Gabriel Jorge Chahine

Mineração de Dados para modelagem de risco de Metástase em Tumor de Próstata

Campinas
2013

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

Gabriel Jorge Chahine

Mineração de Dados para modelagem de risco de Metástase em Tumor de Próstata

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Matemática Aplicada.

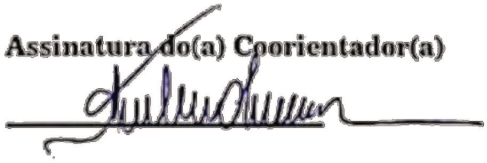
Orientador(a): Prof. Dr. Laercio Luis Vendite

Coorientador(a): Prof. Dr. Stanley Robson de Medeiros Oliveira

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO Gabriel Jorge Chahine, E ORIENTADA PELO Prof. Dr. Laercio Luis Vendite

Assinatura do(a) Orientador(a)

Assinatura do(a) Coorientador(a)



CAMPINAS
2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

C347m

Chahine, Gabriel Jorge, 1982-
Mineração de dados para modelagem de risco de metástase em tumor de próstata / Gabriel Jorge Chahine. – Campinas, SP : [s.n.], 2013.

Orientador: Laercio Luis Vendite.
Coorientador: Stanley Robson de Medeiros Oliveira.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

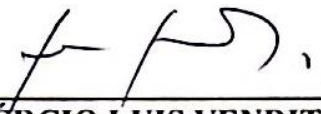
1. Mineração de dados (Computação). 2. Próstata - Tumores. 3. Árvores de decisão. 4. Modelagem de dados. 5. Máquina de vetores de suporte. I. Vendite, Laercio Luis,1954-. II. Oliveira, Stanley Robson de Medeiros. III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Data mining for the modeling of metastasis risk on prostate tumor
Palavras-chave em inglês:
Data mining
Prostate - Tumors
Decision trees
Data modeling
Support vectors machine
Área de concentração: Matemática Aplicada e Computacional
Titulação: Mestre em Matemática Aplicada e Computacional
Banca examinadora:
Laercio Luis Vendite [Orientador]
Ubirajara Ferreira
Rodney Carlos Bassanezi
Data de defesa: 22-10-2013
Programa de Pós-Graduação: Matemática Aplicada e Computacional

Dissertação de Mestrado defendida em 22 de outubro de 2013 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.



Prof.(a). Dr(a). LAÉRCIO LUIS VENDITE

Prof.(a). Dr(a). UBIRAJARA FERREIRA

Prof.(a). Dr(a). RODNEY CARLOS BASSANEZI

Abstract

Of all the cancers of the urinary tract, the most common are the Prostate and Bladder. The first being the most common cause of death by cancer and the most common carcinoma in men. Our goal in this work is to develop predictive models to determine whether a given tumor will grow and invade other organs or, if it doesn't present this risk and will remain constrained. To do this, we collected data from patients with prostate cancer and assessed which variables were the most responsible for the occurrence of metastasis. Hence, we built predictive models that, with the data of a given patient, are able detect whether or not a distant metastasis would occur in. In this work we present models to predict the occurrence of metastasis in prostate cancer. The simulations were made with the data given by prof. Dr. Ubirajara Ferreira, responsible for the disciplines of Urology from Unicamp's Faculty of Medical Sciences.

Resumo

Dos canceres do trato urinário, os mais comuns são os de Próstata e de Bexiga, sendo o primeiro a causa mais comum de morte por câncer e o carcinoma mais comum para homens. Nosso objetivo nesse trabalho é desenvolver modelos para determinar se um dado tumor irá aumentar e invadir outros órgãos ou se não apresenta esse risco e permanecerá contido. Para isso, coletamos dados de pacientes com câncer de próstata e analisamos quais variáveis mais impactam para ocorrência de metástase. Com isso construímos modelos de classificação, que, com os dados de um determinado paciente, detectam se naquele caso haverá ou não metástase a distância. Nesse trabalho apresentamos modelos para predição de ocorrência de metástases em câncer de próstata. As simulações foram feitas com dados cedidos pelo prof. Dr. Ubirajara Ferreira, responsável pela disciplina de Urologia da FCM da Unicamp, do Hospital das Clinicas – UNICAMP.

Sumário

Em memória de Patrício Letelier e Pietro Manes

When unsureness reigns,
Born from the innocence of youth.
Love alone could not hold together
the inconstancy between two callow lovers.

Cloaked behind a velvet veil of aloof lies,
There was a desire that never died.
For the time between us,
made eternal by our souls,
ripped apart by innocence lost on rancor.

After the farewell of that day
I met the agony of nevermore
be reflected in the light of your eyes.

Para Adriana Brandão e meus pais

Agradecimentos

Pode já ser de praxe dizer isso, mas não diminui a significância e a verdade, de que muitas pessoas, demais, tiveram contribuição em minha vida para que eu viesse a concluir esse trabalho, e por mais que eu queira, nunca conseguiria listar todos aqui. E como para mim é impossível listar tantas queridas pessoas em qualquer forma de importância, tento colocar em uma ordem razoavelmente cronológica.

Aos meus pais, obviamente pois sem eles nada existiria, mas também pela família de cientistas e professores de onde vim. Agradeço aos meus pais por, desde bebê, me imergirem no mundo acadêmico e me mostrarem a graça da pesquisa cientifica. Agradeço também comicamente pela peculiaridade de ter sido uma criança cuja uma das primeiras palavras foram "*Álgebra*" e "*Cálculo*".

Talvez, em seguida, agradeço à gerente da área onde trabalhei no Itaú-Unibanco com modelagem e CRM, por quase cinco anos após me formar. Minha chefe Carol Parra. Não fosse por nossos desentendimentos talvez eu estivesse ainda hoje fazendo os mesmos modelos para vender os produtos do banco e preenchendo planilhas de dados.

Na Unicamp, em busca de uma área de pesquisa; agradeço meu orientador e amigo Laercio Vendite. Pelo companheirismo e carisma desde meus tempos de graduação. Pelo suporte, enquanto perambulava pelo instituto procurando orientação. Pelo reconhecimento em depositar tamanha confiança em mim nessa tese. Pela oportunidade de trabalhar com um assunto tão interessante e cativante.

Agradeço ao professor Stanley Oliveira, por sumir com a minha presunção de achar que sabia "muito bem, obrigado" sobre mineração de dados e modelagem; e me mostrar um mundo de novos conhecimentos. Agradeço pelas aulas e palestras que me ensinaram tanto sobre modelagem. Agradeço ao tamanho conhecimento que me proporcionou.

Agradeço de coração ao professor Doutor Ubirajara Ferreira, responsável pela Disciplina de Urologia na Faculdade de Ciências Médicas da Unicamp, sem o qual esse trabalho teria sido imensamente mais árduo, senão

quase impossível. Infelizmente não tive a oportunidade de conhece-lo melhor pessoalmente e nosso pouco relacionamento direto foi breve, porém de imensa importância. Agradeço a todos os membros de sua equipe pela cessão dos dados e pela colaboração para que esse trabalho pudesse ser realizado. Saliento uma de suas orientandas, a Marcela Duran, uma das grandes responsáveis por eu conseguir acesso ao banco de dados necessário e um doce de pessoa. Agradeço e admiro todo o trabalho que ela tem realizado. Devo salientar que tal projeto faz parte da parceria estabelecida entre o IMECC e o Grupo de UroOncologia da Unicamp.

Agradeço ao meu amigo virtual *Doubzstep,* quem nunca conheci pessoalmente, por toda ajuda com programas e auxilio com computador. Sem o qual esse trabalho poderia ter demorado meses a mais.

A minha amiga de mestrado Kelly pela ajuda com toda burocracia e sem a qual teria perdido muitos prazos.

Agradeço a CAPES pelo suporte financeiro.

E claro, por todo esse período, desde meu tempo fora da área acadêmica, até o final desse artigo; agradeço de todo meu coração e alma à Adriana Brandão. Pela infinita fonte de força e inspiração que sempre foi pra mim; por toda dedicação e carinho com a gente e pelo suporte em tudo que fiz. Por aliviar minha mente nos momentos de agonia. Agradeço pelos maravilhosos anos em que passamos juntos.

Lista de Figuras

Lista de Tabelas

1 Introdução

Modelos matemáticos tem sido aplicados nas mais diversas áreas do mundo real, às mais diversas finalidades; e, nisso temos incluído de ampla forma a medicina. Trabalhos anteriores com o intento de modelar tumores através da metodologia Fuzzy já foram desenvolvidos no grupo da Unicamp, resultantes da parceria estabelecida entre o IMECC e a área de UroOncologia do hospital das clínicas.

O trabalho de *SILVEIRA, Graciele Paraguaia; VENDITE, Laercio Luis, Aplicação da teoria de conjuntos fuzzy na predição do estadiamento patologico do cancer de prostata* [1] consiste no desígnio de construir um modelo baseado em regras Fuzzy com o intuito de predizer o estadiamento patológico do câncer de próstata, utilizando como variáveis o estado clínico, nível de PSA e score de Gleason.

Um trabalho similar é o de *CASTANHO, Maria Jose de Paula; YAMAKAMI, Akebo, Construção e avaliação de um modelo matematico para predizer a evolução do cancer de prostata e descrever seu crescimento utilizando a teoria dos conjuntos fuzzy* [2] , que também utiliza a metodologia de conjuntos Fuzzy baseada na análise de ROC (Receiver Operating Characteristic) para discriminar pacientes com o tumor confinado à próstata daqueles com o tumor não confinado.

Ainda na mesma linha de modelagem Fuzzy, outros trabalhos similares que buscam diferenciar pacientes com tumores confinados daqueles com tumores tendentes a metástase são encontrados em *SILVEIRA, Graciele Paraguaia et al, A metodologia ROC na avaliação de um modelo fuzzy de predição do estádio patológico do tumor de próstata* [3] , *CASTANHO, M. J.P. et al, Fuzzy Receiver Operating Characteristic Curve: An Option to Evaluate Diagnostic Tests* [4], *MARIA JOSÉ DE PAULA CASTANHO, Laécio Carvalho de Barros, Fuzzy expert system: An example in prostate cancer* [5]

1.1 Conceitos

As células do corpo humano se reproduzem pelo processo de divisão celular, processo responsável pela formação, crescimento e regeneração dos tecidos do corpo. Em condições normais, essa reprodução celular é controlada e saudável (e indispensável) ao corpo; no entanto, há situações nas

quais essas células crescem sem controle e podem até invadir tecidos vizinhos.[6]

Quando as células crescem, mas não invadem os tecidos vizinhos, ocorre o chamado de *hiperplasia,* um crescimento que pode até vir a causar menores complicações, mas ainda é considerado normal. A *Neoplasia,* no entanto, é a situação onde os mecanismos regulatórios da proliferação da célula deixam de funcionar corretamente. É uma anormalidade na qual as células crescem contínua e desenfreadamente. O acúmulo de células neoplásticas pode vir a se tornar uma massa volumosa e é denominado *tumor*, o qual pode comprometer o órgão de origem e até mesmo infiltrar em outros órgãos.[7]

Os tumores podem ser divididos em dois tipos, Benigno e Maligno. No *Tumor Benigno* as células crescem mais lentamente, (ainda) não invadiram outros tecidos e tem estruturas semelhantes às do tecido original. Geralmente tem tratamento completo e satisfatório, não voltando a ocorrer. O *Tumor Maligno é* uma anormalidade na qual as células crescem continuamente, sem freio. Elas crescem muito mais rápidas que o normal, de forma desordenada, além disso, suas características como tamanho, forma e até mesmo funções se alteram com as mutações e tendem a invadir órgãos vizinhos, podendo se espalhar por diversas regiões do corpo. É também chamado de câncer. [8]

Quando nos referimos a um tumor como maligno, queremos dizer que o tumor está crescendo desenfreadamente, as células se diferenciaram do tecido original e (se ainda não o fizeram) ameaçam invadir tecidos e órgãos vizinhos, um fenômeno chamado de metástase a distância. Metástase então, é a propagação do câncer de um órgão (ou parte dele) para outro, via corrente sanguínea, sistema linfático ou mesmo por extensão direta devido ao tamanho do tumor.

1.2 Próstata

A próstata é uma glândula que se situa encostada na base da bexiga, o órgão muscular onde a urina é armazenada e da qual sai a uretra, a qual passa pelo interior da próstata e do pênis. É um órgão de tamanho insignificante quando comparado ao tamanho do corpo e pesa aproximadamente 15 gramas em adolescentes; porém, com o passar dos anos, começa a crescer e pode gerar problemas que comprometem a qualidade de vida de homens mais velhos.

Quando essa glândula cresce de forma razoável, sem se descontrolar (hiperplasia) e não invade tecidos vizinhos, ocorrem sintomas como aumento na frequência das micções e esforço para urinar. Essa situação é bastante comum em adultos acima dos 50 anos e não apresenta maiores riscos. A *Neoplasia,* no entanto, é a situação onde os mecanismos regulatórios da proliferação da célula deixam de funcionar corretamente, ocasionando o tumor e até câncer.

Nos casos de câncer, além do aumento do volume da próstata, surgem nódulos que podem ser sentidos no exame de toque. Diferentemente do crescimento benigno, o tumor pode não ficar restrito à glândula e invadir os tecidos vizinhos. O câncer de próstata é o tumor mais comum do trato urinário e a neoplasia mais comum em homens. O ministério da saúde estima que em 2013 haverá 60 mil novos casos de câncer de próstata.[9]

1.3 Objetivos

Quando se trata de taxas de sobrevivência de câncer, o diagnóstico e o tratamento precoce são fundamentais. Se o paciente é levado ao caminho errado por um erro de diagnóstico, os custos de tratamento junto como as chance de morte aumentam consideravelmente. Quanto mais avançado for o estágio de um tumor, maiores as doses de quimioterapia e radioterapia necessárias para o tratamento. Este árduo tratamento é muito pior do que as opções de tratamento para uma fase mais inicial do câncer; além de poderem causar efeitos colaterais graves e acabar levando o paciente a ter uma

qualidade de vida ruim. Assim, quanto mais cedo for possível detectar um câncer, maior as chances de um controle mais eficaz e com isso os resultados para os pacientes são melhores.[10] Logo, ferramentas que possam auxiliar nos exames, diagnóstico ou na tomada de decisões, beneficiam diretamente o tratamento.

Nesse trabalho temos como objetivo desenvolver modelos preditivos para determinar se um tumor de próstata irá crescer ou não. I.e. se há o risco de metástase a distância e o tumor irá aumentar e invadir outros órgãos ou se não apresenta esse risco e permanecerá contido.

