UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ANDRÉ LUIZ ABREU SANTOS

O CLASSIFICADOR NAÏVE BAYES NO CONTEXTO DA ANÁLISE DE CRÉDITO

São Luís
2013

ANDRÉ LUIZ ABREU SANTOS

O CLASSIFICADOR NAÏVE BAYES NO CONTEXTO DA ANÁLISE DE CRÉDITO

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Msc. Ivo José da Cunha Serra

São Luís

2013

ANDRÉ LUIZ ABREU SANTOS

O CLASSIFICADOR NAÏVE BAYES NO CONTEXTO DA ANÁLISE DE CRÉDITO

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Aprovada em 13 de dezembro de 2013

BANCA EXAMINADORA

Prof. Ivo José da Cunha Serra (Orientador)
Mestre em Engenharia de Eletricidade
Universidade Federal do Maranhão

Prof. Carlos Eduardo Portela Serra de Castro
Mestre em Informática
Universidade Federal do Maranhão

Prof. Samyr Béliche Vale
Doutor em Informática
Universidade Federal do Maranhão

Santos, André Luiz Abreu.

O classificador Naïve Bayes no contexto da análise de crédito/ André Luiz Abreu Santos. – São Luís, 2013.

73 f.

Impresso por computador (fotocópia).

Orientador: Ivo José da Cunha Serra.

. Monografia (Graduação) – Universidade Federal do Maranhão, Curso de Ciência da Computação, 2013.

1. Mineração de dados. 2.Classificação – Naïve Bayes. 3. KDD. I. Título.

CDU 004.052.42

AGRADECIMENTOS

Agradeço a Deus por me possibilitar a conclusão deste trabalho, sem sua ajuda nada disso seria possível.

Ao Professor Ivo, com sua paciência, sabedoria e interesse depositados na orientação deste trabalho.

A minha mãe Lúcia e meu pai Arnaldo, por não medirem esforços para me oferecer o máximo em educação, me apoiando e suportando em toda a minha vida com amor, paciência e dedicação.

A minha vó, Maria José, por sempre estar presente, me acompanhando e aconselhando, me transmitindo tranquilidade e paz.

A minha irmã, Renata, pelos conselhos e companheirismo, sempre torcendo pelo meu sucesso.

Aos meus amigos do curso de Ciência da Computação: Welsson, Nanderson Wagner (*in memorian*), Higo, Genilson, Alessandro, Keila, Felipe Aragão, Filipe Hiluy, Leonardo, Jefferson, Luis Felipe, Rubem, Márcio, Antônio e Fábio por contribuírem em minha jornada como pessoa e estudante ao longo destes anos de estudo.

Aos meus amigos Anderson, Mábio e Paulo que tanto me incentivaram e apoiaram na conclusão deste trabalho.

Aos professores da UFMA, que me auxiliaram na busca por conhecimento, sempre com interesse e dedicação aos alunos.

A todos que sempre torceram por mim e me ajudaram a realizar este trabalho.

RESUMO

O crescimento do número de informações, justificado pelo aumento da utilização de equipamentos computacionais, em meados da década de 80, foi responsável por uma considerável baixa no custo de armazenamento de dados. As empresas, em geral, passaram a buscar uma maneira de procurar informações válidas em meio a enormes conjuntos de dados armazenados, na tentativa de obter vantagem em mercados de trabalho tão competitivos. Em meio a mercados tão acirrados, empresas ligadas a análise de crédito ganharam destaque na exploração de dados objetivando a extração da maior quantidade de informações ocultas, visando a realização de boas escolhas entre clientes adimplentes e inadimplentes, operação chave para garantir a sobrevivência de empresas concedentes de crédito. Este exemplo de operação ilustra uma das mais comuns fases da metodologia de mineração de dados: a classificação**,** que objetiva realizar associações de novos exemplos a classes presentes em bases de dados históricas conhecidas. O foco deste trabalho recai sobre um classificador em particular, o Naïve Bayes. Tal técnica, baseada em fórmulas probabilísticas, atua no contexto de classificar novos clientes de instituições financeiras como sendo bons ou maus pagadores. Neste trabalho são apresentados dois estudos de caso: o primeiro, no qual é discutido o funcionamento e execução do classificador Naïve Bayes, e o segundo, através do qual são comparados os resultados do classificador Naïve Bayes com classificadores baseados em redes neurais e árvores de decisão. Os resultados finais foram apresentados e discutidos, bem como as contribuições acadêmicas oferecidas por este trabalho.

Palavras-Chave: Classificação, Naïve Bayes, Mineração de dados, KDD

ABSTRACT

The growth of information, justified by the increased use of computer equipment in the mid- 80s, was responsible for a considerable decrease in the cost of data storage. Companies in general began to seek a way to search for valid information in the midst of huge data sets stored in an attempt to gain an advantage in the competitive labor markets. In the middle of fierce markets, companies related to credit analysis gained spotlight in data exploration, aiming the extraction of the greatest amount of hidden information, seeking the good choices between good and bad payers customers, key operation to ensure the survival of companies grantors credit. This example illustrates the operation of one of the most common phases of the methodology of data mining: the classification that aims to make new associations of examples to classes present in databases of known historical data. The focus of this work lies precisely on a classifier called Naïve Bayes .This technique, based on probabilistic formulas , operates in the context of classifying new financial institution clients as being good or bad payers. In this work, two study cases are presented, the first one, in which is discussed the implementation and operation of the classifier Naïve Bayes, and the second, by which the results of the Naïve Bayes classifier with classifiers based on neural networks and decision trees are compared. The final results were presented and discussed, as well as the academic contributions made by this work.

Key words: Classification, Data Mining, KDD, Naïve Bayes

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O mercado financeiro hoje em dia, é bastante competitivo. Empresas buscam formas de encontrar saídas na tentativa de obter vantagens na extração de informações essenciais para a sobrevivência de seus patrimônios. Na área de análise de crédito, ferramentas que auxiliem na classificação de clientes como bons ou maus pagadores se tornam um elemento chave que gera grande vantagem competitiva.

Com o maior acesso equipamentos computacionais, a partir da década de 80, o custo de armazenamento de dados caiu consideravelmente. Logo, as empresas começaram a notar que poderiam estar perdendo informações importantes dentre o enorme estoque de dados armazenados. Então se passou a buscar uma forma de extrair o máximo de informações válidas escondidas na imensidão dos dados que por si só não transmitem qualquer informação que possibilite o entendimento ou interpretação de uma determinada situação. Com a transformação dos dados em informações, o processamento destes dados se torna possível, possibilitando sua interpretação e posterior qualificação.

Neste contexto, a mineração de dados surge como uma metodologia que visa extrair conhecimento útil a partir de bases de dados históricas auxiliando na tomada de decisões estratégicas e melhorando a qualidade e eficiência destas decisões.

Segundo Fayyad et al. (1996) a Mineração de Dados implica na aplicação de algoritmos para análise e descoberta de conhecimento e na produção de padrões ou modelos a partir de grandes bases de dados. E na análise de crédito, a utilização adequada do conhecimento para corretas decisões de crédito é imprescindível pois segundo Steiner (1999, p. 56) “a correta decisão de crédito é essencial para a sobrevivência das empresas financeiras” afirma ainda que “qualquer erro na decisão de concessão pode significar que em uma única operação haja a perda do ganho obtido em dezenas de outras bem sucedidas”.

Dentre as fases contidas na mineração de dados, podemos destacar a fase de classificação, responsável por associar ou classificar novos dados de acordo com determinada classe ou função.

O objetivo deste trabalho é apresentar através da fase de classificação o classificador Naive Bayes (Hand, 2009), e suas aplicações em relação a domínios financeiros com o auxílio da ferramenta de mineração de dados WEKA (Abernethy,

2010). Espera-se contribuir para a evolução do processo de descoberta de conhecimento bem como a tarefa de classificação de dados.

1.1 Motivação

No processo de classificação, existem diversas técnicas com várias particularidades e funções, como redes neurais, árvores de decisões, algoritmos genéticos e SVM's (*Support Vector Machines*). Mas o que realmente nos interessa neste trabalho é o classificador Naïve Bayes (Hand, 2009) (Gonçalves, 2013) baseado em probabilidades, grande motivação para o desenvolvimento deste trabalho, pois apesar de sua simplicidade, produz resultados bastante satisfatórios, acompanhado de baixos custos computacionais e viabilidade notável, tornando-se uma boa opção para domínios financeiros.

Também é necessário citar, que técnicas de classificação são bastante relativas, cada uma podendo garantir desempenhos satisfatórios em determinados problemas de acordo com suas particularidades e métodos únicos. É importante ressaltar que esta variação em relação ao desempenho dos classificadores em geral é que motiva o interesse pela busca do método mais eficaz para cada aplicação em área específica. Neste trabalho, o foco está em conhecer a apresentação da técnica bayesiana de classificação.

1.2 Objetivos do Trabalho

O objetivo deste trabalho está na apresentação da técnica bayesiana de classificação sob o contexto da análise de crédito, buscando resultados em bases de dados de domínios financeiros.

Este trabalho possui como objetivos:

1) Apresentar a técnica de classificação Naïve Bayes, destacando seus aspectos positivos e limitações

2) Apresentar resultados do classificador Naïve Bayes no contexto de análise de crédito, a partir de estudos de casos com bases de dados voltadas ao domínio financeiro.

3) Avaliação qualitativa entre o Naïve Bayes e outras técnicas de classificação, como árvores de decisão e redes neurais, na tentativa de

notar os pontos positivos e negativos do Naïve Bayes no contexto financeiro.

1.3 Estrutura do trabalho

O capítulo 2 trata sobre o processo de classificação, seu conceito, suas características, discutindo métricas avaliativas de modelos classificadores, bem como as técnicas avaliadoras de classificadores.

O capítulo 3 aborda a técnica utilizada no algoritmo de classificação Naïve Bayes, definindo seu conceito, sua origem, seu desenvolvimento, suas particularidades, sua atuação no processo de classificação, demonstrando seus pontos positivos e limitações.

No capítulo 4, temos os estudos de casos baseados em bases de dados financeiras onde são explorados os princípios e métodos de execução do Naïve Bayes, além de outro estudo de caso através dos qual é realizada uma comparação da técnica de classificação bayesiana com a técnica de árvore de decisão e a técnica de redes neurais. Neste capítulo ainda são apresentados resultados obtidos mediante tal comparação.

No capítulo 5 são feitas as conclusões finais sobre este trabalho, apresentando expectativas sobre trabalhos futuros e contribuições acadêmicas geradas por este trabalho.

2 CLASSIFICAÇÃO DE DADOS

Desde que nascemos somos impelidos a classificar. Classificamos, reorganizamos, priorizamos e reclassificamos nossos próprios pensamentos, atitudes e ações. A Classificação é um processo indispensável e até mesmo elementar no cotidiano humano. Classificar induz a associar, a ligar uma coisa a outra, uma pessoa a outra, uma informação, um conhecimento.

A tarefa de classificação é uma das mais conhecidas e utilizadas no contexto de mineração de dados e possui caráter importante na tarefa de analisar grandes bases de dados com o objetivo de se extrair o máximo de conhecimento válido possível.

Na prática, a classificação realiza a associação ou classificação de um objeto a uma classe já existente, prevendo assim a qual classe um novo dado será associado ou classificado. Por exemplo, temos uma base de dados com os dados de clientes junto a instituições financeiras, a partir das características baseadas em suas transações anteriores ou até mesmo em suas características pessoais como renda e idade, podemos classificar estes clientes em categorias para posterior análise de crédito. Além deste último exemplo, outros tipos de aplicações práticas interessantes podem ser aplicadas em áreas distintas como *marketing*, finanças, comércio, diagnósticos médicos bioinformática, segurança de informações.

Na área de bioinformática podemos prever a classe de determinadas proteínas, para se determinar a função delas, utilizando um algoritmo classificador. Na área de finanças, podemos detectar fraudes a partir de filtragens de transações financeiras, as classificando como legais ou suspeitas. Também utilizamos a classificação em filtragem de spam de e-mails onde estes últimos são classificados como spam ou não.

Este capítulo aborda o conceito de classificação, bem como os processos e métodos utilizados nesta fase de mineração de dados, apresentando um apanhado sobre as técnicas gerais na determinação da eficiência dos classificadores.

2.1 Definição

Classificação, segundo HAN e KAMBER (2000), é o processo de encontrar um modelo ou função que descreve e distingue classes de dados e conceitos, com o objetivo de usar este modelo para predizer objetos ainda não classificados e consiste em examinar as características de um objeto (ou situação) e atribuir a ele uma classe pré-definida, ou seja, esta tarefa tem o propósito de construir modelos que permitam o agrupamento de dados em classes. A classificação é considerada também preditiva, pois uma vez que as classes são definidas, ela pode prever automaticamente a classe de um novo dado.

Uma técnica estatística interessante no método de classificação é a análise discriminante. Os objetivos desta técnica são definidos por uma descrição algébrica ou gráfica das características que diferenciam das de outras diversas populações, classificando estas características em uma ou mais classes predeterminadas. O foco principal é derivar uma regra para ser utilizada na classificação de uma nova característica ligada a uma classe já rotulada. Segundo MATTAR (1998), a análise discriminante permite que dois ou mais grupos possam ser comparados, com o objetivo de determinar se diferem uns dos outros de maneira que, a partir de uma base em um conjunto de variáveis, seja possível classificar indivíduos ou objetos em duas ou mais categorias mutuamente exclusivas.

Ao analisar um conjunto de dados de treinamento e construir um modelo para cada classe baseado nas características dos dados, um conjunto de regras de classificação é gerado pelo processo de classificação, que pode ser usado para entender melhor cada classe no banco de dados e para classificação de futuros dados.

Os estudos de caso, apresentados no capítulo 4, apresentam o trabalho envolvendo classes de bancos de dados voltados a aplicações de crédito, onde temos um histórico relacionado às características dos clientes em relação a possíveis empréstimos ou pagamentos junto a instituições financeiras. Os clientes foram classificados em “Bons Pagadores” ou “Maus Pagadores”. E é a partir das características utilizadas para se determinar a classe destes clientes, que descobrimos uma função definidora capaz de associar corretamente os clientes, através de seus dados e características, às classes supracitadas.

Segundo STEINBACH e KUMAR (2009), os dados de entrada da tarefa de classificação são um conjunto de registros que por sua vez possui um conjunto de atributos e outro conjunto denominado de rótulo de classe. A função da tarefa de classificação é aprender uma função que seja capaz de mapear cada conjunto de atributos a um dos rótulos de classes pré-determinados. Essa função é conhecida como modelo de classificação, como mostrado na figura 2.1.

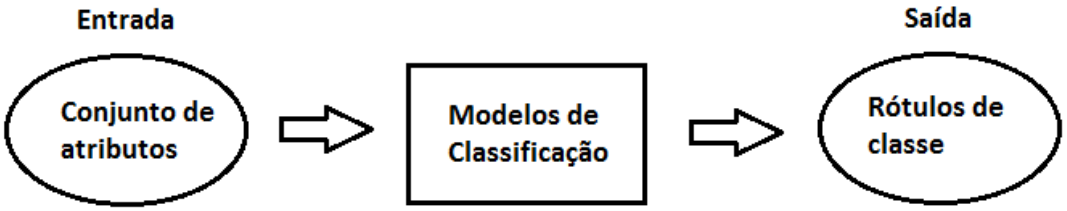


Figura 2.1 – Esquema da tarefa de classificação

Os modelos de classificação se dividem em modelo descritivo e modelo preditivo, como definido abaixo:

1) **Modelo Descritivo –** É aquele que nos ajuda a entender a realidade já existente em determinada situação e parte do pressuposto que devemos esclarecer o campo de conhecimento para que tenhamos termos claros e consistentes para uma compreensão mais clara. O modelo descritivo é utilizado para se distinguir características e objetos de classes diferentes. A partir da tabela 2.1, podemos gerar um modelo descritivo, para diferir os clientes bons pagadores dos maus pagadores, o que seria bastante útil e interessante no entendimento de como os atributos influenciam na determinação da classe final como bom ou mau pagador.

