UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC
CURSO DE CIÊNCIA DA COMPUTAÇÃO

BRUNA BALDINI DIAS

OS MÉTODOS BAYESIANOS DE APRENDIZADO DE MÁQUINA PELOS ALGORITMOS NAÏVE BAYES E REDES DE CRENÇA NA PREDIÇÃO DE PERÍODO CHUVOSO NA CIDADE DE BLUMENAU

CRICIÚMA
2018

BRUNA BALDINI DIAS

OS MÉTODOS BAYESIANOS DE APRENDIZADO DE MÁQUINA PELOS ALGORITMOS NAÏVE BAYES E REDES DE CRENÇA NA PREDIÇÃO DE PERÍODO CHUVOSO NA CIDADE DE BLUMENAU

Trabalho de Conclusão de Curso, apresentado para obtenção do grau de Bacharel no curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientadora: Profª. Dra. Merisandra Côrtes de Mattos Garcia

CRICIÚMA

2018

BRUNA BALDINI DIAS

OS MÉTODOS BAYESIANOS DE APRENDIZADO DE MÁQUINA PELOS ALGORITMOS NAÏVE BAYES E REDES DE CRENÇA NA PREDIÇÃO DE PERÍODO CHUVOSO NA CIDADE DE BLUMENAU

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de Bacharel, no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC, com Linha de Pesquisa em Inteligência Artificial

Criciúma, 30 de novembro de 2018.

BANCA EXAMINADORA

Profa. Merisandra Côrtes de Mattos Garcia - Doutora - UNESC - Orientador

Prof. Cristian Cechinel - Doutor - UFSC (Araranguá)

Prof. Luciano Antunes - Mestre - UNESC

Ao meu eterno companheiro, Lucas.

AGRADECIMENTOS

Gostaria de agradecer ao meu namorado, que sempre esteve ao meu lado nos momentos difíceis e que me apoiou durante toda essa jornada.

Agradeço aos meus amigos Agatha, Cristiano, Simone, Cinthia, Maria Luíza e Nicole por estarem sempre comigo e me proporcionarem bons momentos.

Agradeço à minha mãe que me proporcionou esta oportunidade.

A minha orientadora, Merisandra Côrtes de Mattos Garcia, por toda a ajuda disponibilizada para a realização deste trabalho.

.

“Aquilo que não está morto pode jazer eternamente
e com eras estranhas, até a morte pode morrer.”

H. P. Lovecraft

RESUMO

Com a crescente quantidade de dados, surgiu a demanda de novos métodos para a análise destes. Para resolver este problema, originou-se o *data mining*, que consiste em várias tarefas e técnicas para analisar grandes quantidades de dados e descobrir novas informações e conhecimento ocultados nestes. Uma das tarefas de data mining é a classificação, uma forma supervisionada de aprendizado de máquina que atribui um objeto à uma classe pré-definida dentro de um conjunto. A tarefa de classificação consiste em duas etapas: a fase de treinamento, onde um modelo é construído a partir de um conjunto de dados reservados para esta etapa; e a fase de teste, onde ocorre a validação do modelo obtido na etapa anterior em um conjunto de dados desconhecido. Há vários algoritmos disponíveis para a realização da classificação. Esta pesquisa consiste na identificação de um modelo de predição de período chuvoso na cidade de Blumenau por meio da tarefa de classificação bayesiana. Os algoritmos empregados são o Naive Bayes e o de Redes de Crença. O algoritmo Naïve Bayes determina a probabilidade condicional das classes assumindo que os atributos sejam condicionalmente independentes uns dos outros. O algoritmo de rede de crenças apresenta os resultados de forma gráfica, por meio de um grafo acíclico direcionado. A ferramenta de *data mining* escolhida para a realização desta pesquisa foi o Weka 3.8, por possuir licença gratuita e apresentar os algoritmos necessários. Os dados empregados foram disponibilizados pela Agência Nacional das Águas na ferramenta Hidroweb. O conjunto é constituído de dados pluviométricos coletados na estação do bairro Itoupava Central na cidade de Blumenau. O modelo gerado pelo algoritmo *Naïve Bayes* obteve uma acurácia de 85,36% na fase de teste, enquanto o algoritmo de rede de crenças obteve uma acurácia de 89,57%. Após a aplicação das medidas de qualidade para classificadores em *data mining* e a comparação dos mesmos, conclui-se que o modelo gerado pelo algoritmo de rede de crenças possui melhor acurácia e resultados superiores nas demais medidas de qualidade para classificadores empregadas, como a taxa de erro, *Area Under Curve*, estatística *kappa*, proporção de falsos positivos, proporção de falsos negativos, sensibilidade, especificidade, precisão e *F*-score.

Palavras-chave: Aprendizado de máquina. *Data mining*. Classificação. Naïve Bayes. Rede de Crenças Bayesiana.

ABSTRACT

With the growing data volume, the need of new methods to better analyse this data was born. To solve this problem, the concept of data mining was created. It consists in multiple tasks and techinques to analyse huge quantities of data to discover new information and knowledg hidden in them. One of the data mining tasks is classification, a supervisioned machine learning form that gives a pre-defined class label to an object inside a dataset. The classification task has two steps: the training phase, where a classification model is built from a dataset reserved for this step; and the testing phase, where the model previously generated is tested in a set of unknow data. There are various algorithms available to perfom the classification task. This research consists in the identification of rainy season in the city of Blumenau with the bayesian classification task. The algorithms applied in this research are Naïve Bayes and Bayesian Belief Network. The Naïve Bayes algorithm determines the conditional probability of the classes assuming that the atributes aren't conditionaly dependentes of each other.The Bayesian Belief Network algorithm presents the results in the form of a directed acyclic graph. The chosen data mining tool to perform this research was Weka 3.8, since it's an open source software and it has the needed algorithms. The dataset used in this research were disponibilized by the National Water Agency in their Hidroweb tool. The dataset is constituted by rainfall data colected by the station of the Itoupava Central (bairro) in the city of Blumenau. The generated model by Naïve Bayes scored na accuracy of 85.36% in its test phase, while the Bayesian Belief Network scored and accuracy of 89.57%. After the application of the quality measures for classifiers in data mining and the comparission of them, it was conclued that the model generated by the Bayesian Belief Networks algorithm has better accuracy and superior results in the other applied quality measures, like error rate, Area Under Curve, kappa statistic, false positive rate, false negative rate, sensibility, specificity, recall, and F-score.

Palavras-chave: Machine learning. Data Mining. Classfication. Naïve Bayes.

Bayesian Belief Network.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
AUC	*Area Under Curve*, no português, Área Abaixo da Curva
BBN	*Bayesian Belief Networks*, no português, Redes de Crenças Bayesianas
CPT	*Conditional Probability Table*, no português, Tabela de Probabilidade Condicional
GAD	Grafo Acíclico Direcionado, do inglês, *Directed Acyclic Graph*
DP	Desvio Padrão
FN	Falso Negativo
FP	Falso Positivo
IA	Inteligência Artificial
KDD	*Knowledge Discovery in Databases*, no português, Descoberta de Conhecimento em Bancos de Dados
VN	Falso Negativo
VP	Falso Positivo

SUMÁRIO

1 INTRODUÇÃO

O rápido crescimento e integração das bases de dados proveem a cientistas, engenheiros e empresários um vasto recurso que pode ser utilizado para fazer novas descobertas científicas, aperfeiçoar sistemas industriais e descobrir novos padrões nos dados. *Data mining* é a análise de bases de dados, geralmente grandes, para a descoberta de relações entre esses dados e a sua organização em novas formas, tornando-os compreensíveis e úteis ao usuário (HAND; MANNILA; SMYTH, 2001, tradução nossa). Segundo Han, Kamber e Pei (2012, tradução nossa), para que se possa aproveitar os dados e melhorar os resultados das análises, o *data mining* apresenta diversas tarefas que são usadas para especificar os tipos de padrões que são encontrados, tais como: agrupamento (*clustering*), associação, regressão e classificação.

A classificação é uma forma supervisionada de aprendizado de máquina (MONARD; BARANAUSKAS, 2003) que consiste em analisar um objeto e atribuí-lo a uma classe pré-definida dentro de um conjunto. O processo de classificação é dividido em duas etapas: na primeira etapa, um modelo de classificação é construído a partir da análise de um conjunto de dados selecionados para o aprendizado; na segunda etapa, o modelo resultante da primeira etapa é usado para a classificação dos dados (BERRY; LINOFF, 2004).

Há uma vasta quantidade de métodos que podem ser utilizados para realizar a tarefa de classificação, entre eles pode-se citar os bayesianos, que são algoritmos que se baseiam no Teorema de Bayes, um princípio estatístico para combinar informações já conhecidas com novas informações extraídas dos conjuntos de dados. Dentre os algoritmos que utilizam o método bayesiano tem-se o *Naïve Bayes*, que determina a probabilidade condicional das classes assumindo que os atributos são condicionalmente independentes; e as Redes de Crença, também conhecidas como Redes Bayesianas, que permitem especificar quais atributos são condicionalmente independentes (HAN; KAMBER; PEI, 2012; TAN; STEINBACH; KUMAR, 2009).

Data mining vem se tornando uma ferramenta eficiente nos campos da ciência, engenharia, indústria, medicina, assistência à saúde e outros serviços sociais (KUMAR, 2011). Recentemente, uma nova onda de pesquisa em *data mining*

tem sido conduzida em séries temporais. Mineração de dados de séries temporais consiste em analisar uma sequência de pontos de dados que contém medidas sucessivas feitas durante um determinado período de tempo. No presente, vários domínios depositam expectativas em dados de séries temporais, como a área financeira, mercado de ações, alterações climáticas e outros (ZAINUDIN; JASIM; BAKAR, 2016, tradução nossa).

Ainda de acordo com Zainudin, Jasim e Bakar (2016, tradução nossa), a análise de alterações climáticas visa estudar o comportamento do clima durante um período de tempo específico. A característica chave por trás da mudança climática se encontra na natureza de seus dados, que são capturados em formato de pontos temporais. Uma tarefa de alteração climática é a previsão de chuva, onde atributos específicos como umidade e vento são usados para prever chuvas em uma localização específica. Várias técnicas como máquina de vetores de suporte, *Naïve Bayes*, redes neurais e outras vêm sendo usadas para a previsão de chuvas, sendo que a maioria das técnicas empregadas consiste em aprendizado supervisionado.

Assim, esta pesquisa propõe a análise deste conjunto de dados por meio de métodos bayesianos de aprendizado de máquina. No entanto, segundo Monard e Baranauskas (2003), apesar do aprendizado de máquina ser uma forma importante para descoberta automática de conhecimento, deve-se atentar para o fato de que não existe um único algoritmo que apresente o melhor desempenho em variadas problemáticas. Dessa forma, esta pesquisa compreenderá a aplicação dos algoritmos *Naïve Bayes* e Redes de Crença Bayesianas em dados pluviométricos. Os resultados originados são comparados pelo uso de medidas de qualidade em data mining, comumente empregadas pela comunidade de aprendizado de máquina, a fim de identificar um modelo de predição de período chuvoso na cidade de Blumenau.

1.1 OBJETIVO GERAL

Identificar um modelo de predição de período chuvoso na cidade de Blumenau por meio da tarefa de classificação bayesiana.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos desta pesquisa consistem em:

a) descrever os conceitos de pluviometria, *data mining*, classificação e os algoritmos empregados;
b) empregar os algoritmos *Naïve Bayes* e Redes de Crença;
c) analisar por meio de medidas de qualidade em *data mining* os modelos gerados;
d) identificar por meio de medidas de qualidade o melhor modelo de predição para o conjunto de dados da pesquisa.

1.3 JUSTIFICATIVA

O uso de análise estatística é tido como o principal método para a análise de dados na maior parte dos campos científicos na história recente. No presente, acredita-se que, enquanto a análise estatística é vista como um método primário de análise de dados, *data mining* pode ser visto como um método secundário. No entanto, existem diferenças significativas entre os dois métodos. Primeiro, enquanto a análise estatística tende a usar estratégias tradicionais, *data mining* é mais flexível quanto aos métodos a serem utilizados; segundo, a análise estatística usa uma amostra de dados coletados de uma população, *data mining* emprega dados que englobam uma população inteira; terceiro, a análise estatística é uma abordagem matemática e lida apenas com dados numéricos, *data mining* pode lidar com vários tipos de dados, como por exemplo, imagens, texto e sons; por último, a análise estatística é hipotético-dedutiva1, *data mining* é indutivo[2]. Na análise estatística, primeiramente se levanta uma hipótese, e então há a coleta e análise dos dados. Em *data mining*, dados anteriormente coletados são explorados sem a necessidade de uma hipótese, descobrindo-se padrões escondidos nos dados analisados (YOO et al., 2011).

No que se refere a um modelo de classificação, este pode ser usado como uma ferramenta exploratória para distinguir objetos de classes diferentes, podendo também ser usado para predizer a classe de objetos desconhecidos. O

[1] O método hipotético-dedutivo caracteriza-se pela formulação de hipóteses para um problema ou lacuna no campo científico, seguida pelo teste de predições da ocorrência dos fenômenos abrangidos pela dita hipótese (PRODANOV; FREITAS, 2013).
[2] O método indutivo é caracterizado pela generalização, ou seja, parte-se de um caso particular e o toma como geral para todos os casos semelhantes (PRODANOV; FREITAS, 2013).

processo de classificação, independente do algoritmo usado, apresenta uma fase de aprendizado, que permite identificar um modelo de classificação que melhor se adequa à relação entre o atributo determinado e a classe dos dados inseridos. O uso de métodos bayesianos para realizar a classificação se dá pelo fato de que o Teorema de Bayes permite estimar com facilidade a probabilidade a posteriori a partir de um conjunto de treinamento, calculando a fração dos dados deste conjunto que pertencem a cada classe (TAN; STEINBACH; KUMAR, 2009).

Nesta pesquisa, a escolha do algoritmo *Naïve Bayes* deu-se em função de ser resistente a pontos de ruído e atributos irrelevantes, não os considerando no cálculo probabilístico. O *Naïve Bayes* também pode lidar com valores ausentes, ignorando o exemplo na construção e classificação do modelo. No que se refere as Redes de Crença estas fornecem uma abordagem para capturar conhecimento anterior e apresentá-lo de forma gráfica, além de serem adequadas quando se trabalha com dados incompletos. Embora ambos os algoritmos façam uso do Teorema de Bayes, eles lidam com os dados de formas diferentes. Enquanto o algoritmo Naïve Bayes assume que os atributos de uma base de dados são condicionalmente independentes entre si, as Redes de Crença permitem especificar quais atributos possuem dependências condicionais (TAN; STEINBACH; KUMAR, 2009).

Os estudos referentes à precipitação pluvial, domínio de aplicação desta pesquisa, vêm se dando em diferentes regiões do mundo, em razão de importância que esta representa para o ciclo hidrológico e para a manutenção da vida no planeta, assim tornando as secas um enorme problema para a sociedade e os ecossistemas (SILVA et al, 2011). Entretanto, chuvas em demasia também apresentam efeitos prejudiciais para a sociedade. De acordo com Borges e Thebaldi (2016), a precipitação máxima – definida como a ocorrência extrema com duração e distribuição temporal e espacial crítica para uma área ou bacia hidrográfica – pode agravar a erosão do solo, assim como também pode causar inundações em áreas rurais e urbanas e danificar obras hidráulicas.

Zainudin, Jasim e Bakar (2016, tradução nossa) dizem que a seleção de uma técnica apropriada para uma duração específica de chuva é uma tarefa crucial, portanto, sendo necessário investigar múltiplas técnicas a fim de determinar qual delas possui melhor desempenho na predição de chuvas.

Os modelos descritivos apenas sumarizam os dados da forma conveniente ou esperada, levando a maior compreensão de como estes dados funcionam. Modelos preditivos, por sua vez, têm o objetivo específico de predizer valores desconhecidos a partir de valores já conhecidos de outras variáveis. Identificar um modelo de predição por *data mining* se torna interessante, visto que se provou eficiente na descoberta de novos padrões e informações até então desconhecidas a partir de vastos conjuntos de dados (HAND; MANNILA; SMYTH, 2001; KUMAR, 2011).

Baseando-se nas informações apresentadas, aplicar estes algoritmos em uma base de dados de pluviometria pode revelar resultados potencialmente novos, e comparar seus resultados se torna interessante pela forma distinta que ambos os algoritmos lidam com valores desconhecidos, que interferem diretamente no modelo resultante.

1.4 ESTRUTURA DO TRABALHO

Este trabalho é composto por seis capítulos, onde o primeiro descreve a pesquisa proposta, os objetivos e a justificativa para a realização do trabalho.

No capítulo 2 são abordados os conceitos básicos de *data mining* e outros temas relacionados, como pré-processamento dos dados, aprendizado de máquina e a tarefa de classificação.

No capítulo 3 é introduzido o conceito básico de algoritmos bayesianos, algoritmos que se baseiam no Teorema de Bayes e constroem modelos de classificação com base em cálculos probabilísticos. O funcionamento de dois algoritmos bayesianos – *Naïve Bayes* e Redes de Crenças Bayesianas – é descrito, e são apresentados métodos de comparação entre os algoritmos e modelos obtidos a partir dos mesmos, a fim de determinar qual deles obteve maior desempenho no problema proposto.

No capítulo 4, são descritas algumas pesquisas que se relacionam com os temas abordados nesta pesquisa, com a aplicação de classificadores em dados pluviométricos e a aplicação dos algoritmos bayesianos em outras bases de dados.

No capítulo 5, são descritos a metodologia empregada para a realização desta pesquisa e os resultados obtidos a partir da análise dos dados com os algoritmos propostos. No capítulo 6, encontra-se a conclusão deste trabalho.

2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Avanços rápidos na tecnologia de acúmulo e armazenamento de dados permitiram que organizações acumulassem vastas quantidades de dados. Contudo, extrair informações úteis é extremamente desafiador. Frequentemente, técnicas e ferramentas tradicionais de análise de dados não podem ser usadas devido ao tamanho massivo do conjunto de dados ou de sua natureza não tradicional (TAN; STEINBACH; KUMAR, 2009), da mesma forma que se carece de profissionais capacitados em transformar estes dados em conhecimento (LAROSE, 2005, tradução nossa).