Esses modelos podem servir como auxílio à um médico na agilização de tomada de decisões, assim como também permitir que em locais sem a presença de um especialista, ou mesmo em emergências, um rápido diagnóstico possa ser dado e uma decisão mais assertiva possa ser tomada

Para a construção dos modelos, coletamos dados de pacientes já diagnosticados com câncer de próstata, analisamos as variáveis cadastrais e de diagnóstico para entender quais delas mais impactam para que o tumor progrida, ocorrendo a metástase. Fizemos uma análise de Chi-Quadrado e análises de correlação para visualizar e entender melhor a relação entre as variáveis e quais as mais significativas para serem usadas.

Com essas informações, criamos modelos de classificação que utilizam essas variáveis como entrada e assim detectam se naquele caso haverá ou não metástase a distância. Utilizamos os métodos de classificação por Árvore de Decisão e Support Vector Machine (SVM). Comparamos os índices de performance para cada modelo e escolhemos o melhor classificador para cada caso.

1.4 Terminologias

Para que possamos analisar as variáveis e desenvolver os modelos, é fundamental que entendamos alguns conceitos e terminologias básicos usados na Oncologia. Um desses conceitos é o de *Estadiamento*, que é essencial para a compreensão e prognóstico do paciente. No Estadiamento

são levados em consideração dados referentes a extensão (tamanho) do tumor, a presença de linfonodos (gânglios linfáticos) afetados e metástases.

O sistema atual mais usado é o TNM, de 2002 da União Internacional Contra o Câncer (UICC).[11] O sistema utiliza as letras TNM para designar o tamanho do tumor primário (T), a extensão em linfonodos regionais (N) e a presença de Metástases (M) (se o tumor se alastrou para outros tecidos).

Existem pequenas diferenças descritivas nas variáveis de Estadiamento dependendo do tipo de câncer pois apresentam sintomas e comportamentos diferentes, mas o fundamento é similar. 'T' varia de T0 a T4; 'N' vai de Nx a N1 e 'M' de Mx a M1. A descrição completa pode ser vista na tabela 1.1.

Outro conceito muito importante, usado somente em tumores de próstata, é o Score de Gleason (ou escala ou pontuação de Gleason). É uma pontuação dada a um câncer de próstata baseada em sua aparência microscópica. O score de Gleason é importante porque scores maiores estão associados a piores prognósticos.[12]

T0:	Sem evidência de Tumor
T1:	Tumor Presente, mas não detectável com imagem
T1a:	Tumor encontrado incidentalmente em menos de 5% do tecido da próstata
T1b:	Tumor encontrado incidentalmente em mais de 5% do tecido da próstata
T1c:	Tumor encontrado em uma biópsia feita devido a alto nível de PSA
T2:	Tumor pode ser sentido no exame de toque, mas não se espalhou para fora da próstata
T2a:	Tumor engloba metade de um dos lobos ou menos
T2b:	Tumor engloba mais da metade de um dos lobos, mas não os dois.
T2c:	Tumor engloba ambos lobos
T3:	Tumor se espalhou para além da cápsula prostática
T3a:	Tumor não envolve a vesícula seminal
T3b:	Tumor envolve a vesícula seminal
T4:	Tumor invadiu estruturas vizinhas (colo vesical, esfíncter, reto, músculos elevadores ou parede)
Nx:	Linfonodos regionais não avaliados
N0:	Não houve invasão a linfonodos regionais
N1:	Houve invasão a linfonodos regionais
Mx:	Metástase a distância não foram avaliadas
M0:	Não há metástase a distância
M1:	Houve metástase a distância
M1a:	Tumor se espalhou para linfonodos além dos regionais
M1b:	Tumor invadiu o osso
M1c:	Tumor invadiu outros órgãos

Tabela 1.1: Classificação TNM

Para determinar o score de Gleason, uma peça de tecido prostático deve ser obtida por meio de biópsia. Um patologista examina a amostra da biópsia e fornece um score baseado em dois padrões, indo de 1 a 5 para cada padrão. O primeiro, chamado de grau primário, representa a maior parte do tumor. O segundo, grau secundário, está relacionado com a menor parte do tumor. Estes escores são então somados para se obter o escore final de Gleason. Dessa forma, o score de Gleason pode variar de 2 (score 1 para ambos graus) a 10 (score 5 para ambos). O primeiro valor de um score Gleason é mais

significativo na avaliação da gravidade do tumor do que o segundo; assim, um score de 7 = 4+3 é mais grave do que um 7 = 3+4.

O Antígeno Prostático Específico (PSA do inglês Prostate-Specific Antigen) é uma enzima com algumas características de marcador tumoral ideal, sendo utilizado para diagnóstico, e controle da evolução do carcinoma da próstata.[13] O PSA é uma proteína produzida pela glândula prostática[14] e o teste de PSA mede os níveis de PSA no sangue. Como essa enzima é produzida pelo corpo e pode ser usado para detectar doenças, é chamado de marcador biológico.

Visto que inúmeras condições adversas como hipertrofia benigna da próstata, inflamação, infecção, idade, vida sexual e até mesmo etnia podem influenciar no nível de produção da enzima; níveis isolados de PSA não dão informações suficientes para distinguir entre condições benignas e malignas de um possível tumor de próstata, sendo muito mais importante acompanhar a tendência dos níveis de PSA e assim avaliar se há aumento ou diminuição da enzima com o tempo.

Há controvérsias em relação a eficácia do uso do teste do PSA na detecção de câncer de próstata, apesar de estudos já terem apontado sua eficácia. O European Randomized Study of Screening for Prostate Cancer publicou em Março de 2009, um estudo que comprovou uma redução de 20% nos casos de mortalidade por câncer de próstata devido a análise do PSA.[15] Enquanto detecção de câncer não é o escopo desse artigo, mostraremos que níveis altos de PSA estão altamente relacionados com o avanço e aumento do tumor.

1.5 Metodologia Pré-Modelagem

Utilizamos a classificação por Árvore de Decisão e Support Vector Machine (SVM) para predizer se em cada caso haverá ou não metástase, i.e., se o tumor irá progredir e invadir outros órgãos do corpo. Comparamos os modelos e escolhemos o de melhor performance.

A priori, fizemos uma análise de correlação e uma pelo teste do Chi-Quadrado para identificar se havia relação entre as variáveis e se algumas

poderiam ser descartadas. Também rodamos os modelos com todas as variáveis iniciais para compararmos.

A base de dados provém do Hospital das Clínicas da Unicamp, da área de Urooncologia, e conta com dados de 1061 pacientes até o final de 2011, distribuídos pelo tipo de diagnóstico como a seguir:

Diagnostico	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	53	5.00%	53	5.00%
01- NEOPLASIA DE PROSTATA	648	61.07%	701	66.07%
02- NEOPLASIA DE BEXIGA	118	11.12%	819	77.19%
03- NEOPLASIA DE PELVE RENAL	8	0.75%	827	77.95%
05- NEOPLASIA RENAL	87	8.20%	914	86.15%
06- NEOPLASIA DE PÊNIS	31	2.92%	945	89.07%
07- NEOPLASIA DE TESTÍCULO	38	3.58%	983	92.65%
08- NEOPLASIA DE URETRA	1	0.09%	984	92.74%
09- TUMOR DE ADRENAL	5	0.47%	989	93.21%
10- HPB	31	2.92%	1020	96.14%
11- ANGIOMIOLIPOMA	9	0.85%	1029	96.98%
12- OUTROS	23	2.17%	1052	99.15%
13- ONCOCITOMA	4	0.38%	1056	99.53%
14- CISTO RENAL	5	0.47%	1061	100.00%

Tabela 1.2: Distribuição de Diagnósticos do HC

Dessa forma, temos 648 pacientes com tumor de Próstata, correspondendo a 61% do total de pacientes.

No banco de dados temos informações sobre o Estadiamento do Tumor, último nível de PSA e score de Gleason, assim como o sexo e a idade do paciente. Outras variáveis incluem Obesidade, Hipertensão, Problemas Cardíacos, etc. As variáveis completas são apresentadas a seguir.

Hc	Gleason	Tabagismo
sexo	PSA	Etilismo
Diagnostico	Anato_Pato	Obesidade
D_Nasc	Qualls	Colesterol
idade	Hipertensao	Outros
Estadiamento	Diabetes	Radioterapia
Tumor	Cardiopatia	Hormonioterapi
Nodo	Prob_Pulmao	Quimioterapia
Metastase	Insuf_Renal	Prostatectomia

Tabela 1.3: Variáveis do banco de dados

Aqui, a variável *Estadiamento* foi aberta em *Tumor*, *Nodo* e *Metástase* para que possamos trabalhar separadamente com o tamanho do tumor, se houve difusão para linfonodos locais e se houve ou não metástase. Assim, *metástase* representa a variável alvo para qual queremos fazer a previsão.

1.6 Análises Primárias e Outliers

Uma análise estatística inicial das variáveis numéricas Idade, PSA e Gleason nos mostra os outliers e os pontos de corte.

Variable	N	N Miss	Minimum	Maximum	Mean	Mode	1st Pctl	5th Pctl	25th Pctl	Median	75th Pctl	95th Pctl	99th Pctl
PSA	518	130	0	1713	27.209768	100	0.57	2.8	6.08	10.81	24.9	99.99	99.99
Gleason	551	97	0	10	6.8439201	7	4	6	6	7	7	9	9
idade	638	10	-7	99	74.299373	76	50	59	69	76	81	87	91

Tabela 1.4: Estatísticas de PSA, Gleason e Idade

Queremos ressaltar a importância de se assegurar a qualidade dos dados de um paciente, aqui vemos pacientes com a data de nascimento depois de 2012, a base completa contém ainda mais erros de idade. Para a variável idade, faremos o corte de outlier no primeiro e último percentil na tabela geral, deixando assim o mínimo e o máximo para idade igual a 50 e 91 respectivamente. Para o score de Gleason cortaremos o mínimo no 5º percentil e o máximo será deixado em 10, pois representa um valor aceitável. Para o nível de PSA também faremos o corte no 5º e no 99º percentil.

1.6.1 Tumor

Quando analisamos a frequência dos diferentes tamanhos de tumor (Estadiamento 'T'), referente ao tamanho do tumor primário, encontramos os seguintes resultados.

Tumor	Frequency	Percent	Cumulative Frequency	Cumulative Percent
x	166	25.62%	166	25.62%
1	189	29.17%	355	54.78%
2	196	30.25%	551	85.03%
3	93	14.35%	644	99.38%
4	4	0.62%	648	100.00%

Tabela 1.5: Frequência dos Diferentes tamanhos de Tumores

Aqui '*x*' representa tanto um tumor que não pôde ser detectado quanto a falta de informação. Devido a baixa frequência de T3 e T4 e a semelhança de serem os dois mais agressivos, agruparemos os dois em um só.

1.6.2 Linfonodos

Aqui analisamos as evidências do Estadiamento '*N*', do quanto um tumor atinge linfonodos regionais.

Nodo	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	512	79.01%	512	79.01%
0	122	18.83%	634	97.84%
1	11	1.70%	645	99.54%
2	1	0.15%	646	99.69%
3	2	0.31%	648	100.00%

Tabela 1.6: Frequência de linfonodos

Devido ao baixo número de N1, N2 e N3 iremos agrupa-los como um só. Dessa forma, dividiremos apenas entre, se linfonodos regionais foram afetados ou não (e a falta de informação ou não-detecção).

1.6.3 Metástases

Frequência de ocorrência de metástase.

Metastase	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	235	36.27%	235	36.27%
0	344	53.09%	579	89.35%
1	69	10.65%	648	100.00%

Tabela 1.7: Frequência de Metástases

Temos apenas 69 evidencias de metástases; e infelizmente muitos dados faltantes, o que compromete o estudo e os modelos, pois teremos de excluir os registros que não tem a informação.

1.6.4 Demais Variáveis

A ocorrência das demais variáveis não diretamente relacionadas com a análise do tumor, como Tabagismo, Diabetes, Hipertensão, etc., são apresentadas a seguir

Variable	N	%
Hipertensao	286	44.14%
Tabagismo	149	22.99%
Diabetes	73	11.27%
Etilismo	45	6.94%
Cardiopatia	40	6.17%
Obesidade	21	3.24%
Prob_Pulmao	19	2.93%
Insuf_Renal	7	1.08%

Tabela 1.8: Demais variáveis

Aqui, novamente temos o problema de falta de dados, devido tanto a falhas no preenchimento da ficha dos pacientes, quanto ao fato de muitos pacientes não admitirem de beber ou fumar em excesso.

1.7 Balanceamento das classes

Nosso objetivo nesse artigo é determinar se um dado tumor irá ou não aumentar e progredir, i.e., se deixará de estar contido apenas na próstata e chegará a atingir tecidos e órgãos vizinhos; e, como vimos anteriormente em nossa base de dados, temos apenas 15% de casos de tumor de próstata que resultaram em metástase a distância (desconsiderando os dados faltantes). Enquanto gostaríamos muito que esse número fosse o menor possível na realidade (preferencialmente zero), para uma análise de dados isso é um percentual pequeno, resultando em um desbalanceamento de classe que acaba por prejudicar os modelos.

Para cuidar desse problema aplicamos o método SMOTE de over-sampling para a classe minoritária (metástase = 1). O método consiste em inserir novas instâncias por interpolação. Para cada instância da classe minoritária, um novo exemplo é introduzido ao longo da reta que a liga com seus k vizinhos (da mesma classe) mais próximos.[16] Maiores detalhes sobre o método SMOTE podem ser vistos no apêndice.

Utilizamos o software WEKA, com a opção de recriar 100% novos dados por interpolação – efetivamente dobrando o número de ocorrências de

metástases – utilizando os 4 vizinhos mais próximos. A tabela 1.9 a seguir ilustra a distribuição da nossa nova base de dados.

Metastase	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	235	32.78%	235	32.78%
0	344	47.98%	579	80.75%
1	138	19.25%	717	100.00%

Tabela 1.9: Frequência de metástases após Over-sampling

Como pode ser visto, passamos de 69 para 138 casos de metástase a distância.

1.8 Análise de Correlação

Uma análise de correlação consiste em identificar a relação entre uma variável dependente e uma ou mais variáveis independentes, é uma ferramenta importante para diferentes áreas do conhecimento, muitas vezes não como resultado final, mas principalmente como uma das etapas para a utilização de outras técnicas de análise. A análise de correlação nos permite visualizar se um par de variáveis são correlacionadas (linearmente ou não), i.e., se o aumento de uma implica na variação da outra (aumentando ou diminuindo).

Fazer uma boa análise de correlação antes da modelagem pode impactar fortemente na construção e qualidade do modelo. Uma das principais contribuições de análise de correlação antes da modelagem é de nos permitir enxergar quais variáveis são mais correlacionadas com a variável-alvo (no caso de uma regressão ou outro modelo preditivo por exemplo), e assim servir tanto como um primeiro filtro como também para ordenar a entrada de variáveis.

