UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

STEPHANIE LURI KACUTA

UTILIZAÇÃO DA METODOLOGIA KDD PARA DESCOBERTA DE CONHECIMENTO EM DADOS RELACIONADOS A TOXICODEPENDÊNCIA E IDEAÇÃO SUICIDA

LONDRINA

2021

STEPHANIE LURI KACUTA

UTILIZAÇÃO DA METODOLOGIA KDD PARA DESCOBERTA DE CONHECIMENTO EM DADOS RELACIONADOS A TOXICODEPENDÊNCIA E IDEAÇÃO SUICIDA

THE USE OF KDD METHODOLOGY TO DISCOVER KNOWLEDGE IN DATA RELATED TO DRUG DEPENDENCE ANDSUICIDE IDEATION

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador: Dr. Bruno Samways dos Santos.

LONDRINA
2021

STEPHANIE LURI KACUTA

UTILIZAÇÃO DA METODOLOGIA KDD PARA DESCOBERTA DE CONHECIMENTO EM DADOS RELACIONADOS A TOXICODEPENDÊNCIA E IDEAÇÃO SUICIDA

Trabalho de Conclusão de Curso de Graduação para obtenção do título de Bacharel em Engenharia de produção da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 02 de dezembro de 2021

_____

Bruno Samways dos Santos
Doutor
Universidade Tecnológica Federal do Paraná

_____

Rafael Henrique Palma Lima
Doutor
Universidade Tecnológica Federal do Paraná

_____

Pedro Rochavetz de Lara Andrade
Doutor
Universidade Tecnológica Federal do Paraná

RESUMO

A ideação suicida é uma questão de extrema importância que afeta todas as raças e países. Atualmente, há lacunas de estudos de aplicação da metodologia KDD relacionando à influência da toxicodependência na ideação suicida. O seguinte trabalho tem como objetivo a descoberta de conhecimento em dados utilizando técnicas para a predição de ideação suicida. Para o aprendizado de máquina aplicaram-se as técnicas de classificação Árvore de Decisão e Floresta Aleatória, e para partição de dados, o método de *k-fold* com 10 subconjuntos. A base de dados utilizada é uma base americana pertencente à Pesquisa Nacional sobre Uso de Drogas e Saúde (*National Survey on Drug Use and* Health - NSDUH), foram exploradas 5.055 instâncias e 19 atributos pertencentes às dimensões demográficas, saúde mental e toxicodependência, dados do ano de 2019 e 2020. Três experimentos foram realizados, todos aplicando as duas técnicas de classificação a fim de analisar os diferentes comportamentos do modelo, alternando as dimensões dos atributos. O melhor desempenho foi encontrado no experimento 1 com o conjunto de dados contendo todas as dimensões, aplicando-se a técnica de Árvore de decisão, a qual apresentou 63,7% de acurácia, 55,2% de precisão e 49,1% de *recall*. Não foi possível observar uma influência considerável relacionada à toxicodependência no conjunto de dados. Em maior parte dos resultados dos experimentos, a técnica que teve melhor desempenho foi a Árvore de decisão. Porém, constatou-se que a porcentagem de predição correta ainda foi baixa, abrindo oportunidades para futuros trabalhos na área com a análise de novos parâmetros, técnicas e pré-processamento das informações que possam melhorar a predição.

Palavras-chave: ideação suicida; toxicodependência; aprendizado de máquina; classificação.

ABSTRACT

Suicidal ideation is an extremely important issue that affects all races and countries. Currently, there are gaps in studies on the application of the KDD methodology relating to the influence of drug addiction on suicidal ideation. The following work aims to discover knowledge in data using techniques for the prediction of suicidal ideation. For machine learning, the Decision Tree and Random Forest classification techniques were applied, and for data partition, the k-fold method with 10 subsets. The database used is an American database belonging to the National Survey on Drug Use and Health (NSDUH), 5,055 instances and 19 attributes belonging to the demographic dimensions, mental health and drug addiction were explored, data from the year 2019 and 2020. Three experiments were carried out, all applying the two classification techniques in order to analyze the different behaviors of the model, alternating the dimensions of the attributes. The best performance was found in experiment 1 with the dataset containing all dimensions, applying the Decision Tree technique, which showed 63.7% accuracy, 55.2% precision and 49.1% recall. It was not possible to observe a considerable influence related to drug addiction in the dataset. In the majority of the experiments results, the technique that performed best was the Decision Tree. However, it was found that the percentage of correct prediction was still low, opening up opportunities for future work in the area with the analysis of new parameters, techniques and pre-processing of information that can improve prediction.