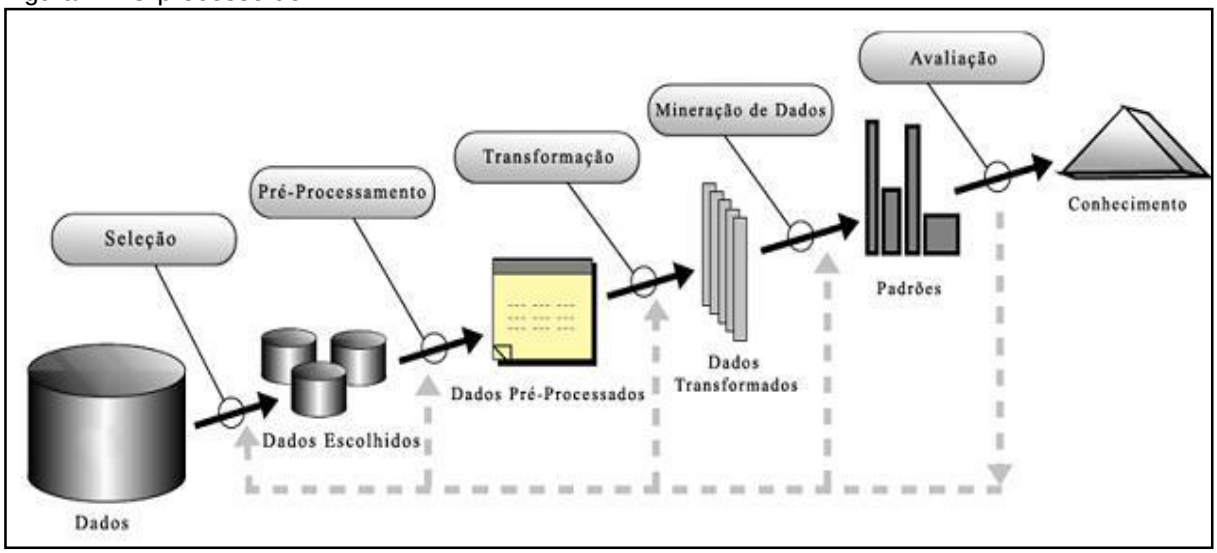
Para atender essa nova necessidade, surge um conceito chamado Descoberta de Conhecimento em Bases de Dados, tradução de *Knowledge Discovery in Databases* (KDD) (GOLDSCHMIDT; PASSOS, 2005). O termo KDD foi cunhado em 1989 por Piatetsky-Shapiro, para enfatizar que o conhecimento é o produto final de uma descoberta baseada em dados. KDD se refere ao processo geral de descobrir conhecimentos úteis a partir de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

De acordo com Han, Kamber e Pei (2012, tradução nossa), muitas pessoas tratam *data mining* como sinônimo para KDD, embora, na realidade, *data mining* é um passo essencial no processo de descoberta de conhecimento. O processo de KDD, conforme descrito na figura 1, consiste em:

a) **limpeza dos dados:** remoção de ruídos e dados inconsistentes;
b) **integração dos dados:** várias fontes ou conjuntos de dados podem ser combinadas;
c) **seleção de dados:** os dados mais importantes para a tarefa de análise são selecionados;
d) **transformação dos dados:** os dados são transformados e consolidados em formatos apropriados para a mineração por meio das operações de resumo ou agregação;
e) ***data mining*:** processo essencial, onde métodos inteligentes são aplicados para extrair padrões dos dados;
f) **avaliação dos padrões:** identificação dos padrões que melhor representem o conhecimento;

g) **apresentação do conhecimento:** técnicas de visualização e representação de conhecimento são usadas para apresentar o conhecimento obtido para os usuários.

Figura 1 – O processo de KDD



Fonte: Adaptado de Fayyad, Piatetsky-Shappiro e Smyth (1996).

2.1 PRÉ-PROCESSAMENTO DE DADOS

Aplicações de mineração de dados, por muitas vezes, são utilizadas em dados que foram coletados para outro propósito ou para usos futuros que ainda não foram especificados (TAN; STEINBACH; KUMAR, 2009).

Comumente, os dados não se encontram em um estado que seja apropriado para processamento. Por exemplo, os dados podem estar codificados em logs complexos ou documentos de formato livre. Em muitos casos, diferentes tipos de dados podem estar aleatoriamente misturados em um documento de formato livre. Para fazer com que os dados se tornem apropriados, é essencial transformá-los em um formato que seja adequado para algoritmos de *data mining*, como multidimensional, séries temporais ou semiestruturado (AGGARWAL, 2015, tradução nossa).

A qualidade dos dados se refere à característica de corresponder de forma aproximada ao uso para o qual eles foram coletados ou, alternativamente, se eles refletem apropriadamente o contexto real do qual foram obtidos. Assegurar essa qualidade é particularmente importante quando os dados são usados para

propósitos específicos, como modelagem de tempo (GORUNESCU, 2011, tradução nossa).

É irrealista esperar que os dados sejam perfeitos. Pode haver problemas decorridos de erro humano, limitações por parte das ferramentas de medição ou falhas no processo de coleta de dados. Em outros casos, há a possibilidade de existir dados ilegítimos ou duplicados (TAN; STEINBACH; KUMAR, 2009). Segundo Gorunescu (2011, tradução nossa), os principais problemas encontrados no processo de coleta de dados são:

a) **ruído:** componente aleatório de um erro de medição. Pode envolver a distorção de um valor ou a adição de valores ilegítimos;
b) ***outlier*:** um *outlier* é um objeto distante do resto dos dados, cujas características são consideravelmente diferentes da maior parte dos objetos no conjunto de dados;
c) **valores faltantes;**
d) **dados duplicados.**

Outros problemas que podem ser encontrados na coleta de dados, de acordo com Tan, Steinbach e Kumar (2009), são:

a) **artefatos:** distorções determinísticas nos dados, por exemplo, uma faixa que se repete em um conjunto de fotografias;
b) **precisão:** proximidade de medidas repetidas entre si. É geralmente medida pelo desvio padrão de um conjunto de valores;
c) **foco:** variação sistemática de medições a partir da quantidade sendo medida. É medido tirando a diferença entre a média do conjunto de valores e o valor conhecido da quantidade sendo medida;
d) **valores inconsistentes.**

Outros fatores que afetam a qualidade dos dados são credibilidade, que reflete o quanto os usuários confiam nos dados, e interpretabilidade, que representa a facilidade de compreensão dos dados (HAN; KAMBER; PEI, 2012, tradução nossa).

A fase de pré-processamento dos dados possivelmente é a mais importante no processo geral de *data mining*. Esta fase consiste nos seguintes passos (GORUNESCU, 2011, tradução nossa):

a) **extração de características:** nesta fase ocorre a abstração das características mais relevantes para a aplicação em questão. Por

vezes, o analista pode ser confrontado pela grande quantidade de dados brutos. Portanto, esta fase requer conhecimento na área específica da aplicação;

b) **limpeza dos dados:** os dados extraídos na fase anterior podem possuir entradas errôneas ou faltantes, alguns registros podem precisar ser descartados, valores faltantes podem precisar ser estimados e inconsistências podem precisar ser removidas. Esta fase visa resolver esses problemas;

c) **seleção e transformação de características:** quando os dados são de grande dimensão, muitos algoritmos de *data mining* não funcionam devidamente. Ademais, muitas das características de grande dimensão são ruidosas e podem acrescentar erros ao processo de *data mining*. Nesta fase, características irrelevantes são removidas, ou o atual conjunto de características é transformado em um novo espaço de dados que é mais acessível para análise.

2.2 *DATA MINING*

Data mining consiste na extração de informações implícitas, anteriormente desconhecidas, e potencialmente úteis a partir de um conjunto de dados, utilizando softwares que examinam minuciosamente bases de dados em busca de regularidades e padrões (WITTEN; FRANK; HALL, 2011, tradução nossa). É um campo multidisciplinar e abrange conceitos e conhecimentos de áreas similares como sistemas de bases de dados, estatística, reconhecimento de padrões e aprendizado de máquina (ZAKI; MEIRA JÚNIOR, 2013, tradução nossa).

De acordo com Witten, Frank e Hall (2011, tradução nossa), os conceitos de Aprendizado de Máquina (AM) são amplamente usados em *data mining*. AM é uma área da Inteligência Artificial (IA) que visa o desenvolvimento de técnicas computacionais sobre aprendizado, assim como a construção de sistemas que sejam capazes de adquirir conhecimento de forma automática (MONARD; BARANAUSKAS, 2005).

Os principais tipos de aprendizado, de acordo com Russel e Norvig (2013), são:

a) **aprendizado não-supervisionado:** o algoritmo aprende padrões a partir dos valores de entrada, mesmo que nenhuma observação seja fornecida. São característicos desse tipo de aprendizado os algoritmos de *cluster*, ou agrupamento;
b) **aprendizado por reforço:** o algoritmo adquire conhecimento a partir de uma série de reforços – sejam recompensas ou punições;
c) **aprendizado supervisionado:** o algoritmo observa alguns exemplos de pares de entrada e saída e aprende uma função que mapeia a partir da entrada para a saída. São característicos desse tipo de aprendizado os algoritmos de classificação.

Segundo Monard e Baranauskas (2005), um sistema de aprendizado pode ser caracterizado por um algoritmo que toma decisões baseando-se em experiências acumuladas por meio da solução bem sucedida de problemas prévios. Witten, Frank e Hall (2011, tradução nossa) dizem que *data mining* trata-se de obter conhecimento e descobrir padrões a partir da análise automática – ou semiautomática – de dados já armazenados, e é dessa forma que as duas disciplinas estabelecem uma conexão.

Ainda de acordo com Witten, Frank e Hall (2011, tradução nossa), *data mining* é um campo que envolve aprendizado em um sentido prático e não-teórico. O maior interesse está em técnicas para encontrar e descrever padrões estruturais nos dados, como uma ferramenta para ajudar a dar sentido aos dados e realizar predições a partir destes.

Para Hand, Manilla e Smyth (2001, tradução nossa), torna-se conveniente dividir *data mining* em diferentes tipos de tarefas, correspondendo aos diferentes objetivos daqueles que estão analisando os dados. De acordo com Tan, Steinbach e Kumar (2009), as tarefas de data mining podem ser classificadas como:

a) **tarefas preditivas:** usam variáveis existentes para prever valores futuros de outras variáveis (GORUNESCU, 2011, tradução nossa);
b) **tarefas descritivas:** agrupam dados medindo a similaridade entre os objetos e descobrem padrões ou relações desconhecidos entre os dados, de forma que os usuários possam compreender uma grande quantidade dados (YOO et. al., 2011, tradução nossa).

As tarefas de *data mining* são:

a) **estimativa:** semelhante à classificação, exceto que a variável alvo é numérica ao invés de categórica (LAROSE, 2005, tradução nossa);
b) **classificação:** consiste na atribuição de um objeto recém-apresentado a um grupo de classes predefinidas (BERRY; LINOFF, 2011, tradução nossa);
c) **agrupamento (*clustering*):** é o processo de divisão de objetos (ou observações) de um conjunto de dados em subconjuntos (HAN; KAMBER; PEI, 2012, tradução nossa);
d) **associação:** usada para descobrir padrões que descrevem características associadas nos dados (TAN; STEINBACH; KUMAR, 2009, tradução nossa).

2.3 A TAREFA DE CLASSIFICAÇÃO

Uma das mais comuns tarefas de *data mining*, classificação parece ser um imperativo humano. Para compreendermos e nos comunicarmos sobre o mundo, nós estamos constantemente classificando, categorizando e nivelando. Nós dividimos os seres vivos em sub-reinos, espécies e gêneros; classificamos matéria por elementos, cães por raças, pessoas por etnias (BERRY; LINOFF, 2011, tradução nossa).

Em *data mining*, a classificação consiste em classificar dados em etiquetas categóricas pré-definidas. A classe é o atributo ou característica em que os usuários estão mais interessados em um conjunto de dados (YOO et al., 2011, tradução nossa). Tan, Steinbach e Kumar (2009), definem a tarefa de classificação como uma tarefa de aprendizado, onde uma função f mapeia cada conjunto de atributos *x* e os atribui a um rótulo de classe pré-definido *y* (figura 2).

Figura 2 – Atribuição de um conjunto de dados de entrada x a um rótulo de classe y

Fonte: Adaptado de Tan, Steinbach e Kumar (2009)

A maioria dos algoritmos de classificação possuem duas etapas. Na primeira etapa, um modelo de classificação é construído a partir de dados de treinamento. Na segunda etapa, o modelo obtido anteriormente é usado para classificar dados ainda não analisados (AGGARWAL, 2015, tradução nossa).

Durante a primeira etapa, também conhecida como fase de aprendizado, um modelo de classificação é construído por meio da análise de um conjunto de treinamento, constituído por tuplas de dados e seus rótulos de classe associados. Pode-se representar as tuplas como um vetor *n*-dimensional de atributos $X = (x_1, x_2, \dots, x_n)$, onde assume-se que cada tupla pertença a uma classe pré-definida, como denominado por seu atributo de rótulo de classe. Por ser necessário fornecer os rótulos, esta etapa também é conhecida como aprendizado supervisionado, contrastando com o aprendizado não supervisionado, onde os rótulos de cada tupla não são conhecidos e o número ou conjunto de classes a serem aprendidas podem não ser previamente conhecidos (HAN; KAMBER; PEI, 2012, tradução nossa).

Uma vez identificada a função de classificação, verifica-se a sua acurácia por meio do conjunto de testes, comparando os resultados esperados com aqueles que foram observados para validar, ou não, o modelo (GORUNESCU, 2011, tradução nossa).

De acordo com Zaki e Meira Júnior (2014, tradução nossa), vários tipos de métodos de classificação já foram propostos. Dentre eles pode-se destacar:

a) **árvores de decisão:** estrutura de árvore semelhante a um fluxograma em que cada nó interno denota um teste em um atributo, cada galho representa um resultado do teste, e cada folha possui um rótulo de classe (HAN; KAMBER; PEI, 2012, tradução nossa);

b) **classificadores bayesianos:** utilizam o teorema de Bayes para predizer a classe como aquela que maximiza a probabilidade *a posteriori*. O objetivo principal é estimar a função de densidade de probabilidade conjunta para cada classe, a qual é modelada via distribuição normal multivariada (ZAKI; MEIRA JÚNIOR, 2013, tradução nossa);
c) **redes neurais artificiais:** modelo que busca simular o sistema nervoso humano, onde os neurônios são representados por unidades de computação. A função de um neurônio é definida pelos pesos de conexões de entrada do mesmo. Se os pesos forem configurados apropriadamente, a função poderá ser aprendida (AGGARWAL, 2015, tradução nossa).

3 ALGORITMOS BAYESIANOS

Dado um conjunto de objetos, onde cada um destes pertence a uma classe conhecida e possui um vetor de variáveis também conhecido, o objetivo é construir uma regra que permitirá atribuir objetos futuros a uma classe, dados apenas os vetores de variáveis para descrevê-los. Problemas desse tipo, chamados de problemas de classificação supervisionada, são universais e muitos métodos para a construção dessas regras já foram desenvolvidos (HAND, 2009, tradução nossa).

Os classificadores bayesianos são classificadores estatísticos baseados no Teorema de Bayes. Eles podem predizer a probabilidade de associações de classes, assim como a probabilidade de uma determinada tupla pertencer a uma classe em particular (HAN; KAMBER; PEI, 2012, tradução nossa).

De acordo com Barbetta, Reis e Bornia (2010), o Teorema de Bayes pode ser definido da seguinte forma:

$$P(E_i|F) = \frac{P(E_i)\ .\ P(F|E_i)}{P(F)} \quad (1)$$

Onde, segundo Han, Kamber e Pei (2012, tradução nossa), $\mathrm{P(E_i|F)}$ é a probabilidade posterior, ou *a posteriori*, de $\mathrm{E_i}$ condicionado em *F*; $\mathrm{P(E_i)}$ é a probabilidade anterior, ou *a priori*, de $\mathrm{E_i}$; $\mathrm{P(F|E_i)}$ é a probabilidade posterior de *F* condicionado em $\mathrm{E_i}$; e $\mathrm{P(F)}$ é a probabilidade isolada de *F* ocorrer.

Alguns exemplos de classificadores bayesianos são:

a) ***Naïve Bayes*:** este classificador assume que os valores dos atributos são condicionalmente independentes uns dos outros (HAN; KAMBER; PEI, 2012, tradução nossa);
b) **regressão logística:** assume-se que a variável alvo é extraída de uma distribuição de Bernoulli cuja média é definida por uma função *logit* parametrizada nas variáveis de características (AGGARWAL, 2015, tradução nossa);
c) **redes de crenças bayesianas:** semelhante ao Naïve Bayes, no entanto, este classificador permite ao usuário decidir se há ou não dependências entre os atributos (HAN; KAMBER; PEI, 2012, tradução nossa);

3.1 ALGORITMOS *NAÏVE BAYES*

O algoritmo de classificação *Naïve Bayes*, também conhecido como classificador Bayes simples, avalia a probabilidade condicional de um objeto pertencer a uma determinada classe supondo que seus atributos sejam condicionalmente independentes (TAN; STEIBACH; KUMAR, 2009).

Refere-se a este algoritmo como “ingênuo” devido à suposição de independência condicional entre os objetos e as classes. Essa suposição claramente não é real em prática, pois as características em conjuntos reais de dados estão quase sempre correlacionadas até mesmo estando condicionadas a uma classe específica. Ainda assim, apesar desta aproximação, o classificador Naïve Bayes pode possuir um bom desempenho em vários domínios (AGGARWAL, 2015, tradução nossa).

Segundo Tan, Steinbach e Kumar (2009), algumas características do algoritmo *Naïve Bayes* são:

a) é um algoritmo robusto para pontos de ruídos isolados, pois calculam a média destes pontos ao avaliar probabilidades condicionais a partir de dados. Também são capazes de lidar com valores faltantes, ignorando o exemplo durante a construção e classificação do modelo;
b) são robustos para atributos irrelevantes;
c) atributos correlacionados podem prejudicar o seu desempenho, uma vez que a suposição de independência condicional deixa de ser verdadeira para tais atributos.