A outra grande contribuição é que como enxergamos a correlação de cada variável com as demais (não somente com a variável-alvo), podemos eliminar

o número de variáveis de entrada no modelo. Se duas variáveis forem muito correlacionadas, ambas contribuiriam de forma muito semelhante para a predição do alvo, e o erro gerado por uma variável adicional acaba sendo maior do que a contribuição de predição da variável. Assim, um par de variáveis muito correlacionadas muitas vezes acaba por introduzir mais ruído do que qualidade no modelo.

Existe uma variedade de testes numéricos que quantificam a dependência estatística entre duas variáveis, muitas vezes podemos ficar em dúvida de qual teste usar para qual situação E dessa forma é importante conhecer a teoria dos diferentes métodos e as suposições básicas requeridas de cada para que sejam usados de forma adequada. Temos de analisar os prós e contras de cada um sob a consideração e especificidades do caso em questão.

Um dos testes mais comuns, é o teste de correlação de Pearson,[17] que mede a relação linear entre um par de variáveis, i.e., quantifica (com valor absoluto de 0 a 1) a dependência linear dentre as duas variáveis.[18] Outro teste bastante comum é o teste de correlação de Spearman, que mede a relação entre duas variáveis usando uma função monotônica. Em comum com o teste de Pearson, o coeficiente de Spearman é um valor (absoluto) entre 0 e 1.

1.8.1 Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida paramétrica que mede o grau de correlação linear[19] (e a direção dessa correlação - se positiva ou negativa) entre duas variáveis. Este coeficiente, normalmente representado por ρ assume valores apenas entre -1 e 1. Se uma variável y tende a crescer conforme a variável x cresce, então o coeficiente é positivo; se y decresce conforme x cresce então o coeficiente é negativo. Um coeficiente igual a zero indica que não há tendência alguma para y crescer ou diminuir conforme x cresce.

O coeficiente de correlação de Pearson entre duas variáveis x e y é dado pela fórmula

$$\rho_{x,y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)\sigma_x\sigma_y} = \frac{\sum(xy)-n\overline{xy}}{(n-1)\sigma_x\sigma_y}$$

Onde σ_x e σ_y representam o desvio padrão das variáveis x e y respectivamente

Se $\rho_{x,y} > 0 \implies x$ e y são Positivamente Correlacionados

Se $\rho_{x,y} < 0 \implies x$ e y são Negativamente Correlacionados

Se $\rho_{x,y} = 0 \implies$ Significa que as duas variáveis não dependem linearmente uma da outra. Podendo, no entanto, existir uma correlação não-linear entre elas; e assim cabendo a cada situação avaliar se há necessidade de se investigar a relação por outro métodos.

Quanto mais $|\rho_{x,y}|$ se aproximar de 1, maior a correlação entre x e y; e enquanto cada caso deve ser analisado individualmente, normalmente uma correlação forte é dada para $|\rho| > 0{,}7$; uma correlação moderada para $|\rho|$ entre 0,3 e 0,7 e uma correlação fraca para $|\rho| < 0{,}3$.

A matriz de correlação gerada completa é dada a seguir na tabela 1.10.

Pearson Correlation Coefficients														
	idade	tumor	nodo	Gleason	Hipertensao	Diabetes	Cardiopatia	Prob_Pulmao	Insuf_Renal	Tabagismo	Etilismo	Obesidade	Metastase	psa
idade	1	0.125	0.073	0.135	0.066	0.008	0.016	0.015	-0.009	0.054	-0.061	-0.054	0.221	0.042
tumor	0.125	1	0.103	0.312	-0.001	-0.020	0.074	0.047	-0.024	0.104	0.019	-0.083	0.213	0.019
nodo	0.073	0.103	1	0.156	0.089	0.017	-0.025	0.171	-0.013	-0.035	0.032	0.070	0.061	0.050
Gleason	0.135	0.312	0.156	1	0.001	-0.033	-0.033	0.012	-0.031	0.055	0.019	-0.083	0.322	0.202
Hipertensao	0.066	-0.001	0.089	0.001	1	0.144	0.228	0.031	-0.003	0.114	0.089	0.117	-0.018	-0.039
Diabetes	0.008	-0.020	0.017	-0.033	0.144	1	0.021	-0.010	0.041	0.034	0.039	0.114	-0.114	-0.019
Cardiopatia	0.016	0.074	-0.025	-0.033	0.228	0.021	1	0.034	-0.020	-0.028	-0.004	0.160	-0.071	-0.027
Prob_Pulmao	0.015	0.047	0.171	0.012	0.031	-0.010	0.034	1	-0.018	0.077	-0.045	-0.031	-0.028	-0.062
Insuf_Renal	-0.009	-0.024	-0.013	-0.031	-0.003	0.041	-0.020	-0.018	1	-0.009	-0.027	0.100	-0.020	0.018
Tabagismo	0.054	0.104	-0.035	0.055	0.114	0.034	-0.028	0.077	-0.009	1	0.329	-0.016	0.055	0.161
Etilismo	-0.061	0.019	0.032	0.019	0.089	0.039	-0.004	-0.045	-0.027	0.329	1	0.002	-0.035	-0.044
Obesidade	-0.054	-0.083	0.070	-0.083	0.117	0.114	0.160	-0.031	0.100	-0.016	0.002	1	-0.087	-0.107
Metastase	0.221	0.213	0.061	0.322	-0.018	-0.114	-0.071	-0.028	-0.020	0.055	-0.035	-0.087	1	0.522
psa	0.042	0.019	0.050	0.202	-0.039	-0.019	-0.027	-0.062	0.018	0.161	-0.044	-0.107	0.522	1

Tabela 1.10: Matriz de Correlação de Pearson

À primeira vista não há correlação suficiente entre duas ou mais variáveis para que possamos eliminar alguma. Podemos melhorar a organização da matriz para facilitar a visualização, assim, para cada variável, ordenamos as cinco outras mais correlacionadas.

Pearson Correlation Coefficients						
idade	Metastase	Gleason	tumor	nodo	Hipertensao	Etilismo
	0.22054	0.13503	0.12511	0.07262	0.06568	-0.06081
tumor	Gleason	Metastase	idade	Tabagismo	nodo	Obesidade
	0.31217	0.21307	0.12511	0.10409	0.10273	-0.08301
nodo	Prob_Pulmao	Gleason	tumor	Hipertensao	idade	Obesidade
	0.17093	0.15593	0.10273	0.08896	0.07262	0.07024
Gleason	Metastase	tumor	psa	nodo	idade	Obesidade
	0.32175	0.31217	0.20212	0.15593	0.13503	-0.08338
Hipertensao	Cardiopatia	Diabetes	Obesidade	Tabagismo	Etilismo	nodo
	0.22802	0.14433	0.11665	0.1139	0.08934	0.08896
Diabetes	Hipertensao	Metastase	Obesidade	Insuf_Renal	Etilismo	Tabagismo
	0.14433	-0.11432	0.11408	0.04097	0.03895	0.03416
Cardiopatia	Hipertensao	Obesidade	tumor	Metastase	Prob_Pulmao	Gleason
	0.22802	0.16005	0.07411	-0.07134	0.0339	-0.03276
Prob_Pulmao	nodo	Tabagismo	psa	tumor	Etilismo	Cardiopatia
	0.17093	0.07709	-0.06156	0.0467	-0.04535	0.0339
Insuf_Renal	Obesidade	Diabetes	Gleason	Etilismo	tumor	Cardiopatia
	0.09957	0.04097	-0.03058	-0.02684	-0.02354	-0.01958
Tabagismo	Etilismo	psa	Hipertensao	tumor	Prob_Pulmao	Gleason
	0.32946	0.16102	0.1139	0.10409	0.07709	0.0552
Etilismo	Tabagismo	Hipertensao	idade	Prob_Pulmao	psa	Diabetes
	0.32946	0.08934	-0.06081	-0.04535	-0.04393	0.03895
Obesidade	Cardiopatia	Hipertensao	Diabetes	psa	Insuf_Renal	Metastase
	0.16005	0.11665	0.11408	-0.10666	0.09957	-0.08708
Metastase	psa	Gleason	idade	tumor	Diabetes	Obesidade
	0.52164	0.32175	0.22054	0.21307	-0.11432	-0.08708
psa	Metastase	Gleason	Tabagismo	Obesidade	Prob_Pulmao	nodo
	0.52164	0.20212	0.16102	-0.10666	-0.06156	0.05032

Tabela 1.11: Cinco melhores correlações lineares para cada variável

Pela tabela 1.11 podemos ver claramente que não há correlação significativa entre duas variáveis. Com a variável-alvo *Metástase*, uma relação mais razoável existe com o nível de *PSA* e o score de *Gleason*, algo até de se esperar dentro do que conhecemos, mas ainda assim não significativo para

afirmamos que elas estariam correlacionadas linearmente. Dentre as outras variáveis, os pares *Tabagismo* e *Etilismo*; e *Gleason* e *Tumor* também tem uma correlação um pouco acima da média entre si.

Nenhuma dessas correlações, no entanto, são suficientemente significativas. Apesar de estar acima da média das demais, PSA sozinha não consegue explicar a variável *Metástase*. Etilismo e Tabagismo são conceitos que afetam o organismo de forma bem diferente, por essa razão e pelo coeficiente não suficientemente significativo, serão ambas mantidas e inalteradas.

1.8.2 Correlação de Spearman

Como não encontramos relação linear significativa para nenhum par de variáveis pelo coeficiente de Pearson, usamos a correlação de Spearman, que é uma medida de relação não paramétrica entre duas variáveis. Similarmente ao modelo de Pearson, ela avalia quão bem duas variáveis são relacionadas por um fator de -1 a 1, porém usando uma função monotônica ao invés de uma relação linear; e dessa forma pode encontrar relações não lineares entre duas variáveis. O teste de Spearman é apropriado tanto para variáveis contínuas quanto discretas, incluindo e o teste recomendado para variáveis ordinais.

Quando dizemos que o coeficiente de correlação de Spearman é não-paramétrico, significa tanto que uma amostragem da distribuição entre as duas variáveis pode ser obtida sem o conhecimento dos parâmetros da distribuição de cada uma; como também que uma correlação é encontrada quando duas variáveis são correlacionadas por qualquer função monotônica, ao invés de apenas uma linear, sem que precisemos saber sua equação.

O coeficiente de correlação de Spearman $\boldsymbol{r}$ entre duas variáveis x e y é dado por

$$r_{a,b} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Assim como no teste de Pearson, o coeficiente fica entre -1 e 1 e seu resultado pode ser interpretado de forma similar.

Aa tabela 1.12 mostra a matriz de correlação de Spearman ordenada com as cinco variáveis mais significativas para cada outra.

Como podemos ver, novamente não temos nenhuma correlação significativa, pelo menos acima de 0,6. Semelhante a análise anterior, a variável que mais se relaciona com *Metástase* é *PSA*, a segunda mais correlacionada é *Gleason*, porém, não são suficientes para explicar *Metástase* por elas só. Também revemos a mesma correlação os pares *Tabagismo* e *Etilismo; e Gleason e Tumor;* também uma pequena correlação entre *Hipertensão* e *Cardiopatia.*

Spearman Correlation Coefficients						
idade	Metastase	Gleason	tumor	psa	Tabagismo	nodo
	0.21831	0.11117	0.09763	0.08263	0.08125	0.07356
tumor	Gleason	Metastase	Tabagismo	nodo	idade	psa
	0.33596	0.22416	0.1064	0.10387	0.09763	0.0836
nodo	Prob_Pulmao	Gleason	tumor	Hipertensao	idade	Obesidade
	0.17093	0.13371	0.10387	0.08896	0.07356	0.07024
Gleason	tumor	Metastase	psa	nodo	idade	Obesidade
	0.33596	0.30649	0.18234	0.13371	0.11117	-0.08263
Hipertensao	Cardiopatia	Diabetes	Obesidade	Tabagismo	Etilismo	nodo
	0.22802	0.14433	0.11665	0.1139	0.08934	0.08896
Diabetes	Hipertensao	Metastase	Obesidade	Gleason	Insuf_Renal	Etilismo
	0.14433	-0.11432	0.11408	-0.04843	0.04097	0.03895
Cardiopatia	Hipertensao	Obesidade	tumor	Metastase	Prob_Pulmao	Gleason
	0.22802	0.16005	0.07286	-0.07134	0.0339	-0.03003
Prob_Pulmao	nodo	psa	Tabagismo	Etilismo	tumor	Cardiopatia
	0.17093	-0.09593	0.07709	-0.04535	0.0442	0.0339
Insuf_Renal	Obesidade	Diabetes	Gleason	Etilismo	tumor	Cardiopatia
	0.09957	0.04097	-0.0319	-0.02684	-0.02275	-0.01958
Tabagismo	Etilismo	Hipertensao	tumor	psa	idade	Prob_Pulmao
	0.32946	0.1139	0.1064	0.09872	0.08125	0.07709
Etilismo	Tabagismo	Hipertensao	Prob_Pulmao	Diabetes	Metastase	idade
	0.32946	0.08934	-0.04535	0.03895	-0.03509	-0.03363
Obesidade	Cardiopatia	psa	Hipertensao	Diabetes	Insuf_Renal	Metastase
	0.16005	-0.15162	0.11665	0.11408	0.09957	-0.08708
Metastase	psa	Gleason	tumor	idade	Diabetes	Obesidade
	0.41813	0.30649	0.22416	0.21831	-0.11432	-0.08708
psa	Metastase	Gleason	Obesidade	Tabagismo	Prob_Pulmao	tumor
	0.41813	0.18234	-0.15162	0.09872	-0.09593	0.0836

Tabela 1.12: Cinco melhores correlações de Spearman para cada variável

1.8.3 Resultado da Análise de Correlação

Com os dois testes de correlação, não vimos nenhum par de variáveis que se correlacionem de forma significativa o suficiente para justificar a retirada de alguma, ou mesmo uma combinação das duas. Temos agora no entanto, uma boa ideia do que esperar do modelo.

Sabemos que como as variáveis não são correlacionadas entre si; temos, no próximo passo – onde faremos a escolha de variáveis por Chi-Quadrado –

de olhar quais melhor contribuem para a determinação da variável-alvo sem nos preocupar com o ruído gerado por um par de variáveis correlacionadas.

Também importante, agora temos uma boa ideia do que esperar no modelo final. Em ambos os testes podemos ver quais variáveis tem melhor correlação (linear ou não) com a variável-alvo, assim, esperamos que a maioria das variáveis mais correlacionadas com *Metástase* apareçam no modelo final.

