ANÁLISE COMPARATIVA ENTRE ALGORITMO J48, NAIVEBAYES E APRIORI PARA BASE DE DADOS DE SUICÍDIO

Laís Rodrigues Barros Maciel
Graduação em Ciência da Computação pela Universidade Federal do Tocantins - UFT.
Servidora do Governo do Estado de Tocantins, lotada na Secretária da Fazenda no setor de Desenvolvimento de Projetos Tecnológicos.
e-mail: laisrbmaciel@hotmail.com

Revista UniScientiae | UNIVIÇOSA
Volume 2, Número 2, p.137-153 jul./dez. 2019

RESUMO

Esse presente trabalho visa extrair informações na classificação de padrões das características de pessoas que cometem suicídio, através das regras de associação geradas pelo uso do algoritmo apriori, bem como uma análise comparativa entre os algoritmos J48 e NaivaBayes na predição da taxa de pessoas suicidas em relação aos atributos sexo e idade. Ambos os algoritmos são constantemente utilizados na aplicação da mineração de dados, parte da computação que estuda a extração de conhecimento utilizando metodologias como knowledge discovery in databases – KDD.

Palavras-chave: Mineração de dados, Knowledge Discovery in databases, Algoritmo J48, Algoritmo Apriori e Algoritmo NaiveBayes.

ABSTRACT

This work aims to extract information in the classification of patterns of the characteristics of people who commit suicide, through the association rules generated by the use of the a priori algorithm, as well as a comparative analysis between the J48 and NaivaBayes algorithms in predicting the rate of suicidal people in relation to gender and age attributes. Both algorithms are constantly used in the application of data mining, part of computing that studies the extraction of knowledge using methodologies such as knowledge discovery in databases - KDD.

Keywords: *Data mining, Knowledge Discovery in databases, Algorithm J48, Algorithm Apriori and Algorithm NaiveBayes.*

1. INTRODUÇÃO

Sistemas baseados em conhecimento manipulam as informações de forma inteligente, sendo desenvolvidos com o objetivo de serem usados na resolução de problemas que requerem um alto conhecimento humano sobre o problema. O ser humano busca compreender o seu pensamento há milhares de anos. Visando isso, a inteligência artificial tenta construir sistemas inteligentes que abrangem uma grande variedade de temas como, por exemplo, reconhecimento de padrões, robótica, previsão, dentre outros (CASTRO E FERRARI, 2016) (REZENDE, 2003).

O termo mineração de dados foi utilizado em comparação com processo de mineração de metais, uma vez que se explora uma base de dados usando algoritmos adequados para obter conhecimento. O conhecimento é algo que permite a tomada de decisão para a agregação de valor, por exemplo, saber que vai chover pode influenciar na decisão de viajar sim ou não. A mineração faz parte de um processo de descoberta de conhecimento conhecida como KDD (Knowledge Discovery in Databases) (CASTRO E FERRARI 2016).

O KDD é composto de fases, que serão descritas na sessão 3 (metodologia), que se dividem em seleção e integração das bases de dados, limpeza da base de dados, seleção e transformação dos dados, mineração e avaliação dos resultados. Sendo uma metodologia multidisciplinar envolvendo conhecimento de diversas áreas diferentes, como banco de dados, estatística, inteligência artificial, dentre outros, (CASTRO E FERRARI 2016).

Com isso, surgiram várias técnicas de mineração de dados, aonde cada uma possui um objetivo diferente de conhecimento. Predição e classificação avaliam a classe de um atributo e estima-se a valor de um ou mais atributos. Cluster realiza o agrupamento dos objetos para se ter conhecimento dos grupos (rótulos). Associativa faz-se a relação entre os objetos da base (CASTRO E FERRARI, 2016).

Nesse artigo a aplicação de mineração de dados se dará em uma base de dados de taxa de índice de suicídio entre os anos de 1985 e 2016 em um grupo selecionado de 100 países. Isso porque, segundo Mello et al (2005), os dados disponibilizados pelo Departamento de Informática do Sistema Único de Saúde (DATASUS) a taxa de crescimento de suicídios entre 1980 e 2000 foi de 21% para a população brasileira em geral, sendo que, os homens concretizaram mais suicídios que mulheres, cerca de 2,3 a 4 vezes mais ao longo desses 20 anos.

Com base na estimativa fornecida pela Organização Mundial da saúde – OMS, revela que cerca de uma pessoa morre por suicídio a cada 20 segundos, sendo que ocorre uma tentativa a cada 1 a 2 segundos (BERTOLOTI, M. E FLEISCHMENN, 2002).

Um dos primeiros autores a escrever sobre o suicídio foi o Karl Marx, em 1846, onde ele argumenta que o suicídio é um sintoma de uma organização social deficiente, se

tornando mais evidente em momentos de crise, embora todas as classes sociais estejam sujeitas às deficiências da sociedade, a miséria é maior causa de suicídio (GUIMARÃES, 2012).

Esse fato pode ser observado na figura 1, que foi gerado de acordo com a base de dados utilizada nesse estudo, onde no intervalo entre 1985 e 2015, a taxa de suicídio diminuiu de acordo com o crescimento do PIB (Produto Interno Bruto) dos países utilizados na amostragem.

