho de informação.

Uma tabela foi feita resumindo o nível de complexidade e quão frequente um atributo foi selecionado por métodos de seleção objetivos. A frequência foi baseada na quantidade de seleções, sendo que este valor seria alto se estivesse presente na metade dos 22 modelos gerados por métodos de seleção de atributos objetivos, se estivesse presente em 10 a 5 destes modelos (acima de 20%) ela seria média, e para valores inferiores a isso seria baixa (Tabela 21). A complexidade foi subjetiva, de acordo com o mencionado na análise individual dos atributos feita anteriormente. Toda esta análise, juntamente com as medidas de avaliação (seção 6.2.1), foi utilizada para determinar qual o modelo indicado para representar a taxa de progresso da ferrugem do cafeeiro em cada um dos cenários de indução.

Tabela 21: Índices de complexidade e representatividade dos atributos do conjunto de dados.

	Avaliação	
Atributos	Complexidade	Representatividade
ACDINF_PINF	Alta	Alta
DMFI_PINF	Alta	Alta
DFMFI_PINF	Alta	Alta
DDI_PINF	Alta	Alta
NHUR90_PINF	Média	Média
NHNUR90_PINF	Média	Média
THUR90_PINF	Média	Média
SMT_NHUR90_PINF	Média	Baixa
SMT_NHNUR90_PINF	Média	Baixa
MED_INDPLUVMAX_PINF	Média	Baixa
TMAX_PI_PINF	Média	Baixa
TMIN_PI_PINF	Média	Baixa
TMED_PI_PINF	Média	Baixa
TMIN_PINF	Baixa	Alta
TMED_PINF	Baixa	Média
UR_PINF	Baixa	Média
MED_PRECIP_PINF	Baixa	Média
LAVOURA	Baixa	Baixa
TMAX_PINF	Baixa	Baixa
PRECIP_PINF	Baixa	Baixa
DCHUV_PINF	Baixa	Baixa
VVENTO_PINF	Baixa	Baixa
SMT_VVENTO_PINF	Baixa	Baixa

6.2.3 Escolha do melhor modelo em cada cenário

Esta seção trata da escolha do melhor modelo para cada um dos 12 cenários de indução. A seleção foi baseada na avaliação das medidas de desempenho (seção 6.2.1) e dos conjuntos de dados utilizados para gerar cada modelo (seção 6.2.2). Um modelo poderia ser descartado por apresentar uma taxa de acerto muito inferior aos demais, uma alta discrepância entre valores de sensitividade e especificidade ou um conjunto de dados muito complexo.

Sempre que possível, além de um modelo ter sido escolhido, um outro foi recomendado para casos específicos, como, por exemplo, onde haja disponibilidade de dados para usar este modelo.

Cenário Tudo-alta-tx5:

Este cenário apresentou seis modelos (4, 10, 11, 23, 27 e 30/32) no envelope convexo de acordo com a Figura 27. Em quatro destes modelos foi constatada uma grande discrepância na diferença de sensitividade e especificidade, fazendo com que os modelos 4, 23, 27, e 30/32 fossem descartados. Os dois modelos restantes (10 e 11) apresentaram medidas de avaliação praticamente iguais, entretanto, o modelo 11 foi gerado pela seleção de atributos M3, considerada mais simples do que a seleção de atributos M2, utilizada pelo modelo 10. Assim, o modelo 11 foi o recomendado para ser utilizado neste cenário, sendo que o modelo 10 também poderia ser usado caso não ocorra uma indisponibilidade de dados utilizados em seu conjunto de treinamento.

Cenário Tudo-alta-tx10:

Para este cenário apenas, dois modelos (28 e 59) foram selecionados (Figura 28). O modelo 28 apresentou menor medida de taxa de acerto e maior discrepância nas medidas de sensitividade e especificidade, além de que seu conjunto de dados ter sido mais complexo, utilizando atributos especiais, enquanto que o modelo 59 utilizou uma seleção de dados mais simples, tornando-o o mais indicado para este cenário.

Cenário Tudo-baixa-tx5:

Este cenário mostrou que três (26, 30 e 57) modelos foram selecionados para o envelope convexo (Figura 29). O modelo 57 foi o único a ser descartado, pela grande diferença entre valores de sensitividade e especificidade quando comparado aos outros dois.

Os modelos 26 e 30 apresentaram medidas de desempenho muito próximas, sendo que o segundo obteve uma taxa de acerto levemente superior e uma diferença ligeiramente menor nos parâmetros de sensitividade e especificidade do que o primeiro. Entretanto, o conjunto de dados do modelo 30 foi bem mais complexo, utilizando mais atributos e todos os atributos especiais, enquanto que o modelo 26 apresentou um conjunto mais simples, composto pela seleção de atributos M2. Assim, mesmo com medidas de avaliação levemente inferiores, o modelo 26 seria o mais adequado para este cenário. No caso de uma disponibilidade ampla de dados, o modelo 30 poderia ser usado, o que tenderia a trazer mais predições corretas, dado o melhor desempenho deste modelo com relação ao 26.

Cenário Tudo-Novo-alta-tx5:

Este cenário obteve quatro modelos selecionados no envelope convexo (11, 12, 25 e 28 - Figura 30). Dois modelos foram inicialmente descartados, sendo que o modelo 25 foi descartado pois obteve taxa de acerto muito inferior aos demais, enquanto que o modelo 28 foi descartado devido à alta diferença de sensitividade e especificidade em relação aos modelos 11 e 12.

Estes dois modelos apresentaram medidas de avaliação muito próximas, sendo que para o modelo 12 elas foram sutilmente superiores. Quando o conjunto de dados destes dois modelos foram analisados, notou-se que o modelo 12 foi gerado a partir de apenas 2 atributos, sendo um destes especial e outro de temperatura mínima, enquanto que o modelo 11 foi gerado pela seleção de atributos M3. Mesmo com apenas dois atributos, o conjunto utilizado para gerar o modelo 12 é mais complexo do que o utilizado para gerar o modelo 11. Mesmo com as medidas de desempenho levemente inferiores, o modelo recomendado para este cenário é o modelo 11.

Cenário Tudo-Novo-alta-tx10:

Cinco modelos (25,30,56,57 e 59) ficaram presentes no envelope convexo para este cenário (Figura 31). Dois destes modelos (56 e 57) foram descartados pela alta diferença entre as medidas de sensitividade e especificidade com relação aos demais. Já o modelo 25 utilizou todo o conjunto de dados para ser gerado, tornando-se muito mais complexo do que os outros dois modelos (59 e 30), que não utilizaram atributos especiais (mais complexos).

Estes dois modelos apresentaram algumas vantagens e desvantagens quando comparados entre si. A taxa de acerto do modelo 30 foi maior, a medida que o modelo 59 apresentou uma menor diferença de sensitividade e especificidade. Os dois modelos utilizaram atributos de média complexidade no conjunto de dados, o modelo 30 com atributos como NHUR90_PINF E THUR90_PINF e o modelo 59 com os atributos de temperatura durante o período de incubação do fungo.

Qualquer um destes modelos poderia ser utilizado para representar a taxa de progresso da ferrugem do cafeeiro, entretanto o modelo 30 reúne condições mais importantes para o desenvolvimento da ferrugem do que o modelo 59. Assim a recomendação feita neste trabalho é o modelo 30, mas deixando como opção o modelo 59.

Cenário Tudo-Novo-baixa-tx5:

Três modelos (27, 57 e 59) foram selecionados no envelope convexo para este cenário (Figura 32), sendo que apenas um destes foi descartado. Este foi o modelo 57, pois apresentou uma alta diferença entre os valores de sensitividade e especificidade quando comparado aos demais, além de que foi gerado pelo conjunto M1, com atributos mais complexos.

Os modelos 27 e 59 foram gerados pelo mesmo conjunto de dados (M3) e o modelo 27 apresentou maior valor de taxa de acerto, enquanto que o modelo 59 apresentou maior equilíbrio nas medidas de sensitividade e especificidade. Pelas baixas diferenças entre as medidas, qualquer um destes modelos poderia ser utilizado para representar a taxa de progresso da ferrugem.

Cenário Varginha-alta-tx5:

Para este cenário, dois modelos (22 e 28) estiveram presentes no envelope convexo (Figura 24). Destes, o modelo (22) apresentou diferenças entre medidas de sensitividade e especificidade elevadas quando comparados ao modelo 28, além do que a taxa de acerto deste último foi maior do que a dos demais. Para este caso, não houve necessidade de uma análise no conjunto de dados para determinar qual modelo seria mais adequado, sendo este o modelo 28.

Cenário Varginha-alta-tx10:

Para este cenário, seis modelos (15, 26, 52, 59, 61, 62/64) foram selecionados no envelope convexo (Figura 25) e, ao serem analisados, um destes (15) já foi descartado pelo seu baixo valor de taxa de acerto. Quanto aos 5 modelos restantes, apenas um se destacou, o modelo 26, com melhor taxa de acerto e valores de sensitividade e especificidade praticamente iguais, enquanto os demais mostraram diferenças superiores a 6 p.p.

O modelo 26 foi gerado pela seleção de atributos M2, ou seja, um conjunto sem a presença de atributos complexos, fazendo com que este modelo fosse o recomendado para este cenário. Um outro modelo (59) foi gerado por um conjunto de dados mais simples (M3) e poderia ser utilizado no caso de dados insuficientes para gerar o modelo 26, lembrando que, neste caso, haveria uma perda de desempenho pois as medidas de avaliação do modelo 59 foram um pouco inferiores às do modelo 26.

Cenário Varginha-baixa-tx5:

Este cenário selecionou apenas um modelo no envelope convexo, o modelo 25/32 (Figura 26), não possibilitando a comparação com outros. De qualquer forma, este modelo obteve um alto valor de taxa de acerto, além de valores próximos de sensitividade e especificidade. O ponto desfavorável deste modelo é que este foi gerado pela seleção de atributos M1, a qual inclui atributos especiais (mais complexos).

Cenário Varginha-Novo-alta-tx5:

Dois modelos (12/14/16 e 28) foram selecionados no envelope convexo para este cenário (Figura 33). Em termos de medidas de desempenho, taxa de acerto, sensitividade e especificidade, o que se destacou foi o modelo 12/14/16. Entretanto, ao avaliar seu conjunto de dados, nota-se que apenas um atributo (TMIN_PINF) foi utilizado para gerá-lo, que embora tenha sido de alta representatividade, reúne poucas condições de desenvolvimento da doença. Já o modelo 28 foi gerado por 4 atributos, incluindo um atributo especial, o que tornou o conjunto de dados mais complexo. A escolha de um modelo para representar este cenário recai sobre as medidas de desempenho, uma vez que um conjunto de dados é pouco representativo e o outro conteve um atributo especial, assim o modelo escolhido para representar esta situação é o modelo 12/14/16.

Cenário Varginha-Novo-alta-tx10:

Este cenário apresentou quatro modelos (4/6/8, 28, 54 e 59) no envelope convexo (Figura 34). Dois destes modelos (54 e 59), além de apresentarem menor taxa de acerto, também apresentaram maior diferença entre valores de sensitividade e especificidade que os outros dois (4/6/8 e 28), dentre os quais o que apresentou melhores valores destas medidas foi o modelo 4/6/8.

Uma comparação nos conjuntos de dados revela, que mesmo com as melhores medidas de desempenho, o modelo 4/6/8 não é o mais indicado para este cenário. Seu conjunto de dados conteve apenas 2 atributos e estes foram atributos especiais, enquanto que o modelo 28 contou com 4 atributos e nenhum especial. Assim, o modelo 28, mesmo com medidas de avaliação inferiores, seria o indicado para este cenário. Entretanto, o modelo 4/6/8 poderia ser utilizado no caso de disponibilidade de dados para gerar os atributos complexos e, caso este modelo fosse utilizado, ele teria melhor desempenho que o 28, justamente pelas melhores medidas de avaliação.

Cenário Varginha-Novo-baixa-tx5:

Este cenário apresentou cinco (25, 26, 27, 48 e 63) modelos no seu envelope convexo, como mostra a Figura 35. O modelo 48 foi o primeiro a ser descartado, por ser o modelo com a menor taxa de acerto com relação aos demais, além de ter sido o modelo que apresentou menor valor desta medida dentre todos os modelos selecionados nos diversos cenários.