1.9 Análise de Chi-Quadrado

Nesse ponto, faremos um filtro de variáveis, escolhendo apenas aquelas que obedecerem a um valor mínimo de significância com a variável-alvo pelo teste de Chi-Quadrado. Dessa forma, teremos duas bases para modelagem: Uma com todas as variáveis originais; e essa de agora onde eliminaremos aquelas que não chegarem a um valor mínimo de Chi-Quadrado com a variável-alvo.

Maiores detalhes sobre o teste de Chi-Quadrado e seu cálculo podem ser encontrados no Apêndice.

As estatísticas de Chi-Quadrado para cada variável com relação a metástase é dada a seguir na tabela 1.13

Chi-Square Statistics Target=Metastase			
Input	Chi-Square	Df	Prob
psa	130.6641	5	<.0001
Gleason	70.1652	4	<.0001
tumor	33.0952	2	<.0001
idade	22.8533	4	0.0001
Diabetes	6.2991	1	0.0121
Obesidade	3.6551	1	0.0559
Cardiopatia	2.4531	1	0.1173
nodo	1.818	1	0.1776
Tabagismo	1.4393	1	0.2302
Etilismo	0.5935	1	0.4411
Prob_Pulmao	0.366	1	0.5452
Insuf_Renal	0.1842	1	0.6678
Hipertensao	0.1634	1	0.686

Tabela 1.13: Análise de Chi-Quadrado

Como é de comum em estatística, podemos fazer o corte mínimo com um p-valor de 0.05,[20] ou seja, a probabilidade máxima para se recusar a hipótese nula de que a variável não é significativa (com 1 grau de liberdade, seria Chi-Square = 3.8). Com isso, as variáveis significativas para o modelo seriam: *PSA*, *Gleason*, *tumor*, *idade*, *diabetes* e *obesidade*. Para a próxima variável, *Cardiopatia*, o p-valor de 0.11 é muito alto para considerarmos a variável como significativa (rejeitarmos a hipótese nula), assim como as demais; e então serão rejeitadas.

2 Modelagem

Fizemos um modelo de Árvore de Decisão e um SVM para os dois casos: com todas as variáveis iniciais e, rejeitando as menos significativas pelo teste de ChiQuadrado. A tabela 2.1 apresenta nossa base de dados original e a base com as variáveis rejeitadas para comparação.

NAME	Role	New Role
psa	INPUT	DEFAULT
Gleason	INPUT	DEFAULT
tumor	INPUT	DEFAULT
idade	INPUT	DEFAULT
Diabetes	INPUT	DEFAULT
Obesidade	INPUT	DEFAULT
Metastase	Target	DEFAULT
Cardiopatia	INPUT	REJECTED
nodo	INPUT	REJECTED
Tabagismo	INPUT	REJECTED
Etilismo	INPUT	REJECTED
Prob_Pulmao	INPUT	REJECTED
Insuf_Renal	INPUT	REJECTED
Hipertensao	INPUT	REJECTED

Tabela 2.1: Base de Dados original e base com as variáveis rejeitadas

Ressaltamos aqui, antes da apresentação dos modelos, a importância da distinção entre os erros de classificação. Uma classificação pode errar de duas formas: Classificar que um evento irá ocorrer quando na realidade não ocorre, assim como dizer que não irá ocorrer quando ocorre. Esses erros são chamados de Falso-Positivos e Falso-Negativos respectivamente.

A importância de se distinguir os dois tipos de erro depende da situação, do contexto para qual um modelo é feito. Para um modelo de predição de resultados de partidas de futebol por exemplo, preferimos apenas que ele tenha a menor taxa de erro possível, não importando se ele tende a favorecer um time ou outro. Agora, em nosso caso, o modelo classifica se um tumor irá

crescer ou não; e portanto, a distinção entre um falso-negativo e falso-positivo é de grande importância.

Dado que houve um erro de diagnóstico, é bem razoável assumirmos que preferimos um falso-positivo a um falso-negativo. É melhor dizer que um tumor irá crescer, para então mais tarde descobrir que ele está estável; a se tranquilizar em ouvir que um tumor está contido para depois ele ter se alastrado pelo corpo. Por pior que seja ouvir que se tem câncer e por mais doloroso e complicado que seja o tratamento, (ainda mais descobrindo que não era nada); ainda é muito melhor do que ouvir "é um pequeno tumor benigno, não se preocupe" e meses depois estar em risco de vida.

Assim como tumores são diferenciados entre "benignos" e "malignos", mesmo que, no fundo, não exista tumor benigno; mas apenas "ruim" e outro "muito ruim", podemos dizer de forma análoga, que um falso-positivo seria um "erro benigno" enquanto um falso-negativo um "erro maligno".

Com esse conceito em mente, a análise de performance de um modelo se torna um pouco mais delicada, não bastando somente melhores índices de ajuste e menores taxas de erro, agora temos também um conceito subjetivo a ser julgado.

2.1 Árvore de Decisão

Uma árvore de decisão é um fluxograma com a estrutura de uma árvore usado para explicitar visualmente tomadas de decisões, listando as decisões e suas possíveis consequências. Cada nó representa um teste sobre um atributo, cada ramo um resultado e as folhas representam as classes. Uma das vantagens de uma árvore de decisão é a facilidade de entender e interpretar seus resultados, assim como a relação entre as variáveis.

O primeiro modelo de Árvore de Decisão foi feito com todas as variáveis de entrada, cabendo apenas ao algoritmo escolher as mais significativas. Para a segunda Árvore as variáveis menos significativas foram rejeitadas. Da base original, 80% foi usado como treino e o restante 20% para teste.

Para essa modelagem utilizamos o SAS Miner. O algoritmo de árvore de decisão do SAS busca regras que maximizam a medida associada com o

critério de separação especificado na opção CRITERION.[21] A seguir seguem alguns dos parâmetros mais significativos na criação do modelo junto com sua explicação e justificativa (alguns exemplos com diferentes parâmetros e métodos de corte podem ser vistos no Apêndice):

2.1.1 Parâmetros da Árvore de Decisão

Nominal Criterion = Gini

Esse parâmetro especifica o método para pesquisar e avaliar as regras de candidatos a split para uma variável-alvo nominal. A opção PROBCHISQ usa o teste estatístico de Chi-Quadrado, ENTROPY usa a *Entropia* como medida de impureza do nó e GINI usa o índice *Gini* (maiores detalhes sobre cada método podem ser visto no Apêndice). O uso do índice Gini nos forneceu uma árvore maior do que os demais, enquanto o de Chi-Quadrado teve a menor árvore. Em qualidade numérica ambos ficaram próximos, mas, na qualidade de acerto e proporção de Verdadeiros-Positivo para Falso-Positivos (ROC) ficamos com método de split pelo indice Gini, deixando a simplicidade e estética da árvore de Chi-Quadrado.

Os valores em negrito na tabela abaixo destacam os melhores valores encontrados para cada teste relativos ao primeiro modelo de árvore.

Split Method	Train			Test		
	Gini	Entropy	ProbChisq	Gini	Entropy	ProbChisq
Misclassification Rate	0.157	0.154	0.178	0.141	0.162	**0.131**
False Negative	**11.23%**	12.53%	11.49%			
False Positive	4.44%	2.87%	**6.27%**			
Maximum Absolute Error	98.503	0.905	0.941	1.000	1.000	**0.941**
Sum of Squared Errors	79.907	80.134	98.503	22.508	23.708	**21.849**
Average Squared Error	0.104	0.105	0.129	0.114	0.120	**0.110**
Kolmogorov-Smirnov Statistic	0.651	0.636	0.574	0.654	0.619	**0.707**
Roc Index	0.914	0.911	0.848	**0.896**	0.880	0.871
Gini Coefficient	0.828	0.828	0.695	**0.793**	0.761	0.742
Lift	3.021	3.170	2.768	**3.414**	**3.414**	3.072
Cohen's Kappa	**0.586**	0.579	0.538			

Tabela 2.2: Comparação dos Critérios de Corte para primeira árvore de decisão

À primeira vista, a dúvida sobre qual critério usar recai entre o índice Gini e o Chi-Quadrado, Entropia não teve nenhum índice significativamente melhor. Como podemos ver, a árvore que usou o critério de Chi-Quadrado teve menores taxas de erro na base de Teste, com Gini melhor no índice ROC e no Lift – O índice Gini não cabe como justa comparação nesse momento visto que a primeira árvore exatamente o utiliza como critério para *split*. Agora, quando olhamos os tipos de erro pela base de Treino, vemos que ambos os critérios resultaram aproximadamente no mesmo número de falso-negativos, enquanto a árvore que usou o índice Gini teve menos falso-positivos.

Nesse caso, a escolha pelo critério do índice Gini é clara, pois o número de falso-negativos é igual. Se a terceira árvore tivesse significativamente menos falso-negativos, poderíamos ter de (subjetivamente) considerar o que é melhor para um paciente.

Significance Level = 0.08

Especifica o maior p-valor aceitável para que uma regra de split seja aceita. Valores muito altos resultam em desnecessárias ramificações assim como valores muito baixo comprometem a qualidade do modelo por restringir as ramificações possíveis da árvore. Por padrão esse valor é um pouco menor, da ordem de 10^{-2} (o padrão do SAS especificamente é 0,02), mas como nossa base é pequena, poucas observações podem rapidamente representar um percentual alto, dessa forma aumentamos o limiar.

Missing Values = Use in Search

Especifica o programa para utilizar valores faltantes (missing) como valores próprios nas definições para o split ao invés de joga-los dentro de outros grupos. Observações com a variável-alvo (*metástase)* missing foram retiradas da base; e variáveis missing nas especificações do Tumor significam que o estadiamento não foi ou não pode ser classificado por alguma razão,

dessa forma um valor missing tem um significado próprio e deve ser considerado.

Maximum Branch = 3

MAXIMUM BRANCH define o máximo de subgrupos que uma regra de split pode produzir. Um valor de 2 por exemplo resulta em uma árvore binária (dois ramos por nó). A maioria das variáveis na base são binárias, porém, estadiamento tem três grupos (0, 1 e o agrupamento de 2 e 3), Gleason é uma variável discreta de 2 a 10 e PSA é contínua; com isso é interessante permitir mais de 2 divisões. Testes com mais ramos foram feitos e obtivemos resultados mistos em cada teste estatístico para diferentes valores. Alguns exemplos podem ser encontrados no apêndice.

Maximum Depth = 8

Especifica o tamanho máximo da árvore, o maior número de "gerações" de níveis. Oito é grande o suficiente dado o número de variáveis para não interferir na qualidade do modelo e evitar uma árvore muito grande. Em apenas um teste o tamanho da árvore chegou a oito.

Minimum Categorical Size = 4

Mínimo número de vezes que uma categoria deve aparecer em uma folha para que seja considerado um split. Testes com valores entre 2 e 8 não fizeram diferença significativa, mantemos o valor padrão de 4.

Leaf Size = 3

Mínimo número de observações que uma folha deve ter. Testes com diferentes valores podem ser vistos no apêndice.

Parâmetros da Sub-Arvore (subtree node)

Method = Largest

Esse parâmetro especifica qual método usar para selecionar uma sub-árvore da árvore inteira para cada possível número de folhas. A opção ASSESSMENT (default do programa) escolhe a menor sub-árvore com o melhor *valor de avaliação (*Assessment Measure) (definido a seguir) possível. As outras escolhas são: LARGEST, que escolhe a maior árvore; e N, que escolhe a maior árvore com no máximo N folhas. LARGEST provou ser a melhor escolha nesse caso, como mostra a tabela 2.3

Subtree Method	Train			Test		
	Largest	N	Misclass	Largest	N	Misclass
Misclassification Rate	0.157	0.193	0.193	0.141	**0.131**	**0.131**
False Negative	**11.23%**	14.36%	14.36%			
False Positive	4.44%	**4.96%**	**4.96%**			
Maximum Absolute Error	0.917	0.864	0.864	1.000	**0.823**	**0.823**
Sum of Squared Errors	79.907	115.192	115.192	**22.508**	23.607	23.607
Average Squared Error	0.104	0.150	0.150	**0.114**	0.119	0.119
Kolmogorov-Smirnov Statistic	0.651	0.426	0.426	**0.654**	0.592	0.592
Roc Index	0.914	0.720	0.720	0.896	0.801	0.801
Gini Coefficient	0.828	0.439	0.439	0.793	0.603	0.603
Lift	3.021	3.035	3.035	**3.414**	**3.414**	**3.414**
Cohen's Kappa	**0.586**	0.473	0.473			

Tabela 2.3: Comparação dos Métodos para Sub-Árvore

A escolha aqui não apresentou dificuldades, o método de escolher a maior árvore se mostrou claramente superior. O método por MISCLASSIFICATION teve apenas a taxa de erros levemente menor no Teste, o que é esperado visto que cada sub-árvore é escolhida baseada nesse único parâmetro. LARGEST no entanto teve todos os outros índices melhores; o ajuste pelo índice K-S foi melhor, o ROC e o Gini. Além de que tivemos significativamente menos erros falso-negativo, nosso “erro maligno”; com a mesma taxa de falso-positivo.

Assessment Measure = -

Quando o parâmetro anterior, METHOD é estabelecido como ASSESSMENT, essa opção especifica qual método de avaliação usar para selecionar a melhor árvore. A opção DECISION (default do programa) escolhe a árvore que tenha o maior ganho e menor perda caso uma matriz de custo seja definida. Caso contrário, a avaliação é definida como AVERAGE SQUARE ERROR (Erro Quadrado Médio) para uma variável-alvo contínua, e MISCLASSIFICATION no caso de categórica, que é o nosso caso. Essas opções selecionam a árvore com menor erro quadrado médio ou que tenha a menor taxa de erros de classificação respectivamente. Como mostramos anteriormente, o método escolhido foi o LARGEST, e portanto esse parâmetro não cabe.

2.1.2 Resultados dos Modelos de Árvore

Foram criados dois modelos de árvore de decisão, um com a base inteira, cabendo apenas ao algoritmo fazer a seleção das variáveis pelo método escolhido; e outro onde as variáveis foram filtradas a priori pelo resultado do teste de Chi-Quadrado. Os dados estatísticos de cada modelo estão apresentados na tabela 2.4.