3.1.1 Funcionamento do algoritmo *Naïve Bayes*

Inicia-se com um conjunto *D* de tuplas de treinamento e seus rótulos de classe associados, onde cada tupla é representada por um vetor $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n})$. Cada um desses vetores representa *n* medidas feitas na tupla a partir de *n* atributos, respectivamente, $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_\mathrm{n}$. Suponha que há *m* classes, $\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{m}$. Dada uma tupla *X*, o classificador irá prever que *X* pertence à classe com a maior probabilidade *a posteriori* condicionada em *X*. Isto é, o classificador *Naïve Bayes* prevê que a tupla *X* pertence à classe C_i se e somente se:

$$P(C_i|X) > P(C_j|X)\ para\ 1 \leq \mathrm{j} \leq \mathrm{m}, \mathrm{j} \neq \mathrm{i}$$

2)

Deste modo, $\mathrm{P}(\mathrm{C_i}|\mathrm{X})$ é maximizado. A classe $\mathrm{C_i}$ para a qual $\mathrm{P}(\mathrm{C_i}|\mathrm{X})$ é maximizada é chamada de hipótese máxima *a posteriori* (HAN; KAMBER; PEI, 2012, tradução nossa). De acordo com o Teorema de Bayes:

$$P(C_i|X) = \frac{P(C_i)\ .\ P(X|C_i)}{P(X)} \quad (3)$$

Como $\mathrm{P}(\mathrm{X})$ é fixo para todas as classes, é preciso apenas escolher a classe que irá maximizar o termo numerador. Esta abordagem é mais prática, pois não demanda um conjunto muito extenso de treinamento para obter uma boa estimativa da probabilidade (TAN; STEINBACH; KUMAR, 2009). Se as probabilidades *a priori* da classe não forem conhecidas, então comumente se assume que as classes são igualmente prováveis, ou seja, $\mathrm{P}(\mathrm{C_1}) = \mathrm{P}(\mathrm{C_2}) = \ldots = \mathrm{P}(\mathrm{C_m})$, e, portanto, maximizaríamos $\mathrm{P}(\mathrm{X}|\mathrm{C_i})$ (HAN; KAMBER; PEI, 2012, tradução nossa).

Tratando-se de conjuntos de dados com muitos atributos, calcular $\underline{\mathrm{P}(\mathrm{X}|\mathrm{C_i})}$ se tornaria altamente custoso em um ponto de vista computacional. Para reduzir a computação na avaliação de $\mathrm{P}(\mathrm{X}|\mathrm{C_i})$, a suposição ingênua de independência condicional entre as classes é feita. Esta assume que os valores dos atributos são condicionalmente independentes uns dos outros, dado o rótulo de classe da tupla (HAN; KAMBER; PEI, 2012, tradução nossa). De acordo com Han, Kamber e Pei (2012), Hand (2009), e Tan, Steinbach e Kumar (2009), a equação que define a suposição de independência condicional é:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (4)$$

Ou seja:

$$P(X|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times (x_n|C_i) \quad (5)$$

Onde $\mathrm{x_k}$ se refere ao valor do atributo de $\mathrm{A_k}$ da tupla *X*. Para cada atributo, observa-se se o atributo é categórico ou de valor contínuo. Se $\mathrm{A_k}$ for categórico, então $\mathrm{P}(\mathrm{x_k}|\mathrm{C_i})$ é o número de tuplas da classe $\mathrm{C_i}$ em *D* que possui o

valor x_k para A_k, dividido por $|C_{i,D}|$, o número de tuplas da classe C_i em *D* (HAN; KAMBER; PEI, 2012, tradução nossa).

No entanto, se os atributos possuírem valor contínuo, há duas formas de calcular a probabilidade condicional das classes. Uma delas é particionar cada atributo contínuo e então substituir estes valores por seus intervalos discretos correspondentes, assim transformando os atributos contínuos em atributos ordinais. A probabilidade condicional de $P(x_k|C_i)$ é analisada pelo cálculo da fração de registros de treinamento pertencentes à classe C_i que estão inclusos no intervalo que corresponde a x_k. O erro de análise depende da estratégia de particionamento empregada e do número de intervalos discretos. Se o número de intervalos for grande, haverá poucos registros de treinamento em cada intervalo para que se obtenha uma avaliação confiável. Em contrapartida, se houver um número pequeno de intervalos, alguns intervalos podem associar registros de diferentes classes, desta forma perdendo os limites de decisão corretos (TAN; STEINBACH; KUMAR, 2009).

Outra forma de calcular a probabilidade condicional nesse caso é por meio de uma distribuição Gaussiana com uma média μ e desvio padrão (DP) σ, definida da seguinte forma:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (6)$$

De modo que:

$$P(x_k|C_i) = g(x_k, \mu_{Ci}, \sigma_{Ci}) \quad (7)$$

Onde μ_{Ci} é a média e σ_{Ci} é o DP dos valores do atributo A_k para as tuplas de treinamento da classe C_i. Em seguida, ligam-se essas duas quantidades à equação (6), junto com x_k, para estimar $P(x_k|C_i)$. Para prever o rótulo de classe de *X*, $P(C_i)P(X|C_i)$ é avaliado para cada classe C_i. O classificador então prevê que o rótulo de classe da tupla *X* é a classe C_i se e somente se:

$$P(C_i)P(X|C_i) > P(C_j)P(X|C_j) \; para \; 1 \leq j \leq m, j \neq i \quad (8)$$

Em outras palavras, o rótulo de classe previsto é a classe C_i para o qual $P(C_i)P(X|C_i)$ é o máximo (HAN; KAMBER; PEI, 2012, tradução nossa).

De acordo com Hand (2009, tradução nossa), este método é importante pelas seguintes razões: construí-lo é uma tarefa fácil, sem necessidade de qualquer esquema interativo de estimação de parâmetros. Isto significa que o algoritmo pode ser facilmente aplicado em grandes conjuntos de dados. Sua interpretação é fácil, deste modo usuários sem experiência em tecnologias de classificadores podem entender a classificação gerada. Este classificador possui um desempenho surpreendentemente bom: embora seja um classificador simples, pode-se confiar que será robusto e que seu desempenho será satisfatório.

O classificador *Naïve Bayes* se torna o mais preciso em comparação a outros classificadores quando sua suposição de independência condicional entre as classes for verdadeira. Contudo, na prática, podem existir dependências entre as variáveis (HAN; KAMBER; PEI, 2012, tradução nossa).

3.2 REDE DE CRENÇA BAYESIANA

Uma rede de crença Bayesiana, do inglês *Bayesian Belief Network* (BBN), é uma abordagem mais flexível, que não requer que todos os atributos sejam condicionalmente independentes dada uma classe. Uma BBN permite especificar quais pares de atributos são condicionalmente independentes e fornece uma representação gráfica dos relacionamentos probabilísticos entre um conjunto aleatório de variáveis (TAN; STEINBACH; KUMAR, 2009).

Uma BBN é formada por um Grafo Acíclico Direcionado (GAD), cujos nós representam variáveis aleatórias associadas com medidas de incerteza e os arcos refletem a existência de uma influência causal direta entre as variáveis ligadas, e a força destas influências é mensurada por probabilidades condicionais (HRUSCHKA, 2003).

Algumas características de uma BBN ressaltadas por Tan, Steinbach e Kumar (2009), são:

a) a captura de conhecimento anterior de um determinado domínio apresentado em um modelo gráfico. Também é possível usar a rede para codificar dependências causais entre variáveis;

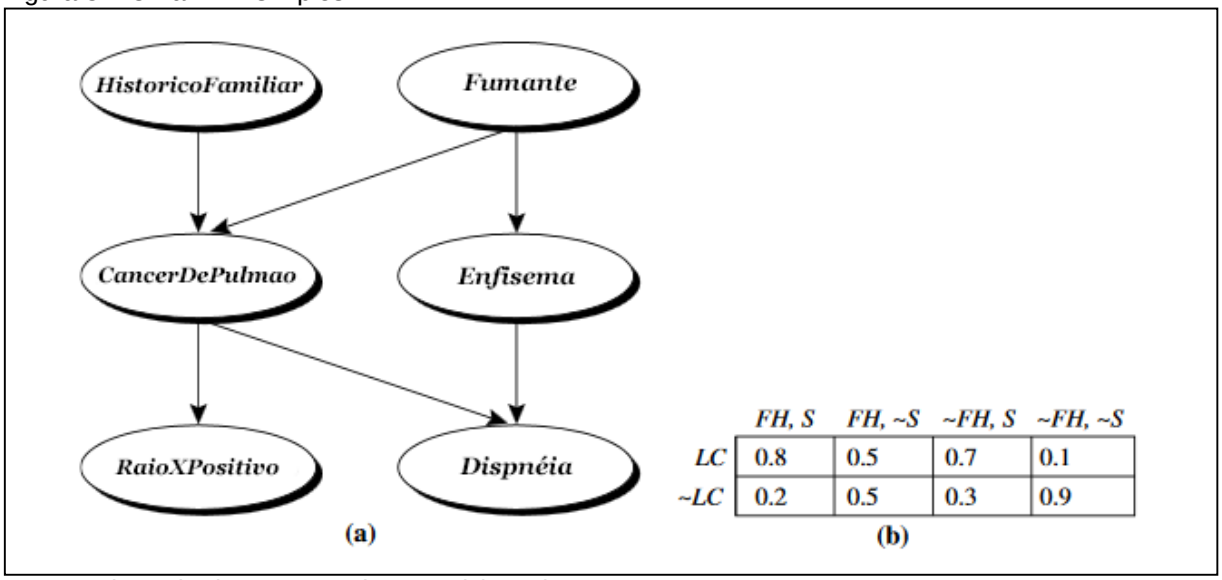
b) embora a construção da rede possa ser custosa em termos de tempo e esforço, adicionar uma nova variável à rede pronta é bastante direto;
c) é um método muito apropriado para lidar com dados incompletos. Instâncias que possuem atributos faltantes podem ser manipuladas através de soma ou integração de probabilidades por todos os valores possíveis do atributo;
d) esse método é bastante robusto para lidar com *overfitting*, já que os dados são combinados probabilisticamente com dados anteriores.

3.2.1 Grafo Acíclico Direcionado em uma Rede de Crenças Bayesiana

Cada nó no GAD representa uma variável aleatória. As variáveis podem possuir valores discretos ou contínuos. Elas podem corresponder a atributos reais presentes nos dados ou a “variáveis escondidas”, acreditando-se que formem um relacionamento (HAN; KAMBER; PEI, 2012, tradução nossa).

Cada arco representa uma variável e as arestas declaram o relacionamento de dependência entre os pares de variáveis. Se uma aresta estiver direcionada de *X* para *Y*, então *X* é o pai de *Y* e *Y* é o filho de *X*. Em adição, se existir um caminho direcionado na rede de *X* para *Z*, então *X* é ancestral de *Z*, enquanto *Z* é descendente de *X* (TAN; STEINBACH; KUMAR, 2009). Este processo é descrito pela figura 3, onde (a) mostra um modelo causal proposto representado por um GAD e (b) mostra a tabela de probabilidades condicionais para os valores da variável *CancerDePulmao* (LC), mostrando cada combinação possível dos valores de seus nós pais, *HistoricoFamiliar* (FH) e *Fumante* (S).

Figura 3 – Uma BBN simples.



Fonte: Adaptado de Han, Kamber e Pei (2012).

Cada nó é condicionalmente independente de seus avós, bisavós e qualquer outro conjunto de não descendentes, se seus pais forem conhecidos (WITTEN; FRANK; HALL, 2011, tradução nossa).

Cada variável da rede possui uma Tabela de Probabilidade Condicional, do inglês *Conditional Probability Tabl*e (CPT). A CPT da variável Y especifica a distribuição $P\left(Y\middle|Pais(Y)\right)$, onde $Pais(Y)$ são os pais de Y (HAN; KAMBER; PEI, 2012, tradução nossa). Para cada nó da rede, procura-se a probabilidade do valor do atributo do nó baseado na linha determinada pelos valores de atributos de seus pais. A multiplicação de todos estes valores juntos resultam na distribuição de probabilidade conjunta existente, que pode ser decomposta na seguinte equação (WITTEN; FRANK; HALL, 2011, tradução nossa):

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} P(x_i|x_{i-1}, \dots, x_1) \tag{9}$$

Que, segundo Han, Kamber e Pei (2012, tradução nossa), pode ser resumida da seguinte forma:

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P\left(x_i\middle|Pais(Y_i)\right) \tag{10}$$

Onde $(x_1, \dots, x_n)$ é uma tupla de dados X descrita pelos valores $Y_1, \dots, Y_n$, respectivamente, $P(x_1, \dots, x_n)$ é a probabilidade de uma combinação particular de valores de *X*, e os valores de $P\left(x_i \middle| Pais(Y_i)\right)$ correspondem às entradas da CPT para *Y*.

3.2.1 Algoritmos de aprendizado para uma rede de crenças

Algoritmos utilizados no aprendizado de uma rede de crenças possuem dois componentes definidos: uma função para estimar uma determinada rede baseado em dados e um método para pesquisar através do espaço de possíveis redes (WITTEN; FRANK; HALL, 2011, tradução nossa).

Assumindo-se que a topologia da rede – a forma como nós e arcos são organizados – é conhecida e as variáveis são visíveis, o treinamento se torna um processo direto. Basta calcular as entradas das CPTs, de forma similar ao cálculo de probabilidades realizado pelo Naïve Bayes. Quando a topologia não é conhecida e não se dispõe de um especialista humano, há algoritmos que podem construir a topologia da rede a partir de dados de treinamento (HAN; KAMBER; PEI, 2012, tradução nossa).

Figura 4 – Pseudocódigo para a generalização da topologia de uma rede Bayesiana

1: $T = (X1, X2, \dots, Xd)$ denota uma ordem total das variáveis.
2: **para** *j* = 1 até d **faça**
3: $X_{T(j)}$ denota a variável de ordem *j* em *T*.
4: $\pi X_{T(j)} = \{X_{T(1)}, X_{T(2)}, \dots, X_{T(j-1),}\}$ denotam o conjunto de variáveis precedendo a $X_{T(j)}$.
5: Remova as variáveis de $\pi X_{T(j)}$ que não afetem X_j (usando conhecimento anterior).
6: Crie uma aresta entre $X_{T(j)}$ e as variáveis restantes em $\pi X_{T(j)}$.
7: **fim do para**

Fonte: Adaptado de Tan, Steinbach e Kumar (2009).

Um exemplo de algoritmo para o aprendizado de redes bayesianas é o algoritmo K2, de funcionamento simples e aprendizado rápido. Ele inicia com uma

ordenação de atributos (nós) já estabelecida, então processa cada nó sucessivamente e considera avidamente adicionar arestas de nós já processados ao nó atual. Quando não há mais melhorias, a atenção do algoritmo é voltada para o próximo nó. Como um mecanismo adicional para evitar *overfitting*, o número de pais de cada nó pode ser restrito com um número máximo pré-definido. Devido a apenas vértices de nós previamente processados serem considerados e por ter uma ordenação fixa, este processo não pode inserir ciclos. Todavia, o resultado depende da ordenação inicial, então é recomendado executar o algoritmo múltiplas vezes com diferentes ordenações (WITTEN; FRANK; HALL, 2011, tradução nossa).

Ainda de acordo com Witten, Frank e Hall (2011, tradução nossa), outros algoritmos para o aprendizado de redes bayesianas são: uma versão mais sofisticada, porém mais lenta, do K2, que não ordena nós, mas avidamente considera adicionar ou excluir vértices entre pais arbitrários de nós; e o algoritmo *tree-augmented Naïve Bayes*, cujo funcionamento é semelhante ao do *Naïve Bayes*, mas com a adição de vértices.

Redes de crenças têm sido usadas para modelar um número de problemas bem conhecidos. Um exemplo é a análise de ligação genética (por exemplo, o mapeamento de genes de um cromossomo). Por moldar o problema de ligação genética em termos de inferência em BBN e usar algoritmos em estado-da-arte, a escalabilidade desta análise tem avançado consideravelmente. Outras aplicações que se beneficiaram do uso de uma BBN incluem visão computacional, análise de documentos e textos, sistemas de suporte a decisão e análise sensitiva. A facilidade com a qual muitas aplicações podem ser reduzidas à inferência de uma BBN é vantajosa na medida em que restringe a necessidade de criar algoritmos especializados para cada uma dessas aplicações (TAN, STEINBACH, KUMAR; 2009, tradução nossa).

As redes de crenças podem reduzir significativamente o número de parâmetros exigido pelo uso completo da inferência Bayesiana e mostram como os dados de um domínio – ou até mesmo a falta de dados – são capazes de dividir e focar o raciocínio. Ademais, a estrutura modular de um domínio de problema, por muitas vezes, permite que o projetista do programa faça muitas suposições de independência que não são permitidas pelo modelo Bayesiano completo (LUGER, 2014).

3.3 MEDIDAS DE QUALIDADE EM *DATA MINING* PARA CLASSIFICADORES

Embora AM seja uma ferramenta eficiente para adquirir conhecimento de forma automática, deve-se observar que não há um único algoritmo que apresente o melhor desempenho para todas as situações. Portanto, é importante compreender a capacidade e as limitações dos diferentes algoritmos de AM através de alguma metodologia de avaliação que permita comparar algoritmos (MONARD; BARANAUSKAS, 2005).

Muitas vezes é útil realizar a comparação entre o desempenho de diferentes classificadores, com o objetivo de determinar qual funciona melhor em certo conjunto de dados (TAN; STEINBACH; KUMAR, 2009).

A qualidade de um classificador é medida pela sua taxa de erro, ou seja, a frequência com que este realiza classificações incorretas, avaliado com base no conjunto de testes. Os exemplos classificados incorretamente são denominados erros de treinamento e os classificados corretamente são chamados de precisão (SCHAPIRE; FREUND, 2012 tradução nossa).

O erro de treinamento é definido por Zaki e Meira Junior (2014, tradução nossa) da seguinte forma:

$$Taxa\ de\ Erro = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{11}$$

E a precisão é apresentada na seguinte equação:

$$Precisão = \frac{1}{n}\sum_{i=1}^{n} (y_i \neq \hat{y}_i) = 1 - Taxa\ de\ erro \tag{12}$$

A base para analisar o desempenho de um classificador é uma matriz de confusão, onde os resultados classificados corretamente estão na diagonal principal da matriz (verdadeiro positivo e verdadeiro negativo) e os incorretos se encontram nos pontos inversos (falso positivo e falso negativo) (BOROVICKA et. al., 2012).

Figura 5 – Matriz de confusão 2x2

	Classe Verdadeira	
Classe Predita	Positivo	Negativo
Positivo (c_1)	VP	FP
Negativo (c_2)	FN	VN

Fonte: Adaptado de Zaki e Meira Junior (2014).