Já na figura 2, foi-se gerado o mesmo gráfico acima, sendo que nesse, encontra-se apenas os dados do Brasil, no entanto, o mesmo padrão se repete, quando há diminuição da renda per capita, aumenta-se da taxa de suicídio.

Segundo Durkheim (2004), a pobreza é um fator protetor ao suicídio, pois a situação econômica faz com que os indivíduos não tenham grandes aspirações e já estão acostumados aos limites externos, durante uma crise

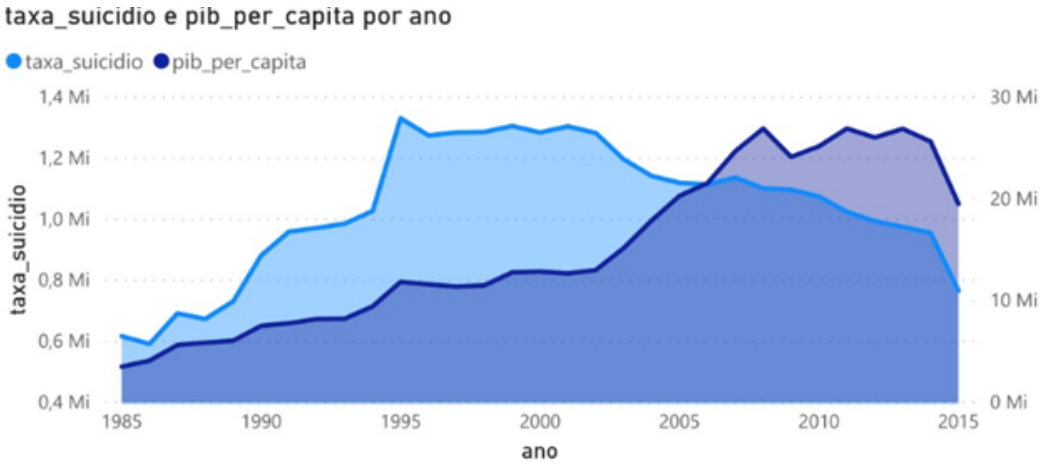

Figura 1: Taxa de Suicídio e Renda Per Capita entre 1985 e 2015

Figura 2: Taxa de Suicídio (Brasil) e Renda Per Capita entre 1985 e 2015

econômica os indivíduos ricos são mais afetados por terem os seus limites alterados, trazendo um desequilíbrio emocional.

Em comparação das taxas entre os sexos, Durkheim (2004) afirma que eles reagem de formas diferentes aos estímulos e expectativas sociais, sendo as mulheres menos afetadas pela sociedade e o homem necessitar de ponto de equilíbrio moral externo. Alguns indicadores possíveis para diferença entre as taxas de suicídio são: alcoolismo, religiosidade, divórcio e não busca de apoio mental e/ou físico.

Na figura 3, é demostrado a diferença da taxa de suicídio entre homens e mulheres, no intervalo entre 1985 e 2015, sendo a linha em vermelho a taxa do PIB per capita.