Sexo	Renda	Idade	Filhos	Bom_Pagador(Classe)
F	≥3000	≥40	Não	Sim
F	1500-3000	20-39	Sim	Não
F	≤1500	≥40	Não	Não
M	≤1500	≥40	Não	**Sim**
F	1500-3000	≥40	Sim	Não
M	≥3000	20-39	Sim	Sim
F	≥3000	≥40	Não	Sim
F	≤1500	≥40	Não	Não
M	≥3000	≥40	Sim	Sim
M	≤1500	20-39	Não	Não

Tabela 2.1 – Histórico de dados de uma instituição financeiras

2) **Modelo Preditivo** - Neste modelo, são previstos os rótulos de classes de registros não conhecidos. Neste trabalho, por exemplo, a partir das características (definidas como atributos preditivos) de clientes conhecidos, realizamos a previsão da classe de clientes junto a empresas financeiras, classificando estes clientes como bons e maus pagadores.

A maioria do modelos de classificação é construída em quatro etapas. O primeiro passo é identificar um conjunto de indivíduos com um comportamento comum e conhecido, posteriormente as entradas são conhecidas, assim como as classes de destino. A segunda etapa consiste na preparação dos dados, incluindo a limpeza dos dados, seleções de recursos e transformação dos dados. No terceiro passo, ocorre o treinamento dos modelos. Neste processo se utiliza em média 80% dos indivíduos identificados no primeiro passo para identificar a relação das entradas e dos dados de destino, esta relação depende diretamente do algoritmo utilizado. Na quarta etapa, acontece o teste do modelo, neste os 20% dos indivíduos restantes são utilizados para testar as relações identificadas na terceira etapa e é onde a precisão do modelo também é testada. Na quinta fase, ocorre o aprimoramento e aperfeiçoamento do modelo como a eliminação de *overfitting*, regularização e validação.

Os modelos de classificação são diferentes das técnicas de classificação (classificadores) num aspecto fundamental. Os modelos de classificação geram informalmente uma função que associa cada registro de uma base a uma classe de

rótulos categóricos, tal função é criada por meio de abordagens específicas conhecidas por técnicas de classificação (classificadores). Tomamos, por exemplo, o classificador Naïve Bayes baseado em probabilidades abordado neste trabalho que gera um modelo probabilístico de classificação, ou seja, a partir de uma tabela probabilística gerada no modelo de classificação apoiado na regra de Bayes é que a técnica de classificação (classificador) bayesiana trabalha.

2.2 Avaliando modelos de classificação

Após a construção do modelo de classificação, este é testado por diversas métricas de avaliação, nas seções 2.2.1, 2.2.2, 2.2.3 e 2.2.4 são discutidas e formalmente definidas algumas destas métricas para melhor estudo.

2.2.1 Acurácia

Acurácia, também chamada *accuracy* (do inglês), mede a porcentagem de acertos em problemas de classificação. É um dos métodos mais utilizados para se avaliar modelos classificadores e é denotada pela fórmula 2.1.

$$Acurácia(\%) = \frac{QC}{QT} \times 100 \qquad (2.1)$$

Onde ***QC*** se refere à quantidade de registros classificados de maneira correta e ***QT*** aponta a quantidade total de registros presentes em um dado conjunto.

Partindo para um exemplo prático para ilustração da acurácia, tomamos por base, a tabela 2.1 tratando sobre um histórico de dados voltados a características de clientes de uma financeira quanto a serem bons ou maus pagadores. A partir desta tabela, geramos um modelo classificador “A” capaz de classificar corretamente sete das dez instâncias contidas na base de dados da tabela 2.1. Então, temos o cálculo da acurácia na fórmula 2.2.

$$Acurácia(\%) = \frac{7}{10} \times 100 = 70\% \qquad (2.2)$$

A acurácia calculada a partir de um conjunto de dados utilizado para construir um classificador é chamada de acurácia de treinamento, chamada de $Acurácia_{treina}$.

Agora que temos um modelo classificador "A", podemos testá-lo em uma base de dados desconhecida, para que possamos submetê-lo a um melhor teste de desempenho. A tabela 2.2 mostra o conjunto de dados para o teste do modelo de classificação "A":

Sexo	Renda	Idade	Filhos	Bom_Pagador(Classe)
F	≥3000	≥40	Sim	Não
M	1500-3000	20-39	Não	Sim
M	≤1500	≥40	Não	Sim
F	1500-3000	20-39	Não	Sim
F	≤1500	≥40	Sim	Não
M	≥3000	20-39	Sim	Não
F	≥3000	≥40	Não	Sim
F	≤1500	≥40	Não	Sim
F	≥3000	20-39	Não	Sim
M	≤1500	20-39	Não	Sim

Tabela 2.2 - Conjunto de dados fictício do histórico de clientes de uma financeira

Suponhamos então, que a partir do teste com o conjunto desconhecido de dados da tabela 2.2, o modelo classificador "A" tenha classificado corretamente seis das instâncias presentes. A acurácia será denotada pela fórmula 2.3.

$$Acurácia(\%) = \frac{6}{10} \times 100 = 60\% \qquad (2.3)$$

A acurácia medida a partir de um conjunto de testes é chamada de acurácia de teste ou $Acurácia_{teste}$.

Como notado, o modelo classificador obteve um desempenho razoável na acurácia de teste e treinamento. É importante enfatizar, porém, que medidas de acurácia de treinamento altas não implicam necessariamente que temos um bom modelo classificador, assim como baixas acurácias de treinamento não implicam em modelos ruins. Se os modelos apresentam altas taxas de acurácia de treinamento, podemos notar, somente, que este apresenta uma boa síntese do conjunto de dados

original (de treinamento). Porém o que se pode levar em conta é a taxa da acurácia de teste, que servirá de grande valia para a construção da acurácia de execução correspondente à taxa de previsão do classificador. O que realmente importa é que ambas acurácia de treinamento e de teste sejam elevadas para que tenhamos um bom modelo classificador.

2.2.1.1 Problemas com acurácia

Consideremos agora que tenhamos um conjunto de instâncias com as seguintes distribuições de classes:

- Distribuição (Classe 1, Classe 2) = (99%, 1%)

Um modelo classificador que venha a classificar sempre novos exemplos como pertencentes à classe majoritária 1, teria uma acurácia de 99%, à princípio sendo um ótimo modelo, obtendo quase o máximo de acurácia possível. Porém, se as classes 1 e 2 fossem definidas da seguinte maneira:

- Classe 1 = Bom pagador
- Classe 2 = Mau Pagador

Teríamos um grande problema, pois a classe 2, minoritária, com um índice de ocorrência de 1%, guardaria informações totalmente relevantes para uma instituição financeira que viesse a analisar os dados dos clientes para possíveis empréstimos já que bons pagadores seriam maioria em um modelo com eficácia de 99%, deixando de classificar maus pagadores como maus pagadores e os classificando como bons pagadores, causando prejuízos às instituições.

Exemplo:
Classe 1: Bom Pagador (450 clientes)
Classe 2: Mau Pagador (5 clientes)

Suponhamos que um modelo “M”, tenha classificado corretamente 410 dos clientes como bons pagadores de um total de 455 clientes analisados, não tendo

acertado, porém, nenhum dos 5 clientes maus pagadores. Mesmo assim sua acurácia seria de 90,1%.

Acurácia(M) = 90,1% = Bom Classificador?

Problemas como este, envolvendo classes majoritárias e minoritárias são chamados de classes desbalanceadas. Neste tipo de problema é aconselhável utilizar outra métrica de avaliação diferente da acurácia, pois os sistemas envolvendo mineração de dados são projetados para otimizar a acurácia, o que favoreceria uma precisão enganosa e consequentemente classificadores de péssimo desempenho quando utilizados em conjuntos de dados com classes desbalanceadas.

Algumas técnicas foram desenvolvidas para lidar com o problema das classes desbalanceadas como a detecção de exemplos de borda e com ruído, introdução de custos de classificação incorreta ou até mesmo redução de amostras redundantes ou prejudiciais. Mas a que interessa neste trabalho é a técnica da matriz de custo, que realiza a atribuição de custos por classe para cada erro na tarefa de classificação. Veremos com mais detalhes esta técnica na seção 2.2.4.

2.2.2 Taxa de erro

A taxa de erro é uma medida utilizada de maneira análoga à métrica de acurácia, sendo um complemento desta última. Ou seja, enquanto a acurácia indica o número de instâncias classificadas corretamente para se obter a precisão de acerto que um modelo classificador possui, a taxa de erro é constituída do número de instâncias classificadas erroneamente num conjunto de dados para se chegar à taxa de classificações incorretas. É denotada pela fórmula 2.4:

$$Taxa\ de\ erro = \frac{QI}{QT} \times 100 \qquad (2.4)$$

Onde ***QI*** corresponde ao número de instâncias classificadas incorretamente, e ***QT*** define o número de total de instâncias presentes no conjunto de dados. Como dito anteriormente, a taxa de erro é complementar à acurácia e se define pela fórmula 2.5.

$$Taxa\ de\ erro = 1 - Acurácia \qquad (2.5)$$

2.2.3 Matriz de confusão

A contagem das classificações corretas e incorretas dos modelos de classificação é armazenada em uma matriz chamada de matriz de confusão, ilustrada na tabela 2.3.

Matriz de Confusão		Classe Predita		
		Classe 1	Classe 2	Classe 3
Classe Atual	Classe 1	Classe 11	Classe 12	Classe 13
	Classe 2	Classe 21	Classe 22	Classe 23
	Classe 3	Classe 31	Classe 32	Classe 33

Tabela 2.3 - Modelo geral de uma matriz de confusão com três classes

Na matriz de confusão existem duas divisões, as classes atuais dos registros realmente pertencentes ao conjunto de dados, e as classes preditas que são as classes indicadas pelo modelo classificador. As estimativas de classificação corretas, correspondentes à classe real a que os registros pertencem, são posicionadas na diagonal da matriz, enquanto que as estimativas incorretas, ou seja, classificadas de maneira errônea pelo classificador permanecem nas outras posições ao longo da matriz de classificação.

Considerando então, que um modelo classificador realizou, por exemplo, a classificação da categoria de um grupo de 66 carros em “Sedan”, “SUV” e “Minivan”. O modelo de classificação foi construído para prever a categoria de automóvel que um cliente compraria ao frequentar um salão de automóveis, e obteve uma matriz de confusão como na tabela 2.4:

Matriz de Confusão		Classe Predita		
		Sedan	SUV	Minivan
Classe Atual	Sedan	25	1	0
	SUV	2	30	4
	Minivan	0	2	2

Tabela 2.4 – Exemplo de matriz de confusão contendo categoria de carros

A matriz de confusão, obteve um índice de 25 Sedan's, 30 SUV's e 2 Minivan's classificadas de maneira correta, o restante das classificações, 1 Sedan, 6 SUV's e 2 Minivan's; foram classificadas de maneira incorreta. O modelo teve um bom desempenho. Assim, podemos ter uma boa ideia da performance do modelo classificatório, ao analisarmos a matriz de confusão construída. A partir desta mesma análise, também podemos obter a acurácia do modelo de classificação, sendo calculada pelo número de estimativas corretas divido pelo número de estimativas incorretas, como visto na seção 2.2.1

A acurácia calculada a partir da matriz de confusão acima é como segue, na fórmula 2.6:

$$Acc = \frac{Classe\ 11 + Classe\ 22 + Classe\ 33}{Classe\ (11+22+33) + Classe(11+12+13) + Classe(11+12+33)} \quad (2.6)$$

$$Acc = \frac{25 + 30 + 2}{25 + 1 + 0 + 2 + 30 + 4 + 0 + 2 + 2} = \frac{57}{66} = 0{,}86$$

2.2.4 Matriz de custo

O custo, representado por Cost (C_i, C_j) é um valor que define uma penalidade quando o classificador erra ao classificar rótulos que são pertencentes à classe C_i, como sendo pertencentes da classe C_j, onde i, j = 1, 2, 3, ..., k onde k é o número de classes. Assim, quando temos Cost(C_i, C_j) = 0, para i = j, pois não constitui um erro e Cost(C_i, C_j) > 0 para i ≠ j. Em termos gerais, para i ≠ j, os classificadores assumem que Cost(C_i, C_j) = 1, quando não são definidos valores explícitos (Goldschimdt e Passos , 2005). Como exemplo, temos, na tabela 2.5, um modelo de uma matriz de custo.

Mat. de Custo	**Classe Predita**		
Classe Real	C(i, j)	Sim	Não
	Sim	C(Sim|Sim)	C(Não|Sim)
	Não	C(Sim|Não)	C(Não|Não)

.

Tabela 2.5 – Modelo geral de uma matriz de custo

Em problemas voltados ao domínio financeiro, como o de concessão de crédito, quando um mau pagador é classificado um bom pagador, o custo do prejuízo para o credor é grande e maior do que o custo do prejuízo quando um bom pagador é classificado como um mau pagador. Naturalmente, nas situações em que um bom pagador ou um mau pagador são classificados corretamente, não há custo de prejuízo ao credor. Um exemplo de matriz de custo em problemas de concessão de crédito é ilustrado na tabela 2.6.

Mat. de Custo	**Classe Predita**		
Classe Real	C(i, j)	Bom Pagador	Mau Pagador
	Bom pagador	0	1
	Mau Pagador	5	0

Tabela 2.6 – Matriz de custo associada à tarefa de concessão de crédito

Matrizes de custo também são muito utilizadas para resolver problemas onde temos classes desbalanceadas. Como apresentado na seção 2.2.1.1, estes problemas ocorrem quando os modelos classificadores tendem, ilusoriamente, a classificar novas instâncias como sendo de uma classe majoritária, ignorando, portanto, a classe minoritária e levando ao erro a tarefa de classificação, o que traz alguns problemas. Para ilustrar a solução do problema, tomamos dois modelos de classificação, M_1 e M_2, ao realizarem a tarefa de classificação em um conjunto de dados com 200 bons pagadores e 50 maus pagadores, contidos nas tabelas 2.7 e 2.8.

A seguir, apresentramos a tabela 2.7 e 2.8, respectivamente.

Modelo M_1	**Classe Predita**		
Classe Real	C(i, j)	Bom Pagador	Mau Pagador
	Bom Pagador	100	50
	Mau Pagador	25	75

Tabela 2.7 - Matriz de confusão do modelo M_1

Modelo M_2	**Classe Predita**		
Classe Real	C(i, j)	Bom Pagador	Mau Pagador
	Bom Pagador	115	0
	Mau Pagador	35	100

Tabela 2.8 - Matriz de confusão do modelo M_2

Tomando por base, a matriz de custo da tabela 2.6, calculamos o custo e a acurácia dos dois modelos mencionados nas tabelas 2.8 e 2.9:

Modelo M_1

$$Acurácia = \frac{175}{250} = 0,7 \times 100 = 70\%$$

$$Custo = 100 \times 0 + 50 \times 1 + 25 \times 5 + 75 \times 0 = 50 + 125 = 150$$

Modelo M_2

$$Acurácia = \frac{225}{250} = 0,9 \times 100 = 90\%$$

$$Custo = 115 \times 0 + 0 \times 1 + 35 \times 5 + 100 \times 0 = 50 + 125 = 175$$

Matrizes de custo, como o nome pressupõe, trabalham com base no custo dos modelos classificadores, os quais foram calculados acima. No modelo M_1, tivemos uma acurácia de 70%, com um custo de 150. Já no modelo M_2, 90% de acurácia e 175 de custo. Ao analisarmos estes resultados, apesar da acurácia mais elevada do que no modelo M_1, o M_2 teve um custo maior do que o M_1. Portanto, o modelo 1 é a melhor opção quando consideramos o custo total dos dois modelos.