A definição dos pontos da matriz pode ser dada como (HAN; KAMBER; PEI, 2012, tradução nossa):

a) **Verdadeiro Positivo (VP):** número de instâncias positivas corretamente classificadas;
b) **Verdadeiro Negativo (VN):** número de instâncias negativas corretamente classificadas;
c) **Falso Positivo (FP):** número de instâncias negativas classificadas como positivas;
d) **Falso Negativo (FN):** número de instâncias positivas classificadas como negativas;

A partir da matriz de confusão, pode-se obter as seguintes medidas de qualidade (SOKOLOVA; LAPALME, 2009, tradução nossa; OLSON; DELEN, 2008, tradução nossa):

a) **acurácia:** eficácia geral de um classificador;

$$Acurácia = \frac{VP + VN}{VP + FN + FP + VN} \tag{11}$$

b) **precisão:** concordância dos rótulos de classe com os rótulos positivos dados pelo classificador;

$$Precisão = \frac{VP}{VP + FP} \tag{12}$$

c) **sensibilidade:** eficácia de um classificador ao identificar rótulos positivos;

$$Sensibilidade = \frac{VP}{VP + FN} \quad (13)$$

d) **especificidade:** eficácia de um classificador ao identificar rótulos negativos;

$$Especificidade = \frac{VN}{FP + VN} \quad (14)$$

e) ***F*-score:** relação entre o número de rótulos positivos (representados por β) do conjunto de dados e os rótulos positivos obtidos pelo classificador;

$$F - score = \frac{(\beta^2 + 1).VP}{(\beta^2 + 1).VP + \beta^2.FN + FP} \quad (15)$$

f) ***Area Under Curve* (AUC):** habilidade do classificador de evitar classificações falsas.

$$AUC = \frac{1}{2}\left(\frac{VP}{VP + FN} + \frac{VN}{VN + FP}\right) \quad (16)$$

g) taxa de falsos negativos:

$$Taxa\ de\ FN = \frac{FN}{VP + FN} \quad (17)$$

h) **taxa de falsos positivos:**

$$Taxa\ de\ FP = \frac{FP}{FP + VN} \quad (18)$$

i) **estatística *kappa*:** a estatística *kappa* é uma medida usada para calcular a concordância entre o que foi predito e o que foi observado nas classificações de um conjunto de dados (WITTEN; FRANK; HALL, 2011, tradução nossa).

$$\frac{\left(\frac{correto}{total}\right) - \left(\frac{concordância}{total^2}\right)}{1 - \left(\frac{concordância}{total^2}\right)} \quad (19)$$

4 TRABALHOS CORRELATOS

Nesta seção são abordadas algumas pesquisas envolvendo os classificadores Naïve Bayes e BBN, junto de uma breve descrição sobre o que consiste as pesquisas e os resultados atingidos pelas mesmas. (INSERIR FONTES NO ÚLTIMO PARAGRÁFO)

4.1 UM ESTUDO COMPARATIVOS DE ALGORITMOS DE CLASSIFICAÇÃO PARA A PREVISÃO DE CHUVAS

O artigo *A Comparative Study of Classification Algrithms for Forecasting Rainfall* foi produzido por Deepti Gupta e Udayan Ghose, e publicado na *4th International Conference on Reliability, Infocom Technologies and Optmization*, do Instituto de Engenheiros Eletricistas e Eletrônicos (IEEE) no ano de 2015.

A pesquisa visa realizar uma comparação entre vários algoritmos de classificação, sendo eles: Árvore de Regressão de Classificação (algoritmo CART), *Naïve Bayes*, *K-nearest Neighbour* e Redes Neurais para o reconhecimento de padrões. Os algoritmos foram aplicados em uma base de dados com 2245 amostras de dados de chuvas de Nova Délhi no período de junho a setembro do ano de 1996 até o ano de 2014 (GUPTA, GHOSE; 2015, tradução nossa).

Entre os algoritmos analisados, as Redes Neurais obtiveram o melhor resultado, com 82,1% de precisão. O algoritmo *K-nearest Neighbour* obteve o segundo melhor resultado, com 80,7% de precisão. Os algoritmos CART e *Naïve Bayes* obtiveram 80,3% e 78,9% de precisão, respectivamente (GUPTA, GHOSE; 2015, tradução nossa).

4.2 UMA NOVA ABORDAGEM: USANDO REDES DE CRENÇA BAYESIANA NA RECOMENDAÇÃO DE PRODUTOS

O artigo *A Novel Approach: Using Bayesian Belief Networks in Product Recommendation* é de autoria de S. S. Thakur, Anirban Kundu e J. K. Sing, e foi publicado na *Second International Conference on Emerging Applications of Information Tecnology*, no ano de 2011.

A pesquisa propõe um sistema de recomendação de produtos baseado em uma BBN. Foram coletados dados de vendas de telefones móveis em um intervalo de um ano. Em seguida, uma BBN foi aplicada aos dados usando as probabilidades conhecidas de A e B, onde A representa as vendas em que o cliente pagou com cartão de crédito e B representa as vendas pagas com dinheiro ou cartão de débito. Várias combinações são testadas e somadas e então as novas probabilidades de A e B sendo verdadeiros são calculadas. Um mapa é traçado para os valores obtidos e a probabilidade inicializada é recalculada quando colocados pesos W. O mapa é redesenhado após a aplicação dos pesos e os passos anteriores são repetidos para fins de comparação (THAKUR; KUNDU; SING, 2011, tradução nossa).

Nos resultados foi observado que, se um desconto de 30% ou 50% é dado para produtos eletrônicos como smartphones, recomendar estes produtos tanto para clientes que utilizam dinheiro e cliente que utilizam cartão de crédito fará com que as vendas aumentem continuamente. Caso o desconto ofertado seja de 10%, as vendas podem aumentar ou permanecerem constantes para o mesmo conjunto de clientes. Observou-se também que o desconto não causa tanto impacto nos usuários de cartão de crédito (THAKUR; KUNDU; SING, 2011, tradução nossa).

4.3 DETECÇÃO DE NOTÍCIAS FALSAS UTILIZANDO O CLASSIFICADOR *NAÏVE BAYES*

O artigo *Fake News Detection Using Naïve Bayes Classifier* foi desenvolvido por Mykhailo Granik e Volodymyr Mesyura, e publicado na *First Ukraine Conference on Electrical and Computer Engineering*, da IEEE (Instituto de Engenheiros Eletricistas e Eletrônicos) no ano de 2017.

O trabalho visa mostrar uma aproximação simples para a detecção de noticias falsas usando o classificador Naïve Bayes. O conjunto de dados utilizado para a realização da pesquisa foi extraído a partir do site *BuzzFeed News*. O conjunto contêm posts do Facebook, coletadas de um total de seis páginas famosas do site, assim como notícias dos sites Politico, CNN e *ABC News* (GRANIK; MESYURA, 2017, tradução nossa).

A precisão de classificação para as notícias verdadeiras e as notícias falsas obteve resultados similares, mas a precisão para as notícias falsas foi

levemente pior. Este resultado pode ter sido causado pela obliquidade do conjunto de dados: apenas 4.9% dos dados eram notícias falsas (GRANIK; MESYURA, 2017, tradução nossa).

4.4 SISTEMA PREDITIVO PARA A DOENÇA DE ALZHEIMER NA TRIAGEM CLÍNICA

Este artigo foi desenvolvido por Leonard Barreto Moreira e Anderson Amendoeira Namen, e publicado na revista *Journal of Health Informatics* no ano de 2016.

O objetivo da pesquisa é descrever uma aplicação para auxiliar os especialistas no processo de diagnóstico de pacientes com suspeita clínica de Alzheimer através de técnicas de data mining. Foram realizadas classificações com os algoritmos Naïve Bayes, BBN e árvores de decisão, e seus resultados foram avaliados por meio de uma validação cruzada estratificada *k-fold* (MOREIRA; NAMEN, 2016).

Os resultados numéricos dos modelos foram avaliados de acordo com as métricas: acurácia/precisão, taxa de erro, sensibilidade, taxa de falsos positivos e taxa de falsos negativos, obtendo as seguintes taxas, respectivamente: 73,8%, 26,2%, 76,3%, 27,4%, 23,7%. Dentre os algoritmos utilizados, os classificadores bayesianos, em especial a BBN, apresentaram os melhores resultados para o diagnóstico de Alzheimer (MOREIRA; NAMEN, 2016).

4.5 O TEOREMA DE PROBABILIDADE PELO ALGORITMO *NAÏVE BAYES* PARA A TAREFA DE CLASSIFICAÇÃO NA *SHELL ORION DATA MINING ENGINE*

Esta pesquisa foi desenvolvida por Marcio Novaski e submetido como Trabalho de Conclusão de Curso para curso de Ciência da Computação da Universidade do Extremo Sul Catarinense (UNESC) e tem como objetivo a implementação do algoritmo Naïve Bayes para a tarefa de classificação na *Shell Orion Data Mining Engine*, uma ferramenta gratuita que visa auxiliar no processo de KDD desenvolvida na UNESC.

A implementação do algoritmo no módulo de classificação foi realizada na linguagem de programação Java no ambiente de desenvolvimento *NetBeans* 7.1.1. Para analisar o desempenho do algoritmo, foi usada uma base de dados contendo dados clínicos de pacientes com diagnóstico positivo para câncer de mama. Para a

avaliação de desempenho e qualidade do algoritmo, foram utilizados métodos estatísticos (NOVASKI, 2012).

O algoritmo apresentou bons índices de validação relacionados à precisão e confiabilidade. Observou-se também que o tempo de processamento gasto na fase de treinamento é significativamente menor do que o tempo gasto na fase de classificação. Em comparação com a ferramenta *Weka*, os resultados obtidos em relação ao tempo de processamento não foram excelentes, no entanto, os índices de validação utilizados na avaliação dos resultados da classificação foram ligeiramente melhores que os apresentados pela ferramenta *Weka* (NOVASKI, 2012).

5 OS MÉTODOS BAYESIANOS DE APRENDIZADO DE MÁQUINA PELOS ALGORITMOS NAÏVE BAYES E REDES DE CRENÇA PARA A PREDIÇÃO DE PERÍODO CHUVOSO NA CIDADE DE BLUMENAU

A pesquisa desenvolvida tem como objetivo a aplicação de dois algoritmos de classificação – *Naïve Bayes* e BBN – em uma base de dados e comparar seus modelos resultantes, a fim de determinar qual algoritmo possui melhores resultados.

A base de dados selecionada para a aplicação dos algoritmos consiste em dados pluviométricos. A pluviometria caracteriza-se pelo estudo do regime de precipitação pluviométrica de uma região, tendo esta importância no ciclo hidrológico e na manutenção da vida no planeta (SILVA *et al*, 2011).

5.1 METODOLOGIA

Para o desenvolvimento desta pesquisa, as seguintes etapas metodológicas foram empregadas: levantamento bibliográfico, seleção da base de dados, seleção da ferramenta de *data mining* a ser utilizada, aplicação dos algoritmos *Naïve Bayes* e rede de crenças na base de dados selecionada, análise dos modelos de predição por meio de medidas de qualidade em *data mining* e identificação do modelo final.

Durante o levantamento bibliográfico, fundamentaram-se os conceitos de KDD, *data mining*, a tarefa de classificação, os algoritmos *Naïve Bayes* e BBN, e medidas de qualidade para classificadores em *data mining*.

5.1.1 Seleção da base de dados

A base de dados selecionada para a realização dessa pesquisa foi extraída da Agência Nacional das Águas (ANA) e consiste em dados pluviométricos do bairro de Itoupava Central na cidade de Blumenau, em Santa Catarina.

Optou-se por usar essa base pela mesma apresentar uma série histórica de mais de 30 anos, apresentar boa consistência nos dados e possuir poucos dados

faltantes. A região norte de Blumenau, região em que está localizado o bairro Itoupava Central, constantemente sofre com inundações.

A base possui 738 instâncias, com a remoção de uma instância com valores vazios, e seis atributos contando com o atributo de classe, descritos na tabela 1.

Tabela 1 – Atributos da base de dados de pluviometria

Atributo	**Valor**	**Descrição**
Mês	Nominal (01 a 12)	-
Ano	Nominal (de 1941 a 2000, 2004, 2006)	-
Máxima	Numérico	Precipitação máxima do mês
Total	Numérico	Precipitação total do mês
NumDiasDeChuva	Nominal (1 a 26)	Número de dias de chuva do mês
MesChuvoso	Nominal (verdadeiro e falso)	Classe

Fonte: Do autor.

Os valores da classe foram atribuídos às instâncias de acordo com as seguintes regras:

a) se o número de dias de chuva no mês for menor que dez, atribui-se o valor *falso*;
b) se o número de dias de chuva for maior que dez e a média do mês possuir valor maior que 7mm, atribui-se o valor *verdadeiro*. Caso contrário, atribui-se o valor falso.

O cálculo da média pode ser definido pela seguinte equação:

$$Classe = \frac{Total - Máxima}{NumDeDiasDeChuva - 1} \quad (22)$$

A equação (22) tem como objetivo impedir que meses onde descontando-se a precipitação máxima da precipitação total do mês resulte em uma proporção muito baixa de chuva para os dias restantes, sejam classificados como Verdadeiro.

5.1.2 Pré-processamento dos dados

A fase de pré-processamento abrange técnicas para limpeza, organização e refinamento dos dados, a fim de se obter melhores resultados no processo de mineração.

Para o pré-processamento da base de dados utilizada nesta pesquisa, fez-se o uso de quatro ferramentas: Microsft Access 2016, MDBViewer, Microsoft Excel 2016 e *Waikato Environment for Knowledge Analysis* (WEKA) 3.8.

A base de dados estava originalmente no formato *mdb*, próprio da ferramenta Access. Nela, houve a remoção dos atributos *RegistroID*, *EstacaoCodigo* e *NivelConsistencia*, pois não contribuem para os resultados da classificação. Também uma instância com valores vazios foi removida. Utilizou-se a ferramenta *MDBViewer* para realizar a conversão do arquivo Access para Excel.

No Excel, foram atribuídos os rótulos de classe Verdadeiro e Falso através das regras de classificação descritas no capítulo anterior. Após a atribuição dos rótulos de classe, o arquivo Excel (*xlsx*) foi convertido para um arquivo com valores separados por vígula (*csv*).

No Weka, foi utilizada a função *ARFF Viewer* para realizar a conversão de *csv* para *arff*, formato recomendado para o WEKA. Alguns ajustes menores precisaram ser realizados, como a troca de vírgulas (,) por pontos (.) nos valores numéricos e a troca de ponto e vírgula (;) por vírgulas simples para a separação de valores.

5.1.3 Execução do *data mining*

O software utilizado para a aplicação dos algoritmos *Naïve Bayes* e BBN foi o WEKA em sua versão 3.8 para o sistema operacional Windows, o software pode ser encontrado gratuitamente para download no link < https://www.cs.waikato.ac.nz/ml/weka/downloading.html>.

Optou-se por utilizar a ferramenta WEKA não apenas por ser gratuita, mas por disponibilizar os algoritmos necessários para esta pesquisa e por possuir uma interface simples e de fácil uso.

O software WEKA é uma coleção de algoritmos de aprendizado de máquina voltados para tarefas de *data mining*. Foi desenvolvido na Universidade de Waikato, na Nova Zelândia, utilizando a linguagem Java.

O WEKA apresenta uma miríade de ferramentas para preparação, visualização, classificação, agrupamento e associação. É um software *opensource*, licenciado sobre a *GNU General Public License*, o que permite que vários usuários possam estudar seu código fonte e realizar alterações no mesmo, contribuindo e aprimorando a ferramenta.

O software possui duas linhas de desenvolvimento, que consistem nas versões estáveis e para desenvolvimento. A versão estável mais recente é a 3.8, utilizada neste trabalho, e recebe somente correções de *bugs*, embora novos recursos podem ser disponibilizados em pacotes. A versão de desenvolvimento continua a partir da versão estável 3.8, recebendo correções de *bugs* e atualizações com novos recursos. Ambas as versões estão disponíveis para os sistemas operacionais Windows, Mac OS X, Linux e similares.

Após realizar o estudo do funcionamento da ferramenta e dos algoritmos utilizados nesta pesquisa, pode-se seguir para a aplicação dos mesmos na base de dados.

5.1.3.2 Aplicação do algoritmo *Naïve Bayes*

Primeiramente aplicou-se o algoritmo *Naïve Bayes*, que avalia a probabilidade condicional de um objeto pertencer à uma certa classe, supondo que não há dependência condicional entre os atributos deste objeto. Esta característica é o que lhe confere a alcunha de “ingênuo”.

Embora esta suposição não seja real em prática, o classificador Naïve Bayes pode possuir um bom desempenho em vários domínios, sendo considerado um algoritmo robusto, capaz de lidar com atributos irrelevantes, pontos de ruído isolados e valores faltantes (HAN; KAMBER; PEI, 2012, tradução nossa).

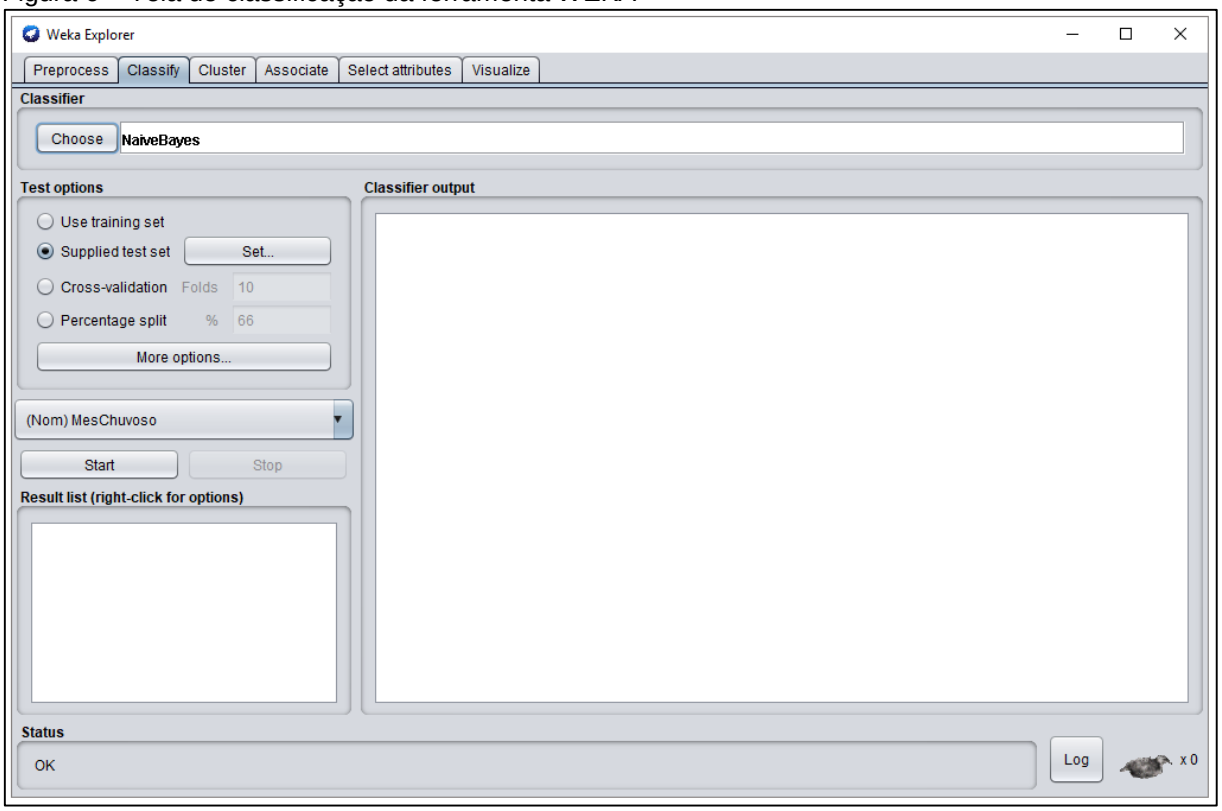
Iniciou-se a aplicação do algoritmo Naïve Bayes no conjunto de dados, onde, a partir deste, foram obtidos o modelo de classificação e sua respectiva matriz de confusão.