Na abordagem da classe de idade, no panorama mundial a taxa

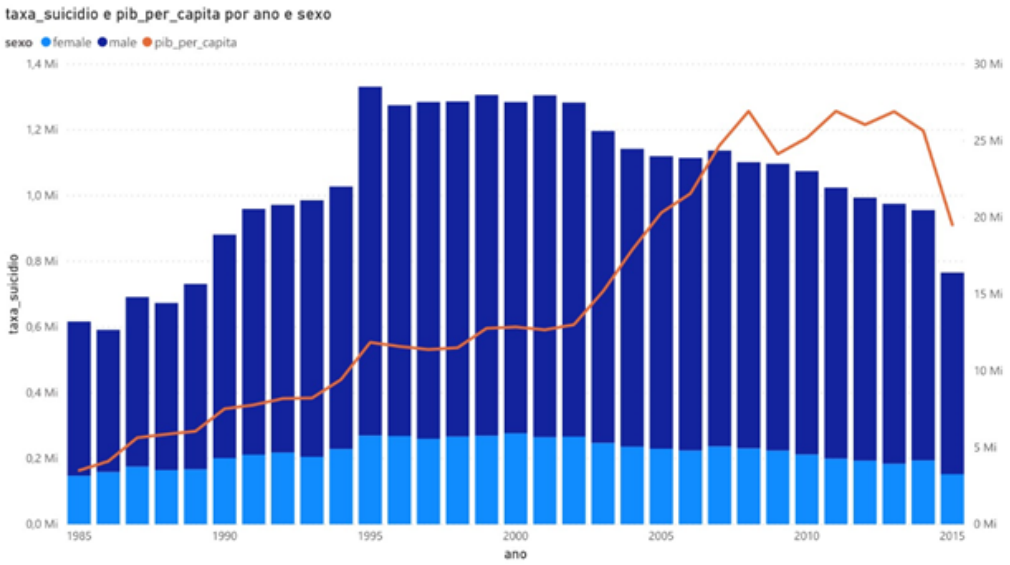

Figura 3: Taxa de Suicídio Agrupada por Sexo entre 2985 e 2015

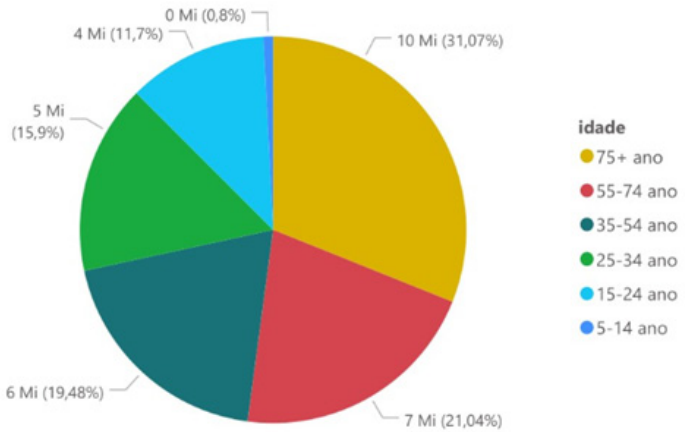

Figura 4: Taxa de Mortalidade por Sexo

de suicídio em idosos é cerca de 6 a 8 vezes maior do que em pessoas jovens, onde pode ser explicado pelo sofrimento e cansaço (característico da idade), acentuado pelas doenças, viuvez, isolamento social, solidão e a dependência física, Guimarães (2012). Na figura 4, demostra-se de acordo com a base de estudo que 31% das pessoas que cometeram suicidio eram maior que 75 anos, sendo a porcenagem regredida de acordo com intervalo de idade diminuir.

De acordo com a OMS, a melhor estratégia contra o suicídio é a prevenção, sendo recomendado como eixo central na assistência à saúde familiar. Assim tomando algumas medidas preventivas, como: limite de acesso a arma de fogo, cuidado especial com os portadores de deficiência física ou mental, alcoólatras, usuários de drogas, entre outros. Com isso, o ato suicida é um grave problema de saúde pública, mas também um grave problema econômico e social, e por tanto político, não podendo sua prevenção ser abordada apenas pelo viés medicamentoso ou psicoterápico (JUNIOR, 2015).

Devido a isso, nesse trabalho será realizado uma análise comparativa entre os algoritmos J48 e Naivebayes, com foco na previsão de suicidas em relação ao grupo de idade e ao sexo, assim obtendo a informação do melhor classificador para essa amostragem de dados e uma análise associativa dos atributos através do uso do algoritmo Apriori, assim obtendo uma análise das caraterísticas com maior relevância. Com isso, esses dados preditivos podem ser usados para o auxílio da classificação do grupo de maior risco e as caraterísticas influenciadoras.

O algoritmo J48 surgiu após a recodificação do algoritmo C4.5, com a finalidade de construir uma arvore de decisão baseado em um conjunto de dados de treinamento, demostra-se adequada para conjuntos de dados de diferentes formas de apresentação, ou seja, envolvendo atributos qualitativos, contínuos e discretos, até mesmo preenchendo algumas faltas de dados (VIEIRA, 2018) (LIBRELOTTO E MOZZAQUATRO, 2013).

Já o algoritmo NaiveBayes foi desenvolvido a partir da teoria de bayes, cujo o qual recebeu o nome do matemático inglês Thomas Bayes (1701 – 1761), que descreve a probabilidade que evento acontece (classe). Assim aprendendo a probabilidade condicional do valor da classe, com base na tupla (atributos) (TOLEDO, 2017).

Um dos algoritmos mais conhecidos pela extração de regras de associação através da análise de suas características é o apriori, onde o mesmo relaciona uma caraterística (transação) em relação à outra, possuindo assim a obtenção de algum conhecimento, o mesmo parte do mesmo princípio da teoria de bayes, no entanto, o seu objetivo é a associação das caraterísticas mais relevantes, enquanto o naivebayes, classifica e/ou prediz uma classe (AGRAWAL, 1994) (TOLEDO, 2017).

2. ALGORITMOS

Nessa sessão serão descritas

as principais caraterísticas dos algoritmos que serão utilizados nesse trabalho.

2.1. J48

O algoritmo J48 foi desenvolvido por Ross J. Quinlan, através da recodificação do C4.5, onde é construído uma arvore de decisão a partir dos dados usados no treinamento. Sendo um dos algoritmos mais populares de inferência, devido a facilidade da extração das regras de "se, então", e pela facilidade em lidar tanto com atributos contínuos e discretos, quanto com valores categóricos e valores ausentes (CAMARGO, 2016).

A arvore de decisão é composta por uma cadeia de nós conectados desde a raiz até as folhas, tendo como estratégia "dividir para conquistar", tendo como conceito a redução da complexidade caso um determinado nó seja escolhido para fazer a partição dos dados (VIEIRA, 2018).

O tratamento dos atributos se dá pela ordenação crescente dos atributos e com isso, a escolha do atributo que favorece a redução da entropia, através da identificação da classe de treinamento, após o cálculo a informação esperada. Por último, o ganho da informação para escolha do nó raiz que normaliza um conjunto de amostra de atributos que apresentam grandes variações, suavizando a escolha, certificando que a melhor escolha foi realizada (QUINLAN, 1993) (VIEIRA, 2018).