Keywords: suicidal ideation; drug addiction; machine learning; classification.

SUMÁRIO

1. INTRODUÇÃO

O avanço da tecnologia levou ao crescimento de diversas áreas, dentre elas a de informações. De acordo com Somasundaram e Shrivastava (2011), devido ao aumento de dispositivos que podem gerar conteúdos, um único indivíduo pode chegar a produzir mais informações do que o volume criado por empresas.

A cada instante, inúmeros bancos de dados são alimentados, o que anteriormente era armazenado em forma física em um caderno de anotações, hoje com a tecnologia se tornou algo prático e dinâmico armazenar dados eletronicamente. Segundo Han *et al*. (2012), a abundância de dados e o seu crescimento rápido acabam se tornando em numerosos repositórios de dados, excedendo a capacidade humana de compreensão e muitas vezes se tornando apenas arquivos de dados sem utilização.

A mineração de dados pode ser vista como resultado da evolução natural da tecnologia da informação, no qual pode transformar um conjunto de dados em “pepitas de ouro” de conhecimento (HAN *et al*. 2012).

A descoberta de conhecimento em bancos de dados (*Knowledge Discovery in Databases* - KDD), é uma metodologia dentro da área de análise de dados que é capaz de identificar padrões não triviais em uma grande quantidade de dados. Bases grandes para serem trabalhados necessitam de um tratamento, um estudo aprofundado para que seja capaz de gerar informações úteis. Bases dos diversos campos existentes como finanças, vendas, saúde podem gerar informações valiosas não disponíveis de forma trivial (FAYYAD *et al.* 1996).

O KDD sendo uma metodologia aplicada para extrair conhecimento em bancos de dados sem restrição, ou seja, de todas as áreas possíveis que possuam um conjunto de dados, torna-se uma ferramenta importante, principalmente ao analisar dados da área da saúde por exemplo, onde estudos com junção de áreas podem trazer benefícios em prol do ser humano.

Segundo o relatório Estimativas Globais de Saúde (2019), realizado pela Organização Mundial da Saúde (OMS), cerca de 800 mil pessoas morrem todos os anos por suicídio no mundo. Embora o número de mortes seja alarmante, o total de países com estratégias para enfrentar o problema ainda é baixo.

A ideação suicida é um assunto delicado e complexo, pois envolve diversas dimensões que devem ser consideradas, e infelizmente o número de pessoas com doenças psicológicas que podem influenciar nesta questão tem aumentado. De acordo com Chachamovich *et al*. (2009), entre os sintomas que podem aumentar a chance de suicídio inclui depressão, dependência química, ansiedade grave, crises de pânico, agitação e insônia. Gonçalves (2015) afirma que quando o assunto tratado é suicídio, a relação que é criada é de que se trata de saúde mental, porém, dados tem mostrado que o comportamento suicida também está fortemente ligado aos distúrbios de dependência química, reconhecidamente para o álcool e outras drogas.

Gonçalves (2015) cita que, por algum motivo, a taxa de suicídio tem aumentado entre os jovens, especialmente em países desenvolvidos. Os dados sobre a quantidade de mortes por suicídio podem incluir mortes acidentais, como por exemplo a overdose por diversos tipos de drogas. Outro motivo da diferença entre um país e outro pode estar relacionado à sua cultura e ao ambiente social a que os cidadãos estão inseridos.

O suicídio tem afetado cada vez mais a população, além disso é um grande problema dentro da saúde pública. Silva et al. (2020), afirma que considerando transtornos de humor e ansiedade, mais de 25% da população em geral será afetada por alguma doença mental em sua vida. Acrescenta dizendo que a saúde mental é uma das chaves para sobreviver a esta última pandemia, considerando tudo que ela acarreta a curto, médio e longo prazo.