Comparação das Árvores	Train		Test	
	Arvore 1	Arvore 2	Arvore 1	Arvore 2
Misclassification Rate	0.131	0.149	0.222	**0.172**
False Negative	**8.36%**	12.79%		
False Positive	**4.70%**	2.09%		
Maximum Absolute Error	0.905	0.947	1.000	**0.947**
Sum of Squared Errors	68.377	79.711	23.874	**22.155**
Average Squared Error	0.089	0.104	0.121	**0.112**
Kolmogorov-Smirnov Statistic	0.683	0.627	0.625	**0.679**
Roc Index	0.934	0.897	0.876	**0.896**
Gini Coefficient	0.868	0.794	0.753	**0.791**
Lift	3.514	3.514	**3.414**	**3.414**
Cohen's Kappa				

Tabela 2.4: Dados Estatísticos das Árvores de Decisão 1 e 2

A segunda árvore, que recebeu o filtro de variáveis, teve um desempenho pior de classificação na base de treino em todos os índices, e resultados melhores na base de teste. A árvore 1 no entanto, com a mesma taxa de erro, aproximadamente quatro pontos percentuais a menos de falso-negativo para dois a mais de falso-positivos. Em outras palavras, menor taxa de erro, menos “erros malignos” e mais “erros benignos”, o que é bom.

Com essas informações, concluímos que para essa situação, nosso filtro manual de variáveis não melhorou a qualidade do modelo. Poderíamos ter esperado esse resultado quando testamos os métodos de split e o teste de Gini se mostrou melhor que o de Chi-Quadrado e de Entropia para esse caso. Seria razoável assumir que provavelmente uma filtragem de variáveis pelos índices de Chi-Quadrado não iriam melhorar o desempenho do modelo.

Enquanto os índices de ajuste, apesar de bons, não estarem fora da margem que temos visto, os erros desse modelo chamaram a atenção. Enquanto a taxa por si só está na margem menor (por volta de 0.13), a proporção de falso-negativos foi significativamente a menor que encontramos, e conforme mostraremos, se provará um importante diferencial perante os outros modelos.

2.2 SVM

O segundo modelo que testamos foi o de SVM (Support Vector Machine ou Máquina de Vetores de Suporte), que é um conjunto de métodos de aprendizado que analisam os dados e reconhecem padrões, assim usado para classificações e análises de regressões.

Em aprendizagem de máquina, Support Vector Machines (ou Máquinas de Vetor de Suporte), são modelos de aprendizagem supervisionados, associadas com algoritmos de aprendizagem e reconhecimento de padrões, usados para a classificação e análise de regressão. Dado um conjunto de exemplos de treino, com cada instância marcada como pertencendo a uma de duas categorias; um algoritmo de SVM constrói um função que atribui os novos exemplos a uma categoria ou outra. Um SVM é um classificador binário e não-probabilístico.[22] Apesar de comumente ser chamado de um classificador linear, um SVM pode fazer classificações não-lineares, basicamente mapeando os dados de entrada em um espaço de maior dimensão.[23]

O SVM basicamente pega um conjunto de dados de entrada e prevê, para cada dado separadamente, a qual de duas classes possíveis ele pertence; baseando sua decisão em aprendizagem anteriores. Um modelo SVM é uma representação dos exemplos como pontos no espaço, mapeados de tal forma que os exemplos das categorias separadas são divididos por um intervalo o mais amplo possível. Novos exemplos são mapeados para o espaço e então preditos como pertencentes a uma categoria baseados em qual lado da fenda eles caem. O SVM busca minimizar a função de erro de classificação, o que equivale a maior margem de separação.

Para ilustrarmos o que querermos dizer, a Figura 3.1 consiste em três funções de decisão lineares que classificam corretamente – i.e. conseguem passar uma reta que separa os pontos brancos dos pretos – um simples conjunto de treinamento em 2D.

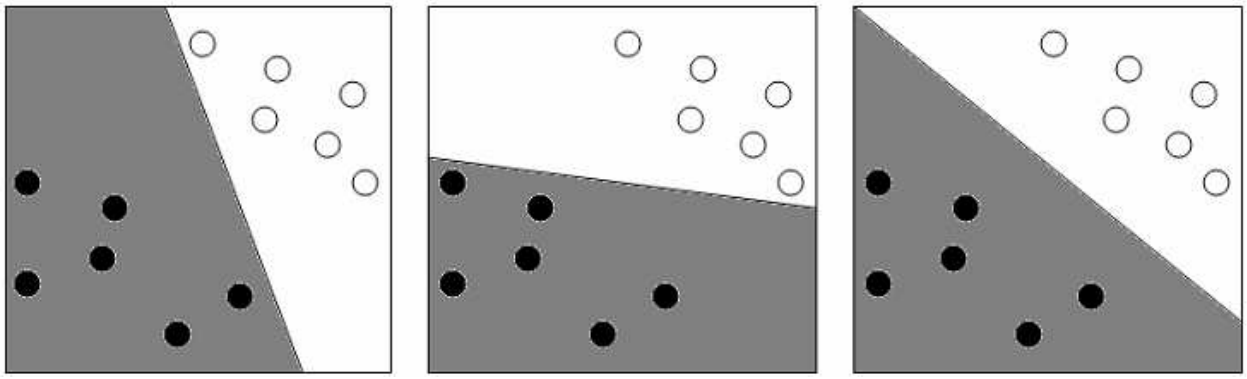


Figura 3.1[6]: Uma simples tarefa de classificação. Separar os pontos pretos dos brancos. O classificador prediz que todos novos pontos na região cinza são pretos e na região branca são brancos. Três possíveis e diferentes funções lineares de decisão são mostradas, A melhor função de decisão nesse caso é a terceira, pois divide os grupos com a maior margem.

Funções de decisão lineares consistem em uma função de decisão de contorno que é um hiperplano (uma linha em 2D, plano em 3D, etc.), que separa as duas regiões diferentes do espaço. Tal função de decisão pode ser expressa por uma função de um vetor x de entrada, com o valor sendo aquilo que é o predito para x (+1 ou -1).

Maiores detalhes sobre a teoria de SVM e o algoritmo Perceptron podem ser vistos no apêndice.

Nesse trabalho, os dois grupos distintos eram os canceres que resultaram em metástase e os que não. Com isso o algoritmo deveria, com os dados de entrada, separar a qual grupo um dado caso pertence.

Da mesma forma, rodamos um modelo de SVM com a base completa e outro com nossa base reduzida pelo critério de Chi-Quadrado. Apresentamos a seguir os parâmetros mais significativos para o modelo, maiores detalhes podem ser vistos no apêndice.

2.2.1 Parâmetros de SVM

Estimation Method = LSVM (Lagrangian SVM)

ESTIMATION METHOD especifica o método que será usado para modelagem de SVM. Pode-se usar *Programação Quadrática Completa (Full Quadratic Programming), Programação Quadrática Decomposta (Decomposed Quadratic Programming), Lagrangiano;* ou *Mínimos Quadrados (Least Squares)*. Maiores detalhes sobre cada método podem ser vistos no apêndice.

A Tabela 2.5 compara os resultado obtidos por cada um dos métodos.

ESTIMATION METHOD	Train				Test			
	FQP	DQP	LSVM	LSSVM	FQP	DQP	LSVM	LSSVM
Misclassification Rate	0.320	0.198	0.191	0.204	0.404	0.192	**0.162**	0.192
False Negative	**7.57%**	17.75%	15.93%	18.54%				
False Positive	24.54%	2.09%	2.61%	1.83%				
Maximum Absolute Error	1.000	0.838	0.789	0.771	1.000	0.789	**0.682**	0.725
Sum of Squared Errors	242.480	112.864	125.114	122.032	78.803	**26.733**	31.122	30.567
Average Squared Error	0.320	0.147	0.163	0.159	0.398	**0.135**	0.157	0.154
Kolmogorov-Smirnov Statistic	0.420	**0.556**	0.544	0.540	0.307	0.599	**0.624**	**0.624**
Roc Index	0.780	0.841	**0.860**	0.859	0.708	**0.892**	0.891	**0.892**
Gini Coefficient	0.560	0.683	**0.720**	0.718	0.417	0.783	0.782	**0.784**
Lift	2.800	3.329	**3.514**	3.144	1.536	**3.414**	2.731	2.731
Cohen's Kappa	0.331	0.416	0.470	0.392				

Tabela 2.5: Comparação dos métodos de modelagem SVM

Como podemos ver, a programação quadrática (FQP) se mostrou a menos eficiente. Apesar do baixíssimo número de falso-negativos que nos chamou a atenção por ser uma resultado favorável, vemos que isso só é devido a uma forte tendência de positivos; que acaba por elevar o erro demais. 32% de erros é muito alto comparado aos outros métodos e nossos modelos de árvore anteriores. Além disso, todos seus índices de ajuste foram significativamente inferiores.

Os outros 3 métodos ficaram muito próximos. A diferença nos índices de ROC e Gini não é significativa, tanto no Treino quanto no Teste. Com isso, olhamos os erros de classificação; o método de Mínimos Quadrados (LSSVM) teve a maior taxa de erros de classificação com maior proporção de falso-negativos, e logo o descartamos. O Lagrangiano se mostrou o melhor nesse caso, com menor taxa de erro e ainda melhor tendência a falso-positivos.

2.2.2 Resultado dos modelos de SVM

Com os parâmetros definidos, fizemos então os dois modelos de SVM, novamente, com a base inteira e com o filtro a priori. O resultado pode ser visto na tabela 2.6

Comparação dos SVMs	Train		Test	
	SVM 1	SVM 2	SVM 1	SVM 2
Misclassification Rate	0.191	0.186	0.162	**0.152**
False Negative	16.19%	16.19%		
False Positive	2.87%	2.35%		
Maximum Absolute Error	0.782	0.771	0.684	0.681
Sum of Squared Errors	125.892	127.324	31.434	31.476
Average Squared Error	0.164	0.166	0.159	0.159
Kolmogorov-Smirnov Statistic	0.543	0.528	0.644	**0.650**
Roc Index	0.858	0.852	0.890	**0.896**
Gini Coefficient	0.717	0.704	0.780	**0.791**
Lift	3.514	3.514	**2.731**	**2.731**
Cohen's Kappa	0.455	0.457		

Tabela 2.6: Comparação entre os dois modelos de SVM

A diferença como podemos ver é pequena, porém significativa, visto que o segundo modelo foi melhor em todos os testes estatísticos, e, com a mesma taxa de falso-positivos, teve um pouco menos de falso-negativos.

Assim, podemos dizer que o filtro de variáveis teve efeito, melhoramos o ajuste do modelo e a taxa de erro. Foi uma melhoria pequena, mas clara.

Quanto ao ganho versus o custo computacional, também vale a pena. Com uma base pequena praticamente não há diferença, o tempo de uma base completa ou filtrada é próximo do mesmo. No caso de um banco de dados grande – o que gostaríamos muito – estaríamos diminuindo o número de variáveis, retirando àquelas que vimos não contribuir significativamente para a melhoria do modelo. Assim acabamos ganhando tempo ao usar uma base com menos variáveis.

2.3 Escolha do Melhor Modelo

Aqui, comparamos o resultado do modelo de árvore com o de SVM – o melhor de cada – para decidirmos se há significativa diferença entre os dois. A tabela 2.6 denota os principais dados dos modelos.

SVM vs Árvore	Train		Test	
	SVM	Árvore	SVM	Árvore
Misclassification Rate	0.186	0.131	**0.152**	0.222
False Negative	16.19%	8.36%		
False Positive	2.35%	4.70%		
Maximum Absolute Error	0.771	0.905	0.681	1.000
Sum of Squared Errors	127.324	68.377	31.476	23.874
Average Squared Error	0.166	0.089	0.159	0.121
Kolmogorov-Smirnov Statistic	0.528	0.683	**0.650**	0.625
Roc Index	0.852	0.934	**0.896**	0.876
Gini Coefficient	0.704	0.868	**0.791**	0.753
Lift	3.514	3.514	2.731	**3.410**
Cohen's Kappa	0.457	0.667		

Tabela 2.7: Comparação entre Árvore de Decisão e SVM

Como podemos ver a árvore teve um ajuste muito bom na base de Teste, superior ou igual ao SVM em todos os índices. No Teste, tivemos o contrário em menor magnitude, o modelo de SVM se sobressaiu em quase todos os índices, e foi levemente pior nos erros.

A figura 3.2 abaixo demonstra a curva ROC comparando os dois modelos.

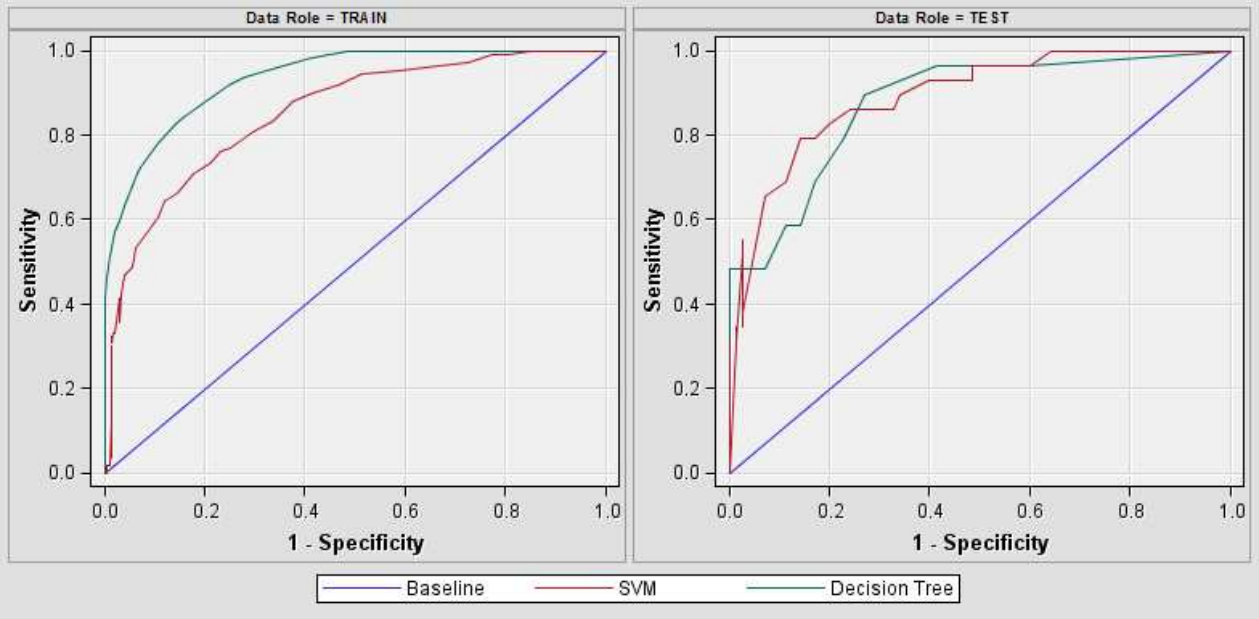


Figura 3.2: Curva ROC entre Árvore de Decisão e SVM

O gráfico ilustra bem os dados da tabela, a árvore foi significativamente superior no Treino e levemente inferior – com alguns resultados mistos – no Teste.