O processo de construção pode ser dividido em duas fases, o crescimento, sub composto pelo treinamento e teste; e a poda. Na fase de crescimento os dados são divididos em grupos, podendo ser direcionados para treinamento e aprendizado da estrutura, e ainda para o teste que identifica a capacidade preditiva da árvore. No processo de construção das arvores podem surgir ruídos (erros) acarretando um problema conhecido como sobre ajuste, um aprendizado muito específico, ou seja, muitos exemplos de uma classe, não permitindo uma generalização adequada (VIEIRA, 2018).

Visando a resolução desse problema, são aplicados os métodos de poda, que podem ser descritas como poda em tempo real ou poda por redução de erros, na poda em tempo real fornecem uma estimativa na presença de conjuntos ruidosos realizado durante o treinamento, na poda por redução de erros é avaliada a taxa de erros da arvore em um conjunto de casos separados (folds), ambos tendo como objetivo melhorar a taxa de acerto (VIEIRA, 2018).

Com isso, pode-se observar que capacidade preditiva dos algoritmos que possui como base a arvore de decisão é baseado na maneira como ela é construída, as características da base de dados podem gerar estruturas inteligíveis e aplicáveis ou gerar resultados desfavoráveis (VIEIRA, 2018).

2.2. *NaiveBayes*

Baseado no teorema de bayes, o teorema foi criado pelo matemático inglês, chamado Thomas Bayes (1701 – 1761), esta categorização tem diversas aplicações na área de Recuperação de Informação, tais como detecção

de SPAM, organização automática de e-mails, identificação de páginas com conteúdo adulto e detecção de expressões multipalavras, (TOLEDO, 2017) (GIANCARLO, 2013).

Segundo Oliveira (2000), o teorema de Bayes é construído utilizando dados de treinamento para estimar a probabilidade que um elemento aparece, equação abaixo representa o algoritmo da probabilidade.

$$P(C = c_i \,|x) = \frac{P(x \,|C = c_i) \times P(C = c_i)}{P(x)}$$

Onde x representa um vetor de termos e c_i representa uma classe.

2.3. Apriori

As organizações têm armazenado uma grande quantidade de dados a respeito seus negócios nas últimas décadas, sendo boa parte não extraído nenhum conhecimento. O algoritmo apriori possui o objetivo de obter o conhecimento através de extração de regras de associação, denominadas como regras ou padrões das transações comerciais de uma organização (AGRAWAL, 1994).

Para obtenção do objetivo o algoritmo visa relacionar as transações (itens) da seguinte forma X (antecedente) => Y (descendente), onde as regras devem atender ao suporte e confiança mínimo especificados, o mesmo pode ser observado na figura 1 abaixo. Define-se o suporte como a frequência que a transação ocorre em toda a base, ou seja, porcentagem em que o item se repete na base. Já a confiança define-se por ser a força usada para determinar a porcentagem de uma transação sobre outra transação, ou seja, para a transação X infere-se que a transação Y aconteça a quantidade mínima de vezes determinada (ROMÃO, 1999) (AGRAWAL, 1994).

O algoritmo é subdivido em duas sub-rotinas, na primeira conhecida como *Apriori_gen*, calcula-se o suporte para cada item individualmente, sendo selecionados para segunda etapa apenas o que atendem ao suporte mínimo. Já na segunda etapa chamada de *Genrules*, são extraídas as regras de associação para os itens mais frequentes, ou seja, produz-se uma regra a ⇒ (l - a) somente se a razão (suporte(l)/suporte(a)) for maior que a confiança mínima estabelecida, a figura 5 representa extração de regras (VASCONCELOS, 2004) (AGRAWAL, 1994).

A seguir apresenta-se o algoritmo apriori na figura 6, aonde Fk - conjunto de itens frequentes de tamanho k (conjunto com k elementos) que atende o suporte mínimo estabelecido e Ck - Conjunto de itens candidatos de tamanho k, (VASCONCELOS, 2004) (ROMÃO, 1999).

Figura 5: Representação do Apriori

```
F1 ←{Conjuntos de itens freqüentes de tamanho 1} /* Na
    primeira passagem k = 1                      */
1  para k = 2; Fk−1 != vazio; k++ faça
   |  /* Na segunda passagem k = 2                */
2  |  Ck ← apriori-gen (Fk−1) /* Novos candidatos */
3  |  para todo transação t ∈ T faça
4  |  |  Ct ← subconjunto (Ck, t) /* Candidatos contidos
   |  |      em t                                 */
5  |  |  para todo candidato c ∈ Ct faça
6  |  |  |  c.contagem++
7  |  |  fim
8  |  |  Fk ← {c ∈ Ck|c.contagem ≥ MinSup}
9  |  fim
10 fim
11 Resposta F ← Reunião de todos os Fk
```

Figura 6: Algoritmo Apriori

Para melhor entendimento o seguinte algoritmo percorre-se todos os itens até o final da lista na linha 1. Na linha 2, ativa-se a função do apriori-gen (cálculo de suporte), da linha 3 a 9 percorre-se todas as transações, sendo na linha 4 a sub-rotina chamada de subconjunto (Genrules) que é extraída as regras de associação e da linha 5 a 7 realiza-se a contagem do conjunto de regras para o filtro usado na linha 8 que selecionara as regras com suporte mínimo superior ao contador. E por último a linha 11 armazena a regra de associação (VASCONCELOS, 2004) (ROMÃO, 1999) (AGRAWAL, 1994).