Utilizou-se a opção “*Explorer*” da interface gráfica de seleção do WEKA, que leva o usuário à tela de pré-processamento dos dados, onde são selecionados o

conjunto de dados utilizados no treinamento e os atributos a serem considerados na classificação.

No menu de classificação, são dados os campos de Classificador, Opções de Teste, Lista de Resultados, Saída do Classificador e Status, como apresentado na figura 6.

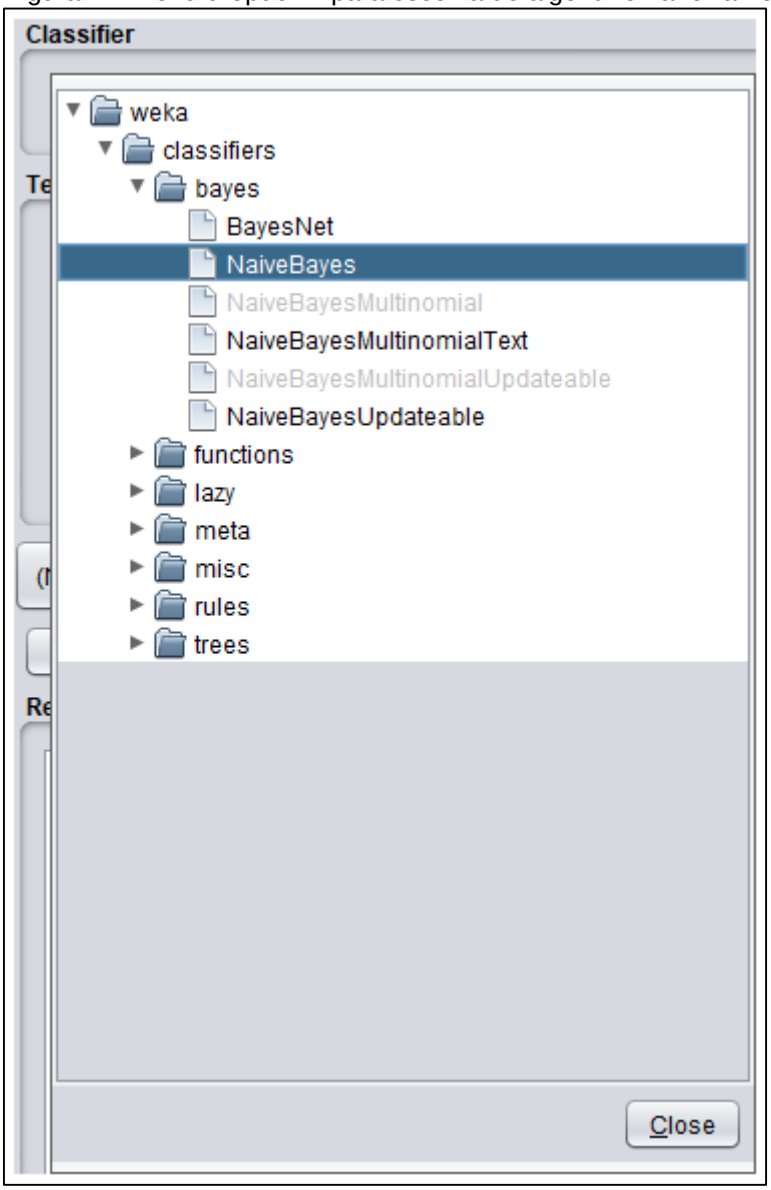
Figura 6 – Tela de classificação da ferramenta WEKA



Fonte: Do autor.

No campo Classificador, é possível escolher o algoritmo de classificação desejado. O botão *Choose* abre um menu *dropdown*, contendo todos os algoritmos de classificação implementados pelo WEKA, como pode ser observado na figura 7.

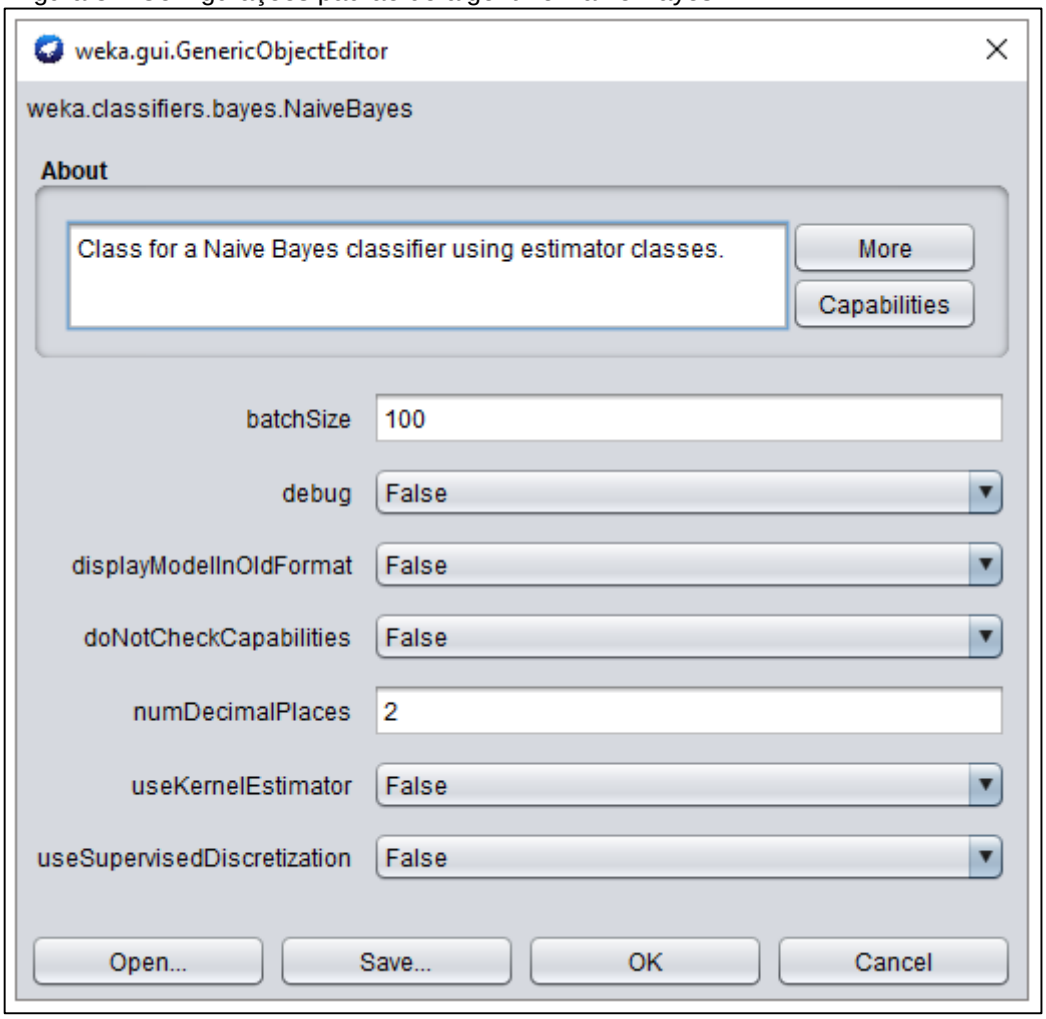
Figura 7 – Menu dropdown para escolha de algoritmo na ferramenta WEKA



Fonte: Do autor.

Após a seleção do algoritmo desejado, neste caso, o algoritmo *Naïve Bayes*, é possível configurá-lo da forma desejada pelo usuário. Ao clicar no campo do classificador escolhido, um menu de configurações será aberto (figura 8). Optou-se por utilizar as configurações padrão do algoritmo *Naïve Bayes* fornecidas pelo WEKA.

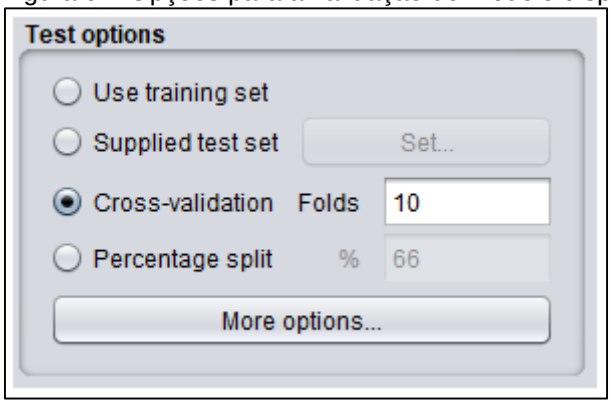
Figura 8 – Configurações padrão do algoritmo Naïve Bayes



Fonte: Do autor.

Em seguida, pode-se escolher o método de validação empregado para geração do modelo (figura 9).

Figura 9 – Opções para a validação do modelo disponibilizadas pelo WEKA



Fonte: Do autor.

Para a realização deste trabalho, foi escolhida a opção “*Cross-validation*”, no português “Validação Cruzada”, por se fazer uso do o método de validação *cross-validation*, com 10 folds. Este método divide o conjunto de treinamento em um número pré-estabelecido de partições (*folds*) onde cada partição é usada para teste uma vez e o restante para treinamento, sendo 10 *folds* considerado o número ideal de partições para se obter a melhor taxa de erro (WITTEN; FRANK; HALL, 2011, tradução nossa).

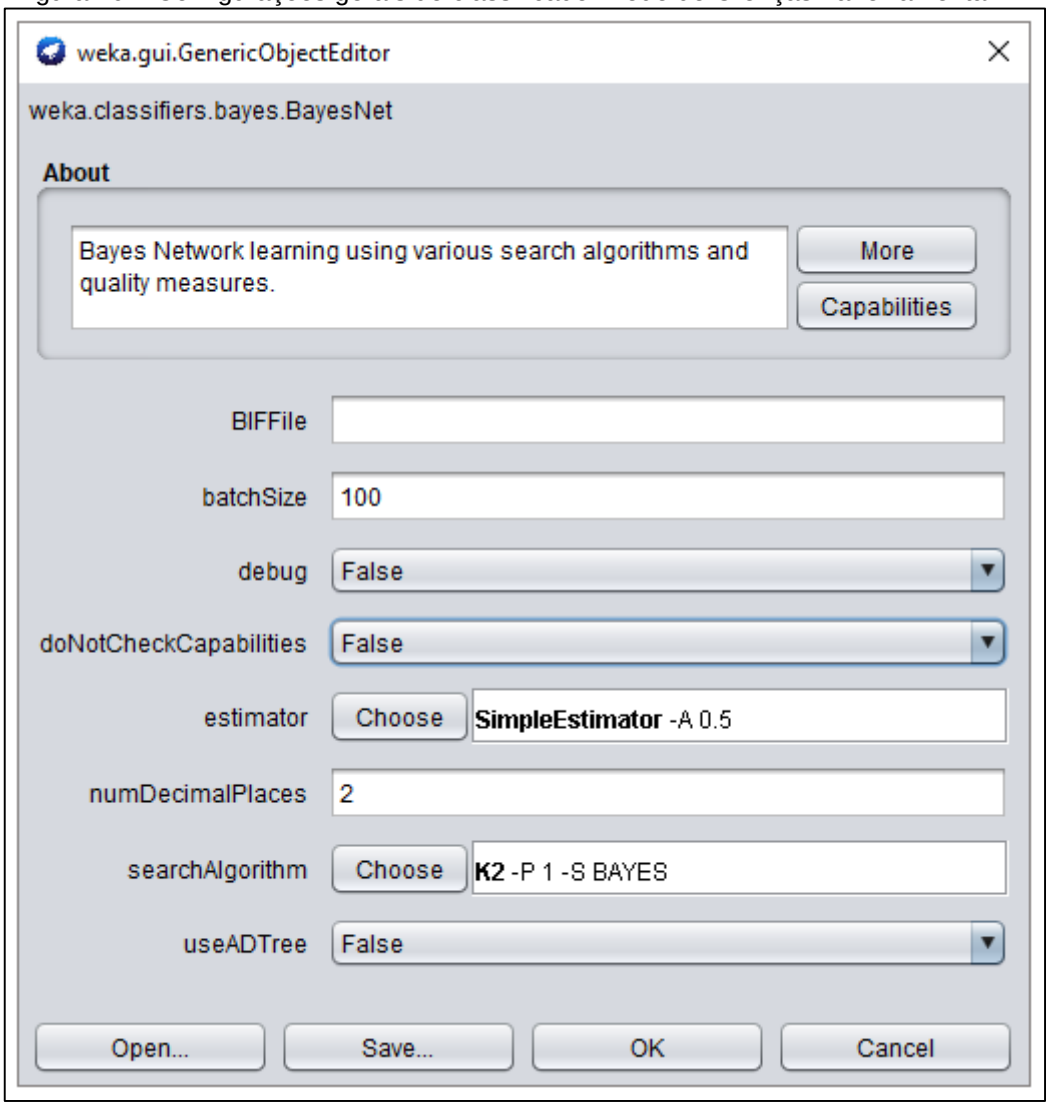
Após configurar o algoritmo de acordo com as necessidades do usuário, pode-se então executar a classificação. Os resultados são exibidos na tela de *output* do classificador, onde apresenta-se uma breve descrição dos atributos da base de dados, o modelo de classificação obtido, os resultados de predição no conjunto de testes fornecido pelo usuário antes de realizar a classificação, a matriz de confusão e os resultados das medidas de qualidade escolhidas pelo usuário.

5.1.3.3 Aplicação do algoritmo rede de crenças

O algoritmo de rede de crenças é semelhante ao algoritmo *Naïve Bayes*, no entanto, este possui uma abordagem mais flexível, não requerendo que os atributos sejam condicionalmente independentes. Ele permite especificar quais atributos são condicionalmente dependentes e apresenta as relações probabilísticas em forma de um GAD, onde cada nó e folha possui uma CPT.

A aplicação do algoritmo de rede de crenças foi similar à aplicação do algoritmo *Naïve Bayes*, seguindo-se os mesmos passos descritos na seção anterior, apenas selecionando a opção *BayesNet* no lugar de *NaiveBayes*. Optou-se por utilizar as configurações padrão do classificador (figura 10).

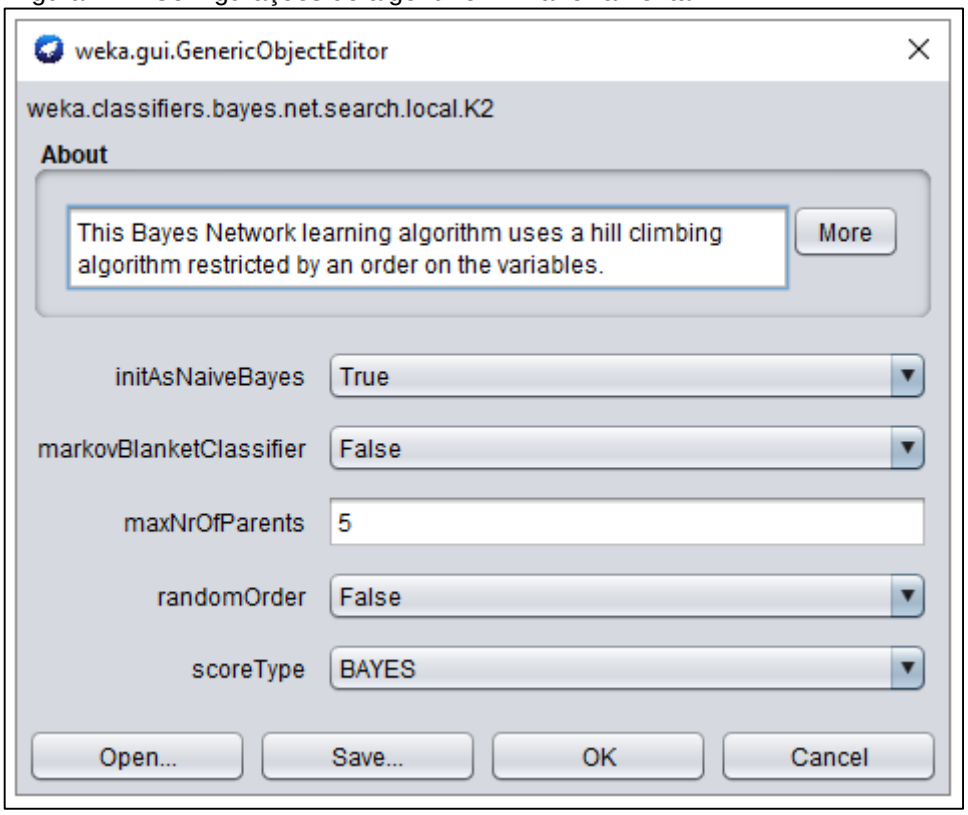
Figura 10 – Configurações gerais do classificador Rede de Crenças na ferramenta WEKA



Fonte: Do autor.

O algoritmo escolhido para a construção da rede foi o algoritmo K2, cujo funcionamento é descrito no capítulo 3.2.1, também utilizando suas configurações padrão, com exceção do número de nós pais que foi configurado como cinco. As configurações podem ser visualizadas na figura 11.

Figura 11 – Configurações do algoritmo K2 na ferramenta WEKA



Fonte: Do autor.

Após a aplicação do algoritmo no conjunto treinamento, foi obtido o modelo de classificação, matriz de confusão e resultados das medidas de qualidade aplicadas.

5.1.4 Aplicação das medidas de qualidade

A partir da geração do modelo de classificação e das matrizes de confusão, torna-se possível aplicar medidas de qualidade em *data mining* para classificadores. As medidas aplicadas nesta pesquisa foram as seguintes: taxa de erro; acurácia; sensibilidade, especificidade; *F-score*; *Area Under Curve* (AUC); taxa de falsos negativos, taxa de falsos positivos e precisão da classe.

As medidas taxa de erro, acurácia, AUC e estatística *kappa* são usadas apenas uma vez em cada modelo, pois visam apenas os resultados gerais do classificador. Já as medidas proporção de falsos positivos, proporção de falsos negativos, sensibilidade, especificidade, precisão e *F-score* necessitam ser aplicadas em cada atributo de classe, pois variam conforme os mesmos. Todas as medidas foram descritas no capítulo 3.3.