2.3 Métodos de avaliação de classificação

Para realizar uma boa avaliação a cerca de classificadores, nas seções 2.3.1 e 2.3.2 são apresentados dois dos principais métodos de avaliação de

classificadores: o *Holdout* e a Validação Cruzada. Serão discutidos suas características e métodos de abordagem.

2.3.1 Holdout

É o método utilizado para dividir os dados originais em dois grupos: os de treinamento e teste. O *holdout* reserva uma parte dos dados para treinamento e o restante, separa para testes. Geralmente é utilizado 2/3 dos dados para treinamento, os 1/3 restantes são voltados para a tarefa de teste. Este método possui algumas deficiências, como a possibilidade de não ocorrência das amostras de determinada classe no grupo de teste por exemplo, já que este geralmente possui a menor quantidade de dados, uma solução poderia ser a utilização de estratificação, o que asseguraria que todas as classes estariam representadas em igual proporção em ambos os grupos (treinamento e teste) .

Outra deficiência encontrada é a grande dependência em torno da composição do grupo de treinamento e teste. Quanto menor for o conjunto de treinamento, maior será a variância do modelo classificador enquanto que se o conjunto de treinamento for muito grande, o conjunto de teste aumentaria bastante o que geraria uma acurácia duvidosa e desconfiável.

2.3.2 Validação cruzada

Neste método, o conjunto de dados é dividido em k subconjuntos, de aproximadamente igual tamanho. Estes k subconjuntos são chamados de *folds*. O algoritmo de classificação é, então, executado k vezes, sendo que em cada rodada de execução, um dos k subconjuntos é tomado como teste, e os restantes k – 1 subconjuntos são tomados como treinamento e assim sucessivamente. As rodadas de execuções só terminam na k-ésima vez quando todos os subconjuntos já foram tomados como teste nas rodadas anteriores.

Por exemplo, consideramos um conjunto de dados D, com 400 exemplos. O conjunto é dividido, aleatoriamente, em quatro subconjuntos, de tamanho igual a 100. Realizamos então, a tomada de três destes quatro subconjuntos como conjunto de treinamento, e um como conjunto de teste. Teremos, portanto, até a quarta execução, quando todos os subconjuntos já foram tomados como conjunto de teste. Esta tarefa é executada ilustrativamente na figura 2.2.

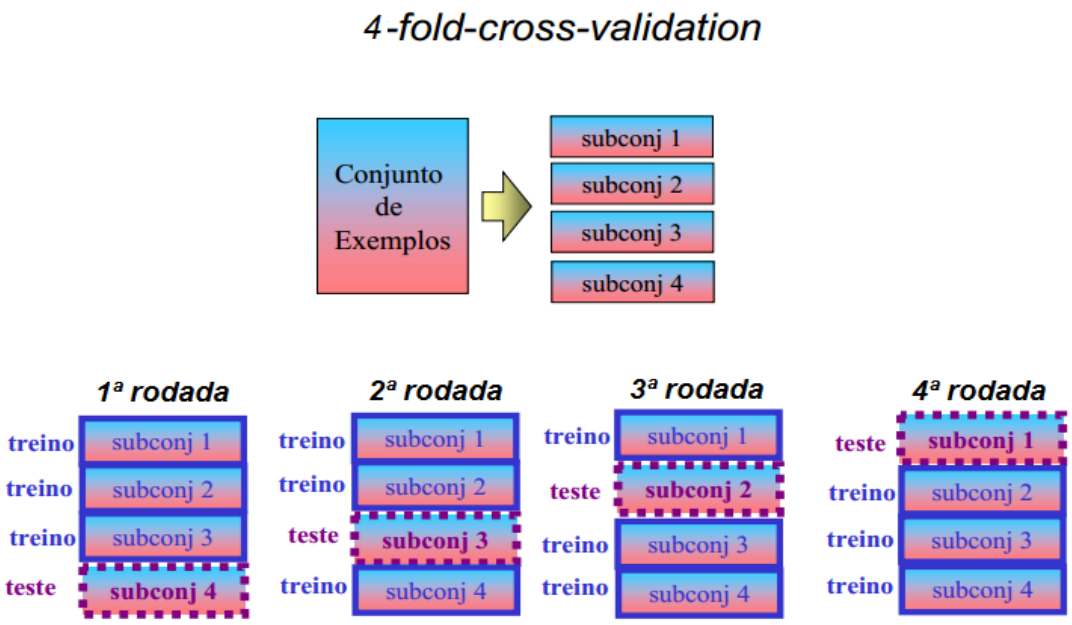


Figura 2.2 – Exemplo de validação cruzada com quatro subconjuntos ou 4-fold-cross-validation (ENGEL, 2008).

A acurácia final é calculada ao final da k-ésima rodada, realizando-se a média da soma das acurácias de cada rodada, ou seja como na fórmula 2.7:

$$Acurácia\ Final = \frac{Acurácia_1 + Acurácia_2 + Acurácia_3 + \ldots + Acurácia_k}{k} \qquad (2.7)$$

Como método padrão de avaliação, a validação cruzada costuma ser realizada até a décima vez (k =10), pois experimentos demonstraram que esta é a melhor escolha para se obter uma estimativa precisa para a maioria das situações. Porém este número (10) serve somente como uma sugestão, não sendo obrigatório e nem impedindo que outra numeração de rodadas venha a obter uma estimativa de avaliação interessante.

Também existem mais dois tipos de validação cruzada: a validação cruzada estratificada e *leave-one-out*. No primeiro, após a geração dos k subconjuntos, mantém-se a mesma proporção de classes nos subconjuntos. Ou seja, se no conjunto original de dados que gerou os subconjuntos, houver duas classes com incidência de 30% em uma e 60% em outra, a mesma proporção de classes deverá ser mantida em cada um dos subconjuntos gerados. Já no método do *leave-one-out*, a quantidade k de subconjuntos é igual à quantidade de exemplos de treinamento, ou seja, haverão de ser construídos k classificadores correspondendo a cada

exemplo de treinamento. Este último método é preciso, realiza o aproveitamento de dados, não envolve sub-amostragem aleatória, porém é computacionalmente muito custoso.

2.4 Considerações Finais

Neste capítulo foi abordada a tarefa de classificação, fundamental no processo de mineração de dados, pois é responsável pela associação de exemplos ou instâncias desconhecidas, a classes provenientes de bases de dados conhecidas. Foram discutidos métricas e métodos que tem o objetivo de avaliar classificadores, dentre eles a acurácia que avalia modelos classificadores através da quantidade de exemplos corretamente classificados em relação ao total de exemplos avaliados, a taxa de erro que é complementar a acurácia, apresentando a quantidade de exemplos incorretamente classificados em um conjunto de dados, a matriz de custo que atribui custos a classificações erroneamente realizadas e a matriz de confusão que armazena a contagem de classificações corretas e incorretas em um conjunto de dados.

Enquanto aos métodos avaliadores, foram definidos o *holdout* e validação cruzada. O *holdout* divide os dados do conjunto de dados em dois grupos: os de treinamento e teste. Já a validação cruzada divide o conjunto de dados original em k subconjuntos, onde enquanto um subconjunto é escolhido como base de teste, os outros k – 1 subconjuntos permanecem como base de treinamento. O mesmo esquema é realizado até que todos os subconjuntos tenham sido escolhidos como base de teste, tendo os outros subconjuntos, obrigatoriamente, permanecido como bases de treinamento até a k-ésima rodada.

3 O CLASSIFICADOR NAIVE BAYES

O classificador Naïve Bayes (Hand, 2009) (Gonçalves, 2013) é provavelmente um dos mais utilizados no que diz respeito a Aprendizado de Máquina. É um dos mais simples, eficientes e populares no processo de classificação, sendo bastante útil em técnicas de mineração de dados. Neste capítulo abordaremos a teoria que fundamenta a construção do teorema de Bayes, assim como tipos de algoritmos deste classificador e como o mesmo atua no processo de classificação, que visa identificar a qual classe um novo objeto pertence para fins de classificar um novo registro no processo de mineração de dados.

3.1 Definição

Naïve Bayes (Hand, 2009) (Gonçalves, 2013) é um classificador estatístico fundamentado em diversos conceitos da teoria da probabilidade, como probabilidade condicional, independência condicional, regra da multiplicação, distribuição conjunta de probabilidades e, especialmente, em uma importante fórmula conhecida como fórmula de Bayes (Hand, 2009). Estes conceitos serão apresentados a partir da seção 3.1.1 até a 3.1.3 próximas.

3.1.1 Probabilidade Condicional

A probabilidade condicional é um importante conceito aplicado à formula de Bayes. A probabilidade condicional é um segundo evento de um espaço amostral que ocorre após a ocorrência do primeiro evento, ou seja, a partir de um espaço amostral Z e um evento A que tenha acontecido em Z, queremos o evento B que também tenha acontecido no espaço amostral Z, a probabilidade condicional será dada então por P(B | A) que indica a probabilidade do evento B em relação ao evento A.

A probabilidade condicional formará um novo espaço amostral contento os elementos do evento B no espaço amostral de A, formando $B \cap A$, como mostra a figura 3.1.

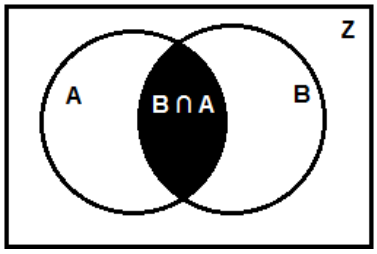


Figura 3.1 – Exemplo do espaço amostral formado pela ocorrência dos eventos A e B

Para calcularmos a probabilidade P (B | A), fazemos:

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

O cálculo de P (B ∩ A) é dado pela multiplicação de P(A) por P(B), como na fórmula 3.1

$$P(B \cap A) = P(A) * P(B) \qquad (3.1)$$

3.1.2 Independência Condicional

O conceito de independência condicional, utilizado pelo classificador Naive Bayes, assume que o efeito do valor de um atributo sobre uma determinada classe é independente dos valores dos demais atributos. Ou seja, consideramos quatro atributos para determinar se uma pessoa possui uma boa renda salarial. Os quatro atributos são: Altura, Peso, Escolaridade e Profissão.

Estes quatro atributos, para o classificador Naïve Bayes, possuem importâncias idênticas para se determinar se uma pessoa possui uma boa renda salarial ou não. O que difere de preceitos reais, que consideram que estes quatro atributos possuem importâncias diferentes em se determinar se um indivíduo possui uma boa renda salarial visto que os atributos ‘escolaridade’ e ‘profissão’ são bem mais relevantes e decisivos do que ‘altura’ e ‘peso’ na determinação da renda de um indivíduo.

O uso da independência condicional no classificador Naïve Bayes simplifica a implementação computacional, porém, não leva em conta fatores reais de dependência entre os atributos na classificação.

3.1.3 Distribuição conjunta de probabilidades

O nome de tal classificador provém da palavra ingênua (naive) devido ao fato de assumir que os atributos contidos no processo de classificação são condicionalmente independentes, ou seja, as informações contidas em um evento não são informativas sobre qualquer outro evento.

Uma característica atraente desse classificador é a sua capacidade de produzir estimativas de probabilidade ao invés de simples classificações. Isto significa que, para cada rótulo de classe, o classificador pode gerar uma estimativa de o novo objeto pertencer à mesma. Por exemplo, considerando a probabilidade de um determinado dia ser nublado ou ensolarado, o classificador Bayes fará a estimativa de o novo objeto, no caso o novo dia, ser nublado ou ensolarado, atribuindo as estimativas de probabilidade adequadas a cada classe representada por “nublado” ou “ensolarado”. Assim, o novo dia poderá ser classificado de acordo com suas probabilidades. A classificação é justificada e reforçada então pela probabilidade numérica que cada novo objeto apresenta perante as opções de classes e não simplesmente classificada aleatoriamente sem qualquer justificativa.

Segundo Hand (2009), o Naïve Bayes é importante por várias razões, dentre elas, a facilidade de construção, não necessitando de quaisquer sistemas de parâmetros iterativos estimativos. Isso significa que pode ser facilmente aplicado a grandes conjuntos de dados já que não é aplicado a tipos específicos e restritos de dados que necessitam de determinadas características, atuando especialmente no processamento de bases de dados bastante volumosas. Ainda segundo Hand (2009), com sua construção simples, o Naïve Bayes possibilita que usuários inexperientes em técnicas de classificação possam entender a essência da classificação de dados e como esta acontece. E completando, Hand (2009) ainda afirma que o Naïve Bayes pode não ser o melhor classificador para determinada aplicação em particular, mas geralmente pode ser utilizado em qualquer aplicação, sendo robusto e funcionalmente excelente, pois é bem aplicado em grandes volumes de bases de dados, realizando o processo de classificação de um ponto de vista computacional simples devido ao uso de técnicas apuradas mas ao mesmo tempo sem grandes complicações ou custos em seu desenvolvimento.

3.2 Teorema de Bayes

Problemas de probabilidades ligados a determinados eventos, em certas condições, já se encontravam bem resolvidos até meados do século XVIII. Tomamos como exemplo um número especifico de bolas pretas e brancas em uma urna, qual a probabilidade de sortearmos uma bola branca? Estes tipos de problemas são chamados de *forward probability*. Mas logo, o problema inverso passou a ser questionado pelos matemáticos da época: Dado o sorteio de uma ou mais bolas, o que pode ser dito sobre o numero de bolas brancas e pretas restantes na urna?

A partir deste problema, Thomas Bayes, ministro britânico do século XVIII, foi o primeiro a idealizar e formalizar um teorema pra solucionar problemas desta natureza vista como revolucionária no meio cientifico da época.

Dado o número de vezes em que um evento desconhecido aconteceu e falhou: É Obrigatório que a chance de que a probabilidade de sua ocorrência em um único teste recai sob quaisquer dois graus de probabilidade de que pode ser conhecido.

O Teorema de Bayes (Hand, 2009) mostra, portanto, a relação entre uma probabilidade condicional e a sua inversa; por exemplo, a probabilidade de uma hipótese dada a observação de uma evidência e a probabilidade da evidência dada pela hipótese. O teorema de Bayes, embasado pela fórmula de mesmo nome, é um corolário do teorema da probabilidade total que permite calcular a probabilidade contida na fórmula 3.2

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \qquad (3.2)$$

- P(A) e P(B) são as probabilidades a priori de **A** e **B**
- P(B|A) e P(A|B) são as probabilidades posteriori de **B** condicional a **A** e de **A** condicional a **B** respectivamente.

A regra de Bayes mostra como alterar as probabilidades a priori tendo em conta novas evidências de forma a obter probabilidades a posteriori.

A ideia principal é que a probabilidade de um evento A *dado um evento* B (e.g. a probabilidade de alguém ter câncer de mama sabendo, ou dado, que a mamografia deu positivo para o teste) depende não apenas do relacionamento entre os eventos A e B (i.e., a precisão, ou exatidão, da mamografia), mas também da probabilidade marginal (ou "probabilidade simples") da ocorrência de cada evento. Por exemplo, se as mamografias acertam em 95% dos testes, então 5% é a probabilidade de termos falso positivo ou falso negativo, ou uma mistura de falso positivo e falso negativo. O teorema de Bayes nos permite calcular a probabilidade condicional de ter câncer de mama, *dado* uma mamografia positiva, para qualquer um desses casos. A probabilidade de uma mamografia positiva será diferente para cada um dos casos.