Outros três modelos (25, 26 e 63) apresentaram uma diferença alta entre sensitividade e especificidade em relação ao modelo 27. Com relação à taxa de acerto, o modelo 63 obteve valor inferior, enquanto que os modelos 25 e 26 obtiveram valores superiores, mas esta superioridade não avançou a marca de 1 p.p.

O modelo 27 foi gerado pelo conjunto de dados mais simples (M3), enquanto que os modelos 25 e 63 contaram com atributos especiais em seus conjuntos de dados e o modelo 26 foi gerado pela seleção M2. Assim o modelo mais indicado para este cenário foi o modelo 27. O modelo 25 poderia ser utilizado no caso de uma disponibilidade ampla de dados, pois obteve taxa de acerto levemente superior ao modelo 27, além de uma diferença de sensitividade e especificidade menor que o modelo 26, que também obteve taxa de acerto melhor que o 27.

6.2.4 Comparação com outros modelos desenvolvidos

Na seção anterior, os modelos foram selecionados acordo com cada cenário da Tabela 9, já nesta seção, foi escolhido apenas um modelo para as combinações de carga e atributo meta (apenas para os modelos criados por dados da cidade de Varginha), o que facilitará a comparação destes com modelos presentes em trabalhos citados na revisão bibliográfica (seção 3.4). Exemplificando: dos cenários que contavam com apenas dados da cidade de Varginha, por exemplo, Varginha-alta-tx5 e Varginha-Novo-alta-tx5, será escolhido apenas um modelo para a combinação de carga alta e atributo meta 5 p.p..

A seleção dos modelos ocorreu de forma idêntica ao realizado na seção 6.2.3, pelos critérios de medidas de desempenho (seção 6.2.1) e atributos do conjunto de dados (seção 6.2.2), mas a comparação entre modelos ocorreu entre dois cenários, com os modelos indicados na seção na seção 6.2.3. O resultado foi que os modelos gerados com dados de 1998

a 2011 obtiveram desempenho melhor do que os gerados com dados de 2007 a 2011. A Tabela 22 apresenta o resultado da escolha dos modelos, bem como alguns dos modelos gerados por Meira (2008) e Cintra et al. (2011), para dados da mesma localidade (Varginha) e mesma combinação de carga e atributo meta.

Como exemplo para a combinação de carga alta e atributo meta 10 p.p., tomou-se o modelo 26 do cenário Varginha-alta-tx10 e o modelo 28 do cenário Varginha-Novo-alta-tx10. O modelo 26 teve uma taxa de acerto de 89,9%, contra apenas 86,8% do modelo 28. A diferença entre valores de sensitividade e especificidade do modelo 26 foi de apenas 0,1 p.p., enquanto que para o modelo 28 ela foi de 7,5 p.p. Com relação ao conjunto de dados, o modelo 26 contou com os atributos da seleção M2, e foi um pouco mais complexo do que o conjunto de dados do modelo 28, o qual contou com quatro atributos filtrados pelo método de seleção Wrapper. Mesmo o conjunto do modelo 26 contendo mais atributos, ele não foi muito mais complexo do que o conjunto do modelo 28, pois ambos não utilizaram atributos especiais (mais complexos). Assim, o modelo 26 foi o mais indicado para representar a situação de carga alta e atributo meta 10 p.p.

Após um modelo ter sido escolhido para representar cada uma das combinações de carga e atributo meta, eles foram comparados com modelos de combinações similares, desenvolvidos por autores diferentes.

Para a combinação de carga alta e atributo meta 5 p.p., o modelo gerado neste trabalho foi comparado com o modelo gerado com os dados de Meira (2008) e com o modelo de Cintra et al. (2011), conforme os dados da Tabela 22. Verificou-se que a taxa de acerto do modelo Varginha-alta-tx5-28 foi a melhor dentre os três modelos desenvolvidos e seus valores de sensitividade e especificidade foram superiores ao modelo tx5altaM2. Apesar de terem obtido melhores valores de acurácia, os modelos Varginha-alta-tx5-28 e M1G5 utilizaram atributos complexos em seus conjuntos de dados, sendo que o primeiro utilizou apenas um e o segundo um total de quatro, tornando-os um pouco mais complexos do que o modelo tx5altaM2. De qualquer forma, mesmo sendo mais complexo, o modelo Varginha-alta-tx5-28 apresentou as melhores medidas de desempenho e apresentou desempenho superior aos demais, mostrando ser o mais adequado para realizar a predição do aumento da taxa de progresso da ferrugem para tal combinação de carga e atributo meta.

Tabela 22: Medidas de avaliação de diversos modelos de alerta.

Nome do Modelo	Atrib. meta	Carga	Taxa de acerto	Erro	Sensitividade	Especificidade
Varginha-alta-tx5-28	5 p.p.	Alta	85,3	14,7	85,4	85,2
tx5altaM2*	5 p.p.	Alta	81,9	18,1	78,5	84,7
M1G5**	5 p.p.	Alta	84,7	15,3	***	***
Varginha-alta-tx10-26	10 p.p.	Alta	89,9	10,1	90,0	89,9
tx10altaM1*	10 p.p.	Alta	78,7	21,3	64,7	84,6
M3G10**	10 p.p.	Alta	83,5	16,5	***	***
Varginha-baixa-tx5-25/32	5 p.p.	Baixa	88,9	11,1	86,3	89,9
tx5baixaM2*	5 p.p.	Baixa	72,1	27,9	38,0	86,0

** Modelo gerado com dados de Meira (2008). ** Modelo gerado por Cintra et al. (2011).*
**** Cintra et al. (2011) não forneceu estas medidas de avaliação.*

A combinação de carga alta e atributo meta 10 p.p. também permitiu comparação do modelo gerado neste trabalho com o modelo gerado com os dados de Meira (2008) e com o modelo de Cintra et al. (2011). A taxa de acerto do modelo Varginha-alta-tx10-26 foi a maior dentre os três modelos desenvolvidos, chegando próximo à casa dos 90%. Já seus valores de sensitividade e especificidade foram bem superiores ao modelo tx10altaM1, principalmente para a sensitividade, a qual foi de 90,0% contra 64,7%, mostrando uma melhora significativa do primeiro modelo ao classificar corretamente exemplos positivos. O modelo Varginha-alta-tx10-26 também se mostrou muito mais equilibrado do que o modelo tx10altaM1, pela diferença de apenas 0,1 p.p. entre as medidas de sensitividade e especificidade. Destes três modelos, o que foi gerado por um conjunto de dados mais simples foi o modelo M3G10, que foi gerado com o conjunto M3, enquanto que o modelo Varginha-alta-tx10-26 foi gerado com o conjunto M2. Entretanto, a diferença de complexidade entre estes dois conjuntos é pequena, uma vez que ambos não utilizam atributos complexos, e aliada às suas melhores medidas de desempenho, o modelo Varginha-alta-tx10-26 se mostrou superior aos demais para esta combinação de carga e atributo meta.

Para a combinação de carga baixa e atributo meta 5 p.p., o modelo gerado neste trabalho foi comparado apenas com um modelo desenvolvido com os dados de Meira (2008). Verificou-se que a taxa de acerto do modelo Varginha-baixa-tx5-25/32 foi 16,8 p.p. superior à do modelo tx5baixaM2, além disso o primeiro modelo também obteve valores de sensitividade e especificidade superiores. O destaque ficou com a sensitividade, que subiu de 38,0% para 86,3%, indicando a melhor capacidade do modelo Varginha-baixa-tx5-25/32 trabalhar com exemplos positivos. A única vantagem do modelo tx5baixaM2 foi que este utilizou o conjunto de dados M2, mais simples do que o utilizado pelo conjunto Varginha-baixa-tx5-25/32, que foi o M1. Todavia, mesmo com um conjunto mais complexo, as medidas de avaliação do modelo Varginha-baixa-tx5-25/32 foram muito superiores às do modelo tx5baixaM2, tornando-o o melhor para representar a taxa de progresso na ferrugem para esta combinação de carga e atributo meta.

Um dos objetivos deste trabalho foi que os modelos desenvolvidos obtivessem melhor desempenho do que modelos já existentes, o que foi comprovado nesta seção. Alguns outros modelos, principalmente os gerados por regressão (seção 3.4), não foram comparados aos desenvolvidos neste trabalho pois utilizaram métricas de avaliação diferentes.

6.2.5 Diferença entre modelos balanceados e não balanceados

Um ponto que merece discussão foi a diferença de desempenho de modelos gerados por arquivos com e sem o balanceamento de classes. Esta diferença começa com a presença de apenas 4 modelos provenientes de arquivos não balanceados nos envelopes convexos para os diversos cenários em que este procedimento foi feito, ou seja, apenas 9%.

Este baixo desempenho foi verificado com mais intensidade nos cenários de carga baixa e atributo meta 5 p.p. (Figura 29, Figura 32, Figura 26 e Figura 35), mas também foi notado nos cenários de carga alta e atributo meta 10 p.p. (Figura 25). A diferença pode ser notada nas respectivas figuras de cada cenário, por meio de linhas paralelas ao eixo das abscissas (1 – especificidade), que dividiram o desempenho dos tipos de modelos.

Para o caso de carga alta e atributo meta 10 p.p., o cenário Varginha-alta-10 foi o que mais apresentou diferença de comportamento entre estes dois tipos de modelo. A diferença de

sensitividade entre o pior modelo proveniente de arquivos balanceados e o melhor modelo proveniente de arquivos não balanceados foi de 5 p.p. (Figura 25), o que mostrou que os primeiro modelos classificaram uma maior porcentagem de exemplos positivos corretamente.

Esta diferença entre os valores de sensitividade foi mais gritante nos cenários de carga baixa e atributo meta 5 p.p., sendo que o menor valor ocorreu para o cenário Varginha-Novo-baixa-5, onde esta diferença foi de 7 p.p. (Figura 35), e o maior valor ocorreu para o cenário Tudo-novo-baixa-5, onde a diferença chegou a 27 p.p. (Figura 32).

Alguns resultados de Meira (2008) indicaram que diversos modelos não apresentaram desempenho interessante para serem utilizados em um sistema de alerta, principalmente modelos de carga baixa e atributo meta 10 p.p. A ferrugem apresenta comportamentos diferentes em anos de carga alta e carga baixa, sendo mais agressiva em anos de carga alta e, consequentemente, trazendo mais ocorrências de aumento da taxa de progresso do que em anos de baixa. Além disso, quanto maior o limite da taxa de progresso, há uma tendência de que ocorram menos registros com este aumento. A baixa quantidade de exemplos de uma classe pode ter sido a responsável pelo desempenho indesejado em alguns modelos desenvolvidos por Meira (2008), mostrando-se que seja efetuado o balanceamento de classes para que se obtenha melhores resultados ao trabalhar com modelos de alerta da ferrugem do cafeeiro.

Antes de encerrar esta seção, vale a pena lembrar que a redução da taxa de progresso para um valor muito baixo, como 1 ou 2 p.p., poderia causar o problema inverso, deixando o conjunto de dados com muitos registros de aumento e poucos de não aumento da taxa de progresso. Assim, por exemplo, um balanceamento poderia ser feito para replicar os exemplos da classe de não aumento da taxa de progresso.

6.2.6 Diferença no comportamento das técnicas de modelagem

A indução dos modelos de alerta por diversas técnicas de modelagem permitiu notar como uma técnica atuou no conjunto de dados, gerando modelos de uma determinada forma e com um conjunto de características específicas.

Um destes casos ocorreu por meio dos modelos gerados pelas redes neurais artificiais, os quais apresentaram altos valores de sensitividade em alguns casos. A Figura 24, a Figura 27 e a Figura 35 ilustram este comportamento por meio de um grupo de modelos que foram destacados na lateral superior direita de cada uma delas, sendo que o maior destes grupos esteve presente na Figura 27, com um total de 5 modelos. Para este caso, estes modelos superaram a faixa de 90% de sensitividade, chegando a valores próximos de 92%. Estes resultados foram melhores do que aqueles obtidos por outras técnicas, como, por exemplo, as florestas aleatórias, que chegaram a obter valores de sensitividade na casa dos 89%. O maior valor de sensitividade registrado em um modelo selecionado foi de 92,6% e ocorreu no modelo 48 da Figura 35.