Como mencionamos anteriormente, a distinção entre Falso-Negativo e Falso-Positivo é importante. Nosso objetivo é "evitar danos ao paciente", i.e. nos certificar ao máximo que o paciente receba um diagnóstico que seja o mais saudável possível. Um Falso-Negativo é condenar alguém com câncer dizendo que ele está saudável; enquanto um Falso-Positivo resultaria apenas em dor de cabeça (e possivelmente dinheiro) enquanto o paciente procura uma cura para uma enfermidade que ele não tem.

O resultado aqui foi delicado, mas acabamos por escolher o modelo de árvore. Com aproximadamente oito pontos percentuais a menos de falso-negativos, e dois desses pontos em falso-positivos, resultando também em uma menor taxa de erro em geral; o modelo se mostrou superior

3 Conclusão

Através da metodologia proposta, mostramos que é possível prever se em um dado paciente com tumor, ocorrerão metástases; i.e., se o tumor irá se alastrar para outros órgãos. Nossa melhor taxa de erro ficou próxima de 13% pelo modelo de árvore de decisão. Com um banco de dados mais completo, com mais registros e menos informações faltantes, esse valor pode melhorar ainda mais. E poderemos aplicar metodologias semelhantes para os demais tipos de tumor e demais estadiamentos.

Uma informação muito importante que não tivemos acesso é o registro de todas as medidas de PSA por paciente. Como vimos, uma medida de PSA por si só, apesar de ter boa relação com o resultado para prever o crescimento do tumor, não é suficiente para tirar uma conclusão significativa. Vários fatores podem causar um valor alto momentâneo – as vezes até mesmo fixos – incluindo mesmo uma predisposição biológica. Um histórico dos níveis de PSA seria de grande importância e melhoraria em muito os modelos; uma análise da tendência pode indicar diretamente se há melhora ou piora do nível da enzima, e consequentemente no avanço ou não do tumor.

Em relação ao problema de classificar a ocorrência de metástases, lidamos com dois grandes problemas: a baixa quantidade de registros e dados faltantes.

Os modelos se beneficiariam bastante com uma base de dados maior, pois mais evidências para treino fortalecem o aprendizado e tornam o resultado de teste mais confiável para ser aplicado na prática. Nossa base de início tinha somente 1061 dados, de onde apenas 648 eram relativos a tumor de próstata. A falta de informações piora ainda mais o problema pois invalida os poucos que sobram.

O estudo proposto tem potencial para ser de utilidade para pacientes e médicos. Poderíamos incluir os resultados no banco de dados de um hospital, por exemplo. Enquanto nada substitui o conhecimento de um médico, tal informação serviria, no mínimo, como um auxílio para a tomada de decisões mais rápidas e confiáveis, e, na ausência de um especialista, permitiria que um outro médico possa tomar uma decisão mais assertiva.

O método proposto também pode ser usado para os demais tipos de tumores, o princípio da metodologia dos modelos preditivos pode facilmente ser aplicado para os demais canceres, bastando o acesso aos dados relativos necessários. Métodos adicionais podem ser usados também, aqui usamos Árvore e SVM, mas uma classificação Bayesiana também pode ser feita, nos dando mais opções para encontrar o melhor modelo.

Um modelo único para diferentes tipos de câncer é complicado e não recomendado de se fazer, as variáveis que impactam a formação e progressão variam com os diferentes tumores. O ideal é a construção de modelos específicos para cada caso.

4 Apêndice

4.1 Teste de Chi-Quadrado

O teste de Chi-Quadrado é um teste de hipóteses que se destina a encontrar um valor de dispersão para duas variáveis de escala ***nominal*** avaliando a associação existente entre elas. O teste nos diz em que medida os valores observados se desviam do valor esperado, caso as duas variáveis não forem correlacionadas.[24] Quanto maior o chi-quadrado, mais significante é a relação entre a variável dependente e a variável independente. É um teste não paramétrico, i.e., não depende dos parâmetros populacionais, como média e variância.[25]

O princípio básico deste método é comparar proporções. Comparar as possíveis divergências entre as frequências observadas e esperadas para certo evento; testando se um conjunto de dados segue uma certa distribuição. O teste de Chi-Quadrado é melhor usado para testar suposições de modelagem do que para comparar modelos.

O teste de Chi-Quadrado tem semelhança com os testes de correlação, onde ambos procuram descrever a relação entre um par de variáveis, e de certa forma são maneiras diferentes de pensar em uma mesma ideia. Os resultados dos testes de significância para Chi-Quadrado e Correlação não são os mesmos, mas normalmente resultam em uma conclusão estatística semelhante.[26]

Esse teste é comumente usado para observar a significância estatística entre uma variável-alvo categórica (se o evento ocorreu ou não ocorreu) e uma variável categórica determinante. Pode-se dizer que dois grupos se comportam de forma semelhante se as diferenças entre as frequências observadas e as esperadas em cada categoria forem suficientemente pequenas.

Para que possamos entender o cálculo e o significado do teste de Chi-Quadrado, precisamos entender o conceito da Hipótese Nula.

4.1.1 A Hipótese Nula

Na ciência, quando formulamos uma hipótese, como a relação existente entre dois fenômenos, temos de testar se ela é verdadeira ou não. Para isso coletamos e analisamos dados relevantes para ver se conseguimos suportar nossa ideia.[27] No entanto, quando obtemos os resultados, é bem possível que quaisquer relações que apareçam entre os dados, sejam apenas por chance. Dessa forma, para conseguirmos suportar nossa hipótese inicial, temos de comparar os resultados contra a situação oposta: Que os fenômenos não tem relação alguma entre si e os resultados foram apenas por chance. Essa última é a *hipótese nula.*

Com dados estatísticos, a única maneira de provar uma hipótese é conseguir negar a *hipótese nula*, i.e., conseguir negar que os resultados foram apenas pela chance. Ou seja, temos de assumir que nossa hipótese inicial está errada até que possamos rejeitar a hipótese nula. "*Inocente até provado culpado*" para *hipótese nula.*

Assim, a *hipótese nula*, é uma formulação inicial, uma declaração qual tentamos negar, provar falsa. Contudo a hipótese nula não pode nunca ser provada, os testes podem apenas rejeitar ou falhar em rejeita-la. Podemos pensar na *hipótese nula* como a ideia "chata" que insiste em dizer: "isso não tem nada a ver". E o que temos de fazer é mostrar que ela está errada.[28]

Por exemplo, estamos desconfiados que uma moeda é viciada, é tendenciosa a cair "cara". Então a hipótese nula diria: "essa moeda não é viciada, só caiu mais caras por chance". Portanto, o experimento consiste em repetidamente jogar a moeda para tentar mostrar que ela não pode ser justa.

Se jogarmos 20 vezes uma mesma moeda, e obtivermos 11 caras e 9 coroas, qualquer um poderia dizer que esses números são resultados de chance e completamente normais, não conseguiríamos provar que ela é viciada. Se no entanto, obtivermos 19 caras e 1 coroa, qualquer pessoa com bom senso diria que as chances disso ocorrer são muito baixas e a moeda com certeza é viciada. Nesse caso conseguimos refutar a hipótese nula – de que a moeda é normal, não viciada. Mas e se tivéssemos obtido 15 caras e 5 coroas? Obviamente bem mais caras, mas seria um resultado tão improvável de acontecer que poderíamos rejeitar a hipótese nula? Para responder isso, precisamos mais do que apenas bom senso, precisamos calcular a probabilidade conseguir um desvio dessa grandeza apenas devido a chance.

O gráfico 4.1 abaixo mostra as chances de conseguirmos todas as possíveis quantidades de caras, de 0 a 20, sob a hipótese nula de que a probabilidade é de 0.5 para sair cara.

A probabilidade de sair n caras em 20 lances é dada por

$$C_{20,n}(0.5^n 0.5^{20-n}) = C_{20,n}(0.5^{20}) = \frac{20!}{n!\,(20-n)!}0.5^{20}$$

Assim probabilidade de sair 15 caras é 0.0148, um valor baixo; mas, essa é a probabilidade de sair *exatamente* 15 caras. O que queremos é a probabilidade de sair 15 *ou mais* caras. Pois se fossemos aceitar 15 caras como evidência de que a moeda é viciada, então também teríamos de aceitar 16 ou mais caras como evidência. Dessa forma, temos de somar todas as probabilidades

individuais para sabermos qual a chance de conseguirmos 15 ou mais caras em 20 lances da moeda.

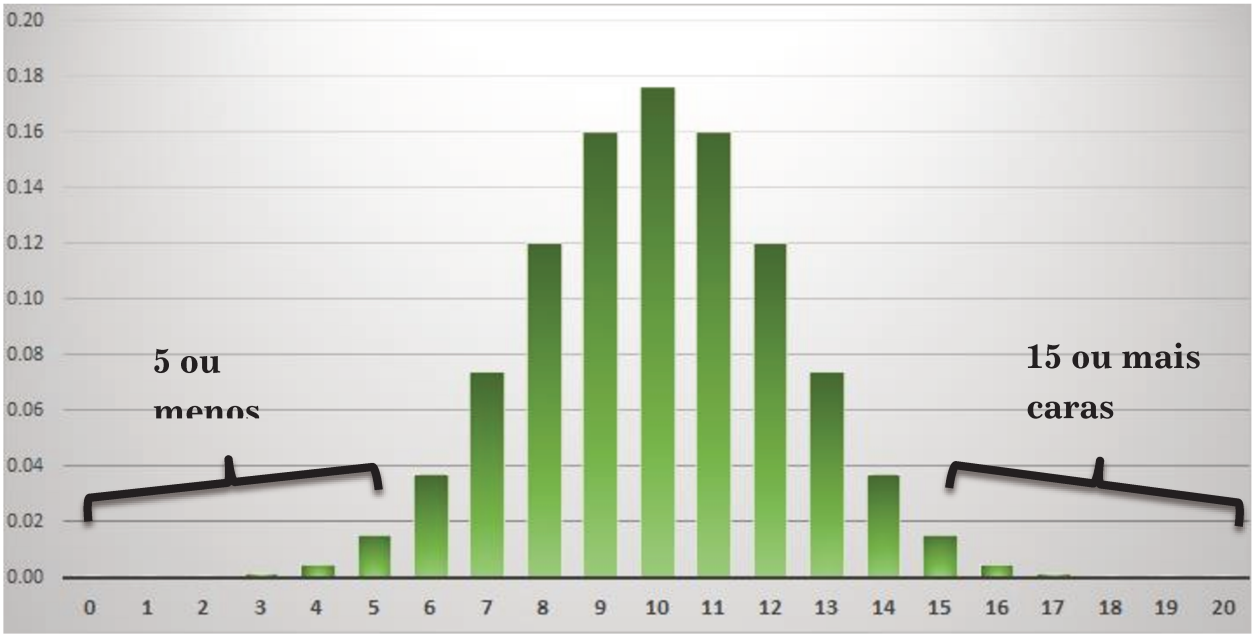


Figura 4.1: Distribuição de probabilidade para cada quantidade possível de caras em uma moeda justa lançada 20 vezes

Somando as probabilidade, temos que, a chance de sair 15 ou mais caras é de 0.021.

Esse número 0.021 é o chamado *P-Valor*, que é definido como sendo a probabilidade de se conseguir *pelo menos* o resultado observado, i.e., de se *conseguir o resultado obtido ou um resultado mais extremo* se a Hipótese Nula for verdade (nesse caso, se a moeda for justa).

A convenção comum em estatística é usar um nível de significância igual a 0.05; ou seja, se o *P-valor* for menor que 0.05 nós rejeitamos a hipótese nula. Lembrando que se o P-valor for maior que 0.05 nós apenas *não refutamos* a hipótese nula, pois como dito anteriormente, a hipótese nula não pode ser provada. O número 0.05 é apenas uma convenção comum e em muitos casos uma pequena variação é feita e pode ser perfeitamente aceitável dentro do contexto.

4.1.2 Determinação de χ^2

O coeficiente Chi-Quadrado é designado por χ^2 (da letra grega "*chi*"), é um valor de dispersão para um par de variáveis *nominais* que indica o quanto os valores observados se desviam do valor esperado, caso as variáveis não forem relacionadas. Quanto maior o valor de χ^2 mais significante é a relação entre as duas variáveis. O teste de não faz suposições sobre a relação entre o par de variáveis ou a forma da distribuição da qual a amostra é retirada.[29] Diferente dos testes de correlação, χ^2 é sempre positivo, com zero indicando nenhuma significância entre as variáveis.

A *Hipótese Nula* H_0 assume que não há relação entre o par de variáveis, e portanto as variáveis são estatisticamente independentes.

O cálculo de χ^2 é dado por

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Onde:

O_i = frequência observada
E_i = frequência esperada teórica, dada pela hipótese nula
n = número de observações da base de dados

O valor de χ^2 junto com o *grau de liberdade* da variável, é então comparado com a função densidade de probabilidade de distribuição Chi-Qudrado para calcular o p-valor. Com isso é possível rejeitar ou não a hipótese nula.

Enquanto um valor de zero (ou muito próximo) pode ser facilmente interpretado como a variável não sendo relacionadas, normalmente, o valor de χ^2 não é descritivo e nem usado por si só; ele deve ser visto como um passo para se calcular o p-valor. Assim, o valor calculado de χ^2 e o grau de liberdade da variável, tem a finalidade somente de calcular o p-valor para saber se a hipótese nula deve ou não ser rejeitada.

O que queremos ressaltar com esse último parágrafo, é que devemos focar no p-valor e não no valor absoluto de χ^2. Em nossa análise de variável, o que nos dirá efetivamente se devemos ou não aceitar uma variável como entrada no modelo é se o p-valor for abaixo de 0.05.

4.2 Balanceamento de Classe e o algoritmo SMOTE

O desbalanceamento de classes ocorre quando existe uma grande desproporção entre o número de exemplos de cada classe – no nosso caso, entre a ocorrência e não de metástase. Isso pode facilmente fazer com que os exemplos da classe minoritária sejam classificados incorretamente.[30]

Existem basicamente duas maneiras – cada qual com diferentes métodos e algoritmos – de se balancear as classes em uma base de dados: A primeira, conhecida como '*under-sampling*' é a mais comum e simples de ser feita, que consiste em extrair um número aproximadamente igual da classe majoritária à classe minoritária; dessa forma, fica-se com um percentual igual de ambas as classes. O problema de se fazer under-sampling ocorre principalmente nos casos onde a base tem poucos dados e a diferença é grande; pois a base usada acaba ficando muito pequena e prejudicando a modelagem devido a insuficiência de dados. Nesses casos, pode-se recorrer à segunda maneira de se balancear as classes: O '*over-sampling*'; que funciona por inserir novos dados da classe minoritária a fim de aumentar sua proporção na base.

O método mais simples – e intuitivo – de over-sampling é o de apenas replicar os dados da classe minoritária, a fim de aumentar sua proporção da base geral. Esse método no entanto, tende a "viciar" os modelos – o modelo só consegue identificar dados da classe minoritária quando são iguais aos da base usada em sua criação.