3. METODOLOGIA

Esse artigo visa realizar uma análise comparativa entre os algoritmos J48 e Naivebayes, assim identificando o algoritmo que possui uma menor margem de erro para o base de dados utilizada nessa pesquisa, sendo assim o mais adequado para realização desse estudo, além de uma análise dos atributos através das regras de associação do algoritmo Apriori.

A base de dados utilizada nessa pesquisa, foi retirada do site Kaggle[1] onde possui 27820 instancias (taxas), de cem países diferentes entre os anos de 1985 a 2016. Fornecendo nelas um total de doze colunas, definidas como: país, ano, sexo, faixa etária, número de suicídios, população, taxa de suicídio, chave composta ano-país, IDH por ano, PIB por ano, PIB per capita, geração (com base na média de agrupamento etário).

No entanto, após serem submetidas nas fases do KDD descritas nessa sessão foram utilizadas para a pesquisa apenas 27660 instancias (taxas), entre o intervalo de data de 1985 a 2015, distribuídas em oito colunas, sendo elas, país, ano, sexo, faixa etária, população, taxa de suicídio, IDH (Índice de desenvolvimento econômico) por ano, PIB (Produto Interno Bruto) per capita.

Outro fator relevante nessa

[1]https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

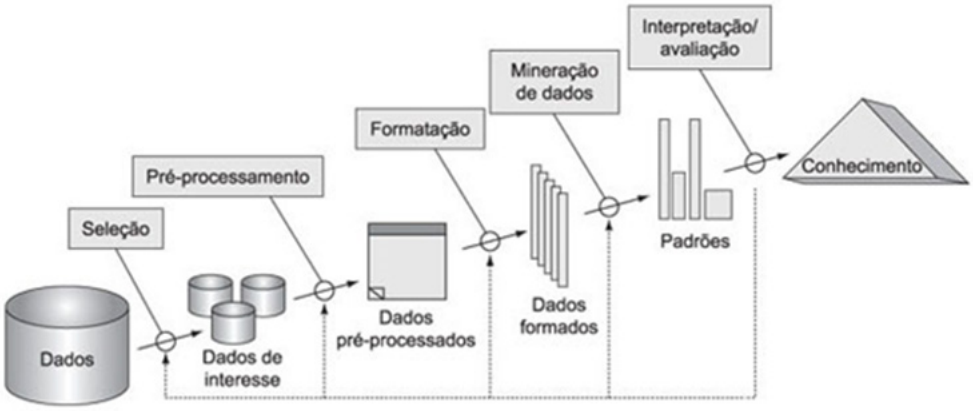

Figura 7: Fases do KDD

amostragem, é que o quantitativo de instancias (dados) para as características do sexo e idade são as mesmas, representando 13830 e 4610 instancias, respectivamente.

Segundo Goldschmidt (2005), o processo de automatização para a descoberta de padrões desconhecidos em uma base de dados se chama KDD (Knowledge Discovery in databases), onde possui as tarefas que podem ser vistas na figura 7, sendo assim consideradas as mais relevantes no processo de mineração de dados, abaixo são descritas as seguintes fases.

Seleção

A fase de seleção dos dados (dados brutos) é a primeira no processo de descobrimento de informação. Nesta fase é escolhido o conjunto de dados, pertencente a um domínio, contendo todas as possíveis variáveis e registros que farão parte da análise, ou seja, é necessário a identificação de quais atributos farão parte do processo do KDD, normalmente encontra-se em uma base de dados transacional (GOLDSCHMIDT E PASSOS, 2005) (CASTRO E FERRARI, 2016).

Nesse trabalho, os dados encontraram-se através de um arquivo ".csv", no site Kaggle, citado acima nessa mesma sessão, no entanto, foram retiradas as colunas: número de suicídios, chave composta ano-país, PIB por ano, geração. Isso porque, esses dados já existiam nas outras colunas, taxa de suicídio, ano, pais, PIB per capita, idade, respectivamente, no entanto representados de maneira diferente, torna-se obsoleto para essa pesquisa.

Pré-processamento

Nesta etapa deverão ser realizadas tarefas que eliminem dados redundantes e inconsistentes, recuperação de dados incompletos e avaliação de possíveis dados discrepantes ao conjunto. Isso porque, a qualidade dos dados possui uma grande influência nos resultados apresentados, Goldschmidt (2005).

O pré-processamento foi

removido os dados do ano de 2016, pois continham um quantitativo menor de dados em relação aos anos anteriores, dando assim um quantitativo menor na taxa de suicídio dos cem países que foram realizados a pesquisa.

Transformação

Após serem selecionados, limpos e pré-processados os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos de aprendizado possam ser aplicados. O principal objetivo é resolver problemas de unidade, escala e padronização de formatos. Sendo parte nessa fase, a normalização dos dados para alguns algoritmos com a rede neural ou métodos baseados em distância (CASTRO E FERRARI, 2016).

Os dados foram ajustados de forma que podem ser abertos na ferramenta WEKA, que na qual irá ser aplicadas as técnicas de mineração de dados. Para os algoritmos J48 e NaiveBayes não foram realizado transformação na base de dados, no entanto, para o algoritmo Apriori foram necessários alguns ajustes.