Como mencionado na seção 6.2.1, a sensitividade indica como o modelo trabalha com exemplos de aumento da taxa de progresso da ferrugem, os altos valores obtidos por estes modelos indicam que eles classificaram corretamente muitos desses exemplos. Entretanto, ocorreu que valores baixos de especificidade estavam atrelados aos altos valores de sensitividade, fazendo com que, em diversos cenários, os modelos gerados pelas redes neurais artificiais não fizessem parte do envelope convexo. De uma forma geral, apenas 3 modelos gerados por RNAs estiveram presentes nos envelopes convexos, cerca de apenas 7% dos casos.

Outra técnica de modelagem que não foi muito presente nos modelos selecionados nos envelopes convexos foram as árvores de decisão, utilizadas em 3 dos 43 modelos selecionados, apenas 7% dos casos. Isto ocorreu pois geralmente as árvores de decisão tiveram desempenho inferior às técnicas como florestas aleatórias e SVM. A Figura 32 e a Figura 33 enfatizam alguns modelos gerados por árvores de decisão que mostraram um desempenho inferior, sempre bem distante do envelope convexo, na parte central da figura. A Figura 32 levou em consideração apenas os modelos gerados por arquivos balanceados para esta comparação. Nenhuma propriedade específica foi notada nas árvores de decisão, elas simplesmente obtiveram medidas de avaliação inferiores aos demais modelos desenvolvidos.

Por sua vez, as técnicas de SVM e florestas aleatórias formaram o restante dos modelos (20 modelos em SVM e 17 em florestas aleatórias), ou seja, estas duas técnicas foram responsáveis por 86% dos modelos selecionados em todos os envelopes convexos. Esta superioridade foi ainda mais evidente quanto aos modelos recomendados para cada cenário

(seção 6.2.3), onde todos os modelos recomendados foram por florestas aleatórias (10 modelos) ou por SVM (2 modelos). A Figura 25, a Figura 28, a Figura 29, a Figura 31 e a Figura 26 evidenciam esta superioridade, sempre destacando diversos modelos próximos ao envelope convexo que foram gerados por algumas estas duas técnicas.

Com estes resultados pode-se concluir que, em termos de medidas de avaliação, as técnicas de SVM e florestas aleatórias foram superiores às demais técnicas utilizadas neste trabalho, sendo recomendadas para novas induções de modelos relacionados a esta área.

7 Conclusões

Os modelos de alerta validados neste trabalho não foram aceitos, estes modelos não estavam se comportando de forma adequada, ou melhor, não estavam tendo um desempenho esperado, mostrando a necessidade de realizar um novo processo de descoberta de conhecimento em bases de dados para gerar novos modelos de alerta.

A aplicação de procedimentos como o balanceamento de classes e o uso de métodos de seleção de atributos nos novos modelos de alerta desenvolvidos elevou o desempenho dos mesmos, além do mais, os métodos de seleção de atributos também auxiliaram na avalição de quais atributos do conjunto de dados foram mais representativos para a taxa de progresso da ferrugem do cafeeiro.

Com relação às técnicas utilizadas para a modelagem, as que apresentaram melhores medidas de desempenho foram Support Vector Machines e Florestas Aleatórias e estas são as mais indicadas para realizar simulações em modelos de alerta da ferrugem do cafeeiro. As redes neurais artificiais apresentaram modelos com altos valores de sensitividade enquanto que as árvores de decisão foram as técnicas que apresentaram menores medidas de avaliação.

Os modelos com dados apenas de Varginha obtiveram, em geral, um desempenho melhor do que os modelos gerados com os dados das três cidades. Dentre os modelos de Varginha, o melhor desempenho ocorreu para os modelos gerados com dados de 1998 a 2011, ao invés dos dados de 2007 a 2011.

Os modelos de alerta selecionados neste trabalho obtiveram, de forma geral, um desempenho superior aos demais modelos de alerta existentes. O reflexo do melhor desempenho é uma maior confiabilidade nas predições feitas por estes modelos, fornecendo condições mais precisas para o monitoramento da ferrugem do cafeeiro em campo.

Existem duas limitações do uso destes modelos de alerta, uma delas está relacionada à 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. A outra restrição é com relação aos atributos utilizados em cada um dos modelos indicados, ou seja, um modelo pode ter um

atributo que irá requerer a medição de uma determinada variável meteorológica que pode não estar disponível. Alguns modelos "alternativos" foram recomendados, utilizando conjuntos de dados mais simples, os quais podem evitar esta situação.

7.1 Sugestões de trabalhos futuros

Diversas sugestões para a continuidade deste trabalho podem ser mencionadas. Algumas destas sugestões representam incrementos e/ou alterações na parte metodológica de execução do processo de descoberta de conhecimento em bases de dados.

- Efetuar uma nova divisão no conjunto de dados separando os cultivares do café, no caso o cultivar Mundo novo do cultivar Catuaí. Estes cultivares tem susceptibilidades diferentes a ocorrência da ferrugem do cafeeiro e esta divisão pode vir a ajudar no processo de aprendizagem dos modelos.
- Levar em consideração o custo de decisões ou classificações incorretas. Para este caso, seria necessário realizar um estudo econômico-ambiental e ver qual o prejuízo das classificações incorretas, como, por exemplo, quanto custa uma aplicação de fungicidas desnecessária, ou qual o impacto na produção no caso de controle da doença quando necessário.
- Utilizar outras técnicas de modelagem para indução dos modelos. O próprio WEKA, além das técnicas utilizadas, possui outras opções de algoritmos para a indução de modelos de classificação, como as redes bayesianas, por exemplo.
- Utilizar regras de exceção (DALY e TANIAR, 2005) para localizar alguns erros do modelo. Pode-se notar que alguns meses do conjunto de dados sempre apresentavam erros, as regras de exceção podem ser utilizadas para tentar reduzir estes erros
- Utilizar uma família ou um conjunto de modelos para realizar a predição da taxa de progresso. Podem ser utilizadas ferramentas com um sistema de votação, por exemplo, onde cada modelo contribui com um voto e o conjunto destes votos determina qual será predição da taxa de progresso.

Outra sugestão é a ampliação do conjunto de dados utilizados para gerar os modelos. A Fundação PROCAFE já ampliou a cobertura e monitoramento das lavouras de café para a cidade de Muzambinho, mais cidades do Sul de Minas (junto com as demais deste trabalho), além de cidades do Triângulo Mineiro, como Araxá, Patrocínio e Araguai. A partir daí, uma nova rodada de modelagem pode ser realizada, criando-se novos cenários de indução.

Por fim, modelos de alerta de outras doenças e pragas do cafeeiro poderiam ser desenvolvidos. A Fundação PROCAFE, 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 tarda*), o bicho-mineiro (*Leucoptera coffeella*) e a broca-do-café (*Hypothenemus hampei*).

8 Referências bibliográficas

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD International Conference on Management of Data, 1993, New York, **Proceedings…** New York: ACM, p.207-216, 1993.

AGRAWAL, R.; MANILLA, H.; SRIKANT, R.; TOIVONEN, H.; VERKAMO, A. I. Fast Discovery of Association Rules. In: FAYYAD, U.M. PIATETSKY-SHAPIRO, G.; SMYTH, P.; VERKAMO, A.I. **Advances in Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press, 1996, p.307-328.

AGRIOS, G. N. **Plant pathology**. 5ed. San Diego: Elsevier Academic Press, 2004.

ALVES, M. C.; CARVALHO, L. G.; POZZA, E. A.; ALVES, L. S. A Soft Computing Approach For Epidemiological Studies of Coffee And Soybean Rusts. **International Journal of Digital Content Technology and its Applications,** v.4, n.1, p.149-154, fev., 2010.

APSNET. APSnet Education Center: **plant disease lessons – coffee rust – disease cyvle and epidemiology.** <http://www.apsnet.org/edcenter/intropp/lessons/fungi/Basidiomycetes/Pages/CoffeeRust.aspx>, 20/03/2013.

APTE, C.; WEISS, S. Data mining with decision trees and decision rules. **Future Generation Computer Systems**. Amsterdam, v. 13, n.2-3, p.197-210, nov., 1997.

BALCI, O. Principles of simulation model validation, verification, and testing. **Transactions of the Society for Computer Simulation International**, San Diego, v. 14, n.1, p.3-12, mar., 1997.

BANKS, J. **Handbook of Simulation:** Principles, Methodology, Advances, Applications and Practices. New York: John Wiley and Sons, 1998.

BATCHELOR, W.D.; YANG, X.B; TSCHANZ, A.T. Development of a neural network for soybean rust epidemics. **Transactions of the American Society of Agricultural Engineers,** v.40, n.1, p.247-252, 1997.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explorations**, v.6, n.1, p.20-29, jun., 2004.

BELLE, V. **Detection and recognition of human faces using random forest for a mobile root.** 104p. Dissertation (Masters on Computer Science) – Rwthaachen University, Germany, 2008.

BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern Recognition**. New York, v.30, n.7, p.1145–1159, jul., 1997.

BRAGA, A. P.; LUDERMIR, T. B.; CARVALHO, A. C. P. L. F. **Redes neurais artificiais:** teoria e aplicações. 2ed. Rio de Janeiro: LTC, 2007.

BREIMAN, L. Random forests. **Machine Learning Journal**. Hingham, v.45, p.5–32, jan. 2001.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, Charles. J. **Classification and regression trees**. Boca Raton: CRC, 1984.

CAMPBELL, C. L.; MADDEN, L. V. **Introduction to plant disease epidemiology**. New York: John Wiley and Sons, 1990.

CAMPBELL, C. L.; REYNOLDS, K. M.; MADDEN, L. V. Modeling epidemics of root diseases and development of simulators. In: KRANZ, J.; ROTEM, J. **Experimental techniques in plant disease epidemiology**. Berlin: Springer-Verlag, 1988. p. 253-265.

CARUANA, R.; KARAMPATZIAKIS, N.; YESSENALINA, A. An empirical evaluation of supervised learning in high dimensions. In: International conference on Machine learning, 25, Helsinki, **Proceedings…** Helsinki: ACM, p.96-103, 2008.

CHALFOUN, S. M. **Doenças do cafeeiro**: importância, identificação e métodos de controle. Lavras: UFLA/FAEPE. 1997.

CHALFOUN, S. M.; CARVALHO, V. L. Controle químico da ferrugem (*Hemileia vastatrix* Berk & Br.) do cafeeiro através de diferentes esquemas de aplicação. **Pesquisa Agropecuária Brasileira**, v.34, n.3, p.363-367, Mar., 1999.

CHANG, C-C.; LIN, C-J. LIBSVM: A library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**. v.2, n.3, Artigo 27 – 27p., abr, 2011.

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.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER,W. P. SMOTE: Synthetic Minority Oversampling Technique. **Journal of Artificial Intelligence Research**, v.16, p.321–357, jun., 2002.

CINTRA, M. E.; MEIRA, C. A. A.; MONARD, M. C.; CAMARGO, H. A.; RODRIGUES, L. H. A. The use of fuzzy decision trees for coffee rust warning in Brazilian crops. In: International Conference on Intelligent Systems Design and Applications, 11, 2011, Córdoba, ES, **Proceedings...** Córdoba: IEEE, p. 1347-1352, 2011.

COAKLEY, S. M. Variation in climate and prediction of disease in plants. **Annual Review of Phytopathology**, v.26, p.163-181, set., 1988.

COHEN, J**.** A coefficient of agreement for nominal scales. **Eductional and Psychological Measurement**, v.20, n.1, p.37 –46, abr., 1960.

CROS, J.; COMBES, M. C.; TROUSLOT, P.; ANTHONY, F.; HAMON, S.; CHARRIER, A.; LASHERMES, P. Phylogenetic analysis of chloroplast DNA Variation in *Coffea* L. **Molecular Phylogenetics and Evolution** v.9, n.1, p.109– 117, fev., 1998.