A identificação de um fator de risco a ideação suicida, que é um preditor do suicídio, pode ajudar a evitar o ser humano a chegar na tentativa ou até mesmo a realizar o ato. Não somente o assunto da base de dados, o processo que será utilizado para analisá-lo também é de extrema importância e a mineração de dados é uma área que tende a ser cada vez mais utilizada, pois ao final do estudo, informações não triviais podem ser encontradas.

A facilidade de coleta e armazenamento de dados citada anteriormente traz oportunidades de estudos em inúmeras áreas. Um destes exemplos recentes é o estudo Gradim *et al.* (2020), o qual utilizou informações de uma rede social pública (*Twitter*) para analisar postagens sobre suicídio relacionadas à comunidade LGBTQ. Para este tipo de estudo, dados não estruturados são pré-processados primeiramente, porém, conjuntos de dados estruturados com um corte transversal também podem ser explorado para este fim.

Alguns países como os Estados Unidos investem em pesquisas para coletar dados da população a respeito da saúde. A Pesquisa Nacional sobre Uso de Drogas e Saúde (*National Survey on Drug Use and* Health - NSDUH) é de origem americana e realiza anualmente suas pesquisas, sendo que fatores como saúde mental e o uso de drogas são o seu foco.

Visto que, a possibilidade de o uso de drogas estar relacionado à ideação suicida é um assunto de suma importância para a saúde pública, metodologias de análises e interpretação de dados, como o KDD, podem ser úteis na extração de informações pertinentes do banco de dados.

1.1 Objetivos

O objetivo geral da presente pesquisa é extrair informações úteis em um conjunto de dados norte americano de ideação suicida, utilizando a metodologia KDD com a tarefa de mineração de dados baseada em classificação, tendo como objetivos específicos:

- Analisar atributos relevantes que podem levar à ideação suicida;
- Pré-processar o conjunto de dados para a etapa de mineração;
- Comparar as técnicas de classificação.

1.2 Justificativas

Uma análise eficaz de dados pode trazer benefícios além do que se espera, podendo tanto resultar em uma informação valiosa como também criar um potencial para novas oportunidades (SOMASUNDARAM; SHRIVASTAVA, 2011).

Breet *et al.* (2018) ao realizar uma revisão sistemática de estudos relacionados entre o uso de substâncias e a ideação e comportamento suicida (108 estudos do ano 2006 a 2016), destaca que maior parte dos estudos se concentrou em substância como álcool e tabaco, enquanto negligenciava o uso de substâncias como *cannabis*, opioides, sedativos, estimulantes, medicamentos prescritos (mas que eram utilizados indevidamente), inalantes e alucinógenos.

A identificação de um fator de risco a ideação suicida que é um preditor do suicídio pode ajudar a evitar o ser humano a chegar na tentativa ou até mesmo a realizar o ato. Não somente o assunto da base de dados, o processo que será utilizado

para analisá-lo também é de extrema importância e a mineração de dados é uma área que tende a ser cada vez mais utilizada, pois ao final do estudo, informações não triviais podem ser encontradas. Desta forma, o presente estudo buscou preencher lacunas encontradas em poucos estudos envolvendo aprendizado de máquina aplicada à ideação suicida a partir da dados de corte transversal.

1.3 Estrutura do trabalho

Após o presente capítulo introdutório, o restante do trabalho está organizado em mais quatro capítulos, de acordo com a descrição a seguir.

O capítulo dois corresponde à fundamentação teórica, o qual enfatiza os conceitos sobre o tema, as técnicas de mineração de dados da pesquisa, o que e como são as métricas que foram utilizadas no estudo, bem como a área em que foi aplicada. Em sequência, o capítulo três aborda a metodologia da pesquisa, onde estão definidos e especificados os métodos que foram seguidos no trabalho. Os resultados e discussões estão apresentados no capítulo quatro, onde todos os pontos relevantes descobertos com o trabalho estão discutidos. Por fim, após a discussão de todos os resultados e o andamento do trabalho, no capítulo cinco apresentam-se as conclusões finais e melhorias para os possíveis futuros trabalhos de extensão ao tema.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo discute os temas sobre a metodologia KDD (incluindo os tipos de aprendizado), as técnicas de classificação utilizadas, as métricas de avaliação, tipo de validação do modelo e a ideação suicida e seus fatores de risco.