5.2 RESULTADOS OBTIDOS

O modelo gerado a partir do Naïve Bayes e as tabelas de probabilidade condicional, do inglês, *Conditional Probability Tables* (CPT) geradas para a rede bayesiana obtida por meio do algoritmo K2 podem ser encontradas nos apêndices A e B, respectivamente.

As medidas de qualidade foram aplicadas na etapa anterior no próprio WEKA, sendo necessário calcular manualmente apenas as medidas AUC e especificidade.

5.2.1 Predição realizada pelo algoritmo Naïve Bayes

A validação do modelo obtido pelo algoritmo *Naïve Bayes* no conjunto de dados de teste obteve uma acurácia de 85,36%, com 630 instâncias classificadas corretamente e 108 classificadas incorretamente. A matriz de confusão resultante é apresentada na tabela 2:

Tabela 2 – Matriz de confusão obtida a partir da validação do modelo com o conjunto de teste

	Classe Verdadeira	
Classe Predita	**Verdadeiro (a)**	**Falso (b)**
Verdadeiro (a)	241	70
Falso (b)	38	389

Fonte: Do autor.

Onde, de um conjunto com 738 instâncias, 241 foram corretamente classificadas como verdadeiras (VP), 38 foram incorretamente classificadas como falsas (FN), 70 incorretamente classificadas como verdadeiras (FP) e 389 corretamente classificadas como falsas (VN).

Também foram aplicadas as medidas de qualidade à matriz de confusão obtida na fase de testes do algoritmo *Naïve Bayes*. Os resultados das medidas gerais podem ser observados na tabela 3:

Tabela 3 – Resumo das medidas de qualidade gerais do modelo de predição obtido pelo algoritmo *Naïve Bayes*

Métrica	**Resultado**
Taxa de erro	14,63%
Acurácia	85,36%
AUC	0,8555
Estatística *kappa*	0,6956

Fonte: Do autor.

A predição realizada a partir do modelo obtido pelo *Naïve Bayes* obteve uma acurácia de 85,36%, com uma taxa de erro de 14,63%. A AUC, ou seja, a capacidade do classificador de evitar classificações falsas, foi de 85,55% e a estatística *kappa* apresentou uma boa concordância entre os dados preditos, ainda que um tanto baixa.

Na tabela 4, são resumidos os resultados das medidas de qualidade aplicadas a cada classe a partir dos valores obtidos na matriz de confusão:

Tabela 4 – Resumo das medidas de qualidade para as classes verdadeiro e falso, obtidas na fase de teste do algoritmo Naïve Bayes

	Classe	
Medida de qualidade	**Verdadeiro**	**Falso**
Proporção de falsos positivos	0,089	0,225
Proporção de falsos negativos	0,225	0,089
Sensibilidade	0,864	0,847
Especificidade	0,847	0,864
Precisão	0,775	0,911
F-score	0,817	0,878

Fonte: Do autor.

A proporção de falsos positivos apresentou-se baixa para a classe Verdadeiro, onde somente 8,9% dos dados classificados como Verdadeiro possuem o rótulo Falso. No entanto, a proporção de falsos negativos foi maior, onde 22,5% dos dados classificados como Falso são, de fato, rotulados como Verdadeiro.

A sensibilidade e especificidade ficaram próximas de 1, denotando uma boa habilidade na identificação correta de objetos com o rótulo Verdadeiro e objetos com o rótulo Falso, respectivamente.

As medidas precisão e *F-Score* também obtiveram bons resultados, demonstrando que o classificador possui uma boa concordância entre a classe predita para cada objeto e a classe verdadeira do mesmo.

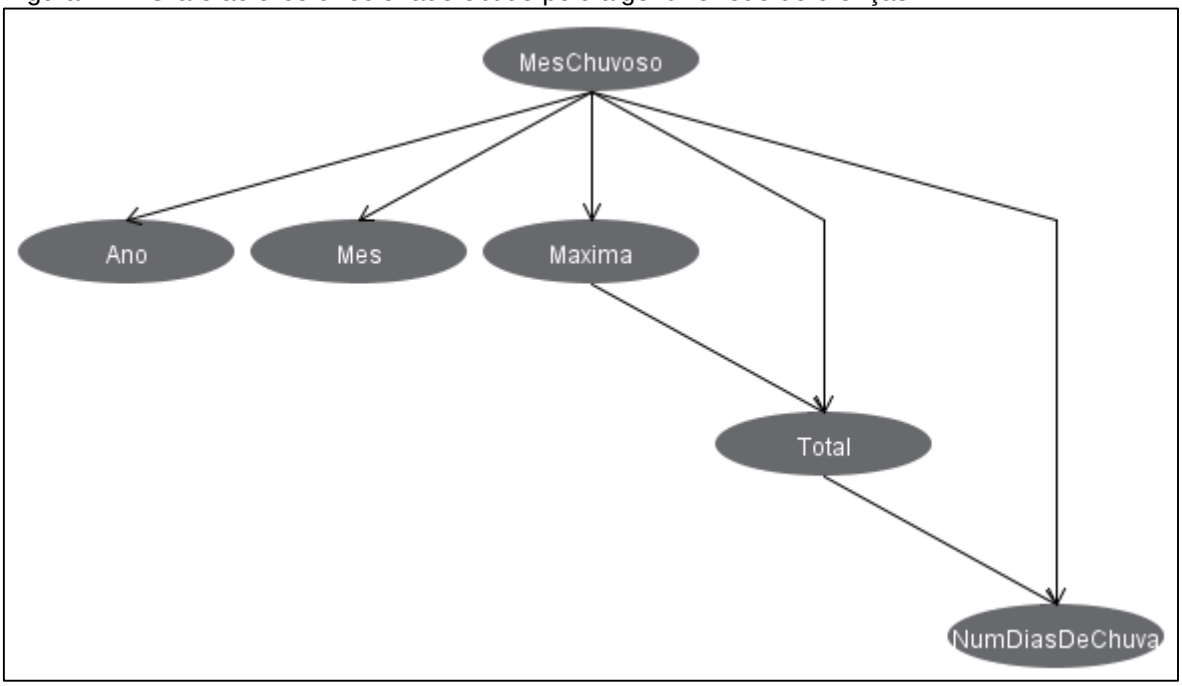
Em conformidade com o que foi observado nos resultados das medidas aplicadas na classe Verdadeiro, a proporção de falsos positivos apresentou-se baixa para a classe Falso, enquanto a proporção de falsos negativos apresentou-se maior, confirmando o que foi observado na classe Verdadeiro.

A sensibilidade e especificidade também estão de acordo com os resultados obtidos pela classe Verdadeiro, e as medidas de precisão e *F-Score* também demonstram uma boa concordância entre a classe predita com o rótulo de classe real do objeto.

5.2.2 Predição realizada pelo algoritmo de rede de crenças

O GAD obtido pela classificação realizada pelo algoritmo de rede de crenças é apresentado na figura 12. Pode-se observar a existência de dependências condicionais entre os atributos *Maxima*, *Total* e *NumDiasDeChuva*, assim como havia sido observado ao atribuir as classes ao conjunto de dados.

Figura 12 – Grafo acíclico direcionado obtido pelo algoritmo rede de crenças



Fonte: Do autor.

Ao que diz respeito à aplicação das medidas de qualidade nos resultados obtidos a partir da validação do modelo no conjunto de testes, a acurácia obtida foi de 89,57%, onde 661 instâncias foram classificadas corretamente e 77 foram classificadas incorretamente. A matriz de confusão gerada para a aplicação do algoritmo de rede de crenças no conjunto de teste é descrita na tabela 5:

Tabela 5 – Matriz de confusão da validação do algoritmo de rede de crenças

	Classe Verdadeira	
Classe Predita	**Verdadeiro (a)**	**Falso (b)**
Verdadeiro (a)	279	32
Falso (a)	45	382

Fonte: Do autor.

Onde 279 instâncias foram corretamente classificadas como verdadeiras, 32 foram incorretamente classificadas como falsas, 45 incorretamente classificadas como verdadeiras e 382 corretamente classificadas como falsas. As medidas de qualidade gerais obtidas para o classificador K2 são resumidas na tabela 6.

Tabela 6 – Resumo das medidas de qualidade gerais para a fase de teste do algoritmo de rede de crenças

Métrica	Resultado
Taxa de erro	10,43%
Acurácia	89,57%
AUC	0,892
Estatística *kappa*	0,7872

Fonte: Do autor.

O modelo gerado pelo algoritmo de rede de crenças apresentou uma taxa de erro de 10,43% – inferior à apresentada pelo modelo do algoritmo *Naïve Bayes*. A acurácia foi de 89,56%, e o classificador apresenta uma habilidade próxima de 90% ao evitar classificações falsas. A estatística *kappa*, com um resultado de 78,72%, indica uma boa concordância entre as classes preditas e as observadas no conjunto. A tabela 7 mostra um resumo dos resultados obtidos ao aplicar as medidas de qualidade nos rótulos de classe.

Tabela 7 - Resumo das medidas de qualidade para as classes verdadeiro e falso, obtidas na fase de teste do algoritmo de rede de crenças

	Classe	
Medida de qualidade	**Verdadeiro**	**Falso**
Proporção de falsos positivos	0,105	0,103
Proporção de falsos negativos	0,103	0,105
Sensibilidade	0,861	0,923
Especificidade	0,923	0,861
Precisão	0,897	0,895
F-score	0,879	0,908

Fonte: Do autor.

Na fase de teste, a classe Verdadeiro apresentou uma baixa proporção de falsos positivos e falsos negativos – cerca de 10% dos dados de cada classe foram classificados erroneamente.

A sensibilidade e especificidade confirmam a boa habilidade do modelo obtido em predizer corretamente as classes dos objetos, enquanto a precisão e *F-score* indicam uma concordância próxima de 100% entre as classes reais e as preditas pelo algoritmo a partir do modelo de classificação.

Os resultados da classe Falso confirmam o que foi observado na classe Verdadeiro: as proporções de falsos positivos e falsos negativos ficaram próximos dos 10%, sensibilidade e especificidade demonstraram boa habilidade do classificador em realizar classificações corretas, e as medidas precisão e *F-score* demonstraram uma concordância entre as classes preditas e as classes reais de quase 100%.

5.2.3 Identificação do modelo final

Para a identificação do modelo final, compararam-se os resultados das medidas de qualidade obtidas por cada algoritmo, a fim de determinar qual dos modelos obteve melhor performance na predição dos dados.

Os resultados das medidas gerais obtidas a partir dos resultados da validação dos algoritmos no conjunto de teste se encontram resumidos na tabela 8.

Tabela 8 – Comparação das medidas de qualidade gerais obtidas na fase de teste dos algoritmos *Naïve Bayes* e rede de crenças

Algoritmo	Métrica	Resultado
Naïve Bayes	Taxa de erro	14,63%
	Acurácia	85,36%
	AUC	0,8555
	Estatística *kappa*	0,6956
Rede de crenças (K2)	Taxa de erro	10,43%
	Acurácia	89,57%
	AUC	0,892
	Estatística *kappa*	0,7872

Fonte: Do autor.

Observando as informações expostas na tabela, é possível perceber que o algoritmo K2 obteve melhores resultados ao realizar a predição das classes durante a fase de teste. O algoritmo apresentou uma taxa de erro pouco maior que 10% e uma acurácia próxima a 90%. O K2 também apresenta uma capacidade melhor de evitar classificações erradas durante a predição, embora o Naïve Bayes apresente um valor próximo. A estatística *kappa* também revela que o algoritmo K2

apresentou melhor concordância entre os rótulos de classe verdadeiros e os preditos pelo classificador.

Na tabela 9 são comparados os resultados das medidas de qualidade aplicadas às classes.

Tabela 9 – Comparação das medidas de qualidade para as classes Verdadeiro e Falso obtidas pelos algoritmos *Naïve Bayes* e rede de crenças

Algoritmo	**Medida**	**Resultado**	
		Classe	
		Verdadeiro	**Falso**
Naïve Bayes	Proporção de falsos positivos	0,089	0,225
	Proporção de falsos negativos	0,225	0,089
	Sensibilidade	0,864	0,847
	Especificidade	0,847	0,864
	Precisão	0,775	0,911
	F-score	0,817	0,878
Rede de crenças (K2)	Proporção de falsos positivos	0,105	0,103
	Proporção de falsos negativos	0,103	0,105
	Sensibilidade	0,861	0,923
	Especificidade	0,923	0,861
	Precisão	0,897	0,895
	F-score	0,879	0,908

Fonte: Do autor.

A partir das informações apresentadas na tabela, pode-se perceber que, embora o algoritmo Naïve Bayes tenha apresentado melhor resultado na proporção de FPs para a classe Verdadeiro e proporção de FNs para a classe Falso, o K2 apresentou resultados melhores para a proporção de FNs para a classe Verdadeiro e proporção de FPs para a classe Falso.

Não se observaram maiores diferenças entre os valores da sensibilidade para a classe Verdadeiro, o que denota que ambos os classificadores possuem boa capacidade para identificar objetos com o rótulo Verdadeiro corretamente. No entanto, o algoritmo K2 apresentou uma sensibilidade mais próxima de 100% para a classe Falso. A situação oposta pode ser observada para a especificidade, onde o K2 obteve melhor desempenho para a classe Verdadeiro, mas um desempenho muito próximo do algoritmo Naïve Bayes para a classe Falso.

Quanto aos resultados da precisão e *F-score*, notou-se que o desempenho do algoritmo K2 foi levemente superior ao desempenho do Naïve Bayes, embora tenha obtido um desempenho inferior no cálculo da precisão para a classe Falso.

Após a análise e comparação dos resultados, conclui-se que o modelo com melhor desempenho foi o modelo gerado pelo algoritmo K2 de redes de crença, devido à sua melhor capacidade de evitar classificações incorretas.

5.2.4 Discussão dos resultados

Analisando as informações obtidas nos modelos a partir dos classificadores, pode-se perceber que a estação chuvosa para a região norte de Blumenau compreende os meses de janeiro a março. Ambos os classificadores apontaram que os meses de outubro e dezembro também possuem alta precipitação.

Não foi possível observar os valores da CPT gerada pelo algoritmo K2 para o atributo ano devido a um problema de visualização advindo do software. No entanto, o modelo obtido pelo Naïve Bayes revela que, com a exceção de alguns anos, os períodos de chuva e seca demonstram-se muito equilibrados por toda a progressão da base. Alguns anos se diferenciam dos demais, como por exemplo os anos de 1941, 1944, 1950, 1955, 1957, 1968 e 1970, que possuem uma quantidade de meses não chuvosos muito maior do que os meses chuvosos, caracterizando um ano de seca. O oposto foi observado nos anos de 1957, 1969 e 1990, onde houve maior ocorrência de meses chuvosos do que não chuvosos.

Um pequeno erro de classificação também foi observado para o atributo *NumDiasDeChuva*, onde pelo menos uma instância para cada número de dias de chuva menor que 10 foi classificada como Verdadeiro. O algoritmo K2 também mostrou uma pequena taxa de erro na classificação do atributo *NumDiasDeChuva*, apresentando valores abaixo de 5% e alguns abaixo de 1% para os dias de chuva menores que 10.

Após as análises finais dos modelos de classificação, comparou-se os resultados obtidos neste trabalho com os resultados de outras pesquisas com temas similares aos que foram abordados, como classificadores, algoritmos bayesianos e

pluviometria. A tabela 10 mostra as pesquisas que foram estudadas, mas apenas os estudos que se assemelham a este trabalho foram abordados.

Tabela 10 – Pesquisas relacionadas ao tema

Autor(es)	Título da pesquisa	Base de dados	Algoritmos	Medidas de qualidade
Deepti Gupta e Udayan Ghose	A Comparative Study of Classification Algrithms for Forecasting Rainfall	Dados pluviométricos	Árvore de Regressão de Classificação (algoritmo CART), Naïve Bayes, K-nearest Neighbour e Redes Neurais	Acurácia
S. S. Thakur, Anirban Kundu e J. K. Singh	A Novel Approach: Using Bayesian Belief Networks in Product Recommendation	Dados de vendas de telefones móveis	Rede de crenças	-
Mykhailo Granik e Volodymyr Mesyura	Fake News Detection Using Naïve Bayes Classifier	Posts do Facebook e notícias dos sites Politico, CNN e ABC News	Naïve Bayes	Acurácia
Leonard Barreto Moreira e Anderson Amendoeira Namen	Sistema Preditivo Para a Doença de Alzheimer na Triagem Clínica	Pacientes com suspeita clínica de Alzheimer	Naïve Bayes, BBN e árvores de decisão	Acurácia, taxa de erro, sensibilidade, proporção de FPs, proporção de FNs, precisão, *F*-score e validação cruzada estratificada *k-fold*
Marcio Novaski	O Teorema de Probabilidade Pelo Algoritmo Naïve Bayes Para a Tarefa de Classificação na Shell Orion Data Mining Engine	Dados de câncer de mama	Naïve Bayes	Sensibilidade, especificidade, acurácia, taxa de erro, confiabilidade positiva e estatística *kappa*

Fonte: Do autor.

Em seu estudo, Gupta e Ghose (2015) concluem que o algoritmo *Naïve Bayes* apresentou a menor acurácia em relação aos outros algoritmos aplicados,

alcançando 78,9% de dados classificados corretamente. Os autores definem o algoritmo Naïve Bayes como um algoritmo de fácil entendimento e de se trabalhar, no entanto, a aplicação de um método de complexidade de custo requer cálculos intensivos, tornando o processo demorado.

Na pesquisa de Granik e Mesyura (2017), o algoritmo *Naïve Bayes* apresentou uma acurácia total de 75,4%. Os autores afirmam que a acurácia obtida poderia ser melhorada se algumas medidas fossem tomadas, como por exemplo, usar uma quantidade maior de dados para realizar o treinamento, usar uma base de dados maior e com notícias mais extensas, ou remover palavras extremamente comuns. Contudo, os autores concluíram que, mesmo sendo um método de IA simples, o algoritmo Naïve Bayes pode desempenhar uma boa performance na identificação de notícias falsas.

Moreira e Namen (2016), concluíram em sua pesquisa que os modelos construídos a partir dos algoritmos bayesianos mostraram melhor desempenho na classificação dos dados de Alzheimer, com acurácias entre 71,7% e 73,8%. No entanto, o algoritmo de rede de crenças apresentou um resultado melhor que o obtido pelo algoritmo *Naïve Bayes*.

6 CONCLUSÃO

Com o crescente aumento dos dados na era da informação, o *data mining* surge como uma alternativa aos métodos tradicionais de análise de dados. O *data mining* vem cada vez mais se mostrando como uma ferramenta poderosa de descoberta de conhecimentos importantes para várias áreas, tais como saúde, ciência, economia e sociedade.