DALY, O.; TANIAR, D. Exception rules in data mining. In: KHOSROW-POUR, M. **Encyclopedia of information science and technology (Volume II)**, Hershey: Idea Group Inc., 2005. p. 1144-1148.

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, jun., 2006.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters.** New York**,** v.27, n.8, p.861-874, jun., 2006.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v.17, n.3, p.37-54, jul., 1996.

FAZUOLI, L. C.; MEDINA FILHO, H. P.; GONÇALVES, W.; GUERREIRO FILHO, O.; SILVAROLLA, M. B.. Melhoramento do cafeeiro: variedades tipo arábica obtidas no Instituto Agronômico de Campinas. In: ZAMBOLIN, L. **O estado da arte de tecnologias na produção de café.** Viçosa: UFV, 2002. p.163-215.

FERNANDES, R. C.; EVANS, H. C.; BARRETO, R.W. Confirmation of the occurrence of teliospores of Hemileia vastatrix in Brazil with observations on their mode of germination. **Tropical Plant Pathology**, v.34,n. 2, p.108-113, mar/abr, 2009.

FINHOLDT, G. Desenvolvimento e avaliação de um sistema automático de alerta de doenças em plantas. 97p. Dissertação (Mestado em Engenharia Agrícola) – Universidade Federal de Viçosa, Viçosa. 2012.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: an overview. **AI Magazine**, v.13, n.3, p.57-70, jul., 1992.

GIROLAMO NETO, C. D.; MEIRA, C. A. A.; RODRIGUES, L. H. A. Seleção de atributos em modelos de alerta da ferrugem do cafeeiro em lavouras com alta carga pendente de frutos. In: Congresso Paulista de Fitopatologia, 35, 2012, Jaguariúna, SP. **Resumos...** Brasília: EMBRAPA, 2012a. CD-ROM.

GIROLAMO NETO, C. D.; MEIRA, C.A.A.; RODRIGUES, L.H.A. Avaliação de modelos de alerta da ferrugem do cafeeiro para lavouras com alta carga pendente de frutos. In: Congresso Brasileiro de Pesquisas Cafeeiras, 38, 2012, Caxambu,MG. Anais... Varginha: PROCAFÉ/MAPA, 2012b.

GOOGLE. **Google Refine**, a power tool for working with messy data. <http://code.google.com/p/google-refine/>. 14/01/2011.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research.** v.3, p.1157-1182, mar., 2003.

HALL, M. A. **Correlation-based feature selection for machine learning.** 178p. Thesis (PhD on Computer Science) – Departament of Computer Science, University of Waikato, Nova Zelândia. 1999.

HALL, M. A.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update; **SIGKDD Explorations**. New York, v.11, n.1, p. 10-18, jun., 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

HARDWICK, N. V. Disease forecasting. In: COOKE, B. M.; JONES, D. G.; KAYE, B. **The epidemiology of plant diseases**. 2 ed. Holanda:Springer, 2006. p. 239-267.

HAYKIN, S. **Neural Networks and Learning Machines**. 3ed., Englewood Cliffs: Prentice-Hall. 2009.

JACOBS, J. 2011. **AI in designing 'intelligent' ship autopilots.** <http://jamjacobs.blogspot.com.br/2011/02/ai-in-designing-intelligent-ship.html>, 21/03/2013.

JAPIASSÚ, L. B.; GARCIA, A.W. R.; FERREIRA, 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, 5p., 2007.

JOHN, G. H.; KOHAVI, R. Wrappers for feature subset selection. **Artificial Inteligence**. v.97, n.1-2, p.273-324, dez., 1997.

KDNUGGETS. **Data Mining Community's Top Resource** for Data Mining and Analytics Software, Jobs, Consulting, Courses, and more. < http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html >, 26/07/2012.

KIRALJ, R.; FERREIRA, M. M. C. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. **Journal of the Brazilian Chemical Society.** São Paulo, v.20, n.4, p. 770-787, 2009.

KLEIJNEN, J. P. C. Validation of models: statistical techniques and data availability. In: Winter Simulation Conference, 31, 1999, Phoenix, **Proceedings...** Piscataway: IEEE p.647-654, 1999.

KÜHN, J. G. **Die Krankheiten der Kulturgewächse: ihre Ursachen und ihre Verhütung** (As doenças de plantas cultivadas: suas causas e sua prevenção) - 1858. <http://books.google.com.br/books?id=Ly4VAAAAYAAJ&printsec=frontcover&hl=pt-BR&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false>, 21/02/2012.

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.

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.; ESKES, A. B. Advances in coffee rust research. **Annual Review of Phytopatology**, v. 27, p.503-531, set., 1989.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v.33, n.1, p.159-174, mar., 1977.

LIU, H.; SETIONO, R. Chi2: Feature selection and discretization of numeric attributes. In:International Conference on Tools with Articial Intelligence, 7, 1995, Herndon, **Proceedings…** Washington: IEEE, p.388-391, 1995.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**. Porto Alegre, v.14, n.2, p.43-67, 2007.

LUACES, O.; RODRIGUES, L. H. A.; MEIRA, C. A. A.; BAHAMONDE; A.. Using nondeterministic learners to alert on coffee rust disease. **Expert systems with applications**, v.38, n.11, p.14276-14283, jan., 2011.

MADDEN, L. V.; ELLIS, M. A. How to develop plant disease forecasters. In: KRANZ, J.; ROTEM, J. **Experimental techniques in plant disease epidemiology**. Berlin: Springer-Verlag, 1988. p. 191-208.

MARTINS, E. M. F. **Sequência de eventos primários do desenvolvimento de *Hemileia vastatrix* em folhas de cafeeiro com suscetibilidade genética, resistência induzida ou resistência genética.** 149 p. Dissertação (Mestrado em Agronomia) – Escola superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba. 1988.

MATIELLO, J. B.; SANTINATO, R.; GARCIA, A. W. R.; ALMEIDA, S. R.; FERNANDES, D. R. Cultura de café no Brasil. Novo manual de recomendações. In: MATIELLO, J. B. **MAPA/PROCAFÉ**. Rio de Janeiro, p.387, 2002.

MEIRA, C. A. A. **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**. 198p. Tese (Doutorado em Engenharia Agrícola) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas. 2008.

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./abr., 2008.

MEIRA, C. A. A.; RODRIGUES, L. H. A.; MORAES, S. A. Modelos de alerta para o controle da ferrugem-do-cafeeiro em lavouras com alta carga pendente. **Pesquisa Agropecuária Brasileira.** v.44, n.3, p.233–242, mar., 2009.

MENDES, L. C.; MENEZES, H. C.; SILVA, M. A. A. P. Optimization of the roasting of robusta coffee (*C. canephora* conillon) using acceptability testes and RSM. **Food Quality and Preference.** v.12, n.2, p.153-162, 2001.

MICHEREFF, S. J. **Fundamentos de Fitopatologia**. Pernambuco: Universidade Federal Rural de Pernanbuco, 2001.

MINISTÉRIO DA AGRICULTURA. **Ministério da Agricultura, Pecuária e abastecimento.** < http://www.agricultura.gov.br/vegetal/culturas/cafe>, 26/01/2013.

MINISTÉRIO DO DESENVOLVIMENTO. Ministério do desenvolvimento, indústria e comércio exterior. <http://www.mdic.gov.br/sitio/interna/interna.php?area=5&menu=1955>, 20/03/2013.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole, 2002. p. 89-135.

MONTOYA, R. H.; CHAVES, G. M. Influência da temperatura e da luz na germinação, infectividade e período de geração de *Hemileia vastatrix* Berk. e Br. **Experientiae**, v.18, n.11, p.239-266, 1974.

MORAES, S. A. **A ferrugem do cafeeiro**: importância, condições predisponentes, evolução e situação no Brasil. Campinas: Instituto Agronômico, 1983.

MORAES, S. A.; SUGIMORI, M. H.; RIBEIRO, I. J. A.; ORTOLANI, A. A.; PEDRO JUNIOR, M. J. Período de incubação de *Hemileia vastatrix* Berk. e Br. em três regiões do Estado de São Paulo. **Summa Phytopathologica**. Piracicaba, v.2, n.1, p.32-38, 1976.

PIMENTEL-GOMES, F. **Curso de Estatística Experimental**. 15ed. São Paulo: FEALQ, 2009.

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, set./out., 2002.

PRATI, R. C.; BATISTA; G. E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina.** v.6, n.2, p.215-222, jun., 2008.

PROVOST, F.; FAWCETT, T.; KOHAVI, J. The case against accuracy estimation for comparing induction algorithms. In: Fifteenth International Conference on Machine Learning, 15, São Francisco, **Proceedings...** San Francisco:Morgan Kaufmann, p.445–453, 1998.

PROVOST, F.; FAWCETT, T.; Robust classifiers for imprecise environments. **Machine Learning**, v.42, n.3, p.203–231, mar., 2001.

RUMELHART, D. E.; McCLELLAND, J. L.; and the PDP Research Group. **Parallel Distributed Processing:** Exploration in the Microstructure of Cognition. Volume 1: Foundations. Cambridge: MIT Press, 1986.

SARGENT, R. G. A tutorial on verification and validation of simulation models. In: Winter Simulation Conference, 39, 2007, Washington, **Proceedings…** Piscataway: IEEE, p.114-121, 2007.

SARGENT, R. G.; Verification and validation of simulation models. **Journal of simulation**, v.7, n.1, p.12-24, fev., 2013

SCHLESINGER, S. Terminology for model credibility. **Simulation,** v.32, n.3,p.103-104, 1979.

SCHÖLKOPF, B.; SMOLA, A. J. **Learning with Kernels:** Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge: MIT Press, 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.

TOMEK, I. Two Modifications of CNN. **IEEE Transactions on Systems Man and Cybernetics**, v.6, n.11, p.769–772, nov., 1976.

USDA. **United States Department Of Agriculture.** <http://www.fas.usda.gov/psdonline/circulars/coffee.pdf>, 15/02/2013.

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, **Resumos...** Brasília: EMBRAPA, p. 171-174, 2000.

VAPNIK, V. N. **The nature of Statistical learning theory**. 2ed. New York: Springer-Verlag, 2000.

WALLER, J. M.; BIGGER, M.; HILLOCKS, R. J. **Coffee Pests, diseases and their management**. Wallingford: CAB International, 2007.

WEISS, G. M.; PROVOST, F. **The effect of class distribution on classifier learnig**: an empirical study. Techinical Report ML-TR-44, Departamente of computer science, Rutgers University, 2001.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3ed. San Francisco: Morgan Kaufmann, 2011.

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. **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.; 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. **Controle de doenças de plantas**: grandes culturas. Viçosa: UFV, 1997. p. 83-139.

Apêndice A – Modelos de alerta para os demais cenários de indução

Este apêndice contém os demais modelos desenvolvidos para cada cenário de indução, complementando o exemplo da seção 6.1.

A.1 Modelos para os cenários "Tudo"

Os modelos para os cenários "Tudo" foram gerados com todos os dados do conjunto (dados das 3 cidades e período de 1998 a 2011) e dispostos nas seções A.1.1 a A.1.3.

A.1.1 Cenário Tudo-alta-tx5

O gráfico ROC da Figura 27 representa o desempenho dos modelos desenvolvidos para o cenário Tudo-alta-tx5, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas na Tabela 24 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 23.