Para aplicação no algoritmo apriori, os atributos numéricos (população, taxa de suicídio, IDH do ano e PIB per capita) da base de dados foram transformados em atributos nominais, divididos em quatro grupos, sendo eles, vazio, baixa, media e grande. Os atributos sem valores foram classificados como vazio, e os que possuíam valores foi-se realizado separação em três grupos de acordo com o maior valor da classe, a primeira um terço, como baixo, o segundo terço, como médio, e o terceiro terço, como grande.

Mineração de dados (data mining)

A mineração de dados é a principal fase do KDD, refere-se à capacidade de determinados algoritmos tem de aprender a partir dos exemplos fornecidos. Ocorrendo a busca efetiva com conhecimento novos e uteis a partir dos dados, procurando abstrair um conhecimento. Para cada tipo de conhecimento que se deseja obter, existe um algoritmo diferente que pode ajudar a alcançar o objetivo final, como por exemplo, a classificação, predição ou clusterização.

Nesse trabalho foi utilizado duas técnicas diferentes de classificação, visando a predição do grupo de pessoa suicida, através do uso da ferramenta WEKA versão 3.8. Para isso, foram realizados dois grupos de experimento, no primeiro grupo a predição será realizada de acordo com o sexo do suicida, ou seja, predizer o sexo da pessoa de acordo com as características da base de dados, e a outra, em relação a faixa etária, sendo esta dívida em seis grupos distintos. Ambos os grupos, foram aplicados nos algoritmos J48 e Naivebayes. Tendo como objetivo a análise do comportamento do algoritmo em relação as duas características do indivíduo.

Para análise das relações entre os atributos, regras de associação, foi-se utilizado o algoritmo apriori, com três experimentos distintos. Sendo geradas 10 regras com a confiabilidade (porcentagem que uma ocorrência se repete) de 1, 0.9 e 0.7.

Avaliação

Onde as regras indicadas pelo processo anterior serão interpretadas e avaliadas. Após a interpretação poderão surgir padrões, relacionamentos e descoberta de novos fatos, que podem ser utilizados para pesquisas, otimização e outros.

4. RESULTADOS

Através da leitura da base na ferramenta de mineração de dados (Weka) pode ser aplicado o algoritmo sobre a mesma e assim avaliar o resultado e obter-se o conhecimento desejado, onde foram realizados dois experimentos diferentes em relação a classificação e três experimentos em relação a regra de associação.

Classificação

Os resultados são apresentados através de matriz de confusão que representam os verdadeiros positivos e negativos e os Negativos positivos e negativos, que são exemplificados na tabela abaixo.

a	b
Positivo Verdadeiro	Positivo Falso
Negativo Falso	Negativo Verdadeiro

Tabela 1: Matriz de Confusão

O primeiro experimento foi realizado levando em consideração o sexo da pessoa, onde para o algoritmo J48 possui uma taxa de acerto de 72,28% e a matriz de confusão de acordo com a figura 8 e o algoritmo NaiveBayes com taxa de acerto de 65,87% e matriz de confusão de acordo com a figura 9, e as porcentagens desses descritas na tabela 2, abaixo.

```
=== Confusion Matrix ===

     a     b   <-- classified as
  6464  7366 |     a = Masculino
   301 13529 |     b = Feiminino
```

Figura 8: Matriz de Confusão (Sexo) J48

```
=== Confusion Matrix ===

     a     b   <-- classified as
  5418  8412 |     a = Masculino
  1028 12802 |     b = Feiminino
```

Figura 9: Matriz de Confusão (Sexo) *NaiveBayes*

SEXO	Taxa de Acerto	Taxa de Erro
J48	72,28%	27,71%
NaiveBayes	65,87%	34,12%

Tabela 2: Resultado Primeiro Experimento

O segundo experimento foi realizado utilizando os mesmos algoritmos, no entanto, o grupo preditivo foi a idade das pessoas suicidas, que na qual, foram divididas em seis grupos distintos. Para o algoritmo J48, houve uma taxa de acerto de 37,45% resultando uma matriz de confusão representada na figura 10, já para o algoritmo NaiveBayes a taxa de acerto foi de 23,3%, com uma matriz de confusão na figura 11. Ambos os resultados, podem ser comparados na tabela 3 a seguir.

```
=== Confusion Matrix ===

    a    b    c    d    e    f   <-- classified as
 1176  517  302  901  301 1413 |    a = 15-24 ano
 1017  962   86  420  979 1146 |    b = 35-54 ano
  778   33 2226  383  197  993 |    c = 75+ ano
 1355  450  319  725  528 1233 |    d = 25-34 ano
  704  975  279  734  828 1090 |    e = 55-74 ano
   77   26   13   43    7 4444 |    f = 5-14 ano
```

Figura 10: Matriz de Confusão (idade) J48

```
=== Confusion Matrix ===

    a    b    c    d    e    f   <-- classified as
  252  273 1775  705   39 1566 |    a = 15-24 ano
 1509  409 1172  190  131 1199 |    b = 35-54 ano
 1183   48 1445  358  129 1447 |    c = 75+ ano
 1702  251 1181    3  125 1348 |    d = 25-34 ano
 1371  296 1433  209   23 1278 |    e = 55-74 ano
  122  144   29    2    0 4313 |    f = 5-14 ano
```

Figura 11: Matriz de Confusão (idade) *NaiveBayes*

IDADE	Taxa de Acerto	Taxa de Erro
J48	37,45%	62,54%
NaiveBayes	23,30%	76,69%

Tabela 3: Resultando Segundo Experimento

Associação

Os resultados são apresentados através de regras que são inferidas por um grau de confiabilidade de cada experimento, onde podem ser alterados através do uso da ferramenta WEKA.