No exemplo dado, há um ponto de grande importância prática que merece destaque: se a prevalência de mamografias resultado positivo para o câncer é, digamos, 5,0%, então a probabilidade condicional de que um indivíduo com um resultado positivo na verdade não tem câncer é bastante pequena, já que a probabilidade marginal deste tipo de câncer está mais perto de 1,0%.

A probabilidade de um resultado positivo é, portanto, cinco vezes mais provável que a probabilidade de um câncer em si. Além disso, alguém pode deduzir que a probabilidade condicional que mamografias positivas realmente tenham câncer é de 20%. Isso poderia ser menor, se a probabilidade condicional que dado um câncer de mama, a mamografia sendo positiva não é de 100% (i.e. falso negativos). Isso serve para mostrar a utilidade do entendimento do teorema de Bayes.

Em termos simples, um classificador Naïve Bayes assume que a presença ou ausência de uma característica particular, está relacionado com a presença ou ausência de qualquer outro elemento, tendo em conta a classe variável. Por exemplo, um fruta pode ser considerada como uma maçã se é vermelha, redonda, e cerca de 3 "de diâmetro. Um classificador Naive Bayes considera cada um desses atributos para contribuir de forma independente para a probabilidade de que esta é um fruta maçã, independentemente da presença ou ausência de outras características.

3.2.1 O princípio do Naïve Bayes

Por conveniência de exposição, assumiremos somente duas classes, demonstradas por i = 0, 1. Nosso objetivo é usar o conjunto inicial de objetos com associações de classe já conhecidos (conjunto de treinamento) para construir contagens da seguinte maneira: contagens maiores estão associadas com a classe de objetos 1 e as menores contagens com a classe de objetos 0. A classificação é então obtida através da comparação desta contagem, com um limiar, t. Se definirmos que P(i |x) é a probabilidade que um objeto com medida de vetor x = (x_1, . . . , x_p) pertencer a classe i, então qualquer função monótona de P(i |x) marcaria uma contagem adequada. Em particular, a relação P(1|x)/P(0|x) já seria adequada. A probabilidade elementar nos diz que podemos decompor P(i |x) proporcionalmente a f (x | i) P (i), onde f (x | i) é a distribuição condicional de x para a classe de objetos i e P (i) é a probabilidade de que um objeto pertença à classe i se não soubermos mais nada sobre isso (probabilidade a priori da classe i). Isso significa que a relação torna-se, como na figura 3.3.

$$\frac{P(1|x)}{P(0|x)} = \frac{f(x|1)P(1)}{f(x|0)P(0)} \qquad (3.3)$$

Para utilizar tal relação para produzir classificações, precisamos estimar o f (x|i) e P (i). Se o conjunto de treinamento for uma amostra aleatória da população em geral, o P(i) pode ser estimado diretamente a partir da proporção da classe de objetos I no conjunto de treinamento. Para estimar o f(x|i), o método do Naïve Bayes assume que os componentes de X são independentes na fórmula de 3.4, e, em seguida, calcula cada uma das distribuições de uni f(x_j | i), j=1,..., P; i = 0, 1; separadamente. Assim, o problema multivariado dimensional p, é reduzido para um problema de estimação univariado p. Estimativas univariadas são simples, familiares e requerem menores conjuntos de treinamento para obter estimativas mais precisas.

Segundo Hand (2009), uma das particularidades e grande qualidade do Naïve Bayes é a estimativa é simples, muito rápida que não requer abordagens estimativas e iterativas complicadas. Se as distribuições marginais f (x_j | i) são discretas, com cada x_j tomando apenas alguns valores, então, a estimativa de f (x_j | i) é um tipo de histograma multinominal do tipo estimador (ver abaixo), simplesmente por contar a

proporção da classe de objetos i que se enquadram em cada célula. Se o $f(x_j \mid i)$ é contínuo, então a melhor estratégia é segmentar cada um deles em um pequeno número de intervalo e novamente se utilizar o estimador multinominal, mas com maiores utilizações de versões baseados em estimações contínuas.

$$f(x|i) = \prod_{j=1}^{p} f\left(x_j \middle| i\right) \qquad (3.4)$$

Assumindo a relação de independência, a partir da fórmula 3.4, obtém-se a fórmula 3.5.

$$\frac{P(1|x)}{P(0|x)} = \frac{\prod_{j=1}^{p} f(x_j|1)P(1)}{\prod_{j=1}^{p} f(x_j|0)P(0)} = \frac{P(1)}{P(0)} \prod_{j=1}^{p} \frac{f(x_j|1)}{f(x_j|0)} \qquad (3.5)$$

Agora, recordando que o nosso objetivo era apenas produzir um resultado monotonamente relacionado à P (i | x), podemos tomar registros de (21)-log como uma função monótona crescente. Isto nos leva a uma razão alternativa presente na fórmula 3.6:

$$\ln \frac{P(1|x)}{P(0|x)} = \ln \frac{P(1)}{P(0)} + \sum_{j=1}^{p} \ln \frac{f\left(x_j \middle| 1\right)}{f\left(x_j \middle| 0\right)} \qquad (3.6)$$

Se definirmos $w_j = \ln\left(f(x_j|1)/f(x_j|0)\right)$ e a constante $k = \ln\left(P(1)/P(0)\right)$ veremos que a fórmula 3.6 tomará a simples forma da fórmula 3.7:

$$\ln \frac{P(1|x)}{P(0|x)} = k + \sum_{j=1}^{p} w_j \qquad (3.7)$$

Assim, o classificador tem uma peculiar simples estrutura. A suposição de independência da x_j dentro de cada classe implícita no modelo do Naïve Bayes pode parecer bastante restritiva. Na verdade, porém, vários fatores podem entrar em jogo, o que significa que a suposição não é tão prejudicial quanto parece ser.

Primeiramente, uma variável a priori de seleção aparece várias vezes, na qual variáveis correlacionadas foram eliminadas em grande número sobre os motivos de que estas são suscetíveis a contribuir de uma forma semelhante à separação de classes. Isso significa que a relação entre as variáveis restantes podem muito bem ser aproximadas por independência.

Em segundo lugar, assumir as interações iguais a zero gera uma etapa implícita de regularização, o que reduz a variância do modelo e leva a classificações mais precisas. Em terceiro lugar, em alguns casos quando as variáveis são correlacionadas, a superfície da decisão ótima coincide com aquela produzida sob a suposição de independência, logo a suposição feita não prejudica seu desempenho de maneira nenhuma. Em quarto lugar, a superfície da decisão produzida pelo Naïve Bayes pode, de fato, gerar uma forma não-linear complicada: a superfície é linear no w_j mas altamente não-linear nas variáveis originais x_j, de modo que possa vir a se encaixar superfícies bem elaboradas.

Segundo Hand (2009), o modelo Naïve Bayes é extremamente atraente por causa de sua simplicidade, elegância e robustez. É um dos mais antigos algoritmos de classificação formal, e ainda assim, mesmo em sua forma mais simples, é surpreendentemente eficaz. É amplamente utilizado em áreas como a classificação de textos e filtragem de spam. Ainda segundo Hand (2009) um grande número de modificações foram introduzidas pela análise estatística, mineração de dados, aprendizado de máquina e padrão de reconhecimento de comunidades, na tentativa torná-lo mais flexível, mas nota-se que tais modificações são necessariamente complicações que prejudicam a sua simplicidade básica.

3.2.2 Execução do classificador Naïve Bayes

Sejam A_1,..., A_k atributos, [a_1,..., a_k] uma tupla do banco de dados, e C uma classe a ser prevista. A previsão ótima é uma classe de valor c tal que:

$P(C = c \mid A_1 = a_1 \ldots A_k = a_k)$ é máxima.

Considerando independência entre os atributos:

$= P (A_1 = a_1 \mid C = c) * P (A_k = a_k \mid C = c) * P(C = c) / P (A_1 = a_1) * P (A_k = a_k)$

Um exemplo prático foi construído com a tabela 3.1, considerando a decisão de querermos jogar futebol tomando por atributos o aspecto do tempo, a temperatura, umidade e velocidade do vento. Nossa classe final será a classe de decisão.

Dia	Aspecto	Temperatura	Umidade	Vento	Decisão
1	Sol	Quente	Alta	Fraco	Não
2	Sol	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chuva	Agradável	Alta	Fraco	Sim
5	Chuva	Fria	Normal	Fraco	Sim
6	Chuva	Fria	Normal	Forte	Não
7	Nublado	Fria	Normal	Forte	Sim
8	Sol	Agradável	Alta	Fraco	Não
9	Sol	Fria	Normal	Fraco	Sim
10	Chuva	Agradável	Normal	Fraco	Sim
11	Sol	Agradável	Normal	Forte	Sim
12	Nublado	Agradável	Alta	Forte	Sim
13	Nublado	Quente	Normal	Fraco	Sim
14	Chuva	Agradável	Alta	Forte	Não

Tabela 3.1 – Base de dados contendo atributos de dias como temperatura, umidade, vento e decisão que influenciam na decisão de se jogar bola ou não

Qual será então a decisão de jogarmos bola se o dia estiver ensolarado, com temperatura fria, alta umidade e ventos fortes?

Os cálculos para duas possibilidades de decisões (“Sim” ou “Não”) serão realizados a seguir. Em caso de decisão **positiva** para se jogar bola em um dia com sol, temperatura fria, umidade alta e ventos fortes, o cálculo levará em conta as seguintes probabilidades:

P(Jogar = Sim / Aspecto = Sol)

P(Jogar = Sim / Temperatura = Fria)

P(Jogar = Sim / Umidade = Alta)

P(Jogar = Sim / Vento = Forte)

E para uma decisão **negativa** para se jogar bola em um dia ensolarado, temperatura fria, umidade alta e ventos fortes, o cálculo será influenciado pelas seguintes probabilidades:

P(Jogar = Não / Aspecto = Sol)

P(Jogar = Não / Temperatura = Fria)

P(Jogar = Não / Umidade = Alta)

P(Jogar = Não / Vento = Forte)

O cálculo para se prever a decisão de se jogar bola em um dia ensolarado, temperatura fria, umidade alta e ventos fortes, seria como na fórmula 3.8.

$$\frac{P(Sol/Sim) * P(Fria/Sim) * P(Alta/Sim) * P(Forte/Sim) * P(Sim)}{P(Sol) * P(Fria) * P(Alta) * P(Forte)} \qquad (3.8)$$

Para se calcular cada uma das probabilidades contidas na fórmula 3.8, é necessário construir uma tabela geral de probabilidades a partir da base de dados contida na tabela 3.1. A tabela 3.2 contém os valores das probabilidades considerando a possibilidade de decisão positiva (Sim) ou negativa (Não) para o fato de se jogar bola ou não em razão dos atributos do dia como: “aspecto”, “temperatura”, “umidade” e “vento”.

Decisão	Aspecto			Temperatura			Umidade		Vento	
	Sol	Nublado	Chuva	Quente	Agradável	Fria	Alta	Normal	Forte	Fraco

Não	3/5	0/5	2/5	2/5	2/5	1/5	4/5	1/5	2/5	2/5
5/14	60,00%	00,00%	40,00%	40,00%	40,00%	20,00%	80,00%	20,00%	40,00%	40,00%
(35,71%)										
Sim	2/9	4/9	3/9	2/9	4/9	3/9	3/9	6/9	4/9	6/9
9/14	22,22%	44,44%	33,33%	22,22%	44,44%	33,33%	33,33%	75,00%	44,44%	75,00%
(64,29%)										

Tabela 3.2 – Valores gerais das probabilidades para a decisão positiva ou negativa considerando individualmente os atributos “aspecto”, “temperatura”, “umidade” e “vento” para o fato de se jogar bola ou não.

De acordo com a tabela 3.2, as probabilidades contidas anteriormente na fórmula 3.8 podem ser calculadas da seguinte forma:

P(Jogar = Sim) = 9/14; P(Jogar = Não) = 5/14;
P(Aspecto = Sol / Jogar = Sim) = 2/9;
P(Aspecto = Sol / Jogar = Não) = 3/5;
P(Temperatura = Fria / Jogar = Sim) = 3/9;
P(Temperatura = Fria / Jogar = Não) = 1/5;
P(Umidade = Alta / Jogar = Sim) = 3/9;
P(Umidade = Alta / Jogar = Não) = 4/5;
P(Vento = Forte / Jogar = Sim) = 4/9;
P(Vento = Forte / Jogar = Não) = 2/5;
P(Aspecto = Sol) = 5/14
P(Temperatura = Fria) = 4/14
P(Umidade = Alta) = 7/14
P(Vento = Forte) = 6/14

Após o cálculo individual das probabilidades contidas na fórmula 3.8, a fórmula em si referente ao cálculo da probabilidade positiva de se jogar bola de acordo com as características com dia ensolarado, temperatura fria, umidade alta e ventos fortes, pode ser calculada. Com as probabilidades calculadas, basta agora que as apliquemos na fórmula 3.8.

Continuando para o cálculo da fórmula 3.8, temos:

$$\frac{P(Sol/Sim) * P(Fria/Sim) * P(Alta/Sim) * P(Forte/Sim) * P(Sim)}{P(Sol) * P(Fria) * P(Alta) * P(Forte)} \quad (3.8)$$

$$\frac{2/9 * 3/9 * 3/9 * 3/9 * 9/14}{5/14 * 4/14 * 7/14 * 6/14} \Rightarrow \frac{0,22 * 0,33 * 0,33 * 0,33 * 0,64}{0,35 * 0,28 * 0,5 * 0,42}$$

$$\frac{0,0052}{0,021} \Rightarrow 0,247$$

Obtemos uma probabilidade de 0,247 de jogar bola num dia ensolarado, frio, úmido e com ventos fortes.

Agora, considerando a probabilidade para a previsão negativa para se jogar bola quando temos um dia ensolarado, temperatura fria, umidade alta e ventos fortes, temos a fórmula 3.9.

$$\frac{P(Sol/Não) * P(Fria/Não) * P(Alta/Não) * P(Forte/Não) * P(Não)}{P(Sol) * P(Fria) * P(Alta) * P(Forte)} \quad (3.9)$$

A partir da fórmula 3.9 e levando em consideração os valores já calculados a partir da tabela 3.3, temos:

$$\frac{3/5 * 1/5 * 4/5 * 2/5 * 5/14}{5/14 * 4/14 * 7/14 * 6/14} \Rightarrow \frac{0,6 * 0,2 * 0,8 * 0,4 * 0,35}{0,35 * 0,28 * 0,5 * 0,42}$$

$$\frac{0,0137}{0,021} \Rightarrow 0,652$$

Obtemos, então, a probabilidade de 0,652 para a decisão negativa de jogar bola em um dia ensolarado, frio, úmido e com ventos fortes. Portanto a probabilidade

de não se jogar bola com essas características em determinado dia, é maior do que a probabilidade de jogar bola.

Com estes exemplos, pudemos constatar a ação do Naïve Bayes. Primeiramente, a partir da uma base de dados inicial, uma tabela de probabilidades contendo os valores determinados para cada classe possível é construída, ou seja, a partir da probabilidade da ocorrência das classes, são calculadas as probabilidades de ocorrências dos atributos contidos na base de dados iniciais. Depois, a partir da fórmula de Bayes, faz-se o contrário, conhecendo somente os atributos, calculamos a probabilidade de uma determinada classe vir a ocorrer. Assim, podemos calcular a probabilidade de qualquer classe a partir do conjunto de atributos conhecidos, considerando, é claro, a base de dados inicial.

3.3 Aspectos positivos e limitações do Naïve Bayes

A seguir enumeramos aspectos positivos e limitações do classificador do algoritmo Naïve Bayes, que serão enumerados da seção 3.3.1 a 3.3.6.