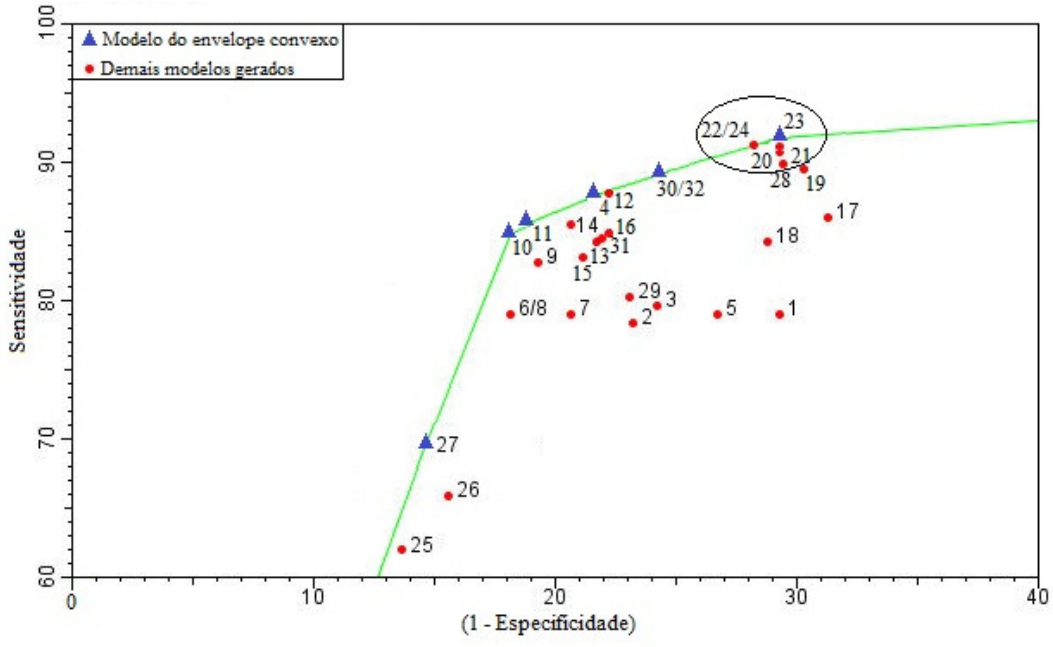


Figura 27: Gráfico ROC para o cenário Tudo-alta-tx5.

Tabela 23: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Tudo-alta-tx5.

	Modelos					
Atributos	4	10	11	23	27	30/32
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				*		*
NHUR90_PINF		*				
SMT_NHUR90_PINF				*		
THUR90_PINF		*		*		*
NHNUR90_PINF		*		*		
SMT_NHNUR90_PINF						
TMAX_PI_PINF		*	*	*	*	
TMIN_PI_PINF	*	*	*	*	*	
TMED_PI_PINF		*	*		*	
VVENTO_PINF						
SMT_VVENTO_PINF						

Tabela 24: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Tudo-alta-tx5.

Modelos	4	10	11	23	27	30/32
Técnica de modelagem	AD	RF	RF	RNA	SVM	SVM
Método de seleção	WRP	M2	M3	GR	M3	Chi^2/IG
Taxa de acerto	82,7	83,2	83,2	80,5	78,0	82,4
Erro	17,3	16,8	16,8	19,5	22,0	17,6
Sensitividade	87,7	84,8	85,4	91,8	69,6	89,2
Especificidade	78,3	81,8	81,3	70,7	85,3	75,7
Confiabilidade Positiva	77,7	80,1	79,8	73,0	80,4	78,0
Confiabilidade Negativa	88,1	86,2	86,6	90,9	76,5	87,9
TP Rate	87,7	84,8	85,4	91,8	69,6	89,2
FP Rate	21,7	18,2	18,7	29,3	14,6	24,3
AUC	0,837	0,877	0,875	0,800	0,775	0,825
Kappa	0,65	0,66	0,66	0,61	0,56	0,65

A.1.2 Cenário Tudo-alta-tx10

O gráfico ROC da Figura 28 representa o desempenho dos modelos desenvolvidos para o cenário Tudo-alta-tx10, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas na Tabela 26 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 25.

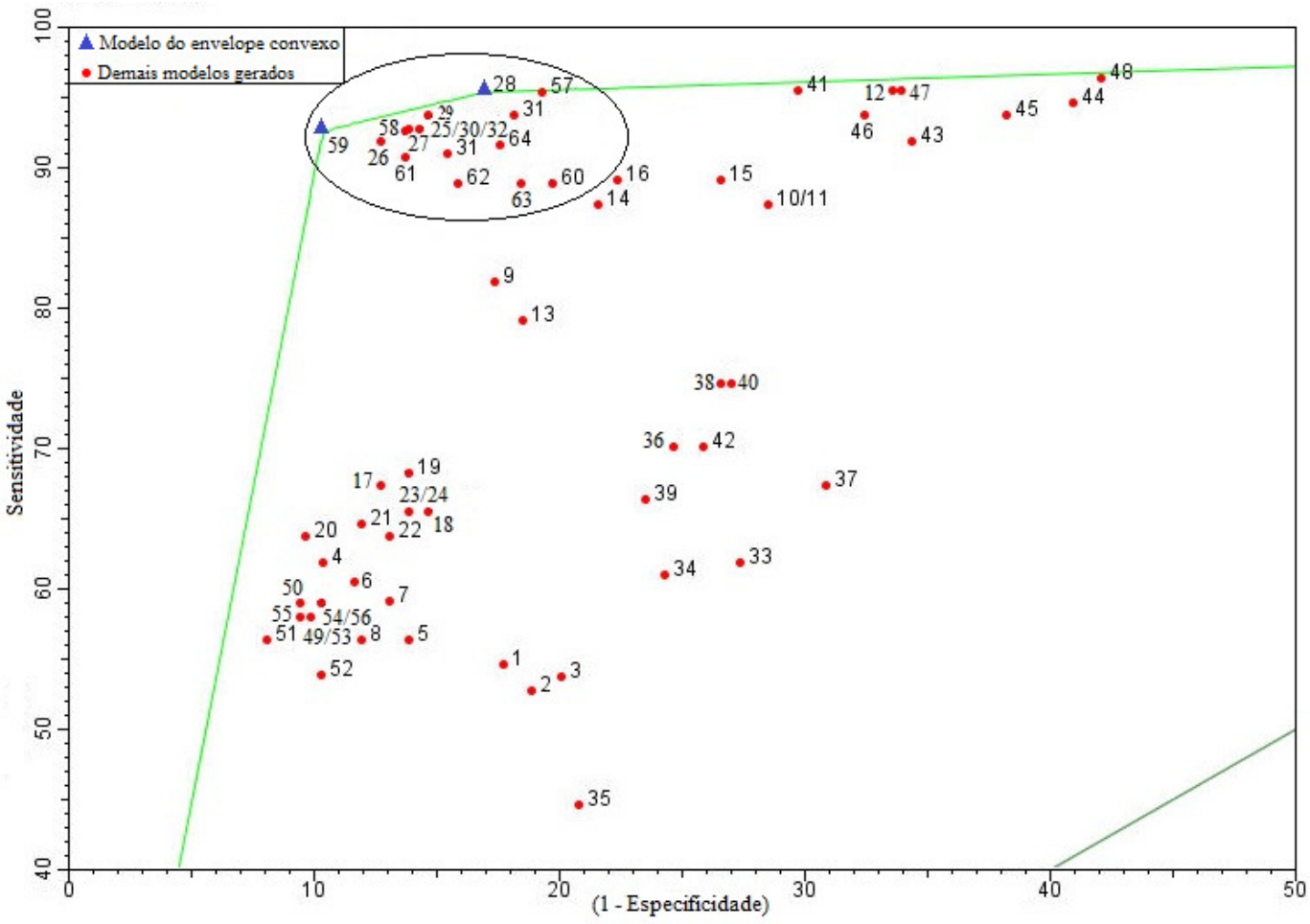


Figura 28: Gráfico ROC para o cenário Tudo-alta-tx10.

Tabela 25: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Tudo-alta-tx10.

	Modelos	
Atributos	28	59
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	*	
NHUR90_PINF		
SMT_NHUR90_PINF		
THUR90_PINF		
NHNUR90_PINF		
SMT_NHNUR90_PINF		
TMAX_PI_PINF		*
TMIN_PI_PINF		*
TMED_PI_PINF		*
VVENTO_PINF	*	
SMT_VVENTO_PINF		

Tabela 26: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Tudo-alta-tx10.

Modelagem	28	59
Técnica de modelagem	RF	SVM
Método de seleção	WRP	M3
Taxa de acerto	86,7	90,5
Erro	13,3	9,5
Sensitividade	95,5	92,7
Especificidade	83,0	89,6
Confiabilidade Positiva	70,5	79,1
Confiabilidade Negativa	97,7	96,7
TP Rate	95,5	92,7
FP Rate	17,0	10,4
AUC	0,936	0,912
Kappa	0,71	0,78

A.1.3 Cenário Tudo-baixa-tx5

O gráfico ROC da Figura 29 representa o desempenho dos modelos desenvolvidos para o cenário Tudo-baixa-tx5, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas Tabela 28 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 27.

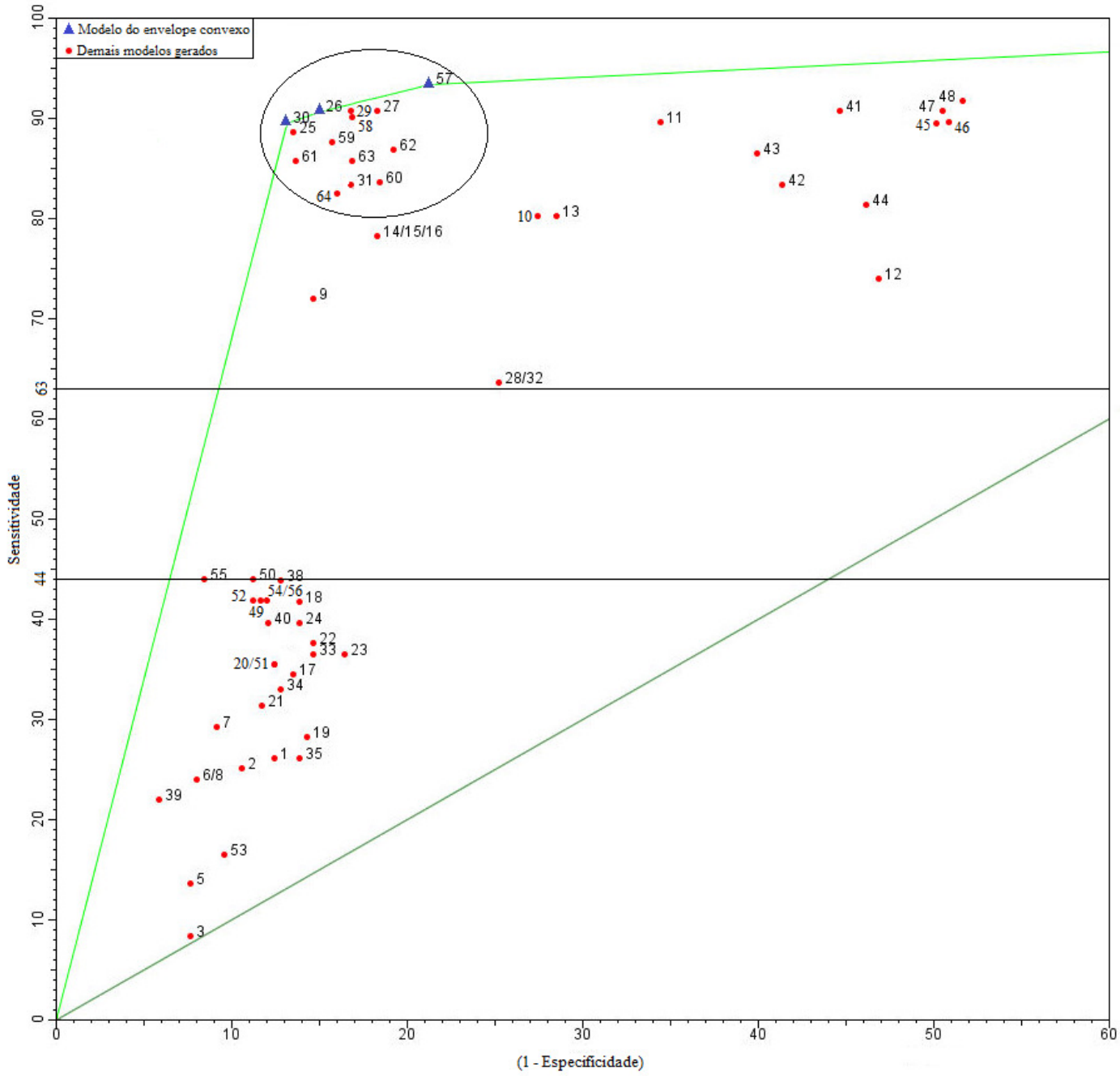


Figura 29: Gráfico ROC para o cenário Tudo-baixa-tx5.