No primeiro experimento, foram geradas as regras da tabela 4, compostas por dez regras de associação com a confiabilidade 1 (100%) e suporte em 0.1, que significa que em um conjunto é inferido 100% do atributo resultante. Como por exemplo, na tabela 4, pode ser observado que todas as vezes em que a idade for maior que 15 anos a população é considerada baixa, ou ainda, quando o sexo for feminino e população for baixa e o índice de desenvolvimento for grande a taxa de suicídio é considerada baixa.

No segundo experimento, regras

Nº	Regras de Associação	Confiança
1	idade=75+ ano 4610 ==> populacao=Baixa 4610	1
2	idade=5-14 ano 4610 ==> taxa_suicidio=Baixa 4610	1
3	idade=5-14 ano populacao=Baixa 4486 ==> taxa_suicidio=Baixa 4486	1
4	idade=75+ ano taxa_suicidio=Baixa 4247 ==> populacao=Baixa 4247	1
5	idade=75+ ano pib_per_capita=Baixa 4114 ==> populacao=Baixa 4114	1
6	idade=5-14 ano pib_per_capita=Baixa 4114 ==> taxa_suicidio=Baixa 4114	1
7	idade=5-14 ano populacao=Baixa pib_per_capita=Baixa 4016 ==> taxa_suicidio=Baixa 4016	1
8	idade=75+ ano taxa_suicidio=Baixa pib_per_capita=Baixa 3772 ==> populacao=Baixa 3772	1
9	sexo=Feiminino idh_ano=Grande 3474 ==> taxa_suicidio=Baixa 3474	1
10	sexo=Feiminino populacao=Baixa idh_ano=Grande 3376 ==> taxa_suicidio=Baixa 3376	1

Tabela 4: Primeiro Experimento

N°	Regras de Associação	Confiança
1	populacao=Baixa 26978 ==> taxa_suicidio=Baixa 26435	0.98
2	taxa_suicidio=Baixa pib_per_capita=Baixa 24150 ==> populacao=Baixa 23632	0.98
3	pib_per_capita=Baixa 24684 ==> populacao=Baixa 24154	0.98
4	populacao=Baixa pib_per_capita=Baixa 24154 ==> taxa_suicidio=Baixa 23632	0.98
5	pib_per_capita=Baixa 24684 ==> taxa_suicidio=Baixa 24150	0.98
6	populacao=Baixa idh_ano=Vazio 18816 ==> taxa_suicidio=Baixa 18406	0.98
7	idh_ano=Vazio 19296 ==> taxa_suicidio=Baixa 18874	0.98
8	taxa_suicidio=Baixa 27105 ==> populacao=Baixa 26435	0.98
9	taxa_suicidio=Baixa idh_ano=Vazio 18874 ==> populacao=Baixa 18406	0.98
10	idh_ano=Vazio 19296 ==> populacao=Baixa 18816	0.98

Tabela 5: Segundo Experimento

N°	Regras de Associação	Confiança
1	populacao=Baixa 26978 ==> taxa_suicidio=Baixa 26435	0.98
2	taxa_suicidio=Baixa pib_per_capita=Baixa 24150 ==> populacao=Baixa 23632	0.98
3	pib_per_capita=Baixa 24684 ==> populacao=Baixa 24154	0.98
4	populacao=Baixa pib_per_capita=Baixa 24154 ==> taxa_suicidio=Baixa 23632	0.98
5	pib_per_capita=Baixa 24684 ==> taxa_suicidio=Baixa 24150	0.98
6	taxa_suicidio=Baixa 27105 ==> populacao=Baixa 26435	0.98
7	pib_per_capita=Baixa 24684 ==> populacao=Baixa taxa_suicidio=Baixa 23632	0.96
8	populacao=Baixa 26978 ==> pib_per_capita=Baixa 24154	0.90
9	populacao=Baixa taxa_suicidio=Baixa 26435 ==> pib_per_capita=Baixa 23632	0.89
10	taxa_suicidio=Baixa 27105 ==> pib_per_capita=Baixa 24150	0.89

Tabela 6: Terceiro Experimento

geradas na tabela 5, é composta por dez regras de associação com a confiabilidade de 0.9 e suporte de 0.65, que normalmente é a configuração padrão da ferramenta. Nessa, foram retornadas regras como por exemplo, quando a população for baixa a taxa de suicídio é baixa.

Já no terceiro experimento, foram geradas dez regras da tabela 6, sendo os parâmetros de confiabilidade informados de 0.7 com suporte de 0.85. Resultando regras iguais até a sexta regra, sendo alteradas as regras posteriores.