A tarefa de classificação demonstra-se aplicável nas mais diversas áreas para a descoberta de padrões e predição de dados novos. Os métodos bayesianos são algoritmos de funcionamento simples e fácil entendimento, no entanto, demonstraram-se eficazes na classificação do conjunto de dados apresentado nesta pesquisa.

As dificuldades encontradas nesta pesquisa foram relacionadas à seleção da base de dados a ser utilizada. Após a decisão de aplicar os algoritmos de uma base de dados de clima para realizar a predição de chuvas, novamente outra dificuldade foi encontrada. Os dados não apresentavam boa consistência, contando com boa parte dos dados vazios, séries temporais muito curtas, ou falta de atributos importantes para realizar a predição. Outra dificuldade observada, em relação à ferramenta de data mining utilizada, foi a visualização dos valores das CPTs obtidas para cada nó do GAD gerado pelo algoritmo de rede de crenças, o que impossibilitou a visualização dos resultados obtidos para o atributo Ano.

Embora as dificuldades tenham sido encontradas, os objetivos foram alcançados e um modelo de predição adequado foi identificado através da aplicação de medidas de qualidade em *data mining* e comparação dos resultados.

O modelo final foi obtido pelo algoritmo de rede de crenças, levando-se em conta seus resultados nas medidas de qualidade. A acurácia demonstrada pelo algoritmo na validação do modelo no conjunto de testes foi de 89,57%, com uma taxa de erro de 10,43%. O classificador obteve uma AUC de 89,2%, denotando uma ótima habilidade em evitar classificações incorretas. As medidas aplicadas às classes confirmam o que foi observado nos resultados das medidas gerais, com índices baixos de classificações erradas e bons índices de classificações corretas e concordância entre os dados.

Com base no conhecimento adquirido durante o desenvolvimento desta pesquisa, sugerem-se algumas possibilidades de trabalhos futuros:

a) realizar a análise do mesmo conjunto de dados utilizando outros algoritmos de classificação;
b) realizar a análise do mesmo conjunto de dados utilizando outra tarefa de *data mining*;
c) aplicar os mesmos algoritmos em um conjunto de dados diferente, tanto na mesma área aplicada neste trabalho, quanto em uma área diferente.

REFERÊNCIAS

AGGARWAL, Charu C.. **Data Mining:** The Textbook. [s.i.]: Springer, 2015.

BARBETTA, Pedro Alberto; REIS, Marcelo Menezes; BORNIA, Antônio Cezar. **Estatística:** para cursos de engenharia e informática. 3. ed. São Paulo: Atlas, 2010. 410 p.

BHIDE, Amarnath; CARIC, Vedrana; ARULKUMARAN, Sabaratnam. Prediction of vaginal birth after cesarean delivery. **International Journal Of Gynecology & Obstetrics**, [s.l.], v. 133, n. 3, p.297-300, 12 fev. 2016. Wiley-Blackwell. http://dx.doi.org/10.1016/j.ijgo.2015.09.031.

BORGES, Giovana Mara Rodrigues; THEBALDI, Michael Silveira. Estimativa da precipitação máxima diária anual e equação de chuvas intensas para o município de Formiga, MG, Brasil. **Ambiente e Agua - An Interdisciplinary Journal Of Applied Science**, [s.l.], v. 11, n. 4, p.891-902, 25 out. 2016. Instituto de Pesquisas Ambientais em Bacias Hidrograficas (IPABHi). http://dx.doi.org/10.4136/ambi-agua.1823.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **Ai Magazine,** [s.i.], v. 17, n. 3, p.37-54, out. 1996.

FIELD, Andy. **Descobrindo a Estatística Usando o SPSS:** Descobrindo a Estatística Usando o SPSS. 2. ed. Porto Alegre: Artmed, 2009. 688 p.

GORUNESCU, Florin. **Data Mining:** Concepts, Models and Techniques. [s.i.]: Springer, 2011.

GRANIK, Mykhailo; MESYURA, Volodymyr. Fake news detection using naive Bayes classifier. **2017 Ieee First Ukraine Conference On Electrical And Computer Engineering (ukrcon),** [s.l.], v. 1, n. 1, p.900-903, maio 2017. IEEE. http://dx.doi.org/10.1109/ukrcon.2017.8100379.

GUILLET, Fabrice; HAMILTON, Howard J. **Quality Measures in Data Mining**. Berlin: Springer, 2007.

GUPTA, Deepti; GHOSE, Udayan. A comparative study of classification algorithms for forecasting rainfall. **2015 4th International Conference On Reliability, Infocom Technologies And Optimization (ICRITO) (trends And Future Directions)**, [s.l.], v. 4, set. 2015. IEEE. http://dx.doi.org/10.1109/icrito.2015.7359273.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining:** Concepts and Techniques. 3. ed. Waltham: Elsevier, 2012. 517 p.

HAND, David J.. Naïve Bayes. In: WU, Xindong; KUMAR, Vipin. **The Top Ten Algorithms in Data Mining.** Minneapolis: Chapman & Hall/crc, 2009. p. 163-177.

HAND, David; MANNILA, Heikki; SMYTH, Padhraic. **Principles of Data Mining**. Massachusetts: Bradford, 2001. 546 p.

KUMAR, A. V. Senthil. **Knowledge Discovery Practices and Emerging Applications of Data Mining:** Trends and New Domains. Hershey: Igi Global, 2011. 390 p.

LAROSE, Daniel T. **Discovering Knowledge in Data:** An Introduction to Data Mining. New Jersey: Wiley Publishing, 2005.

LINOFF, Gordon S; BERRY, Michael J. A. **Data Mining Techniques:** For Marketing, Sales, and Customer Relationship Management. 3. ed. Indianapolis: Wiley Publishing, 2011. 888 p.

LUGER, George F.. **Inteligência Artificial:** Estruturas e estratégias para a solução de problemas complexos. 6. ed. São Paulo: Pearson, 2014. 632 p.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre Aprendizado de Máquina. In: REZENDE, Solange Oliveira. **Sistemas Inteligentes**. 2. ed. Barueri: Manole Ltda., 2005. Cap. 4. p. 89-114.

MOREIRA, Leonard Barreto; NAMEN, Anderson Amendoeira. Sistema preditivo para a doença de Alzheimer na triagem clínica. **Journal Of Health Informatics,** São Paulo, v. 8, n. 3, p.87-94, fev. 2016.

NOVASKI, Marcio. **O Teorema de Probabilidade Pelo Algoritmo Naive Bayes Para a Tarefa de Classificação na Shell Orion Data Mining Engine.** 2011. 103 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, 2012.

OLSON, David L.; DELEN, Dursun. Advanced data mining techniques. Berlim: Springer Science & Business Media, 2008.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do Trabalho Científico:** Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico. 2. ed. Novo Hamburgo: Universidade Feevale, 2013. 277 p.

RUSSELL, Stuart J.; NORVIG, Peter. **Inteligência Artificial**. 3. ed. Rio de Janeiro: Elsevier, 2013.

SILVA, Vicente P. R. da et al. Análise da pluviometria e dias chuvosos na região Nordeste do Brasil. **Revista Brasileira de Engenharia Agrícola e Ambiental**, [s.l.], v. 15, n. 2, p.131-138, fev. 2011. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s1415-43662011000200004.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, [s.l.], v. 45, n. 4, p.427-437, jul. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.ipm.2009.03.002.

TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining**: Mineração de dados. Boston: Pearson, 2009. 769 p.

TAVARES, L.G. ; LOPES, H. S. ; LIMA, C.R.E. . Estudo comparativo de métodos de aprendizagem de máquina na detecção de regiões promotoras de genes de Escherichia coli. **In: Simpósio Brasileiro de Inteligência Computacional**, 2007, Florianópolis. Anais do I SBIC. Florianópolis, 2007.

THAKUR, S.s.; KUNDU, Anirban; SING, J.k.. A Novel Approach: Using Bayesian Belief Networks in Product Recommendation. **2011 Second International Conference On Emerging Applications Of Information Technology,** [s.l.], v. 2, n. 2, p.37-40, fev. 2011. IEEE. http://dx.doi.org/10.1109/eait.2011.21.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A.. **Data Mining:** Practical Machine Learning Tools and Techniques. 3. ed. Burlington: Elsevier, 2011. 629 p.

YOO, Illhoi et al. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. Journal Of Medical Systems, [s.l.], v. 36, n. 4, p.2431-2448, 3 maio 2011. Springer Nature. http://dx.doi.org/10.1007/s10916-011-9710-5.

ZAINUDIN, Suhaila; JASIM, Dalia Sami; BAKAR, Azuraliza Abu. Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction. **International Journal On Advanced Science, Engineering And Information Technology**, [s.l.], v. 6, n. 6, p.1148-1153, 9 dez. 2016. Insight Society. http://dx.doi.org/10.18517/ijaseit.6.6.1487.

ZAKI, Mohammed J.; MEIRA JUNIOR, Wagner. **Data Mining and Analysis:** Fundamental Concepts and Algorithms. Cambridge: Draft, 2013.

APÊNDICES

APÊNDICE A – Modelo de classificação obtido pelo algoritmo *Naïve Bayes*

	Classe	
Atributo	VERDADEIRO	FALSO
	(0.42)	(0.58)
Ano		
1941	3.0	11.0
1942	5.0	9.0
1943	6.0	8.0
1944	3.0	11.0
1945	6.0	8.0
1946	5.0	9.0
1947	7.0	7.0
1948	7.0	7.0
1949	5.0	9.0
1950	4.0	10.0
1951	5.0	9.0
1952	5.0	9.0
1953	5.0	9.0
1954	7.0	7.0
1955	1.0	13.0
1956	5.0	9.0
1957	11.0	3.0
1958	5.0	9.0
1959	7.0	7.0
1960	7.0	7.0
1961	8.0	6.0
1962	5.0	9.0
1963	8.0	6.0
1964	8.0	6.0
1965	6.0	8.0
1966	7.0	7.0
1967	6.0	8.0

1968	2.0	12.0
1969	10.0	4.0
1970	4.0	10.0
1971	7.0	7.0
1972	8.0	6.0
1973	6.0	8.0
1974	5.0	9.0
1975	6.0	8.0
1976	7.0	7.0
1977	7.0	7.0
1978	4.0	10.0
1979	5.0	9.0
1980	8.0	6.0
1981	5.0	9.0
1982	7.0	7.0
1983	8.0	6.0
1984	5.0	9.0
1985	6.0	8.0
1986	7.0	7.0
1987	7.0	6.0
1988	5.0	9.0
1989	3.0	11.0
1990	11.0	3.0
1991	7.0	7.0
1992	7.0	7.0
1993	7.0	7.0
1994	7.0	7.0
1995	5.0	9.0
1996	6.0	8.0
1997	6.0	8.0
1998	9.0	5.0
1999	5.0	9.0
2000	6.0	8.0
2004	3.0	6.0

2006	5.0	9.0
[total]	373.0	489.0

Mes

01	43.0	20.0
02	44.0	19.0
03	41.0	22.0
04	14.0	49.0
05	14.0	49.0
06	15.0	49.0
07	12.0	51.0
08	21.0	43.0
09	27.0	37.0
10	36.0	28.0
11	22.0	42.0
12	34.0	30.0
[total]	323.0	439.0

Maxima

mean	53.6152	32.9792
std. dev.	25.7836	19.6193
weight sum	311	427
precision	0.4727	0.4727

Total

mean	208.3202	87.5046
std. dev.	78.1408	44.1504
weight sum	311	427
precision	0.8251	0.8251

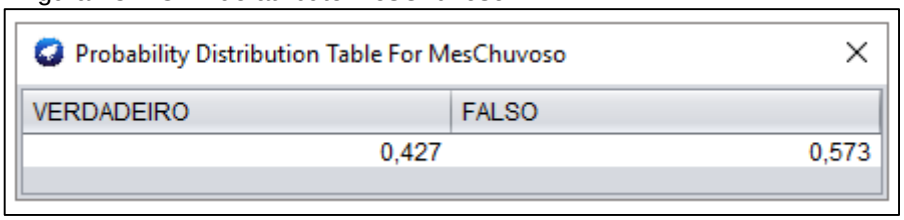
NumDiasDeChuva

1	1.0	5.0
2	1.0	9.0
3	1.0	7.0

4	1.0	22.0
5	1.0	30.0
6	1.0	29.0
7	1.0	36.0
8	1.0	48.0
9	1.0	49.0
10	36.0	29.0
11	31.0	21.0
12	39.0	29.0
13	42.0	25.0
14	26.0	18.0
15	24.0	27.0
16	21.0	18.0
17	22.0	15.0
18	21.0	9.0
19	15.0	7.0
20	20.0	6.0
21	9.0	6.0
22	6.0	2.0
23	2.0	2.0
24	5.0	2.0
25	4.0	1.0
26	5.0	1.0
[total]	337.0	453.0

APÊNDICE B – Tabelas de probabilidade condicional geradas pelo algoritmo de rede de crenças

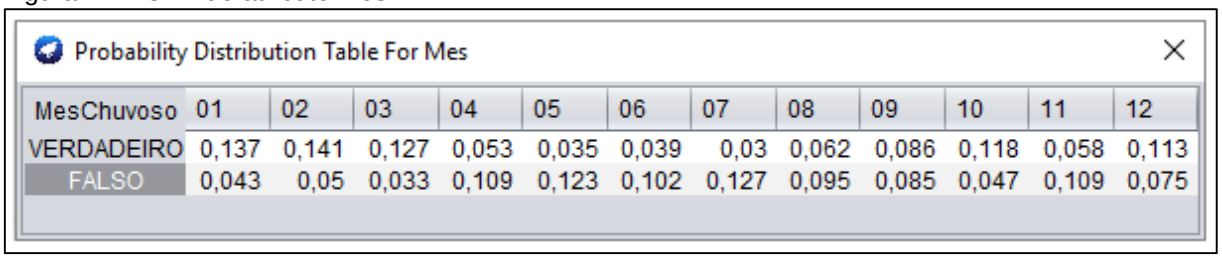
Figura 13 – CPT do atributo *MesChuvoso*



Probability Distribution Table For MesChuvoso

VERDADEIRO	FALSO
0,427	0,573

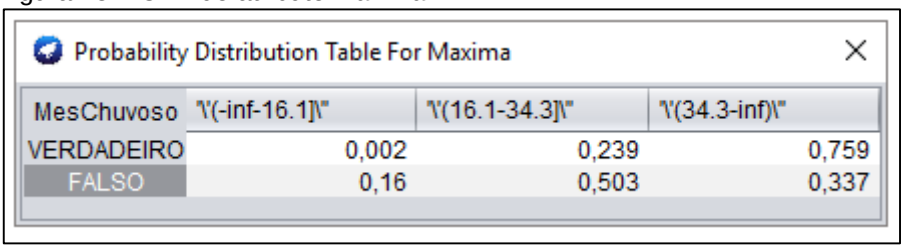
Figura 14 – CPT do atributo *Mes*



Probability Distribution Table For Mes

MesChuvoso	01	02	03	04	05	06	07	08	09	10	11	12
VERDADEIRO	0,137	0,141	0,127	0,053	0,035	0,039	0,03	0,062	0,086	0,118	0,058	0,113
FALSO	0,043	0,05	0,033	0,109	0,123	0,102	0,127	0,095	0,085	0,047	0,109	0,075

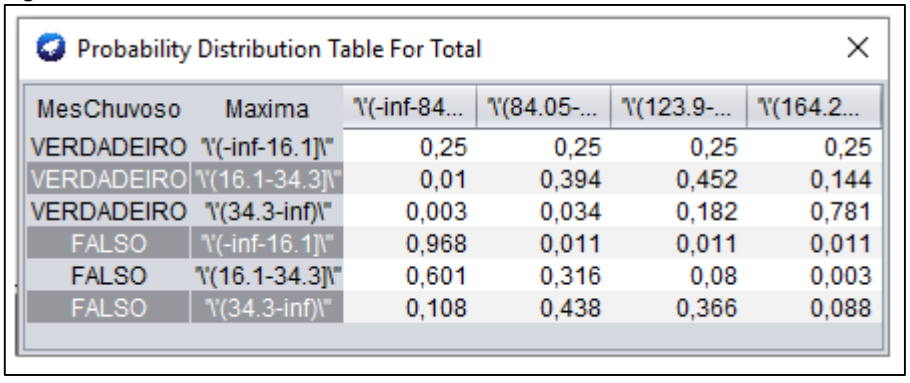
Figura 15 – CPT do atributo *Maxima*



Probability Distribution Table For Maxima

MesChuvoso	'\'(-inf-16.1]\''	'\'(16.1-34.3]\''	'\'(34.3-inf)\''
VERDADEIRO	0,002	0,239	0,759
FALSO	0,16	0,503	0,337

Figura 16 – CPT do atributo *Total*



Probability Distribution Table For Total

MesChuvoso	Maxima	'\'(-inf-84...	'\'(84.05-...	'\'(123.9-...	'\'(164.2...
VERDADEIRO	'\'(-inf-16.1]\''	0,25	0,25	0,25	0,25
VERDADEIRO	'\'(16.1-34.3]\''	0,01	0,394	0,452	0,144
VERDADEIRO	'\'(34.3-inf)\''	0,003	0,034	0,182	0,781
FALSO	'\'(-inf-16.1]\''	0,968	0,011	0,011	0,011
FALSO	'\'(16.1-34.3]\''	0,601	0,316	0,08	0,003
FALSO	'\'(34.3-inf)\''	0,108	0,438	0,366	0,088

Figura 17 – CPT do atributo *NumDiasDeChuva*

Probability Distribution Table For NumDiasDeChuva

MesChuvoso	Total	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
VERDADEIRO	'\'(-inf-84.05]\''	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038	0,038
VERDADEIRO	'\'(84.05-123.9]\''	0,013	0,013	0,013	0,013	0,013	0,013	0,013	0,013	0,013	0,303	0,171	0,118	0,092	0,013	0,039	0,013	0,013	0,013	0,013	0,013	0,013	0,013	0,013	0,013	0,013	0,013
VERDADEIRO	'\'(123.9-164.25]\''	0,008	0,008	0,008	0,008	0,008	0,008	0,008	0,008	0,008	0,162	0,162	0,131	0,131	0,115	0,069	0,069	0,008	0,008	0,023	0,008	0,008	0,008	0,008	0,008	0,008	0,008
VERDADEIRO	'\'(164.25-inf)\''	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,051	0,045	0,079	0,147	0,051	0,065	0,086	0,092	0,086	0,065	0,106	0,031	0,024	0,003	0,017	0,003	0,017
FALSO	'\'(-inf-84.05]\''	0,016	0,036	0,023	0,081	0,088	0,107	0,075	0,12	0,075	0,14	0,062	0,055	0,042	0,023	0,01	0,003	0,016	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003
FALSO	'\'(84.05-123.9]\''	0,005	0,005	0,005	0,005	0,015	0,015	0,065	0,075	0,145	0,015	0,035	0,125	0,095	0,065	0,115	0,055	0,045	0,025	0,025	0,025	0,015	0,005	0,005	0,005	0,005	0,005
FALSO	'\'(123.9-164.25]\''	0,008	0,008	0,008	0,008	0,025	0,059	0,059	0,11	0,144	0,008	0,059	0,025	0,025	0,008	0,059	0,11	0,076	0,059	0,042	0,008	0,025	0,008	0,008	0,025	0,008	0,008
FALSO	'\'(164.25-inf)\''	0,024	0,024	0,024	0,024	0,024	0,024	0,071	0,024	0,071	0,024	0,024	0,024	0,024	0,071	0,024	0,071	0,024	0,024	0,071	0,119	0,071	0,024	0,024	0,024	0,024	0,024