Tabela 27: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Tudo-baixa-tx5.

	Modelos		
Atributos	26	30	57
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		*	*
NHUR90_PINF	*	*	*
SMT_NHUR90_PINF		*	*
THUR90_PINF	*	*	*
NHNUR90_PINF	*	*	*
SMT_NHNUR90_PINF		*	*
TMAX_PI_PINF	*		*
TMIN_PI_PINF	*	*	*
TMED_PI_PINF	*	*	*
VVENTO_PINF			*
SMT_VVENTO_PINF			*

Tabela 28: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Tudo-baixa-tx5.

Modelos	26	30	57
Técnica de modelagem	RF	RF	SVM
Métodos de seleção	M2	Chi^2	M1
Taxa de acerto	86,5	87,5	82,7
Erro	13,5	12,5	17,3
Sensitividade	90,6	89,6	93,4
Especificidade	85,0	86,8	78,7
Confiabilidade Positiva	68,0	70,5	61,6
Confiabilidade Negativa	96,3	96,0	97,0
TP Rate	90,6	89,6	93,4
FP Rate	15,0	13,2	21,3
AUC	0,928	0,926	0,861
Kappa	0,68	0,70	0,62

A.2 Modelos para os Cenários "Tudo-Novo"

Os modelos para os cenários "Tudo-Novo" foram gerados com todos os dados novos (dados das 3 cidades e período de 2007 a 2011) e dispostos nas seções A.2.1 a A.2.3.

A.2.1 Cenário Tudo-Novo-alta-tx5

O gráfico ROC da Figura 30 representa o desempenho dos modelos desenvolvidos para o cenário Tudo-novo-alta-tx5, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas Tabela 30 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 29.

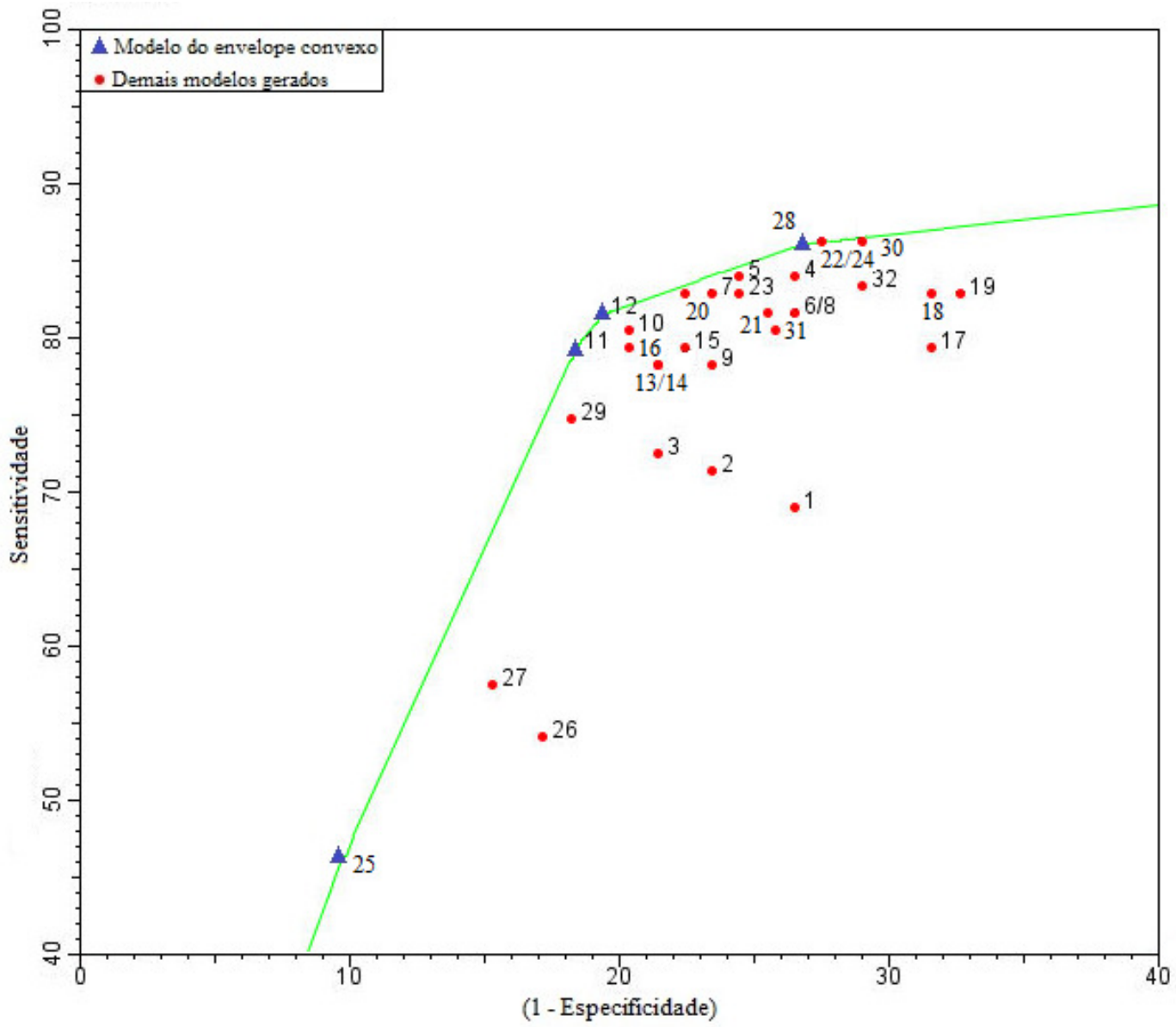


Figura 30: Gráfico ROC para o cenário Tudo-Novo-alta-tx5.

Tabela 29: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Tudo-Novo-alta-tx5.

	Modelos			
Atributos	11	12	25	28
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			*	*
NHUR90_PINF			*	
SMT_NHUR90_PINF			*	
THUR90_PINF			*	
NHNUR90_PINF			*	*
SMT_NHNUR90_PINF			*	
TMAX_PI_PINF	*		*	
TMIN_PI_PINF	*		*	
TMED_PI_PINF	*		*	
VVENTO_PINF			*	
SMT_VVENTO_PINF			*	

Tabela 30: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Tudo-Novo-alta-tx5.

Modelos	11	12	25	28
Técnica de modelagem	RF	RF	SVM	SVM
Método de seleção	M3	WRP	M1	WRP
Taxa de acerto	80,5	81,1	68,9	79,4
Erro	19,5	18,9	31,1	20,6
Sensitividade	79,3	81,6	46,0	86,2
Especificidade	81,6	80,6	90,3	73,1
Confiabilidade Positiva	79,3	78,9	81,6	75,0
Confiabilidade Negativa	81,6	83,2	64,1	85,0
TP Rate	79,3	81,6	46,0	86,2
FP Rate	18,4	19,4	9,7	26,9
AUC	0,860	0,839	0,682	0,780
Kappa	0,61	0,62	0,37	0,59

A.2.2 Cenário Tudo-Novo-alta-tx10

O gráfico ROC da Figura 31 representa o desempenho dos modelos desenvolvidos para o cenário Tudo-novo-alta-tx10, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas Tabela 32 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 31.

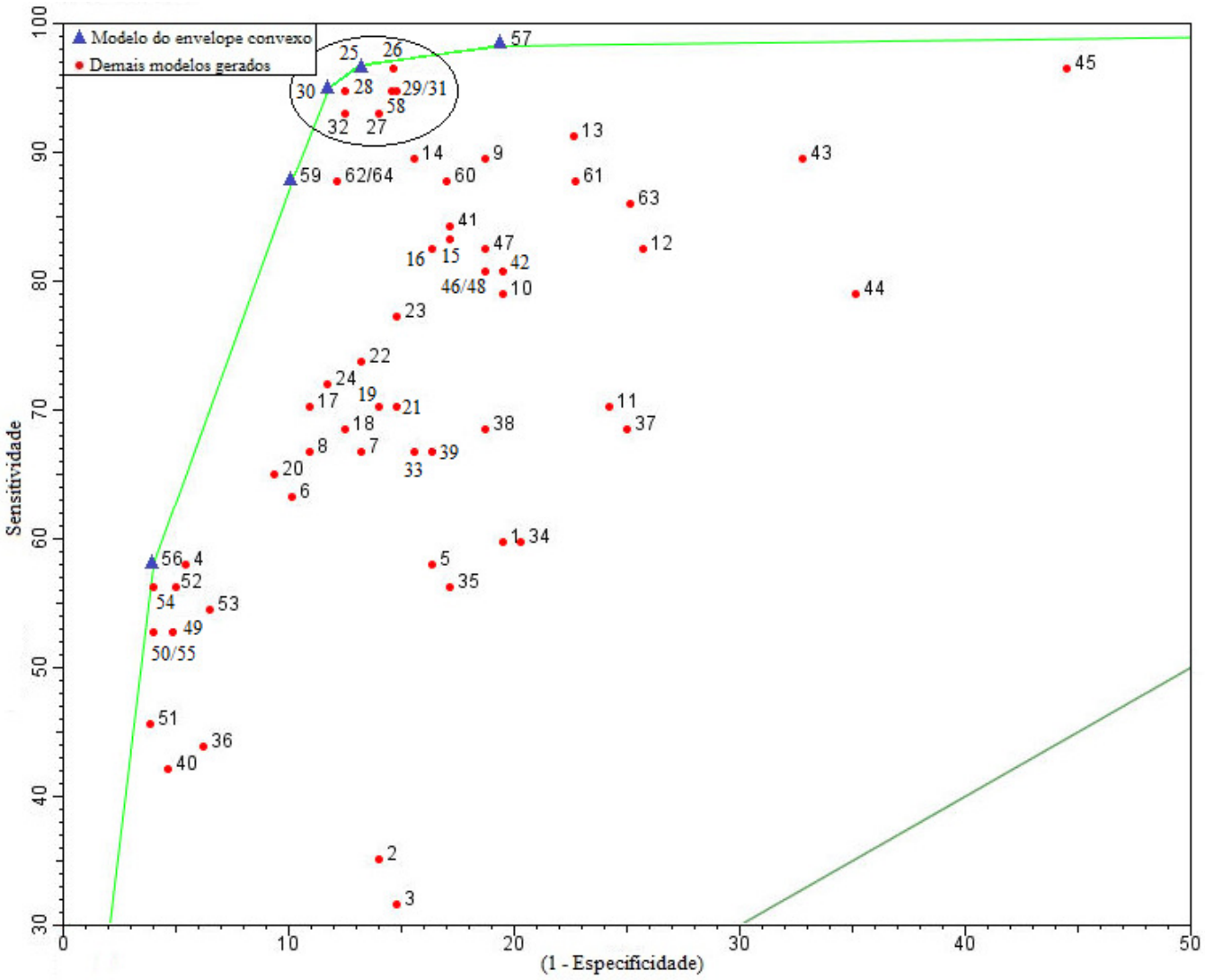


Figura 31: Gráfico ROC para o cenário Tudo-Novo-alta-tx10.

Tabela 31: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Tudo-Novo-alta-tx10.

	Modelos				
Atributos	25	30	56	57	59
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	*		*	*	
NHUR90_PINF	*	*	*	*	
SMT_NHUR90_PINF	*			*	
THUR90_PINF	*	*	*	*	
NHNUR90_PINF	*			*	
SMT_NHNUR90_PINF	*			*	
TMAX_PI_PINF	*			*	*
TMIN_PI_PINF	*			*	*
TMED_PI_PINF	*			*	*
VVENTO_PINF	*			*	
SMT_VVENTO_PINF	*			*	

Tabela 32: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Tudo-Novo-alta-tx10.