5. CONSIDERAÇÕES FINAIS

Com base nesse estudo pode-se observar que o aumento dos atributos classificatórios entre o sexo e a idade reduziram significativamente a taxa de acerto de ambos os algoritmos com redução de cerca de 50% dos acertos. Isso porque, a discrepância entre pessoas de sexo diferentes é muito maior do que a discrepância entre as idades.

Na matriz de confusão da característica sendo o sexo, o algoritmo J48 (figura 6) possui uma margem de erro de classificação (Positivo Falso) do sexo masculino maior do que o feminino, com uma diferença de 27,71% para 34,12%. O NaiveBayes (figura 7), segue no mesmo padrão de erros, no entanto, com uma margem maior de 62,54% para 76,69%.

Já na matriz de confusão da característica da idade, o algoritmo J48 (figura 8) possui uma classificação de Positivo falso e Negativo falso muito maior, sendo os grupos com resultados melhores entre a faixa etária, de 5 a 14 anos e maior de 75 anos, respectivamente. No algoritmo NaiveBayes (figura 9), as melhores faixas etárias são as mesmas do algoritmo J48, só que com uma margem de erro maior.

De acordo com a analisado pode-se inferir que para essa amostragem de dados o algoritmo J48 apresentou um melhor resultado em relação ao NaiveBayes.

A análise das regras de associação com maior confiabilidade, ou seja, confiança de 1.0, pode-se observar que a idade nos intervalos entre 5 a 14 anos e maior que 75 anos são grandes influenciadoras nas regras geradas, sendo exibida 8 das 10 regras, juntamente com a classificação da população e PIB per capita. No experimento, segundo e terceiro, as seis primeiras regras são idênticas, onde os atributos do PIB per capita e população possuem influência maior em relação ao atributo da taxa de suicídio, onde por exemplo quando a confiabilidade de 98%, o PIB é baixo, a renda per capita é baixa, e consequentemente a taxa de suicídio é baixo, tendo em vista que a classificação dos atributos baixa, médio ou alto é em relação aos dados em geral.

As regras de associação que divergem, alteração a quantidade de ocorrências iguais, no entanto, possui os mesmos atributos com alta influência nas regras de associação.

Através do conhecimento obtido, identifica-se os atributos com maiores ocorrências nas regras de associação, onde um deles (idade) foi identificado

um resultado de classificação mais bem apresentado pelo algoritmo J48. Com isso, pode ser realizado um estudo mais profundo nos países e/ou faixa etária que estão classificados com essas caraterísticas e assim realizar medidas preventivas ao combate ao suicídio.

6. REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases. - Santiago, Chile, 12-15 de September de 1994. - pp. 487-499.

BERTOLOTE J. M. E FLEISHMANN A. A global perspective in the epidemiology of suicide. Suicidology. 2002. - pp. 6-7.

CAMARGO, A. et al Mineração de dados eleitorais: descoberta de padrões de candidatos a vereador na região da campanha do Rio Grande do Sul. Revista Brasileira de Computação Aplicada. 2016. - pp. 64-73.

CASTRO, LEANDRO NUNES E FERRARI, DANIEL GOMES Introdução à mineração de dados: Conceitos básicos, algoritmos e aplicações. - São Paulo : Saraiva, 2016.
DURKHEIM, ÉMILE O suicídio. - São Paulo. Martins Fontes, 2004.

GOLDSCHMIDT, RONALDO; PASSOS, EMANUEL Data Mining, um guia prático. 2005.

GUIMARÃES, TATIANA Suicídio e Ocupação: Um Estudo Comparado. 2012.

JUNIOR, AVIMAR FERREIRA O comportamento suicida no Brasil e no mundo. 2015.

LIBRELOTTO S. R.; MOZZAQUATRO P. M. Análise dos Algoritmos de Mineração J48 e Apriori Aplicados na Detecção de Indicadores da Qualidade de Vida e Saúde. 2013.

LUCCA, GIANCARLO, et al Uma implementação do algoritmo Naïve Bayes para classificação de texto. ERBD - Escola Regional de Banco de Dados. 2013.

MELLO, SANTOS C. D.; BERTOLOTE, J. M.; WANG, Y. P. Epidemiology of suicide in Brazil (1980-2000): characterization of age and gender rates of suicide. Revista Brasileira de Psiquiatria. 2005. pp. 131-134.

OLIVEIRA, GRINALDO; MENDONÇA, MANOEL ExperText: Uma Ferramenta de Combinação de Múltiplos Classificadores Naive Bayes. 2000.

QUINLAN, J. R. C4.5: Programing for machine learning, Morgan Kauffmann. 1993.
REZENDE, S. O. Sistemas Inteligentes

Fundamentos e Aplicações. Manole, 2003.

ROMÃO, W. Extração de Regras de Associação Em C&T: O Algoritmo Apriori. 1999.

TOLEDO A. P.; CABRERA, J. D. L.; DOMÍNGUEZ, L. A. Q. Revista Cubana de Ciências Informáticas. 2017.

VASCONCELOS L.; CARVALHO, C. Aplicação de Regras de Associação para Mineração de Dados na Web. 2004.

VIEIRA, et al Avaliação da performance do algoritmo J48 para construção de modelos baseados em árvores de decisão. Revista Brasileira de Computação Aplicada. 2018. - 2 : Vol. 10. - pp. 80-90.