- Aspectos positivos:

  - Treinamento rápido (varredura única).
  - Classificação rápida e simples.
  - Insensibilidade a características irrelevantes.
  - Boa manipulação de dados discretos e contínuos.

- Aspectos Negativos e limitações:

  - Ingenuidade: Assume Independência entre as características.
  - Resultados inferiores em tarefas envolvendo pequenos conjuntos de dados.

3.3.1 Treinamento rápido e varredura única

Através da figura 3.2, notamos que cada classe (C_j) está relacionada a determinado atributo ou característica (d_n), gerando uma probabilidade p ($p(d_1|c_j)$, $p(d_2|c_j)$ ... $p(d_n|c_j)$).

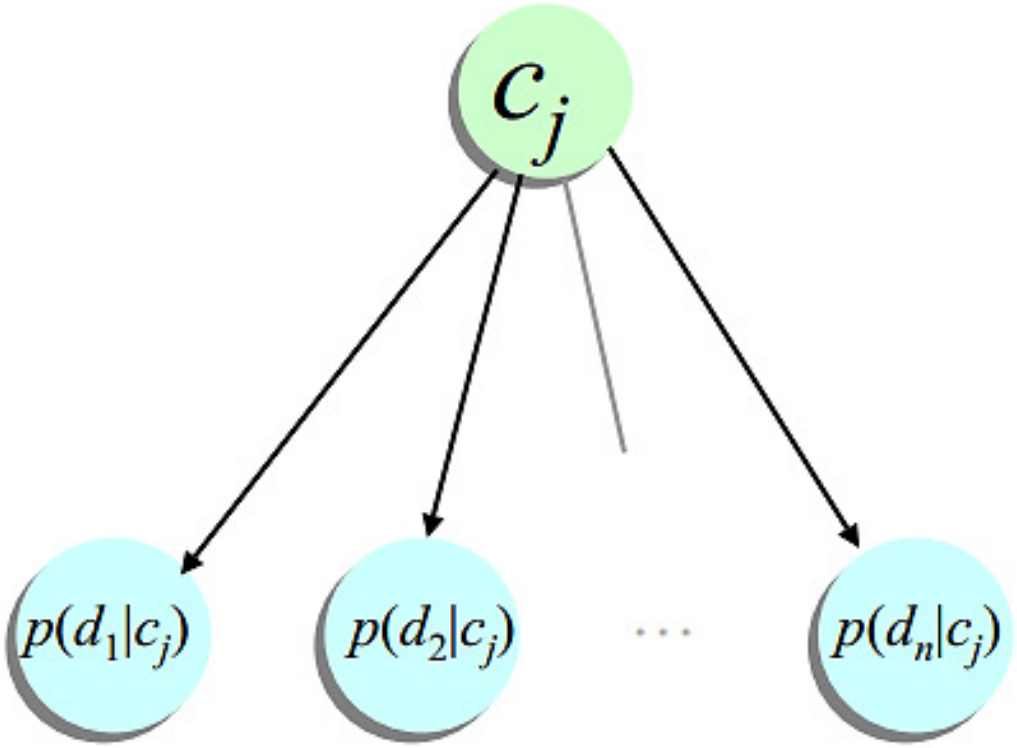
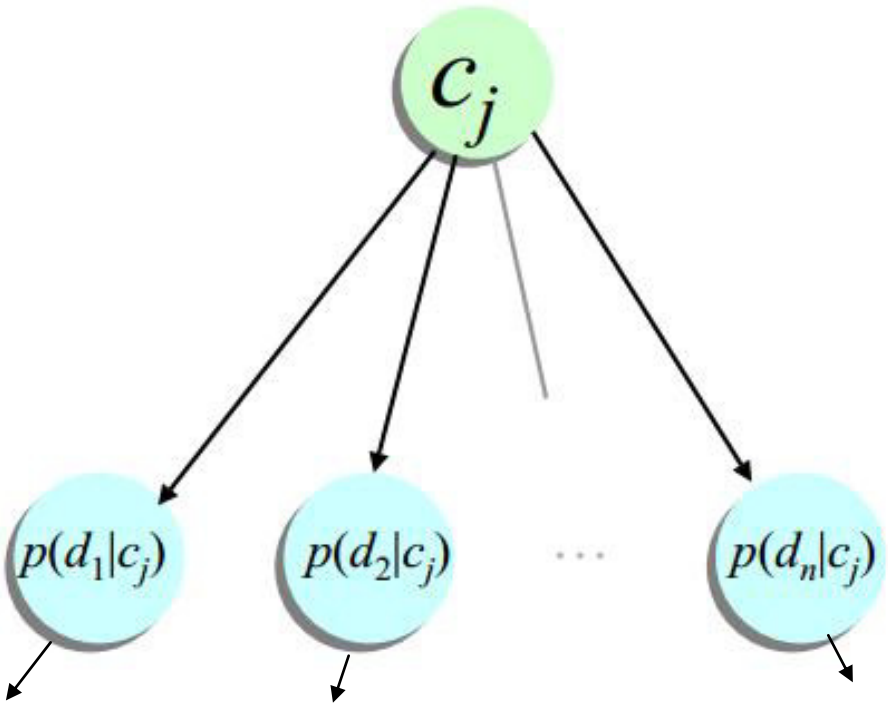


Figura 3.2 – Demonstra a associação que cada classe C_j, tem em relação a determinado atributo (d_1, d_2, d_3... d_n) (GARCIA, 2011).

Podemos então descobrir todas as probabilidades de um determinado evento, varrendo apenas uma vez a base de dados, e posteriormente armazenando-os em tabelas, chamadas de tabelas de probabilidades, como na figura 3.3.



Sexo	Mais de 1.80cm	Prob.
Masc.	Sim	0,85
	Não	0,15
Fem.	Sim	0.01
	Não	0,99

Sexo	Cabelos Longos	Prob.
Masc.	Sim	0,05
	Não	0,95
Fem.	Sim	0.70
	Não	0,30

Sexo	Força Muscular	Prob.
Masc.	Sim	0,75
	Não	0,25
Fem.	Sim	0.10
	Não	0,90

Figura 3.3 – Tabelas de probabilidades de uma pessoa ser homem ou mulher de acordo com características como altura, comprimento do cabelo e força muscular (GARCIA, 2011).

3.3.2 Classificação rápida e simples

A classificação realizada pelo classificador Bayes é realizada de maneira simples, eficiente e rápida através da leitura da base de dados, formando as tabelas de probabilidades, como dito no item anterior. Após a formação das tabelas, são realizados os cálculos relacionados à classificação propriamente dita das classes dos elementos citados, utilizando a própria tabela de probabilidades. O processo de classificação é realizado rapidamente por meio dos cálculos probabilísticos e posteriormente destacado na classificação final do algoritmo.

3.3.3 Insensibilidade a atributos irrelevantes

O Naïve Bayes não é sensível a atributos irrelevantes, ou seja, a classificação final realizada por ele, não sofre influências em relação a características insignificantes.

Por exemplo, suponhamos que desejamos classificar o sexo de uma pessoa através de características como a cor dos olhos, este atributo é totalmente insignificante perante a classificação final.

P(Maria | Cj) = p(cor dos olhos = azuis | Cj) * p(cabelos longos | Cj)*...

P(Maria | Homem) = 9,00/10,00 * 2,00/10,00 * ...

P(Maria | Mulher) = 9,01/10,00 * 8,00/10,00 * ...

Podemos notar que as probabilidades relacionadas à cor dos olhos, para a classificação de Maria como homem ou mulher são quase iguais, entretanto o classificador assume que ambas as probabilidades irrelevantes são totalmente suficientes para a classificação, portanto, quanto mais características, mesmo que irrelevantes, melhor.

3.3.4 Boa manipulação de dados discretos e contínuos

O classificador Bayes é capaz de trabalhar com dados discretos e contínuos de uma maneira versátil e eficiente, não possuindo qualquer dificuldade em relação a estes dados.

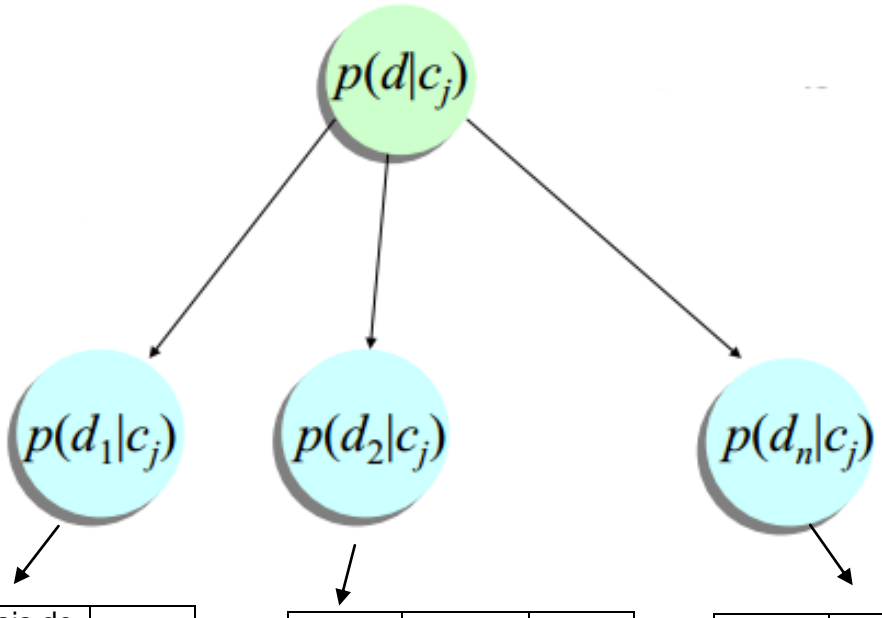
Ferramentas de mineração de dados como o WEKA[1] (Abernethy, 2010), também ajudam a trabalhar com atributos numéricos, atuando de maneira transparente pois utiliza a suposição de que os atributos numéricos possuem distribuição de probabilidade gaussiana, o que possibilita estimar probabilidades condicionais para estes tipos de atributos que podemos encontrar. Também podemos discretizar os atributos numéricos da base de dados, ao transformarmos os atributos numéricos em atributos discretos.

[1] Ferramenta composta por algoritmos voltados à mineração de dados

[2] Tipo de arquivo padrão utilizado pelo WEKA para tarefas de mineração de dados

3.3.5 Ingenuidade: Assume Independência de características

O Naïve Bayes assume uma independência entre os atributos ou características na classificação, sendo considerado um classificador ingênuo, pois considera os atributos e características igualmente importantes entre si e condicionalmente independentes em relação à classe. Assim, em vez de considerar determinados atributos como sendo mais relevantes no processo de classificação, acaba considerando, por exemplo, a característica da cor dos olhos tão importante quanto a característica do comprimento dos cabelos para classificar um indivíduo como homem ou mulher. Na figura 3.4, temos três atributos: peso, altura e comprimento dos cabelos. Estas três características são usadas para se determinar se um indivíduo é homem ou mulher.



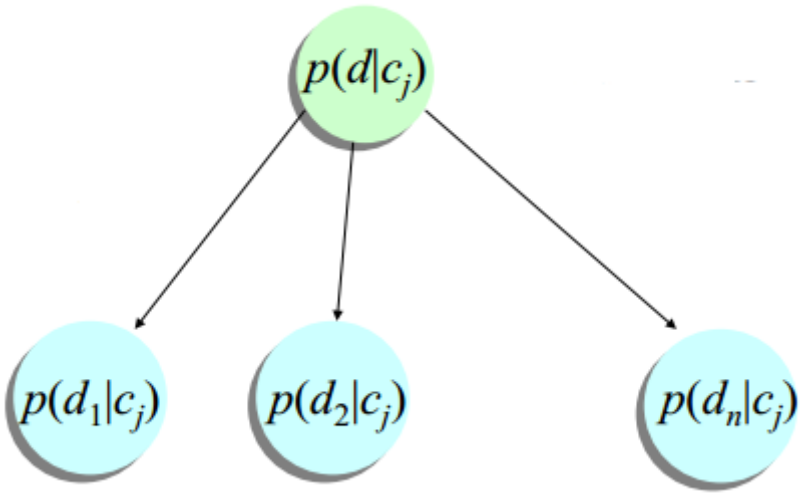
Sexo	Mais de 1.80cm	Prob.
Masc.	Sim	0,85
	Não	0,15
Fem.	Sim	0.10
	Não	0,90

Sexo	Mais de 100kg	Prob.
Masc.	Sim	0,25
	Não	0,75
Fem.	Sim	0.05
	Não	0,95

Sexo	Cabelos Longos	Prob.
Masc.	Sim	0,05
	Não	0,95
Fem.	Sim	0.70
	Não	0,30

Figura 3.4 - Tabelas de probabilidades com atributos relacionados a altura (mais de 1,80cm) e peso (mais de 100kg) para se determinar o sexo de uma pessoa (GARCIA, 2011).

Poderíamos corrigir parcialmente o problema ao correlacionarmos o conjunto de atributos a fim de buscarmos uma junção destes atributos, como na figura 3.5.



Sexo	Mais de 1.80cm	Prob.
Masc.	Sim	0,85
	Não	0,15
Fem.	Sim	0.01
	Não	0,99

Sexo	Mais de 100kg	Prob.
Masc.	Sim, mais de 1,80cm	0,25
	Não e mais de 1,80cm	0,75
Fem.	Sim e Não mais de 1,80cm	0.05
	Não e Não mais do que 1,80cm	0,95

Figura 3.5 – Relaciona a conjunção dos atributos “peso” e “altura”, quebrando assim a independência entre estes atributos (GARCIA, 2011).

3.3.6 Resultados inferiores em tarefas com pequenos conjuntos de dados

Mesmo contendo bons aspectos positivos como robustez a ruídos, insensibilidade a atributos irrelevantes e um treinamento rápido, o Naïve Bayes mostra-se deficiente em problemas envolvendo pequenos conjuntos de dados. Por ser um classificador probabilístico, o Naïve Bayes necessita de uma base de dados contendo grandes volumes de dados e instâncias, já que em pequenos conjuntos de dados, o número de variações geralmente é maior, o que causaria uma inconsistência e inconstância dos cálculos probabilísticos, o que não encontramos em bases de dados com um maior volume. Nestas condições, classificadores como

árvores de decisão e redes neurais levam vantagem e apresentam resultados mais precisos, em conjuntos de dados pequenos.

3.4 Considerações Finais

Como visto, o classificador Naïve Bayes é baseado na fórmula probabilística de Bayes. O nome Naïve, entendido por “ingênuo”, é denominado assim por assumir independência condicional, ou seja, na tarefa de classificação, os atributos presentes na base de dados possuem a mesma importância e não apresentam dependência entre si, mesmo quando esta suposição não é verdadeira. Mesmo assim, o Naïve Bayes obtém bons resultados em tarefas de classificação por ser simples, rápido e insensível a atributos irrelevantes, possuindo uma boa manipulação de dados.

Neste capítulo também foram abordados o princípio de funcionamento do classificador bayesiano, bem como este atua na tarefa de classificação.

4 ESTUDO DE CASO: ANÁLISE DE CRÉDITO

Hoje em dia, cada vez mais ouvimos falar em operações bancárias como concessões de crédito. O mercado financeiro aquecido possibilita um maior número de empréstimos e concessões em um mundo cada vez mais capitalista. E para que se haja equilíbrio e se evite grandes perdas, é necessário que se invista em ferramentas e métodos eficazes para a concessão de crédito.

A tarefa de concessão de crédito é uma das mais arriscadas no que diz respeito a grandes operações bancárias. Nesse sentido é importante que nada fracasse em operações deste tipo. A análise de crédito então é capaz de prever e consequentemente minimizar os riscos que envolvem tais tarefas, gerando cálculos que remetem a futuras concessões a partir de informações e dados financeiros de clientes analisados.