APÊNDICE C – Artigo

Os Métodos Bayesianos de Aprendizado de Máquina Pelos Algoritmos Naïve Bayes e Redes de Crença na Predição de Período Chuvoso na Cidade de Blumenau

Bruna Baldini Dias[1], Merisandra Côrtes de Mattos Garcia[2]

[1]Acadêmica do Curso de Ciência da Computação – Unidade Acadêmica de Ciências, Engenharias e Tecnologias – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma – SC – Brasil

[2]Professora do Curso de Ciência da Computação – Unidade Acadêmica de Ciências, Engenharias e Tecnologias – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma – SC – Brasil

brunabdias12@gmail.com, mem@unesc.net

Abstract. *Rainfall represents a huge importance to the hydrologic cicle, besides being one of the main sustentable irrigation font for agriculture. However, excessive rainfall can presente hazards for society, as such floods and agravate ground erosion. This work proposes the analyis of rainfall data from Blumenau city through Naïve Bayes and Bayesian Belief Networks algorithms to predict rainy season. Quality measures where applies to the models generated to determine wich one was most apropriated for the analysed data.*

Resumo. *A precipitação pluvial representa uma grande importância para o ciclo hidrológico, além de serem uma das principais fontes de irrigação sustentável para a agricultura. No entanto, chuvas em demasia podem apresentar riscos para a sociedade, como causar inundações e agravar a erosão do solo. Este trabalho propõe a análise de dados pluviométricos da cidade de Blumenau através dos algoritmos Naïve Bayes e Redes de Crença para a predição de período chuvoso. Medidas de qualidade foram aplicadas aos modelos obtidos para determinar o melhor dentre os mesmos.*

1. INTRODUÇÃO

O rápido crescimento e integração das bases de dados proveem a cientistas, engenheiros e empresários um vasto recurso que pode ser utilizado descobrir novos padrões nos dados. *Data mining* é a análise de bases de dados, geralmente grandes, para a descoberta de relações entre esses dados e a sua organização em novas formas, tornando-os compreensíveis e úteis ao usuário (HAND; MANNILA; SMYTH, 2001, tradução nossa). Segundo Han, Kamber e Pei (2012, tradução nossa), para que se possa aproveitar os dados e melhorar os resultados das análises, o data mining apresenta diversas tarefas que são usadas para especificar os tipos de padrões que são encontrados, tais como: agrupamento (*clustering*), associação, regressão e classificação.

A classificação é uma forma supervisionada de aprendizado de máquina (MONARD; BARANAUSKAS, 2003) que consiste em analisar um objeto e atribuí-lo a uma classe pré-definida dentro de um conjunto. O processo de classificação é dividido em duas etapas: na

primeira etapa, um modelo de classificação é construído a partir da análise de um conjunto de dados selecionados para o aprendizado; na segunda etapa, o modelo resultante da primeira etapa é usado para a classificação dos dados (BERRY; LINOFF, 2004).

Há uma vasta quantidade de métodos que podem ser utilizados para realizar a tarefa de classificação, entre eles pode-se citar os bayesianos, que são algoritmos que se baseiam no Teorema de Bayes, um princípio estatístico para combinar informações já conhecidas com novas informações extraídas dos conjuntos de dados (HAN; KAMBER; PEI, 2012; TAN; STEINBACH; KUMAR, 2009).

Data mining vem se tornando uma ferramenta eficiente para a descoberta de conhecimento em vários campos (KUMAR, 2011). Um exemplo seria a análise de alterações climáticas (ZAINUDIN; JASIM; BAKAR, 2016, tradução nossa).

De acordo com Zainudin, Jasim e Bakar (2016, tradução nossa), a análise de alterações climáticas visa estudar o comportamento do clima durante um período de tempo específico. A característica chave por trás da mudança climática se encontra na natureza de seus dados, que são capturados em formato de pontos temporais. Uma tarefa de alteração climática é a previsão de chuva, onde atributos específicos como umidade e vento são usados para prever chuvas em uma localização específica. Várias técnicas como máquina de vetores de suporte, *Naïve Bayes*, redes neurais e outras vêm sendo usadas para a previsão de chuvas, sendo que a maioria das técnicas empregadas consiste em aprendizado supervisionado.

Esta pesquisa compreenderá a aplicação dos algoritmos Naïve Bayes e Redes de Crença Bayesianas em dados pluviométricos. Os resultados originados são comparados pelo uso de medidas de qualidade em data mining, comumente empregadas pela comunidade de aprendizado de máquina, a fim de identificar um modelo de predição de período chuvoso na cidade de Blumenau.

2. MATERIAIS E MÉTODOS

O software utilizado para a aplicação dos algoritmos *Naïve Bayes* e BBN foi o WEKA em sua versão 3.8 para o sistema operacional Windows. Optou-se por utilizar a ferramenta WEKA não apenas por ser gratuita, mas por disponibilizar os algoritmos necessários para esta pesquisa e por possuir uma interface simples e de fácil uso. Foi realizado um estudo sobre a ferramenta, sobre o funcionamento dos algoritmos Naïve Bayes e Rede de Crenças e das medidas de qualidade escolhidas para validar os modelos obtidos.

2.1. Base de dados empregada

A base de dados selecionada para a realização dessa pesquisa foi extraída da Agência Nacional das Águas (ANA) e consiste em dados pluviométricos do bairro de Itoupava Central na cidade de Blumenau, em Santa Catarina.

Optou-se por usar essa base pela mesma apresentar uma série histórica de mais de 30 anos, apresentar boa consistência nos dados e possuir poucos dados faltantes. A região norte de Blumenau, região em que está localizado o bairro Itoupava Central, constantemente sofre com inundações.

A base possui 738 instâncias, com a remoção de uma instância com valores vazios, e seis atributos contando com o atributo de classe, descritos na tabela 1.

Tabela 1. Atributos da base de dados de pluviometria

Atributo	Valor	Descrição
Mês	Nominal (01 a 12)	-
Ano	Nominal (de 1941 a 2000, 2004, 2006)	-
Máxima	Numérico	Precipitação máxima do mês
Total	Numérico	Precipitação total do mês
NumDiasDeChuva	Nominal (1 a 26)	Número de dias de chuva do mês
MesChuvoso	Nominal (verdadeiro e falso)	Classe

Fonte: Do autor.

2.2. Pré-processamento dos dados

A fase de pré-processamento abrange técnicas para limpeza, organização e refinamento dos dados, a fim de se obter melhores resultados no processo de mineração.

A base de dados estava originalmente no formato *mdb*, próprio da ferramenta Access. Nela, houve a remoção dos atributos *RegistroID*, *EstacaoCodigo* e *NivelConsistencia*, pois não contribuem para os resultados da classificação. Também uma instância com valores vazios foi removida. Utilizou-se a ferramenta *MDBViewer* para realizar a conversão do arquivo Access para Excel.

No Excel, foram atribuídos os rótulos de classe Verdadeiro e Falso através das regras de classificação descritas no capítulo anterior. Após a atribuição dos rótulos de classe, o arquivo Excel (*xlsx*) foi convertido para um arquivo com valores separados por vígula (*csv*). Por fim, utilizou-se o Weka para realizer a conversão de csv para o format ARFF.

2.3. Execução do *data mining*

Primeiramente aplicou-se o algoritmo *Naïve Bayes* no conjunto de dados. Foram utilizadas suas configurações padrão e o método de validação *cross-validation* com 10 *folds*. Este método divide o conjunto de treinamento em um número pré-estabelecido de partições (*folds*) onde cada partição é usada para teste uma vez e o restante para treinamento, sendo 10 *folds* considerado o número ideal de partições para se obter a melhor taxa de erro (WITTEN; FRANK; HALL, 2011, tradução nossa). Foram obtidos o modelo de classificação, matriz de confusão e resultados das medidas de qualidade escolhidas.

Em seguida, realizou-se a aplicação do algoritmo de rede de crenças. Utilizou-se as configurações padrão da ferramenta. O algoritmo escolhido foi o K2, com o número de nós pais configurado como cinco. O método de validação do modelo foi o mesmo escolhido para o algoritmo *Naïve Bayes*.

3. RESULTADOS

A partir de obtenção das matrizes de qualidade, torna-se possível a aplicação das medidas de qualidade, realizada na etapa anterior. Após a obtenção dos resultados, torna-se possível realizar as análises necessárias para a identificação do modelo final.

3.1. Comparação das medidas de qualidade

Para a identificação do modelo final, compararam-se os resultados das medidas de qualidade obtidas por cada algoritmo, a fim de determinar qual dos modelos obteve melhor performance na predição dos dados.

Os resultados das medidas gerais obtidas a partir dos resultados da validação dos algoritmos no conjunto de teste se encontram resumidos na tabela 2.

Tabela 2. Comparação das medidas de qualidade gerais obtidas na fase de teste dos algoritmos *Naïve Bayes* e rede de crenças

Algoritmo	Métrica	Resultado
Naïve Bayes	Taxa de erro	14,63%
	Acurácia	85,36%
	AUC	0,8555
	Estatística *kappa*	0,6956
Rede de crenças (K2)	Taxa de erro	10,43%
	Acurácia	89,57%
	AUC	0,892
	Estatística *kappa*	0,7872

Fonte: Do autor.

Observando as informações expostas na tabela, é possível perceber que o algoritmo K2 obteve melhores resultados ao realizar a predição das classes durante a fase de teste. O algoritmo apresentou uma taxa de erro pouco maior que 10% e uma acurácia próxima a 90%. O K2 também apresenta uma capacidade melhor de evitar classificações erradas durante a predição, embora o Naïve Bayes apresente um valor próximo. A estatística *kappa* também revela que o algoritmo K2 apresentou melhor concordância entre os rótulos de classe verdadeiros e os preditos pelo classificador.

Na tabela 3 são comparados os resultados das medidas de qualidade aplicadas às classes.

Tabela 3. Comparação das medidas de qualidade para as classes Verdadeiro e Falso obtidas pelos algoritmos *Naïve Bayes* e rede de crenças

Algoritmo	Medida	Resultado	
		Classe	
		Verdadeiro	Falso
Naïve Bayes	Proporção de falsos positivos	0,089	0,225
	Proporção de falsos negativos	0,225	0,089
	Sensibilidade	0,864	0,847
	Especificidade	0,847	0,864
	Precisão	0,775	0,911
	F-score	0,817	0,878
Rede de crenças (K2)	Proporção de falsos positivos	0,105	0,103
	Proporção de falsos negativos	0,103	0,105
	Sensibilidade	0,861	0,923
	Especificidade	0,923	0,861
	Precisão	0,897	0,895
	F-score	0,879	0,908

Fonte: Do autor.

A partir das informações apresentadas na tabela, pode-se perceber que, embora o algoritmo Naïve Bayes tenha apresentado melhor resultado na proporção de FPs para a classe Verdadeiro e proporção de FNs para a classe Falso, o K2 apresentou resultados melhores para a proporção de FNs para a classe Verdadeiro e proporção de FPs para a classe Falso.

Não se observaram maiores diferenças entre os valores da sensibilidade para a classe Verdadeiro, o que denota que ambos os classificadores possuem boa capacidade para identificar objetos com o rótulo Verdadeiro corretamente. No entanto, o algoritmo K2 apresentou uma sensibilidade mais próxima de 100% para a classe Falso. A situação oposta pode ser observada para a especificidade, onde o K2 obteve melhor desempenho para a classe Verdadeiro, mas um desempenho muito próximo do algoritmo Naïve Bayes para a classe Falso.

Quanto aos resultados da precisão e *F-score*, notou-se que o desempenho do algoritmo K2 foi levemente superior ao desempenho do Naïve Bayes, embora tenha obtido um desempenho inferior no cálculo da precisão para a classe Falso.

Após a análise e comparação dos resultados, conclui-se que o modelo com melhor desempenho foi o modelo gerado pelo algoritmo K2 de redes de crença, devido à sua melhor capacidade de evitar classificações incorretas.

4. DISCUSSÃO

Analisando as informações obtidas nos modelos a partir dos classificadores, pode-se perceber que a estação chuvosa para a região norte de Blumenau compreende os meses de janeiro a março. Ambos os classificadores apontaram que os meses de outubro e dezembro também possuem alta precipitação.

Não foi possível observar os valores da CPT gerada pelo algoritmo K2 para o atributo ano devido a um problema de visualização advindo do software. No entanto, o modelo obtido pelo Naïve Bayes revela que, com a exceção de alguns anos, os períodos de chuva e seca demonstram-se muito equilibrados por toda a progressão da base. Alguns anos se diferenciam dos demais, como por exemplo os anos de 1941, 1944, 1950, 1955, 1957, 1968 e 1970, que possuem uma quantidade de meses não chuvosos muito maior do que os meses chuvosos, caracterizando um ano de seca. O oposto foi observado nos anos de 1957, 1969 e 1990, onde houve maior ocorrência de meses chuvosos do que não chuvosos. Também foi observada uma pequena inconsistência de classificação para ambos os algoritmos, onde pelo menos uma instância para cada número de dias de chuva menor que 10 foi classificada como Verdadeiro em ambos os algoritmos.

Após as análises finais dos modelos de classificação, comparou-se os resultados obtidos neste trabalho com os resultados de outras pesquisas com temas similares aos que foram abordados, como classificadores, algoritmos bayesianos e pluviometria.

Em seu estudo, Gupta e Ghose (2015) concluem que o algoritmo *Naïve Bayes* apresentou a menor acurácia em relação aos outros algoritmos aplicados, alcançando 78,9% de dados classificados corretamente. Os autores definem o algoritmo Naïve Bayes como um algoritmo de fácil entendimento e de se trabalhar, no entanto, a aplicação de um método de complexidade de custo requer cálculos intensivos, tornando o processo demorado.

Na pesquisa de Granik e Mesyura (2017), o algoritmo *Naïve Bayes* apresentou uma acurácia total de 75,4%. Os autores afirmam que a acurácia obtida poderia ser melhorada se algumas medidas fossem tomadas, como por exemplo, usar uma quantidade maior de dados para realizar o treinamento, usar uma base de dados maior e com notícias mais extensas, ou remover palavras extremamente comuns. Contudo, os autores concluíram que, mesmo sendo um método de IA simples, o algoritmo Naïve Bayes pode desempenhar uma boa performance na identificação de notícias falsas.

Moreira e Namen (2016), concluíram em sua pesquisa que os modelos construídos a partir dos algoritmos bayesianos mostraram melhor desempenho na classificação dos dados de Alzheimer, com acurácias entre 71,7% e 73,8%. No entanto, o algoritmo de rede de crenças apresentou um resultado melhor que o obtido pelo algoritmo *Naïve Bayes*.

5. CONCLUSÃO

Com o crescente aumento dos dados na era da informação, o *data mining* surge como uma alternativa aos métodos tradicionais de análise de dados. O *data mining* vem cada vez mais se mostrando como uma ferramenta poderosa de descoberta de conhecimentos importantes para várias áreas, tais como saúde, ciência, economia e sociedade.

A tarefa de classificação demonstra-se aplicável nas mais diversas áreas para a descoberta de padrões e predição de dados novos. Os métodos bayesianos são algoritmos de funcionamento simples e fácil entendimento, no entanto, demonstraram-se eficazes na classificação do conjunto de dados apresentado nesta pesquisa.

As dificuldades encontradas nesta pesquisa foram relacionadas à seleção da base de dados a ser utilizada. Após a decisão de aplicar os algoritmos de uma base de dados de clima para realizar a predição de chuvas, novamente outra dificuldade foi encontrada. Os dados não apresentavam boa consistência, contando com boa parte dos dados vazios, séries temporais muito curtas, ou falta de atributos importantes para realizar a predição. Outra dificuldade observada, em relação à ferramenta de data mining utilizada, foi a visualização dos valores das CPTs obtidas para cada nó do GAD gerado pelo algoritmo de rede de crenças, o que impossibilitou a visualização dos resultados obtidos para o atributo Ano.

Embora as dificuldades tenham sido encontradas, os objetivos foram alcançados e um modelo de predição adequado foi identificado através da aplicação de medidas de qualidade em *data mining* e comparação dos resultados.

O modelo final foi obtido pelo algoritmo de rede de crenças, levando-se em conta seus resultados nas medidas de qualidade. A acurácia demonstrada pelo algoritmo na validação do modelo no conjunto de testes foi de 89,57%, com uma taxa de erro de 10,43%. O classificador obteve uma AUC de 89,2%, denotando uma ótima habilidade em evitar classificações incorretas. As medidas aplicadas às classes confirmam o que foi observado nos resultados das medidas gerais, com índices baixos de classificações erradas e bons índices de classificações corretas e concordância entre os dados.

REFERÊNCIAS

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining:** Concepts and Techniques. 3. ed. Waltham: Elsevier, 2012. 517 p.

HAND, David; MANNILA, Heikki; SMYTH, Padhraic. **Principles of Data Mining**. Massachusetts: Bradford, 2001. 546 p.

KUMAR, A. V. Senthil. **Knowledge Discovery Practices and Emerging Applications of Data Mining:** Trends and New Domains. Hershey: Igi Global, 2011. 390 p.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre Aprendizado de Máquina. In: REZENDE, Solange Oliveira. **Sistemas Inteligentes**. 2. ed. Barueri: Manole Ltda., 2005. Cap. 4. p. 89-114.

TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining**: Mineração de dados. Boston: Pearson, 2009. 769 p.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A.. **Data Mining:** Practical Machine Learning Tools and Techniques. 3. ed. Burlington: Elsevier, 2011. 629 p.

ZAINUDIN, Suhaila; JASIM, Dalia Sami; BAKAR, Azuraliza Abu**. Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction.** International Journal On Advanced Science, Engineering And Information Technology**,** [s.l.], v. 6, n. 6, p.1148-1153, 9 dez. 2016. Insight Society. http://dx.doi.org/10.18517/ijaseit.6.6.1487.