Modelos	25	30	56	57	59
Técnica de modelagem	RF	RF	SVM	SVM	SVM
Métodos de seleção	M1	Chi^2	IG	M1	M3
Taxa de acerto	89,7	90,3	83,9	86,1	89,2
Erro	10,3	9,7	16,1	13,9	10,8
Sensitividade	96,5	94,7	57,9	98,2	87,7
Especificidade	86,7	88,3	95,9	80,5	89,8
Confiabilidade Positiva	76,4	78,3	86,8	70,0	79,4
Confiabilidade Negativa	98,2	97,4	83,1	99,0	94,3
TP Rate	96,5	94,7	57,9	98,2	87,7
FP Rate	13,3	11,7	4,1	19,5	10,2
AUC	0,958	0,948	0,753	0,894	0,888
Kappa	0,78	0,78	0,54	0,71	0,75

A.2.3 Cenário Tudo-Novo-baixa-tx5

O gráfico ROC da Figura 32 representa o desempenho dos modelos desenvolvidos para o cenário Tudo-novo-baixa-tx5, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas Tabela 34 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 33.

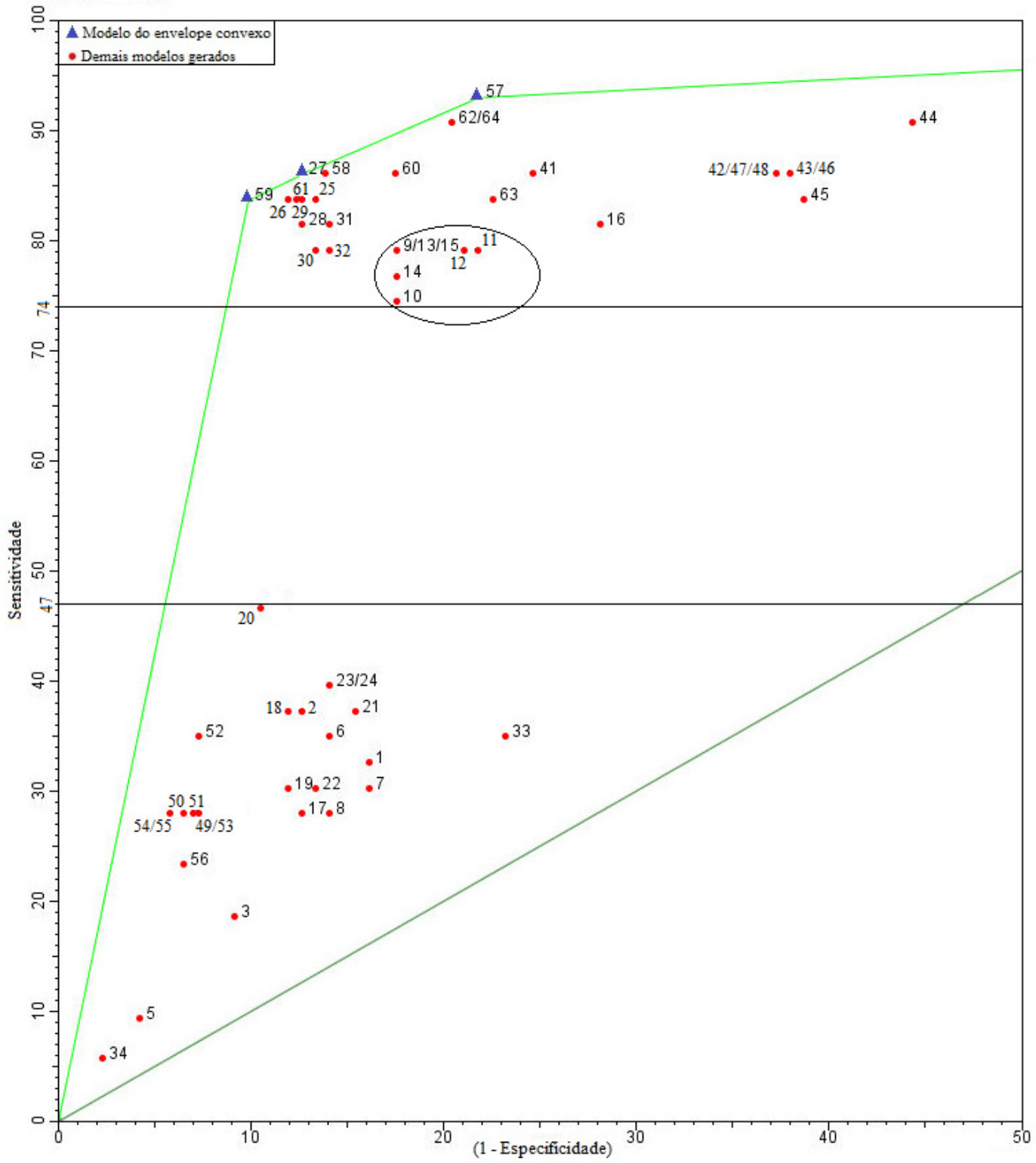


Figura 32: Gráfico ROC para o cenário Tudo-Novo-baixa-tx5.

Tabela 33: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Tudo-Novo-baixa-tx5.

	Modelos		
Atributos	27	57	59
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		*	
NHUR90_PINF		*	
SMT_NHUR90_PINF		*	
THUR90_PINF		*	
NHNUR90_PINF		*	
SMT_NHNUR90_PINF		*	
TMAX_PI_PINF	*	*	*
TMIN_PI_PINF	*	*	*
TMED_PI_PINF	*	*	*
VVENTO_PINF		*	
SMT_VVENTO_PINF		*	

Tabela 34: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Tudo-Novo-baixa-tx5.

Modelos	27	57	59
Técnica de modelagem	RF	SVM	SVM
Método de seleção	M3	M1	M3
Taxa de acerto	87,0	81,7	88,7
Erro	13,0	18,3	11,3
Sensitividade	86,1	93,0	83,7
Especificidade	87,3	78,1	90,1
Confiabilidade Positiva	67,3	57,1	72,0
Confiabilidade Negativa	95,4	97,3	94,8
TP Rate	86,1	93,0	83,7
FP Rate	12,7	21,9	9,9
AUC	0,914	0,856	0,869
Kappa	0,67	0,58	0,70

A.3 Modelos para os cenários "Varginha-Novo"

Os modelos para os cenários "Varginha-Novo" foram gerados com os dados novos (2007 a 2011) da cidade de Varginha e dispostos nas seções A.3.1 a A.3.3.

A.3.1 Cenário Varginha-Novo-alta-tx5

O gráfico ROC da Figura 33 representa o desempenho dos modelos desenvolvidos para o cenário Varginha-Novo-alta-tx5, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas Tabela 35 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 36.

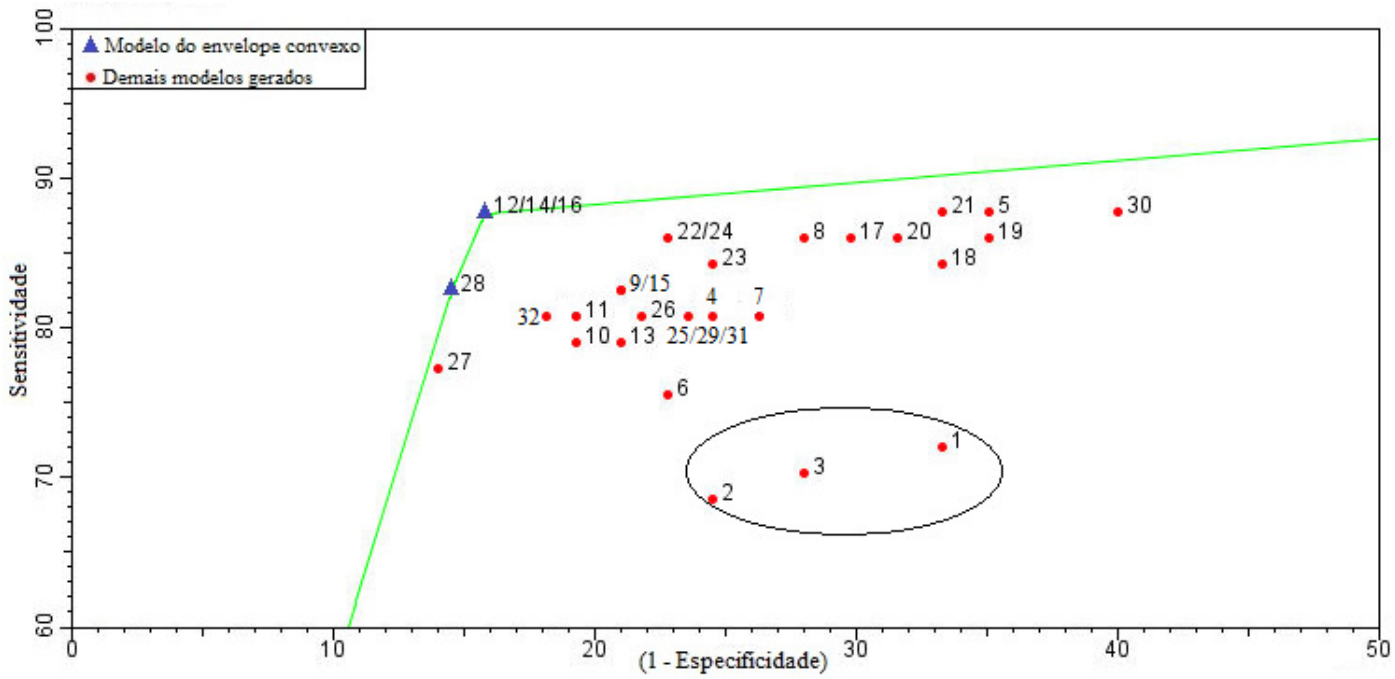


Figura 33: Gráfico ROC para o cenário Varginha-Novo-alta-tx5.

Tabela 35: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Varginha-Novo-alta-tx5.

Modelos	12/14/16	28
Técnica	RF	SVM
Método de seleção	WRP/Chi^2/IG	WRP
Taxa de acerto	86,0	83,9
Erro	14,0	16,1
Sensitividade	87,7	82,5
Especificidade	84,2	85,5
Confiabilidade Positiva	84,8	85,5
Confiabilidade Negativa	87,3	82,5
TP Rate	87,7	82,5
FP Rate	15,8	14,6
AUC	0,881	0,840
Kappa	0,72	0,68

Tabela 36: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Varginha-Novo-alta-tx5.

	Modelos	
Atributos	12/14/16	28
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		
NHUR90_PINF		
SMT_NHUR90_PINF		
THUR90_PINF		*
NHNUR90_PINF		*
SMT_NHNUR90_PINF		
TMAX_PI_PINF		
TMIN_PI_PINF		
TMED_PI_PINF		
VVENTO_PINF		
SMT_VVENTO_PINF		

A.3.2 Cenário Varginha-Novo-alta-tx10

O gráfico ROC da Figura 34 representa o desempenho dos modelos desenvolvidos para o cenário Varginha-Novo-alta-tx10, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas Tabela 38 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 37.