O que entendemos sobre crédito envolve transação comercial e uma relação de confiança entre as partes. Esses dois fatores são essenciais e estruturam o conceito de concessão de crédito que nada mais é do que uma transação comercial estruturada na confiança quando o concessor do crédito, por meio da análise de crédito, tentará diminuir o risco de não receber o valor do crédito, enquanto o cliente o receberá como solução.

Segundo Santos (2000), “o processo de análise e concessão de crédito recorre ao uso de duas técnicas: a técnica subjetiva e a técnica objetiva ou estatística”. A primeira diz respeito à técnica baseada no julgamento humano e a segunda é baseada em processos estatísticos. Podemos então agregar o fator intuitivo e pessoal do analista de crédito com o poder estatístico que dispomos através dos cálculos realizados com a base de dados dos clientes analisados. E é a partir desta segunda técnica, focada na estatística aplicada nos cálculos de concessão de crédito encontrada nas informações dos clientes, que é enfocado o nosso estudo de caso.

Este capítulo é direcionado à aplicação do algoritmo Naïve Bayes em operações de análise de crédito, realizando-se a classificação a partir de bases de dados desenvolvidos para este trabalho e também a partir de bases de domínio financeiros reais, encontrados nas tabelas 2.7 e 2.8. Posteriormente é realizada uma comparação dos resultados esperados com os realmente obtidos com o algoritmo Naïve Bayes. Em seguida é construída uma tabela comparativa entre os resultados

obtidos através do classificador estudado neste trabalho e resultados provenientes de outros classificadores

4.1 Estudo de Caso 1: Base de dados fictícia

Nas próximas sessões será apresentado o estudo de caso baseado em uma base de dados fictícia criada para este trabalho a fim de analisarmos o processo de classificação utilizado com o algoritmo Naïve Bayes. Serão mostrados os pontos chave de utilização do classificador probabilístico, bem como os cálculos realizados ao longo da tarefa de classificação.

4.1.1 Base de Dados

Uma base de dados é um conjunto de dados e informações determinadas para um objetivo específico. Pode ser uma lista de CD's e repertórios, cadastros de clientes ou uma lista de livros. O objetivo de se construir uma base de dados é obter e utilizar dados lá depositados, extraindo o máximo de informações possíveis como se encontrar uma música a partir da lista de repertórios ou se encontrar o endereço de alguém a partir do cadastro de clientes. Ao contexto deste trabalho, foi desenvolvida uma base de dados com quatro atributos: Sexo, Renda, Idade e Filhos. Estes atributos têm como objetivo determinar se o individuo analisado é considerado um bom pagador ou não. Uma amostra das tuplas representadas pelos atributos Sexo, Renda, Idade e Filhos é apresentada na tabela 4.1 e tem a pretensão de demonstrar os possíveis valores assumidos por cada atributo.

Atributos	Valores assumíveis
Sexo	Pode assumir os valores 'M' para indivíduo do sexo masculino ou 'F' para indivíduos do sexo feminino
Renda	Pode assumir os valores ≤1500(menor ou igual a R$ 1500), 1500-3000(entre R$ 1500 e R$ 3000) ou ≥3000(maior ou igual a R$ 3000).
Idade	Pode assumir os valores 20-39(entre 20 e 39 anos) ou ≥40(maior ou igual a 40 anos de idade)
Filhos	Pode assumir os valores 'Sim' para caso de o indivíduo ter filhos e 'Não' para o caso de não ter filhos

Tabela 4.1 – Demonstra os valores que cada atributo pode assumir na base de dados construída neste trabalho

A partir dos valores que cada atributo pode assumir na tabela 4.1, construímos a base de dados na tabela 4.2, com o objetivo de se determinar se um individuo é um bom ou mau pagador. Temos então uma classe chamada de "Bom_Pagador" que assume o valor "Sim" se o indivíduo é considerado um bom pagador ou "Não" se o contrário. A seguir temos a tabela 4.2:

Id	Sexo	Renda	Idade	Filhos	Bom_Pagador(Classe)
1	F	≥3000	≥40	Não	Sim
2	F	1500-3000	20-39	Sim	Não
3	F	≤1500	≥40	Não	Não
4	M	≤1500	≥40	Não	Sim
5	F	1500-3000	≥40	Sim	Não
6	M	≥3000	20-39	Sim	Sim
7	F	≥3000	≥40	Não	Sim
8	F	≤1500	≥40	Não	Não
9	M	≥3000	≥40	Sim	Sim
10	M	≤1500	20-39	Não	Não
11	F	≥3000	≥40	Sim	Sim
12	F	≥3000	20-39	Sim	Não
13	F	1500-3000	20-39	Sim	Não
14	F	≤1500	20-39	Não	Não
15	F	1500-3000	20-39	Não	Sim

Tabela 4.2 – Mostra os atributos Sexo, Renda, Idade e Filhos como características para se determinar se um indivíduo é um bom pagador ou não

Como visto na tabela 4.1 e ratificado na tabela 4.2, o atributo *Sexo* é determinado pelas letras *M* e *F*, que representam os sexos **Masculino** e **Feminino** respectivamente. Já o atributo *Renda* foi divido em três intervalos. O primeiro - menor ou igual a R$1500 (≤1500), o segundo - entre R$1500 e R$3000 (1500-3000) e o terceiro - maior ou igual a R$3000 (≥3000). Já o atributo Idade fora dividido em dois intervalos diferentes: O primeiro – entre 20 e 39 anos e o segundo - maior ou igual a 40 anos. Por fim, o atributo Filhos é definido por "Sim" para o fato de o indivíduo possuir filhos, ou *Não* para o fato de o indivíduo não possui-los.

A base de dados da tabela 4.2, foi construída segundo critérios lógicos. Levamos em consideração atributos reais na constatação do fato de o indivíduo a ter o seu crédito analisado é um bom pagador ou não. O Atributo Sexo é importante

individualmente, pois Homens costumam acumular mais bens e finanças do que as Mulheres num contexto geral. Levamos isto em consideração na projeção desta base de dados.

O atributo Renda, um dos mais importantes, avalia a quantia arrecadada mensalmente pelo individuo, o que influencia diretamente na análise realizada posteriormente. Por exemplo, a renda menor ou igual a R$1500 é considerada uma média razoável, se individualmente analisada, mas em conjunto com outros atributos, pode-se ter um crédito negativado assim como também pode ser positivado, é algo que varia de acordo a presença de outros atributos.

Já o atributo Idade, demonstra à priori em que fase da vida financeira, encontra-se o indivíduo. Aqueles com idade entre 20 e 39 anos estão em faixa etária iniciante em relação ao mercado de trabalho, geralmente apresentam pouca ou média renda de acordo com sua precisa idade, o que já não acontece com o indivíduo em idade acima de 40 anos, cuja renda, geralmente, está totalmente estabilizada, já que exercem atividade trabalhista fixa há longo prazo.

Por último, temos o atributo Filhos, que define se o indivíduo possui filhos ou não. Este atributo tem sua importância no fato de que filhos representam gastos ao longo do mês, o que abate a renda e o crédito do indivíduo a ser analisado, podendo pesar na definição final de ser um bom ou mau pagador.

Como observado, cada atributo tomado de forma individualmente é considerado bastante vago a respeito de se prever ou definir se uma pessoa é boa ou má pagadora. Não podemos definir se uma pessoa possui crédito somente por sabermos que é do sexo masculino ou feminino, ou somente pela renda, ou bastando a idade ou também pelo único fato de se ter filhos ou não. Entretanto tomando-se por consideração, a presença em conjunto destes quatro atributos, podemos definir e prever se um indivíduo é ou não um bom ou mau pagador.

Estes quatro fatores tomados de forma conjunta, representam a informação ideal para se determinar a condição de um indivíduo a ser um bom ou um mau pagador. Como o exemplo abaixo:

Um indivíduo do sexo masculino, com renda menor ou igual a R$1500, idade entre 20 e 39 anos e com filhos, provavelmente será um mau pagador já que não possui uma renda muito elevada, é ainda jovem, considerando sua idade, provavelmente não possuindo muitos bens ou finanças, e ainda possui filhos,

contribuindo para gastos elevados e posterior abatimento de sua renda ao mês. Percebemos então que este candidato à concessão de crédito é deveras inviável, não sendo confiável, podendo assim representar grande prejuízo aos credores.

Já um exemplo tomado com uma indivíduo do sexo feminino, renda maior ou igual a R$3000, idade maior ou igual a 40 anos, sem filhos é um potencial bom pagador, já que além de possuir uma boa renda ao fim do mês, também já possui uma idade avançada, tendo provavelmente um emprego fixo, bens e finanças estabilizadas contribuindo o fato de não possuir filhos, o que não abate a sua renda mensal com os cuidados dos mesmos.

Por fim, uma pessoa do sexo feminino, com renda entre R$1500 e R$3000, idade maior ou igual a 40 anos e com filhos, representa um risco alto quando relacionado à concessão de crédito pois apesar ser considerada estável de acordo com a idade avançada (supondo-se que exerce emprego fixo há longo prazo), a sua renda mensal é arriscada quanto a concessão de crédito, visto que esta senhora possui filhos e deve, portanto, dedicar grande parte de sua renda aos cuidados destes últimos, inviabilizando a classificação desta pessoa como uma Boa Pagadora.

Como vimos com base nestes três exemplos, a tabela 4.2 tomada por base de dados é totalmente lógica e é também considerada como base para um conjunto de treinamento de futuras previsões para a classificação de um individuo como um bom pagador ou não, utilizando o algoritmo Naïve Bayes.

De acordo, então, com nossa base de dados, o algoritmo Naïve Bayes se utilizará de um método surpreendentemente simples para realizar a tarefa de classificação, dividindo esta tarefa em duas fases: (i) construção do modelo classificador, onde será realizada a construção da tabela de probabilidades e (ii) aplicação da Fórmula de Bayes para classificar novos objetos.

4.1.2 O modelo classificador

O modelo classificador é basicamente a construção de uma tabela de probabilidades condicionais contendo um resumo dos dados presentes na base de dados que será o alvo da fase de classificação através do algoritmo Naïve Bayes. As informações contidas neste modelo-resumo da tabela também são tidas como parâmetros do modelo do Naïve Bayes e a construção deste modelo se dá de maneira simples, consumindo pouco e gastando menos do ponto de vista computacional. A seguir, está a tabela 4.3 preenchida com os parâmetros estatísticos representados através da classe e dos atributos preditivos inspirados na tabela 4.2.

Bom_Pagador	Sexo		Renda			Idade		Filhos	
	F	M	≤1500	1500-3000	≥3000	20-39	≥40	Não	Sim
Não	6/11	1/4	4/5	3/4	1/6	5/7	3/8	4/8	4/7
8/15	54,55%	25,00%	80,0%	75,00%	16,67%	71,43%	37,50%	50%	57,14%
(53,33%)									
Sim	5/11	3/4	1/5	1/4	5/6	2/7	5/8	4/8	3/7
7/15	45,45%	75,00%	20,0%	25,00%	83,33%	28,57%	62,50%	50%	42,86%
(46,67%)									

Tabela 4.3 – A partir de uma prévia base de dados, um modelo com os parâmetros estatísticos de cada atributo relacionado às classes Bom e Mau_Pagador

Na Tabela 4.3, a primeira coluna apresenta as probabilidades “a priori” de um cliente ser um bom pagador ou não, para Bom_Pagador = “Não” o valor é 53,33% e para Bom_Pagador = ”Sim” o valor é 46,67%. A partir da segunda célula, as informações tornam-se bem mais importantes e interessantes, representando uma série de valores de probabilidades condicionais. Como explicado anteriormente, o modelo classificador leva em conta fatores como o Sexo, Renda, Idade e Filhos do cliente a ter seu crédito avaliado, para então decidir se o mesmo é um bom pagador em potencial ou não. Por esta razão, as colunas de 2 a 10 apresentam as probabilidades condicionais dos valores dos atributos preditivos “Sexo”, “Renda”, “Idade” e “Filhos” a partir dos dois rótulos possíveis da classe “Bom_Pagador” definidos por “Sim” ou “Não”.

Peguemos por exemplo, os valores de probabilidades condicionais do atributo preditivo “Sexo”. Na terceira coluna, temos a fração 2/4, esta indica que dos quatro homens existentes na base de dados, dois destes não foram considerados bons

pagadores (Bom_Pagador = “Não”). Analogamente com os outros dois restantes, representados na célula logo abaixo, o valor, também de 2/4, corresponde ao fato de que dois dos quatro homens existentes, foram considerados bons pagadores (Bom_Pagador = “Sim”). Podemos tomar desta forma, também, a segunda coluna da tabela que trata do sexo feminino. Das onze mulheres existentes, seis não foram consideradas boas pagadoras (Bom_Pagador = “Não”). Já, na célula abaixo, através da mesma lógica, a fração 5/11 significa que das onze mulheres existentes, apenas 5 foram consideradas boas pagadoras (Bom_Pagador = “Sim”).

Com a criação da tabela de parâmetros, o algoritmo bayesiano está pronto à tarefa de classificar novos objetos. A partir de uma adaptação da Fórmula de Bayes, será realizado o cálculo das probabilidades finais de, no nosso caso, o cliente ser um potencial bom pagador ou não. Vamos então ao um exemplo. Suponhamos que o nosso novo cliente tenha algumas características como as seguintes:

- Sexo: Masculino
- Renda: Entre R$1500 e R$3000
- Idade: Maior que 40 anos
- Não tenha Filhos

A partir dos valores dos atributos acima, a tarefa será a de classificar esse novo cliente como bom pagador ou não. Tomando por consideração as probabilidades condicionais apontadas pela tabela 4.3, que nos fornece um relevantes histórico de dados oferecidos por clientes, em tese, bons pagadores e maus pagadores, é que nos basearemos para estimar as previsões mencionadas anteriormente.

O algoritmo Naïve Bayes realiza a aplicação da fórmula de Bayes ao cálculo de probabilidades tomando por base, as estimativas contidas na tabela de parâmetros, ou seja, o algoritmo utiliza os dados probabilísticos já existentes na tabela 4.3, e de acordo com as características dos novos clientes, realiza o cálculo com as probabilidades representadas pelas características dos novos clientes, estas últimas correspondentes aos atributos preditivos já existentes na tabela.

Este cálculo possibilita a realização de estimativas bastante precisas no que diz respeito aos rótulos das classes. Demonstraremos agora, na prática, como

funcionam os cálculos das estimativas relacionadas primeiramente ao rótulo da classe Bom_Pagador = “Sim” e em seguida para o rótulo Bom_Pagador = “Não” do novo cliente mencionado anteriormente, que será considerado como novo cliente C = (Sexo = “M”, Renda = “1500-3000”, Idade = “≥40 anos”, Filhos = “Não”).

Exemplo 1:

Estimativa positiva para “Bom_Pagador”:

P(Bom_Pagador = “Sim” | C) = P(Sexo = “M” | Bom_Pagador = “Sim”) x
P(Renda = “1500-3000” | Bom_Pagador = “Sim”) x
P(Idade = “≥40” | Bom_Pagador = “Sim”) x
P(Filhos = “Não” | Bom_Pagador = “Sim”) x
P(Bom_Pagador = “Sim”)

P(Bom_Pagador = “Sim” | C) = 0,75 x 0,25 x 0,6250 x 0,50 x 0,4667 = 0,0273

Estimativa negativa para “Bom_Pagador”:

P(Bom_Pagador = “Não” | C) = P(Sexo = “M” | Bom_Pagador = “Não”) x
P(Renda = “1500-3000” | Bom_Pagador = “Não”) x
P(Idade = “≥40” | Bom_Pagador = “Não”) x
P(Filhos = “Não” | Bom_Pagador = “Não”) x
P(Bom_Pagador = “Não”)

P(Bom_Pagador = “Não” | C) = 0,25 x 0,75 x 0,3750 x 0,50 x 0,5333 = 0,0187

Os exemplos acima demonstraram uma chance mais alta de que o novo cliente venha a ser um bom pagador. Para analisarmos os resultados de uma forma mais real, os valores calculados acima serão convertidos em probabilidades através da normalização da soma para 1.