O método SMOTE (Synthetic Minority Over-sampling Technique) consiste em criar novos dados da classe minoritária por interpolação ao invés de apenas repô-los. Para cada instância (da classe minoritária), nós

introduzimos uma nova instância ao longo da linha que a une com seus k vizinhos mais próximos.[31]

Em nosso caso por exemplo, nós dobramos o número de instâncias onde ocorreram metástases, i.e., geramos 100% mais dados. Então, para cada ponto (com metastase = 1) real, pegamos um único vizinho mais próximo e criamos uma nova instância ao longo da reta que os une.

A criação desses exemplos "sintéticos", faz com que o classificador consiga uma região de classificação maior, mais genérica; ao invés de uma região muito específica[25] – vício do modelo – e assim tendo melhor performance em exemplos reais, que se diferem daqueles da base usada na modelagem; ao invés de se "viciar" e conseguir classificar apenas exemplos semelhantes.

4.3 Coeficiente Kappa de Cohen

O coeficiente Kappa – também conhecido como Kappa de Cohen – é uma medida estatística de concordância entre dois classificadores.[32] Dessa forma, se considerarmos os valores reais de ocorrência de metástases como um classificador, – classificador perfeito –, temos um índice de análise para cada uma das árvores que denota sua concordância com os valores reais. Seu cálculo é dado por

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

Onde $P(a)$ é a taxa de concordância entre os dois classificadores, i.e., a fração – do total – de vezes em que eles tiveram a mesma resposta. $P(e)$ é a probabilidade de que eles poderiam ter resultados iguais por chance. κ pode variar de zero a 1, onde zero indica nenhuma concordância entre os dois classificadores – i.e., qualquer resultado igual foi devido apenas a chance – e um indica concordância total entre os classificadores.

O conteúdo referente às seções 5.4 e 5.5 a seguir, é baseado na informação apresentada no suporte e documentação do SAS; encontrado em http://support.sas.com/documentation/index.html

4.4 Critérios de *Split* da Árvore de Decisão

O algoritmo de árvore de decisão do SAS busca regras que maximizam a medida associada com o critério de separação especificado na opção CRITERION. Algumas medidas são baseadas em uma medida de *impureza* do nó e outras no *p-valor de um teste estatístico*.

Uma medida para uma variável-alvo categórica pode incorporar probabilidades anteriores. Uma medida para alvo nominal alvo pode incorporar funções de lucro ou perda, incluindo erros de classificação. Uma medida para alvos ordinais deve incorporar distâncias entre os valores-alvo. O algoritmo cria uma função de distância a partir de uma função de perda especificado pela declaração DECISION

4.4.1 Redução da Impureza do Nó

A Impureza $\mathrm{I}(\eta)$ do nó η é um número não negativo que é igual a zero se todas as observações em η tem o mesmo valor de destino, e é grande se os valores-alvo em η são muito diferente.

A opção CRITERION=VARIANCE especifica como medida de impureza o erro quadrado médio no intervalo. Assim,

$$\mathrm{I}(\eta) = \frac{1}{N(\eta)} \sum_{i=1}^{N(\eta)} (x_i - \bar{x})^2$$

Onde $N(\nu)$ é o número de observações no nó η, x_i é o valor da observação i e $\bar{x}$ a média de x_i em η.

A opção CRITERION=ENTROPY usa a *entropia* como medida de impureza para uma variável categórica. Nesse caso a impureza do nó η é calculada por

$$\mathrm{I}(\eta) = -\sum_{j}^{J} p_j \cdot \log_2 p_j$$

Onde p_j é a proporção de oberservações com valor-alvo j em η.

A opção CRITERION=GINI usa o índice *Gini* como medida de impureza. O coeficiente Gini é uma medida estatística de dispersão. Ele mede a desigualdade entre valores de distribuição de frequência, e vai de 0 a 1. Um valor de zero corresponde a uma perfeita igualdade enquanto 1 corresponde a total desigualdade entre os valores.

O índice Gini é dado por

$$\mathrm{I}(\eta) = 1 - \sum_{j}^{J} p_j$$

A qualidade de um split s é medido como a redução da impureza do nó

$$\Delta I(s, \eta) = I(\eta) - \sum_{b}^{B} p(\eta_b|\eta) I(\eta_b)$$

Onde a soma é sobre os B ramos que o split s define, e $p(\eta_b|\eta)$ é a proporção de observações em η atribuídas ao ramo b.

4.4.2 Testes estatísticos e p-valores

Uma alternativa a utilizar a redução de impureza do nó, é testar uma diferença significativa dos valores-alvo entre os diferentes ramos definidos por uma divisão(*split*)-candidata. O valor da divisão é igual a $-\mathbf{log_{10}}\,\boldsymbol{p}$, onde p é o p-valor do teste.

Para uma variável-alvo numérica contínua, a opção CRITERION=PROBF utiliza a estatística-F que é igual a Variação (**entre** médias amostrais) / Variação (entre indivíduos **dentro** das amostras)

$$F = \frac{S_{\text{entre}}/B - 1}{S_{\text{dentro}}/N(\eta) - B}$$

Onde

$$S_{\text{entre}} = \sum_{k=1}^{B} N(\eta_k)\left(\bar{x}(\eta_k) - \bar{x}(\eta)\right)^2$$

$$S_{dentro} = \sum_{k}^{B} \sum_{i}^{N(\eta_k)} \left(x_{ki} - \bar{x}(\eta_k)\right)^2$$

onde B é o número de ramos (ou galhos) definidos no *split*.

Para uma variável-alvo nominal, a opção CRITERION=PROBCHISQ utiliza a estatística de Chi-Quadrado:

$$\chi^2 = N(\eta) \sum_{k}^{B} \sum_{i}^{J} \frac{\left(p_i(\eta_k) - p(\eta_k|\eta)p_i(\eta)\right)^2}{p(\eta_k|\eta)p(\eta)}$$

Onde J é o número de valores-alvo.

4.5 Diferentes estruturas de árvores geradas pela modificação de parâmetros e métodos de divisão.

4.5.1 Maximum Branch

Com 3 ramos, temos a melhor qualidade do modelo medido pelos índices de ROC, Gini e melhor Lift; na base de Teste. Com o máximo de 6 ramificações tivemos melhor resultado no Treino e menos erros de falso-negativo. Com esses resultados apenas, ficou complicado escolher a melhor opção, então colhemos os resultados da base de Teste individualmente para analisar o que seriam os números de falso-negativo e falso-positivo para as opções com 3 e 6 ramificações.

Felizmente, como podemos ver pela tabela 4.1, essa informação nos permitiu fazer uma escolha com mais segurança. Com 3 ramificações, temos melhores resultados na base de Teste, incluindo melhores taxas de falso-positivo e falso-negativo.

Maximum Branch	Train				Test			
	2	3	4	6	2	3	4	6
Misclassification Rate	0.157	0.157	0.157	0.131	**0.141**	**0.141**	**0.141**	0.192
False Negative	10.44%	11.23%	12.01%	**9.14%**		*10.10%*		*11.11%*
False Positive	5.22%	4.44%	3.66%	3.92%		*4.04%*		*8.08%*
Maximum Absolute Error	0.889	0.917	0.895	0.909	**1.000**	**1.000**	**1.000**	**1.000**
Sum of Squared Errors	86.083	79.907	86.083	66.560	23.843	22.508	**21.788**	28.498
Average Squared Error	0.112	0.104	0.107	0.087	0.120	0.114	**0.110**	0.144
Kolmogorov-Smirnov Statistic	0.627	0.651	0.629	**0.711**	**0.687**	0.654	0.647	0.561
Roc Index	0.892	0.914	0.909	**0.942**	0.866	**0.896**	0.892	0.803
Gini Coefficient	0.783	0.828	0.819	**0.884**	0.731	**0.793**	0.785	0.605
Lift	3.035	3.021	2.938	**3.370**	2.868	**3.414**	**3.414**	2.959
Cohen's Kappa	0.593	0.586	0.578	0.661				

Tabela 4.1: Número máximo de ramificações

4.5.2 Leaf Size

Aqui tivemos uma escolha delicada, pois a primeira vista parece ser aparente que quanto menor o tamanho mínimo da folha, maior o número de divisões possíveis e portanto mais assertivo será o modelo. Porém, testes com diferentes valores mostram que um valor mais alto – e consequentemente menos divisões – pode ter melhor performance em alguns casos específicos.

Como vemos na tabela 4.2 a seguir, a árvore com mínimo de 3 observações por folha tem melhor performance na maioria dos testes com a base de Treino, porém, com a base de Teste, a árvore com mínimo de 5 folhas apresenta melhores índices de ROC e o Gini; e com mínimo de 3 folhas temos os menores erros. Como tivemos uma taxa maior de Falsos-Positivos para Falsos-Negativos com a árvore 3 folhas, junto com os testes estatísticos que mesmo na base de Teste ficaram bastante próximos, escolhemos o mínimo de 3 folhas como o número ótimo.

Essa escolha foi bem razoável, nossa base infelizmente – para modelagem – é pequena, não podemos exigir que cada folha tenha muitas observações.

Leaf Size	Train				Test			
	3	6	8	10	3	6	8	10
Misclassification Rate	0.131	0.157	0.157	0.170	0.2220	0.1410	0.1410	0.1410
False Negative	**8.36%**	11.23%	12.53%	13.58%				
False Positive	4.70%	4.44%	4.18%	3.39%				
Maximum Absolute Error	0.905	0.917	0.917	0.917	1.000	1.000	1.000	1.000
Sum of Squared Errors	68.377	79.907	84.855	86.725	23.874	22.508	23.519	**21.994**
Average Squared Error	0.089	0.104	0.111	0.113	0.121	0.114	0.119	**0.111**
Kolmogorov-Smirnov Statistic	0.683	0.651	0.629	0.625	**0.668**	0.654	0.654	**0.668**
Roc Index	0.934	0.914	0.903	0.899	0.876	**0.896**	**0.896**	0.893
Gini Coefficient	0.868	0.828	0.807	0.799	0.753	**0.793**	**0.793**	0.787
Lift	3.514	3.021	2.985	2.761	**3.414**	**3.414**	2.731	3.072
Cohen's Kappa	**0.667**	0.586	0.550	0.533				

Tabela 4.2: Número Mínimo de Folhas

4.6 SVM e o Algoritmo Perceptron

Em aprendizado de máquina, SVM's são modelos supervisionados, com algoritmos que analisam e reconhecem padrões de dados. Um modelo de SVM recebe um conjunto de dados como entrada, onde cada dado pertence à uma de duas possíveis classes, e então, analisa os padrões dos dados de cada uma das classes. Com isso, para futuros conjuntos de dados, o modelo tenta prever a qual classe cada dado pertence; baseando sua decisão em aprendizagem anteriores.

Um modelo SVM representa os dados como pontos no espaço, mapeados de tal forma que, todas as instâncias de um grupo são separadas do outro por uma margem mais ampla o possível. Novos exemplos de entrada são então mapeados para o espaço e então preditos como pertencentes a um dos grupos baseados em qual lado da margem eles caem. Basicamente, o SVM busca minimizar uma função de decisão de erro de classificação, o que equivale a buscar a maior margem de separação.

A função de decisão pode ser expressa por uma função que toma um vetor x como entrada, cujo resultado – da função – será +1 ou -1. Representando a qual lado o dado deve pertencer.

O classificador linear pode então ser escrito como:

$$g(\boldsymbol{x}) = sign\big(f(\boldsymbol{x})\big)$$

$$\text{onde } f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$$

onde $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ é o produto escalar entre $\boldsymbol{x}$ e o vetor de pesos $\boldsymbol{w}$.

Um dos primeiros algoritmos a tentar resolver o problema de SVM foi o algoritmo Perceptron, que usava simplesmente um procedimento iterativo para incrementalmente ajustar $\boldsymbol{w}$ e b até que a decisão de contorno fosse capaz de separar as duas classes do conjunto de treino. Dessa forma, o algoritmo não dava preferência alguma entre, por exemplo, as três soluções apresentadas da figura 4.2. Isso é bem insatisfatório, já que a decisão a direita é claramente uma decisão superior pois divide as duas classes da melhor forma possível.

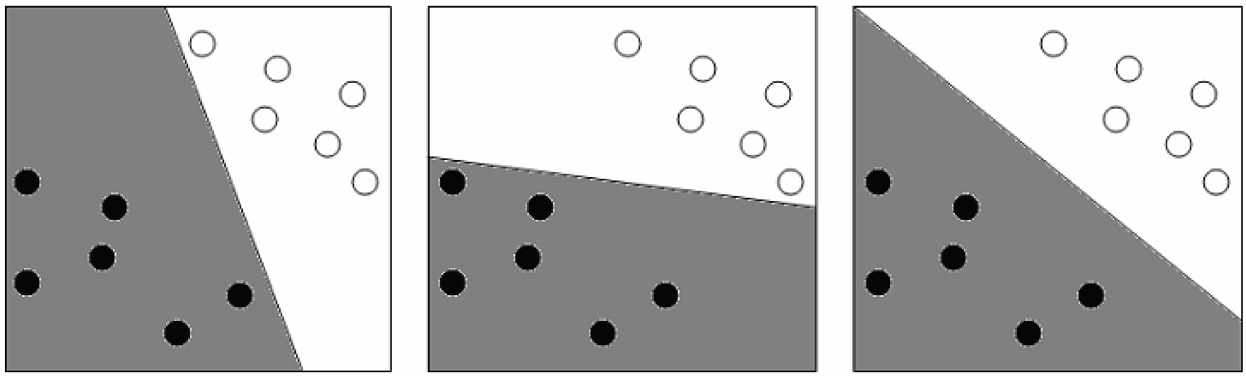


Figura 4.26: Uma tarefa de classificação, separar os pontos pretos dos brancos. O classificador prediz que todos novos pontos na região cinza são pretos e na região branca são brancos. Três possíveis e diferentes funções lineares de decisão são mostradas, A melhor função de decisão nesse caso é a terceira, pois divide os grupos com a maior margem.

Para um classificador SVM é necessário preferir decisões de contorno que não somente separem as duas classes no conjunto de treino, mas que as separem com a maior margem possível. *Margem é a distância do hiperplano ao exemplo mais próximo.*[33]

Para podermos encontrar o hiperplano com a maior margem, é interessante então escrever o problema como um problema de otimização, que consiste de uma função objetiva da qual queremos encontrar o máximo ou mínimo, junto com um conjunto de restrições que devem ser satisfeitas. A figura 4.3 mostra as equações para o hiperplano e sua margem, o que nos ajuda a visualizar os parâmetros para o problema.