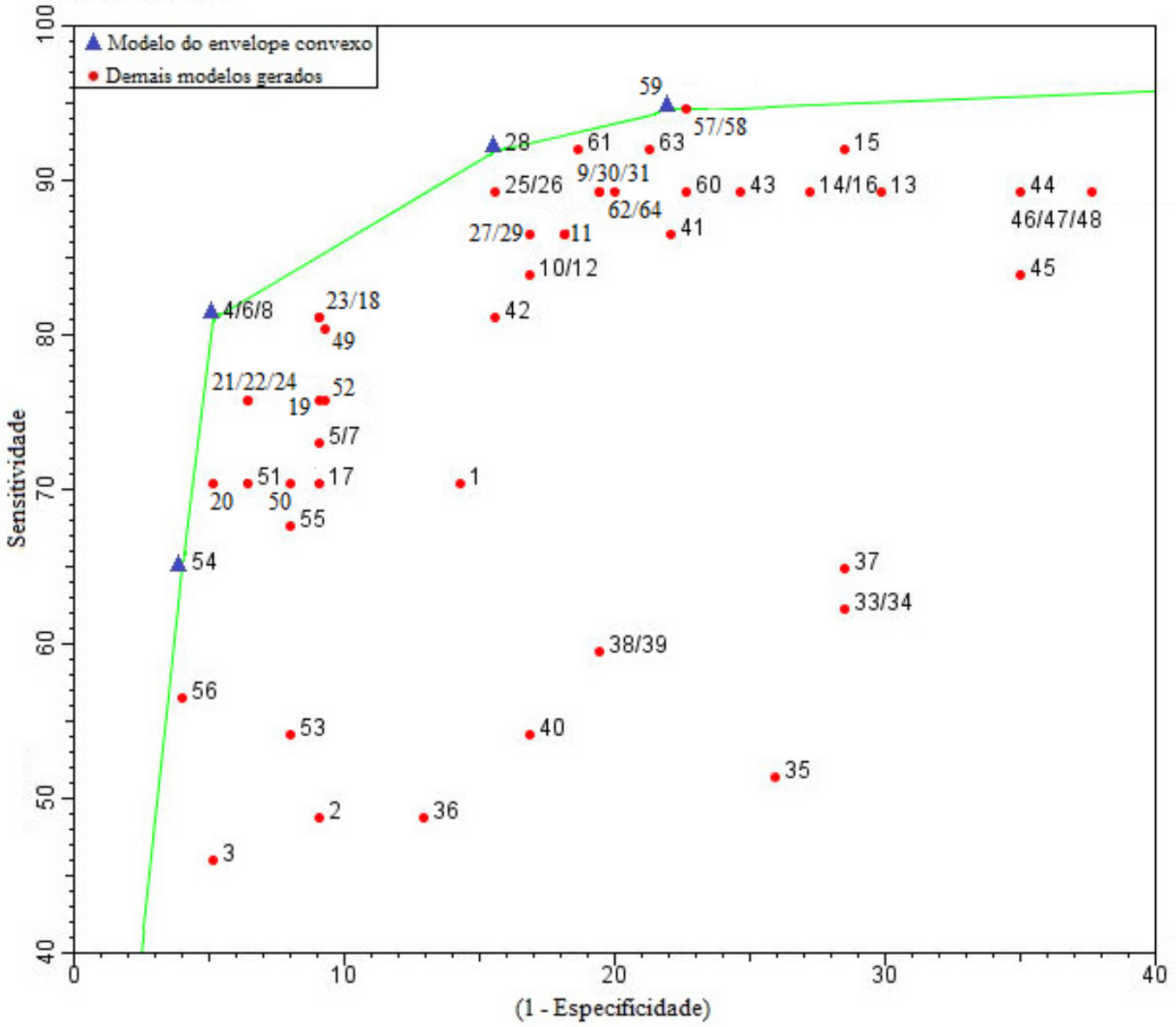


Figura 34: Gráfico ROC para o cenário Varginha-Novo-alta-tx10.

Tabela 37: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Varginha-Novo-alta-tx10.

	Modelos			
Atributos	4/6/8	28	54	59
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				
NHUR90_PINF		*		
SMT_NHUR90_PINF				
THUR90_PINF				
NHNUR90_PINF				
SMT_NHNUR90_PINF				
TMAX_PI_PINF				*
TMIN_PI_PINF				*
TMED_PI_PINF				*
VVENTO_PINF				
SMT_VVENTO_PINF		*		

Tabela 38: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Varginha-Novo-alta-tx10.

Modelos	4/6/8	28	54	59
Técnica	AD	RF	SVM	SVM
Método de seleção	WRP/Chi^2/IG	WRP	Chi^2	M3
Taxa de acerto	90,4	86,8	85,7	83,3
Erro	9,6	13,2	14,3	16,7
Sensitividade	81,1	91,9	64,9	94,6
Especificidade	94,8	84,4	96,0	77,9
Confiabilidade Positiva	88,2	73,9	88,9	67,3
Confiabilidade Negativa	91,3	95,6	84,7	96,8
TP Rate	81,1	91,9	64,9	94,6
FP Rate	5,2	15,6	4,0	22,1
AUC	0,821	0,888	0,804	0,863
Kappa	0,78	0,72	0,65	0,66

A.3.3 Cenário Varginha-Novo-baixa-tx5

O gráfico ROC da Figura 35 representa o desempenho dos modelos desenvolvidos para o cenário Varginha-Novo-baixa-tx5, onde os modelos selecionados no envelope convexo encontram-se destacados. As medidas de avaliação referentes aos modelos do envelope convexo estão dispostas Tabela 40 e os atributos utilizados no conjunto de dados que gerou estes modelos estão na Tabela 39.

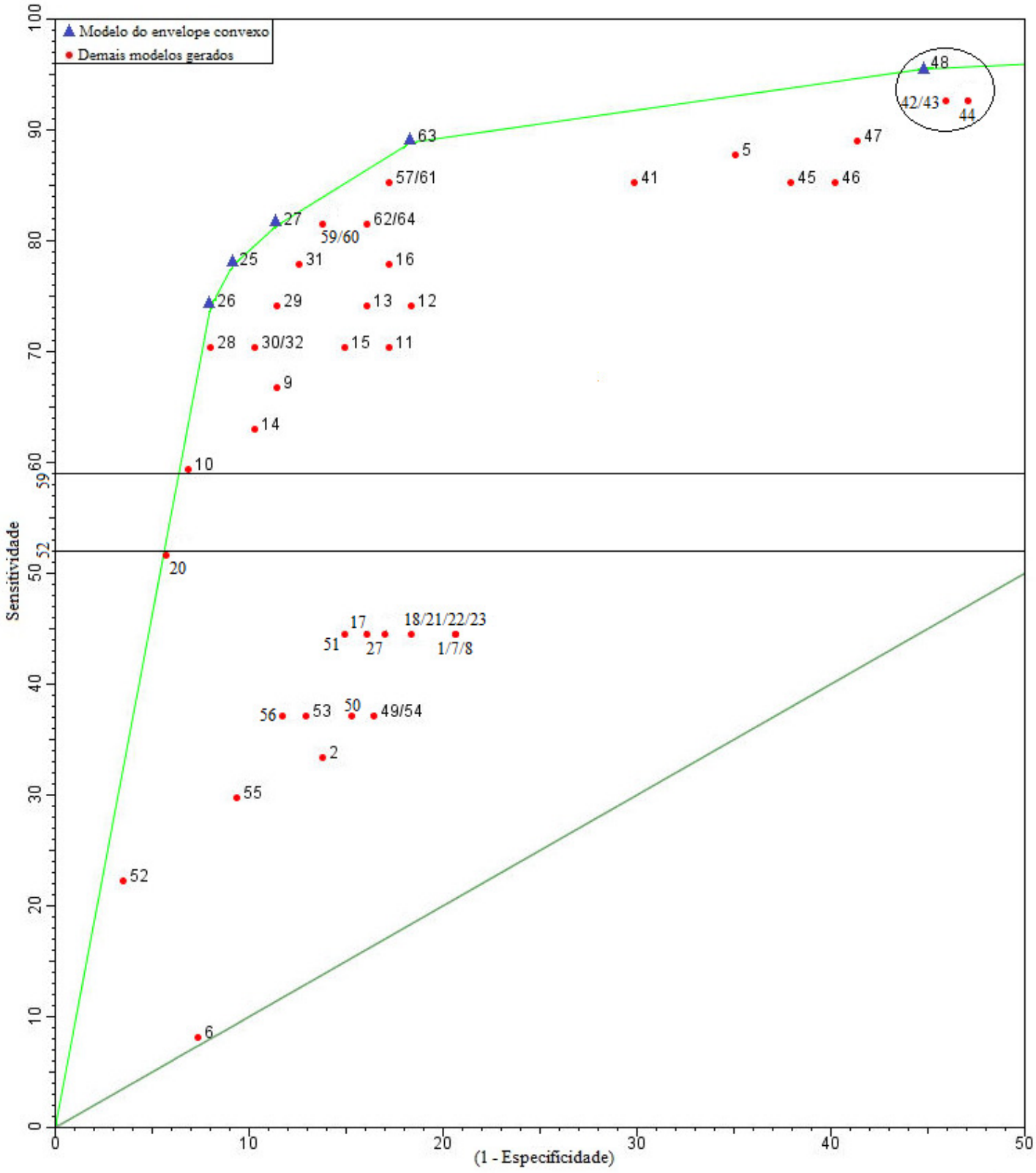


Figura 35: Gráfico ROC para o cenário Varginha-Novo-baixa-tx5.

Tabela 39: Atributos utilizados no conjunto de dados que gerou os modelos selecionados no envelope convexo para o cenário Varginha-Novo-baixa-tx5.

	Modelos				
Atributos	25	26	27	48	63
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	*				
NHUR90_PINF	*	*		*	
SMT_NHUR90_PINF	*				
THUR90_PINF	*	*			
NHNUR90_PINF	*	*		*	
SMT_NHNUR90_PINF	*				
TMAX_PI_PINF	*	*	*		
TMIN_PI_PINF	*	*	*		
TMED_PI_PINF	*	*	*		
VVENTO_PINF	*				
SMT_VVENTO_PINF	*				

Tabela 40: Resultado da avaliação para os modelos selecionados no envelope convexo para o cenário Varginha-Novo-baixa-tx5.

Modelos	25	26	27	48	63
Técnica	RF	RF	RF	RNA	SVM
Método de seleção	M1	M2	M3	IG	GR
Taxa de acerto	87,7	87,7	86,8	64,1	83,3
Erro	12,3	12,3	13,2	35,9	16,7
Sensitividade	77,8	74,1	81,5	92,6	88,9
Especificidade	90,8	91,2	88,5	55,2	81,6
Confiabilidade Positiva	72,4	74,1	68,8	39,1	60,0
Confiabilidade Negativa	92,9	91,2	93,9	96,0	96,0
TP Rate	77,8	74,1	81,5	92,6	88,9
FP Rate	9,2	8,1	11,5	44,8	18,4
AUC	0,893	0,891	0,890	0,683	0,853
Kappa	0,67	0,66	0,66	0,32	0,61

Apêndice B – Quantidade de registros no conjunto de dados

Este apêndice contém informações sobre a quantidade dos registros obtidos no conjunto de dados final utilizado na modelagem. O cálculo destes registros é mostrado de acordo com cada uma das cidades citadas na seção 4.1.1. Devido à dinâmica de funcionamento do algoritmo, sempre os dois últimos meses relativos a cada cidade foram descartados. As falhas mencionadas estão na Tabela 5.

Varginha:

Varginha tem lavouras de alta e baixa carga pendente de frutos, além de ser em espaçamento largo e adensado. Assim, de outubro de 1998 até outubro de 2011, tem-se 12 anos (outubro 1998 até setembro de 2011) mais um mês (outubro de 2011), totalizando 157 registros. Houve sete falhas para Varginha, sendo que uma delas foi para setembro de 2011, assim deve-se descontar as sete falhas, mais o mês de outubro de 2011, totalizando 8 meses descartados. Chega-se ao valor de 149 meses com registros, que deve ser multiplicado por quatro combinações de espaçamento e cargas, totalizando 596 registros.

Carmo de Minas:

Carmo de Minas tem lavouras de alta e baixa carga pendente de frutos, mas somente com espaçamento adensado. Assim, de dezembro de 2006 a outubro de 2011, tem-se 4 anos (dezembro de 2006 até novembro de 2010) mais 11 meses restantes (dezembro de 2010 até outubro de 2011), totalizando 59 meses. Deve-se descartar os meses de setembro e outubro de 2011, além do que este conjunto apresentou uma falha, totalizando 3 meses excluídos. Chega-se ao valor de 56 meses, o qual deve ser multiplicado por duas opções de carga da lavoura, chegando a 112 registros.

Boa Esperança:

Boa Esperança tem lavouras de alta e baixa carga pendente de frutos, mas somente com espaçamento largo. Assim, de junho de 2010 a outubro de 2011, tem-se 17 registros. Deve-se descartar os meses de setembro e outubro de 2011, totalizando dois meses excluídos. Chega-se ao valor de 15 registros, o qual deve ser multiplicado por duas cargas das lavouras, chegando a 30 registros.

Somando-se todos estes, chega-se ao valor de 738, o qual é o número final de registros no conjunto de dados.