O resultado desta conversão é observado a seguir:

- Probabilidade (“Bom_Pagador” = SIM) = 0,0273 / (0,0273 + 0,0187) = 0,0273 / 0,0460 = 59,52%
- Probabilidade (“Bom_Pagador” = NÃO) = 0,0187 / (0,0273 + 0,0187) = 0,0187 / 0,0460 = 40,48%

Utilizaremos agora mais um exemplo da aplicação da fórmula de Bayes no processo de classificação do algoritmo bayesiano. Nosso novo cliente a ser classificado no exemplo 2 será chamado de N e conterá os atributos a seguir:

- Novo Cliente N = (“Sexo” = Feminino, “Renda” = ≥3000, “Idade” = 20-39 e “Filhos” = Sim)

Exemplo 2:

Estimativa positiva para Bom_Pagador:

P(Bom_Pagador = “Sim” | N) = P(Sexo = “F” | Bom_Pagador = “Sim”) x
P(Renda = “≥3000” | Bom_Pagador = “Sim”) x
P(Idade = “20-39” | Bom_Pagador = “Sim”) x
P(Filhos = “Sim” | Bom_Pagador = “Sim”) x
P(Bom_Pagador = “Sim”)

P(Bom_Pagador = “Sim” | C) = 0,4545x 0,8333 x 0,2857 x 0,4286 x 0,4667 = 0,0216

Estimativa negativa para Bom_Pagador:

P(Bom_Pagador = “Não” | N) = P(Sexo = “F” | Bom_Pagador = “Não”) x
P(Renda = “≥3000” | Bom_Pagador = “Não”) x
P(Idade = “20-39” | Bom_Pagador = “Não”) x
P(Filhos = “Sim” | Bom_Pagador = “Não”) x
P(Bom_Pagador = “Não”)

P(Bom_Pagador = “Nao” | N) = 0,5455 x 0,1667 x 0,7143 x 0,50 x 0,5333 = 0,0197

Aplicando-se o complemento de 1 para se obter os resultados covertidos em probabilidades, temos:

- Probabilidade (“Bom_Pagador” = SIM) = 0,0216 / (0,0216 + 0,0197) = 0,0216 / 0,0413 = 52,61%
- Probabilidade (“Bom_Pagador” = NÃO) = 0,0197 / (0,0216 + 0,0197) = 0,0197 / 0,0413 = 47,39%

4.1.3 A ‘ingenuidade’ do Bayes

É a partir dos cálculos realizados acima, que faremos uma breve explanação sobre a adaptação da fórmula de Bayes ao classificador Naïve Bayes.

O algoritmo possui duas suposições peculiares que possibilitam ao classificador realizar a tarefa de classificação a partir da fórmula de Bayes:

- Os atributos da base de dados são igualmente importantes entre si
- Os atributos preditivos são condicionalmente independentes em relação ao tributo classe

Com estas suposições, que fazem do Naïve Bayes, um classificador ingênuo(“Naïve”), realizamos o cálculo anterior para calcular a probabilidade condicional de ocorrência de um evento A, que em nosso caso foi o fato de se ter a probabilidade do cliente ser um Bom Pagador ou não; dado o fato de se ocorrer outro evento B, que para nós significou a ocorrência de vários atributos preditivos determinados por Sexo, Renda, Idade e Filhos. A fórmula 4.1 é utilizada nos cálculos:

- $P(A|B) = (P(B|A) \times P(A)) / P(B)$ (4.1)

O evento “B” é chamado de evidência e é representado por uma instância que não possui rótulo de classe, enquanto o evento “A” é chamado de hipótese, que é representado por um rótulo de classe que pode ou não ser atribuído a uma instância.

A evidência “B”, que nada mais é do que o novo objeto a ser classificado, é uma tupla formada por um conjunto i de atributos $\{B_1, B_2, B_3, B_4..., B_i\}$, cada um com seu valor determinado. Calcula-se então, a probabilidade de B (novo objeto) pertencer a cada um dos j rótulos de classe existentes $\{A_1, A_2, A_3, A_4..., A_j\}$. Portanto,

o algoritmo Naïve Bayes, a partir da consideração ingênua de que cada atributo é independente entre si, nos permite dividir a evidência B, representada pelos atributos preditivos do novo objeto (B_1, B_2, B_3, B_4..., B_i), fazendo-se com que a fórmula de Bayes seja mostrada na fórmula 4.2:

- $P(A_j \mid B) = (P(B_1 \mid A_j) \times P(B_2 \mid A_j) \times ... \times P(B_i \mid A_j) \times P(A_j)) / P(A)$. (4.2)

Nesta fórmula, o cálculo é feito para cada rótulo de classe (A_j). O novo objeto receberá o rótulo de classe que estiver associado à maior probabilidade obtida. Na prática, a divisão por P(A) não é necessária, pois o valor de P(A) é sempre o mesmo em cada cálculo de cada A_j.

Tomando um exemplo prático realizado no exemplo 1, é possível ver a aplicação da fórmula de Bayes junto ao classificador Naïve Bayes. No exemplo em que se previu a probabilidade onde a evidência é dada pela instância representada pelos quatro atributos preditivos determinados por “Sexo”, “Renda”, “Idade” e “Filhos” onde um homem com mais de 40 anos, renda entre R$ 1500 e R$ 3000 e sem filhos é submetido à hipótese de ser um bom pagador (“Bom_Pagador = Sim”) ou não (“Bom_Pagador = Não”) computou-se as chances de 59,52% para ser um bom pagador e 40,48% para não ser um bom pagador.

É importante enfatizar que o Naïve Bayes utiliza suposições ingênuas que muitas das vezes não são verdadeiras na prática. Em bancos de dados com aplicações reais, os atributos possuem variações de importância, por exemplo, o atributo “idade” e “nível de escolaridade” são mais importantes que os atributos “altura” e “peso” para se determinar se uma pessoa tem uma boa renda ou não. Assim como quase sempre atributos como “profissão” e “idade” são condicionalmente dependentes em relação a uma classe chamada de “renda salarial”.

Apesar disso, o algoritmo Naïve Bayes mostra-se bastante eficiente em vários domínios de aplicações e em experiências práticas, sua robustez foi comprovada, pois assume que todos os atributos são independentes entre si quando o valor da classe é conhecido, tendo desempenho excepcional em processamentos de bases de dados bastante volumosas, confrontando à altura algoritmos bem mais sofisticados como redes neurais, SVM’s e árvores de decisão. Além disso, é claro

que o seu baixo custo computacional implica em rentabilidade e economia visto que sua fase de treinamento consiste simplesmente na contagem de frequências cruzadas entre os valores da variável classe e das variáveis preditivas implicando em baixo custo computacional e simplicidade de aplicação, características ausentes em outros algoritmos e técnicas que se utilizam de métodos sofisticados e custosos demais para classificação.

4.1.4 Os resultados da classificação com Naïve Bayes

A partir da base de dados utilizada neste trabalho na tabela 4.2 da seção 4.1, utilizamos uma ferramenta especial voltada especificamente à mineração de dados, chamada WEKA (Abernethy, 2010). Esta ferramenta, desenvolvida pela universidade de Weikato em 1993, fornece um pacote de classes com as quais é possível agregar algoritmos voltados à mineração de dados em inteligência artificial, o que a torna uma ferramenta bastante útil e flexível. Mas dentre o grande número de classes apresentadas no WEKA, a que realmente nos importa neste trabalho é a classe "Naïve Bayes". Esta classe foi desenvolvida especialmente para se demonstrar o funcionamento do Algoritmo Naïve Bayes, demonstrando este último de uma forma melhorada e sofisticada.

Primeiramente, a base de dados utilizada na tabela 4.2 é convertida em formato ARFF[2], somente neste formado que a ferramenta WEKA aceita novas bases de dados. Logo abaixo, na figura 4.1, é apresentada a base de dados convertida no formato ARRF, no bloco de notas.

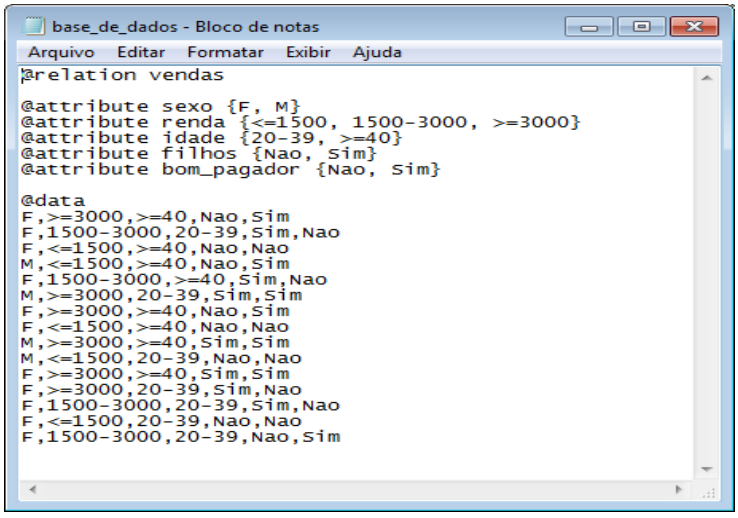


Figura 4.1 – Base de dados representada em arquivo ARFF aceito pelo programa WEKA

[2] Tipo de arquivo padrão utilizado pelo WEKA para tarefas de mineração de dados

Após a conversão da base de dados em ARFF, desenvolve-se um pequeno sistema em Java, para realizar a ponte da base de dados convertida em ARFF com a classe “Naïve Bayes” do WEKA. Esta ponte nada mais faz do que passar a base de dados em ARFF como entrada para que o novo objeto a ser alvo da previsão probabilística seja conhecido e posteriormente, analisado através da classe “Naïve Bayes” do WEKA.

A título de comparação de resultados, utilizaremos como exemplo a previsão calculada anteriormente de que o novo cliente com os seguintes atributos: Sexo Masculino, Renda entre R$1500 e R$3000, Idade maior ou igual a 40 anos e sem filhos; seja um Bom Pagador ou não. Realizando-se a previsão através do WEKA, obtêm-se resultados demonstrados na figura 4.2.

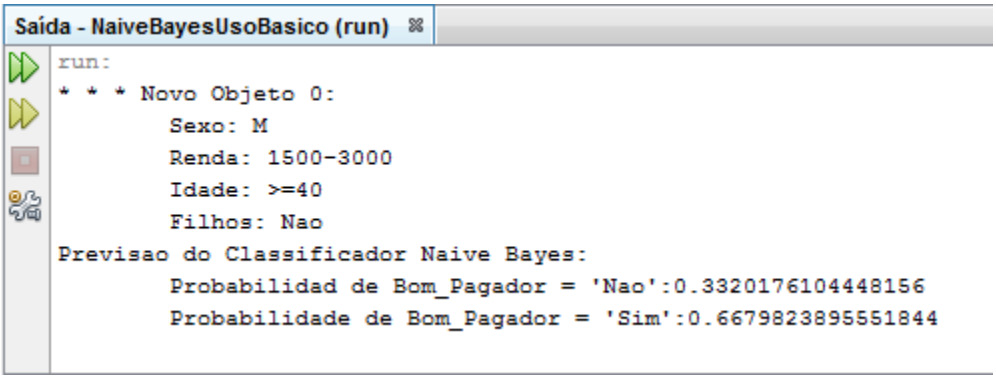


Figura 4.2 – Probabilidades calculadas através do WEKA, considerando um homem, com renda entre R$1500 e 3000, idade maior que 40 anos e sem filhos.

Segundo os resultados provenientes dos cálculos realizados pela classe “Naïve Bayes” originária do WEKA, a probabilidade de o novo cliente de sexo masculino, renda entre R$1500 e R$3000, idade maior ou igual a 40 anos e sem filhos é de 66,44% para ser um bom pagador e de 33,56% para não ser um bom pagador. Enquanto que o cálculo manual realizado mais acima, demonstrando a aplicação da fórmula de Bayes resultou em 59,52% para que o novo cliente seja um bom pagador e 40,48% para que o cliente seja um mau pagador.

Como podemos notar, observamos uma pequena diferença nas porcentagens. Enquanto os resultados, no WEKA, foram de 66,44% para Bom Pagador, o cálculo manual apontou 59,52%, uma pequena diferença de 6,92% considerando a modesta base de dados da tabela 4.2. Esta pequena divergência

deve-se ao fato de que a implementação da classe “Naïve Bayes” no WEKA utiliza uma técnica chamada de correção laplaciana que acrescenta o valor ‘1’ a todas as frequências constantes nas células da tabela de probabilidades condicionais construídas anteriormente para que se evite algum caso em que alguma célula da tabela mencionada apresente alguma probabilidade condicional com o valor ‘0’, o que ocasionaria um cálculo envolvendo o zero na multiplicação do numerador da fórmula de Bayes, causando, consequentemente, uma estimativa de probabilidade para rótulo de classe igual a zero.

Esta técnica é responsável por um leve aprimoramento da classe “Naïve Bayes” do WEKA, e não causa qualquer tipo de alteração significativa em grandes bases de dados, como as de domínios financeiros, pois quanto maior a base de dados for menor será a influência da correção laplaciana em virtude da pouca ocorrência de valores zerados em células de probabilidades condicionais. As leves alterações são imperceptíveis neste tipo vasto de bancos de dados.

4.2 Estudo de caso 2: Bases de dados reais

Realizamos posteriormente experimentos com domínios de bases de dados reais, provenientes de (UCI, 2010). As características de tais bases de dados são enumeradas a seguir:

1. Australian Credit Approval (QUILAN, N/A): Esta base de dados, provida originariamente por Quilan em seus estudos entre 1987 e 1992, para avaliar a técnica de árvores de decisão em aplicações de cartões de crédito, fora pouco modificada pela UCI Repository of Machine Learning Databases. Esta Base de dados consiste de 15 atributos e de um atributo classe. Antes que os atributos fossem validados para uso, seus nomes foram convertidos em símbolos para se preservar o caráter confidencial dos dados originais. O atributo classe pode assumir dois valores: ‘+’ ou ‘-‘, representando respectivamente bons e maus pagadores, e os tipos dos outros atributos podem ser contínuos ou nominais.

Class	No. of Instances
+	218
-	272

Tabela 4.4 - Número de instâncias da classe '+' e '-' relacionados à Australian Credit Approval. São 218 instâncias positivas e 272 instâncias negativas (QUILAN, N/A).

Attribute	Type
A1	nominal
A2	continuous
A3	continuous
A4	nominal
A5	nominal
A6	nominal
A7	nominal
A8	continuous
A9	nominal
A10	nominal
A11	continuous
A12	nominal
A13	nominal
A14	continuous
A15	continuous
Class	nominal

Tabela 4.5 - Tipos dos atributos relacionados à Australian Credit Approval. São divididos em atributos nominais e atributos contínuos (QUILAN, N/A).

A base de dados consiste de 490 instâncias com 44.5% sendo positivas (crédito aprovado) e 55.5% sendo negativas (crédito negado), as restantes 5% são considerados valores perdidos.

2. German credit data(Hoffman, 1994): Esta base de dados provida pelo Prof. Dr. Hans Hoffman, da universidade de Hamburgo, consiste de 1000(mil) instâncias com 20 atributos sendo 7 atributos numéricos e 13 atributos categóricos, e um atributo classe que define cada instância como um bom pagador e um mau pagador. Esta base de dados requer o uso de uma matriz de custo, como na tabela 4.6.