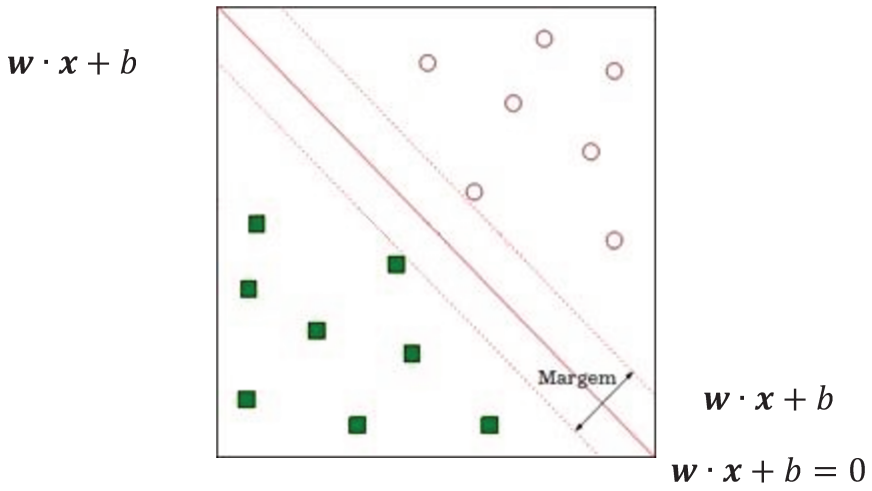


Figura 4.3: O melhor hiperplano (linha nesse caso) que separa os dois grupos. A equação do hiperplano é dada por $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$. $+1$ e -1 representam as duas classes possíveis.

O que precisamos fazer então é maximizar essa margem, a distância entre as duas linhas paralelas em que cada uma toca o ponto mais próximo de cada lado. Essa distância é dada por

$$\frac{2}{\sqrt{\boldsymbol{w} \cdot \boldsymbol{w}}}$$

e maximizar esse valor é o mesmo que minimizar

$$\frac{\boldsymbol{w} \cdot \boldsymbol{w}}{2}$$

Podemos então formalizar o problema de SVM como:

Dado um conjunto de treinamento de vetores $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots \boldsymbol{x_n}$, com classes correspondentes $y_1, y_2, \ldots y_n$ que
assumem os valores $+1$ ou -1, devemos escolher os parâmetros $\boldsymbol{w}$ e b da função de decisão linear que satisfaz:

$$\min_{w,b} \frac{1}{2} \boldsymbol{w} \cdot \boldsymbol{w}$$

$$s.a. \quad y_i(\boldsymbol{w} \cdot \boldsymbol{x} + b) \geq 1$$

Onde a restrição $y_i(\boldsymbol{w} \cdot \boldsymbol{x} + b) \geq 1$ vem do par

$$\begin{cases} \boldsymbol{w} \cdot \boldsymbol{x} + b > 0, \ \ \forall\, y = 1 \\ \boldsymbol{w} \cdot \boldsymbol{x} + b < 0, \ \ \forall\, y = -1 \end{cases}$$

que é o equivalente a dizer que os dados devem cair no lado correto da superfície de decisão

O problema acima engloba toda a base do conceito de SVM, porem podemos nos deparar com um único defeito: E se os dados não forem linearmente separáveis? Em outras palavras, e se não for possível encontrar um hiperplano que separe todos os exemplos de uma classe de todos os exemplos da outra?

Neste caso, não haveria qualquer combinação de $\boldsymbol{w}$ e b que pudesse satisfazer o conjunto de restrições acima. Um exemplo de tal situação está ilustrado na figura 4.4.

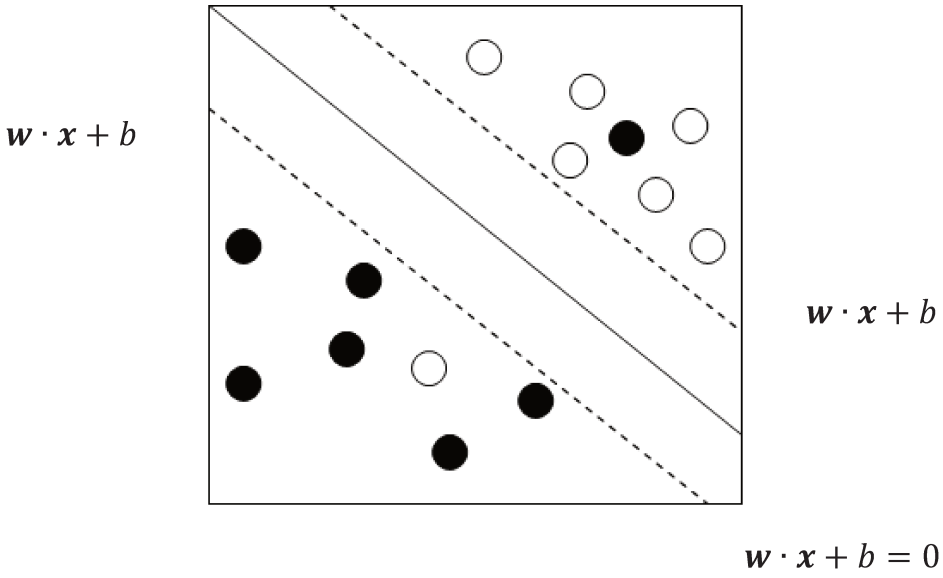


Figura 4.4[27]: Um problema de classificação linearmente inseparável. Aqui temos de separar os pontos pretos dos brancos, veja que não existe reta alguma que possa fazer isso. No entanto, se fôssemos restritos apenas a uma decisão linear, o exemplo mostrado na figura seria o melhor possível, i.e., o que mistura os pontos o mínimo possível.

Para isso, precisamos, de certa forma, amenizar as restrições de que alguns desses pontos estejam no lado +1 ou −1, i.e., temos de permitir que alguns pontos possam violar essas restrições por um pequeno fator. Essa alternativa acaba sendo útil não somente para banco de dados linearmente inseparáveis, mas também como uma forma de melhoria em geral.

A forma para lidar com o problema de dados linearmente inseparáveis é de aumentar a dimensão do problema original, mapeando os pontos da superfície original para dimensões maiores. Diferentemente do caso mostrado na figura 5 acima, onde a solução linear ainda é possível se aceitarmos um pequeno erro em troca de performance computacional, existem dados que não são nem próximos de serem linearmente separáveis; como é o caso da figura 4.5 abaixo; onde apenas uma solução elíptica pode resolver o problema.

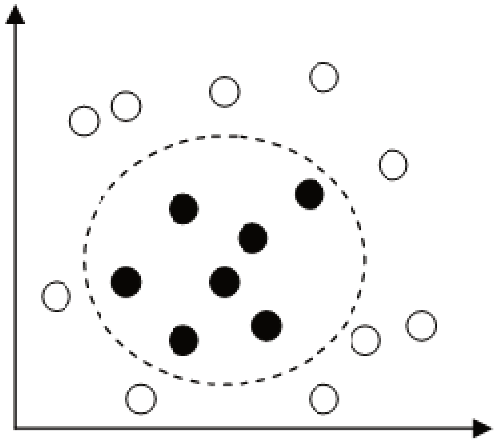


Figura 4.5[27]: Um problema de classificação linearmente inseparável. Nesse exemplo nenhuma reta chega perto de uma solução aceitável, mas uma solução elíptica é claramente visível.

Da forma como estruturamos nosso problema de SVM anteriormente, seria impossível resolver um caso como esse de separar os pontos pretos dos brancos. No entanto, lembrando que a maioria dos problemas reais tem mais de duas dimensões naturalmente, não há razão alguma para hesitarmos em aumentar a dimensão do nosso.

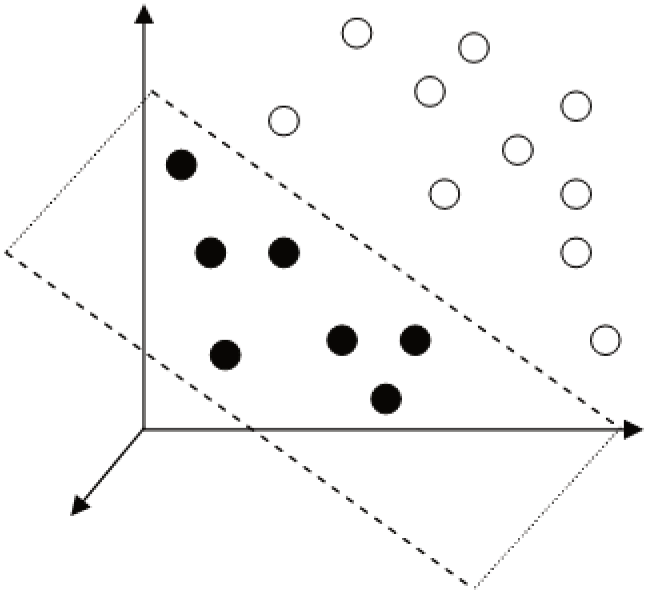


Figura 4.6[27]: Se mapearmos os dados do exemplo anterior em um espaço de maior dimensão podemos encontrar uma solução linear para separar os dados

Imagine que seja possível mapear os dados em um novo espaço de maior dimensão, na qual esses dados sejam linearmente separáveis. Então, se conhecermos esse mapeamento podemos mapear nossos dados para esse novo espaço e assim realizar o SVM normalmente. Se conseguirmos uma boa margem de separação nesse novo espaço, podemos esperar uma boa performance de forma geral apesar do aumento da dimensionalidade. A figura 4.6 ilustra essa situação.

5 Referências Bibliográficas

[1] SILVEIRA, Graciele Paraguaia; VENDITE, Laercio Luis, **Aplicação da teoria de conjuntos fuzzy na predição do estadiamento patologico do cancer de prostata**, disponível em: <http://www.bibliotecadigital.unicamp.br/document/?code=vtls000412772&opt=4>,.

[2] CASTANHO, Maria Jose de Paula; YAMAKAMI, Akebo, **Construção e avaliação de um modelo matematico para predizer a evolução do cancer de prostata e descrever seu crescimento utilizando a teoria dos conjuntos fuzzy**, disponível em: <http://www.bibliotecadigital.unicamp.br/document/?code=vtls000350979&opt=4>,.

[3] SILVEIRA, Graciele Paraguaia *et al*, A metodologia ROC na avaliação de um modelo fuzzy de predição do estádio patológico do tumor de próstata., **Rev. bras. eng. biomed**, v. 26, n. 1, p. 3–9, 2010.

[4] CASTANHO, M. J.P. *et al*, Fuzzy Receiver Operating Characteristic Curve: An Option to Evaluate Diagnostic Tests, **Trans. Info. Tech. Biomed.**, v. 11, n. 3, p. 244–250, 2007.

[5] MARIA JOSÉ DE PAULA CASTANHO, Laécio Carvalho de Barros, Fuzzy expert system: An example in prostate cancer., **Applied Mathematics and Computation**, v. 202, p. 78–85, 2008.

[6] JESS, B, Hyperplasia vs. Neoplasia - Understanding Abnormality, 2010.

[7] **What is Cancer**, American Cancer Society.

[8] *Ibid*.

[9] Ana Paula Roque da Silva, Cláudio Pompeiano Noronha et al, *Incidencia de Cancer No Brasil*, ed. Leticia Casado (Rio de Janeiro, Brasil: Flama, 2012).

[10] SOBIN, Leslie H.; GOSPODAROWICZ, Mary K.; WITTEKIND, Christian (Orgs.), **TNM Classification of Malignant Tumours**, 7. ed. [s.l.]: Wiley-Blackwell, 2009.

[11] *Ibid*.

[12] TANNENBAUM, Myron, **Urologic pathology: The prostate**, [s.l.]: Lea & Febiger, 1977.

[13] PARTIN, A W; OESTERLING, J E, The clinical usefulness of prostate specific antigen: update 1994, **The Journal of urology**, v. 152, n. 5 Pt 1, p. 1358–1368, 1994.

[14] BALK, Steven P; KO, Yoo-Joung; BUBLEY, Glenn J, Biology of prostate-specific antigen, **Journal of clinical oncology: official journal of the American Society of Clinical Oncology**, v. 21, n. 2, p. 383–391, 2003.

[15] SCHRÖDER, Fritz H *et al*, Screening and prostate-cancer mortality in a randomized European study, **The New England journal of medicine**, v. 360, n. 13, p. 1320–1328, 2009.

[16] CHAWLA, Nitesh V., Data Mining for Imbalanced Datasets: An Overview, *in*: MAIMON, Oded; ROKACH, Lior (Orgs.), **Data Mining and Knowledge Discovery Handbook**, [s.l.]: Springer US, 2010, p. 875–886.

[17] MCDONALD, John, **Handbook of Biological Statistics**, 2. ed. Baltimore, Maryland: Sparky House Publishing, 2009.

[18] RODGERS, Joseph Lee; NICEWANDER, W. Alan, Thirteen Ways to Look at the Correlation Coefficient, **The American Statistician**, v. 42, n. 1, p. 59, 1988.

[19] *Ibid*.

[20] GOODMAN, S N, Toward evidence-based medical statistics. 1: The P value fallacy, **Annals of internal medicine**, v. 130, n. 12, p. 995–1004, 1999.

[21] DEVILLE, Barry, **Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner**, [s.l.]: SAS Publishing, 2006.

[22] LOVELL, Brian C.; WALDER, Christian J., Support Vector Machines for Business Applications, p. 267–290, 2006.

[23] CORTES, Corinna; VAPNIK, Vladimir, Support-vector networks, **Machine Learning**, v. 20, n. 3, p. 273–297, 1995.

[24] GREENWOOD, Priscilla E.; NIKULIN, Michael S., **A Guide to Chi-Squared Testing**, 1. ed. [s.l.]: Wiley-Interscience, 1996.

[25] MUSALIA, John, The Chi-Square Test, *in*: , Western Kentucky University: [s.n., s.d.].

[26] NEWSOM, Jason, t-Tests, Chi-squares, Phi, Correlations, **Portland State University**, 2011.

[27] PEIRCE, Charles Sanders, A Neglected Argument for the Reality of God, **Hibbert Journal**, v. 7, p. 90–112, 1908.

[28] MCDONALD, **Handbook of Biological Statistics**.

[29] *Ibid*.

[30] OLIVEIRA, Stanley, Notas de Aula, 2012.

[31] CHAWLA, N. V. *et al*, SMOTE: Synthetic Minority Over-sampling Technique, **arXiv:1106.1813**, 2011.

[32] CARLETTA, Jean, Assessing agreement on classification tasks: the kappa statistic, **Comput. Linguist.**, v. 22, n. 2, p. 249–254, 1996.

[33] LOVELL; WALDER, Support Vector Machines for Business Applications.