Classe Atual	Classificação Prevista	
	1(Bom Pagador)	2(Mau Pagador)
1(Bom Pagador)	0	1
2(Mau Pagador)	5	0

Tabela 4.6 – Matriz de custo formada para se demonstrar o custo na classificação de um bom pagador como mau pagador e vice-versa

As linhas representam a classificação atual ou real, composta pelo atributo classe que pode ser '1' para Bom pagador e '2' para mau pagador, enquanto que as colunas representam a classificação prevista. A matriz então demonstra a possibilidade de se classificar um individuo corretamente ao ratificar sua classe real, ou a possibilidade de se classificar equivocadamente um mau pagador como bom pagador e vice-versa.

Observamos então segundo a matriz de custo da tabela 4.6, que para cada classificação correta de bom pagador ou mau pagador, ou seja, para cada acerto na classificação de um indivíduo como bom ou mau pagador, a matriz aponta um custo 0, não configurando prejuízo ou ganho ao credor. Enquanto que se um bom pagador for classificado erroneamente como mau pagador, o custo será igual a 1, ou seja, haverá prejuízo ao credor, porém menor do que quando um indivíduo mau pagador é classificado como bom pagador com um custo 5. Logicamente então é melhor classificar um bom pagador como mau pagador do que classificar um mau pagador como bom pagador, o que implicaria em um prejuízo bem maior.

Estas duas bases de dados foram formatadas em arquivos ARRF para posterior interpretação, análise e experimentos através do software WEKA. A tabela 4.7 mostra a comparação em relação ao número de atributos, classes e instâncias das duas bases de dados apresentados na tabela 4.4 e 4.5.

Conjunto de Dados Financeiros	Classes	Atributos	Número de Instâncias
Australian Credit Approval	2	15	490
German Credit Data	2	20	1000

Tabela 4.7 – Mostra as bases de dados financeiros com suas respectivas classes, atributos e número de instâncias.

4.3 Redes Neurais e Árvores de decisão

A fim de avaliarmos o desempenho do classificador Naïve Bayes em termos práticos, o compararemos a dois classificadores conhecidos, precisos e vantajosos em mineração de dados: os classificadores baseados em redes neurais e os baseados em árvores de decisão. Os classificadores baseados em redes neurais fornecem um método interessante de classificação já que são estruturados em nós simples (ou neurônios) que são interligados para formar uma rede de nós, ou rede neural. A principal vantagem das redes neurais é a habilidade de se aprender através de seu próprio ambiente, e com isso, melhorar seu desempenho. Para efeito de comparação, neste trabalho decidimos utilizar as redes de múltiplas camadas que são modelos de redes que apresentam uma ou mais camadas de neurônios entre as camadas de dados e de saída de resultados, chamadas camadas intermediárias.

Estas redes neurais artificiais são o modelo mais utilizado na atualidade e são baseadas em um algoritmo de retropropagação (*backpropagation*). Nessas redes, cada camada tem uma função específica. A camada de saída recebe os estímulos da camada intermediária e gera a resposta final. As camadas intermediárias funcionam como extratoras de características, sendo seus pesos uma codificação de características apresentadas nos padrões de entrada e permitem que a rede crie sua própria representação, mais rica e complexa, do problema (CARVALHO, 2000).

Em relação a técnica baseada em árvores de decisão, estas podem ser usadas em conjunto com a tecnologia de indução de regras, e são as únicas a apresentar resultados hierarquicamente representados. Nestas, o atributo de maior importância localiza-se no primeiro nó, enquanto que os atributos de menor importância são apresentados nos nós subsequentes. A grande vantagem das árvores de decisão é a tomada de decisão, já que os atributos mais relevantes são

levados em consideração, o que torna a tarefa de classificação mais compreensível para a maioria das pessoas. Ao apresentar os atributos em ordem de importância, as árvores de decisão possibilitam ao usuário que conheçam com mais clareza os fatores que realmente influenciam o seu funcionamento.

.

4.4 Resultados dos experimentos

Os experimentos foram resultado da aplicação do algoritmo Naïve Bayes junto às bases de dados citadas na seção 4.2, Australian Credit Approval e German Credit Data, com a finalidade de se obter a acurácia do algoritmo estudado. A fim de medir o resultado de nossos experimentos com o algoritmo Naïve Bayes, comparamos este último com outros dois algoritmos também aplicados à mineração de dados: o algoritmo de Redes Neurais e o que utiliza a técnica de árvores de decisão. Na tabela 4.8 demonstramos os resultados através da comparação das porcentagens de acertos obtidos, na fase de classificação, por cada algoritmo.

Base de Dados	Média de Acertos (%)		
	Árvore de decisão	Naïve Bayes	Redes Neurais
Australian Credit Approval	86.087	77.681	83.768
German Credit Data	70.500	75.600	71.600
Média de Acertos Final	**78.293**	**76.640**	**77.684**

Tabela 4.8 – Demonstra comparação entre três métodos de classificadores através de suas respectivas médias de acertos em relação a duas bases de dados de domínios financeiros

Dentre os três classificadores, o Naïve Bayes apresentou a menor acurácia considerando a média das aplicações nos dois bancos de dados mencionados acima. Apesar disso, sua simplicidade, eficiência e facilidade de uso, seu baixo custo computacional, correspondente à simples contagem das frequências cruzadas entre os valores da variável classe e das variáveis preditivas, aliado à superioridade da técnica bayesiana em domínios de dados com atributos diversos, a tornando robusta à ruídos e atributos irrelevantes, fazem do Naïve Bayes uma ótima opção no processo de classificação em domínios de dados financeiros, bem como no processo de classificação de dados como um todo.

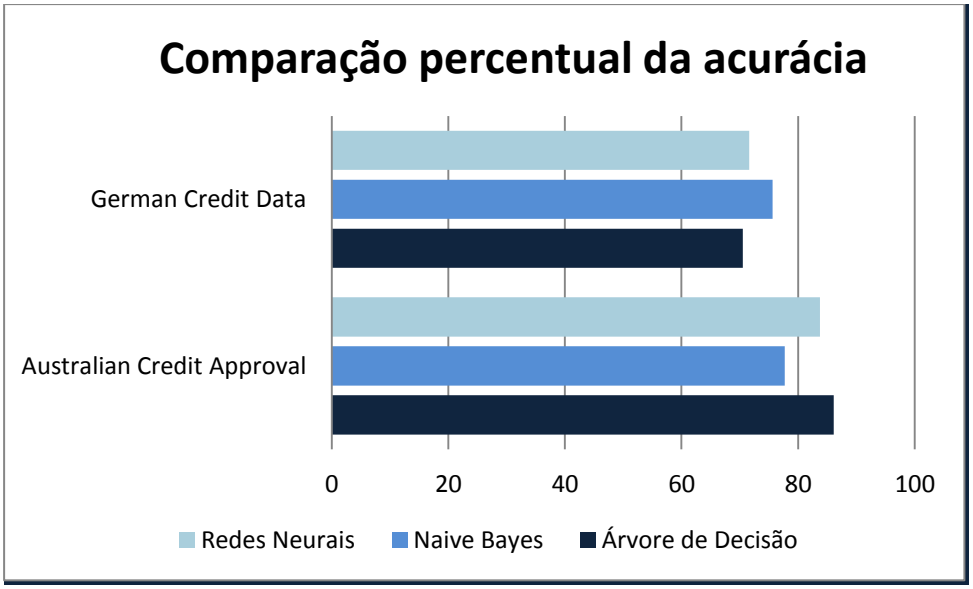


Figura 4.3 – Demonstra um gráfico contento a comparação percentual dos métodos de classificação aplicados a duas bases de dados

Os classificadores envolvendo Redes Neurais e Árvore de Decisões também apresentam contextos interessantes no processo de classificação. Enquanto a primeira tenta construir representações internas de modelos ou padrões perceptíveis nos dados, geralmente invisíveis ao usuário, utilizando um conjunto de elementos de processamento semelhante ao dos neurônios, a segunda é uma boa escolha na predição de saídas e objetivação da categorização dos dados, apresentando resultados fáceis de compreender detalhando quais informações foram mais importantes no processo de classificação. Apesar de processos precisos são custosos computacionalmente na etapa de classificação, apresentando técnicas mais complicadas de classificação.

4.5 Considerações Finais

Como observado neste capítulo, o estudo de casos deste trabalho, baseado em operações envolvendo análise de crédito foi divido em dois. O primeiro demonstrou o comportamento do classificador Naïve Bayes no processo de classificação, realizando a tarefa de classificar novos clientes de uma empresa financeira em bons ou maus pagadores. Sua ação demonstrou a fase de classificação na prática, reforçando o caráter probabilístico do classificador

bayesiano, que efetuou novas classificações a partir da aplicação da fórmula de Bayes.

Já o segundo estudo de casos, direcionado à análise de bases de dados de duas instituições financeiras (German Credit Data e Australian Credit Approval), realizou a comparação entre o classificador Naïve Bayes e outros dois classificadores baseados em redes neurais e árvores de decisão. O classificador baseado em árvores de decisão obteve o melhor resultado seguido de perto pelo de redes neurais. Apesar de ficar com a última posição, em vista da suposição de independência condicional entre atributos e deficiência em bases de dados com poucas instâncias, o Naïve Bayes demonstrou que com sua simplicidade e robustez, rapidez na classificação e boa manipulação de dados, é uma boa opção para operações financeiras de análise de crédito, bem como para tarefas envolvendo classificação, em geral.

5 CONCLUSÃO

Classificar novos objetos quanto a pertencerem a novas classes ou não, é objetivo comum a diversas áreas de conhecimento, como bioinformática, medicina, segurança de dados e principalmente análise de crédito, foco deste trabalho. Este trabalho realizou o estudo sobre o classificador Naïve Bayes mediante a tarefa de classificação aplicada a conjuntos de dados financeiros.

Foram definidos e discutidos métricas e métodos de modelos de classificação e classificadores, no intuito de compreender como se desenvolve a fase de classificação, uma das mais importantes em mineração de dados e bastante aproveitada no contexto de análise de crédito, pois a correta classificação de crédito é vital para a sobrevivência do mercado bancário.

Foi analisado o classificador Naïve Bayes, bem como suas particularidades, sua execução ao logo do processo de classificação, seus cálculos probabilísticos baseados na fórmula de Bayes, seus aspectos positivos e limitações no que diz respeito à tarefa de classificação.

Ao final do trabalho, a partir da análise de bases de dados financeiras realizadas no primeiro estudo de caso (seção 4.1) deste trabalho, o classificador Naïve Bayes foi apresentado na prática, executando previsões para a classe de "bons pagadores" ou "maus pagadores", através da ferramenta voltada para mineração de dados, chamada WEKA, ferramenta esta que possui uma biblioteca de classes totalmente integráveis a programas JAVA, onde foi simulado o nosso primeiro estudo de caso

Posteriormente, a partir das bases de dados estudadas na seção 4.2 (German Credit Data e Australian Credit Approval), foram comparados os desempenhos do Naïve Bayes ao de dois outros classificadores, Redes Neurais e Árvores de decisão. Nesta comparação, apesar do Naïve Bayes ter obtido a menor acurácia dentre os três (76,64% do Bayes, contra 78,29% do classificador referente a árvores de decisão e 77,68% do classificador referente a redes neurais), o classificador bayesiano se mostrou simples, robusto a ruídos, insensível a atributos irrelevantes, com um agradável baixo custo computacional, demonstrando ser uma boa alternativa a problemas envolvendo análise de crédito.

Este trabalho envolveu estudos na área de estatística e probabilidade, focando na técnica de classificação do Naïve Bayes. A partir dele, possibilitou-se

ampliar os conhecimentos sobre a área de mineração de dados enfatizando como esta se tornou objeto de estudo na atualidade, principalmente quando aplicada no contexto de análise de crédito. Ao término do trabalho, foi perceptível os resultados interessantes que tivemos, assimilando a importância da descoberta de conhecimento (KDD) em assuntos atuais.

Espera-se que este trabalho tenha contribuído para os estudos na área de mineração de dados, bem como assuntos afins. A seguir enumerados algumas das contribuições geradas por este trabalho:

- Contribuições deste trabalho:
  - Disponibilização de dois estudos de casos na área de análise de crédito.
  - Avaliação qualitativa sobre o classificador Naïve Bayes por meio de abordagens para classificação de dados
  - Avaliação quantitativa em relação a dois estudos de casos realizados no campo de concessão de crédito.

Como sugestões para trabalhos futuros, propõem-se:

- Utilizar a técnica do algoritmo Naïve Bayes, apresentada neste trabalho, para efetuar diagnósticos em bioinformática com o objetivo de detectar câncer de mama.
- Modificação do Naïve Bayes, apresentado neste trabalho, para problemas que envolvam menos instâncias na base de dados.
- Exploração de redes bayesianas em Inteligência Artificial.

REFERÊNCIAS

ABERNETHY, M. **Mineração de dados com WEKA**. Disponível em <http://www.ibm.com/developerworks/br/opensource/library/os-weka1/> Acesso em 20 de out. 2013

CARVALHO, A.P. **Redes neurais artificiais**. 2000. Disponível em: <http://www.icmc.sc.usp.br/~andre/neural1.html>. Acesso em: 20 de nov. 2013.

ENGEL, P. M. **Avaliação de Modelos.** Disponível em <http://www.inf.ufrgs.br/~alvares/CMP259DCBD/avaliacao.pdf> Acesso em 15 de ago. 2013.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery & data mining**. Menlo Park, CA, USA: AAAI/MIT, 1996.

FREITAS, A. A. **Uma Introdução a Data Mining**. Informática Brasileira em Análise. CESAR - Centro de Estudos e Sistemas Avançados do Recife. Ano II, n. 32, mai./jun. 2000.

GARCIA, J. L. **Aprendizado Probabilístico.** Disponível em <http://wiki.icmc.usp.br/images/2/20/IA12-2011.pdf> Acesso em 15 de out. 2013

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um guia prático**. Editora Campus, Rio de Janeiro: Elsevier. 2005.

GONÇALVES, E. C. **Naïve Bayes: Mineração de dados na prática.** Revista SQL Magazine Nº 110. 2013

HAND, D. J. **The top ten algorithms in data mining.** Disponível em <http://somedocs.googlecode.com/files/Top%2010%20algorithms%20in%20data%20mining.pdf> Acesso em 25 out. 2013

HOFMANN, H. **German Credit Data.** 1994. Disponível em: <http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)>. Acesso em: 10 outubro. 2013.

JAMAIN, A; HAND, D. J. **The Naïve Bayes Mystery: A classification detective story.** Journal Pattern Recognition Letters, 2005

KOHAVI, R. **A study of cross - validation and bootstrap for accuracy estimation and model selection**. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, n.2, p.1137–1143, 1995.

QUINLAN, R. **Australian Credit Approval.** Disponível em: <http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval)> Acesso em: 10 out. 2013.

SANTOS, J. O. **Analise de Crédito**. Editora Atlas, 2000.

SILVA, M. P. **Mineração de dados - conceitos, aplicações e experimentos**. Disponível em: <http://www.inpe.br/>. Acesso em: 20 nov. 2013.

STEINER, M. T. A.; CARNIERI, C.; KOPITTKE, B. H.; NETO, P. J. **Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário**. Revista de Administração da Universidade de São Paulo (RAUSP), São Paulo, v.34, n.3, p.56-67, jul./set. 1999.

TAN, PANG NING; STEINBACH, MICHAEL; KUMAR, VIPIN. I**ntroduction to Data Mining.** Editora Addison-Wesley, 2006

WITTEN, I. H.; FRANK, E. **Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations**. Morgan Kaufmann Publishers. San Francisco, California, 2000.