2.1 KDD (Knowledge Discovery in Databases)

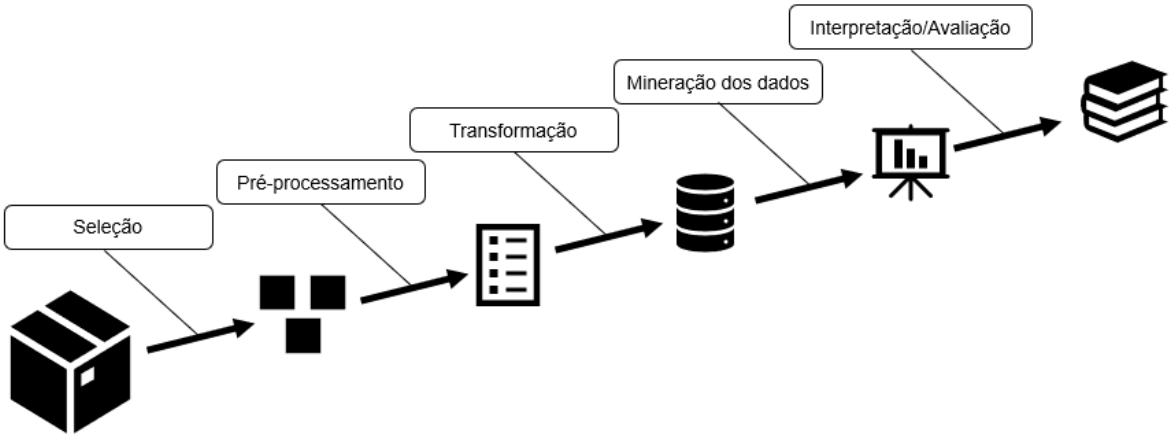
Desde a última década têm sido utilizadas técnicas de aprendizado de máquina em diversas áreas de atuação seja em áreas relacionadas à indústria ou mesmo da saúde, campo socioeconômico e finanças. A criação e utilização de banco de dados têm aumentado, as informações que antes eram apenas armazenadas, atualmente são utilizadas para descobertas com as análises de dados.

Segundo Fayyad et al. (1996, p. 39), o processo de KDD (*Knowledge Discovery in Databases)* surgiu com o objetivo principal de extrair conhecimentos de alto nível de uma grande base de dados, partindo de dados de baixo nível. O KDD consiste em um processo global de descoberta de conhecimento, desde a forma como os dados são armazenados e acessados, a como os algoritmos podem ser dimensionados de acordo com os tipos de conjuntos de dados.

Shapiro et al. (1994), resume o KDD dizendo que a sua essência é a extração de dados não triviais que anteriormente eram desconhecidas, sendo informações potencialmente úteis de dados. No processo de descoberta de conhecimento do KDD existem vários níveis de análise, números de etapas, diversas alternativas de técnicas para aplicação que podem ser utilizadas no processo e as iterações podem ser repetidas em intervalos diferentes com dados atualizados (ZHONG et al. 1997).

Na literatura são encontradas algumas definições que exemplificam as etapas do KDD. De acordo com Fayyad et al. (1996), o processo pode ser descrito nominalmente com as seguintes atividades: (i) Seleção de dados, (ii) Pré-processamento e limpeza de dados, (iii) Transformação de dados, (iv) Mineração de dados e pôr fim, (v) a fase de Interpretação e avaliação dos dados. Segue na Figura 1 a ilustração do sequenciamento de atividades por qual o conjunto de dados passa no processo de KDD e em sequência as descrições das etapas.

Figura 1 - Etapas do processo de KDD



Fonte: Adaptado Fayyad et al (1996)

A primeira etapa seria a seleção de dados, tem como objetivo criar ou selecionar um conjunto de dados de interesse para que seja utilizado no processo de descoberta de conhecimentos. Necessita-se que o conjunto de dados selecionado contenha variáveis influenciáveis para o desenvolvimento da análise, de forma que possam levar a um padrão. É importante que essa fase seja realizada com cuidado, de forma que se garanta uma base de dados de qualidade, seja um conjunto de dados existente ou uma base de dados criada.

A etapa seguinte, é a de pré-processamento que consiste na utilização de técnicas para preparar o banco de dados para a análise. Esta etapa que engloba também a limpeza de dados é responsável por filtrar dados dentro do grupo de dados alvo, a fim de eliminar ruídos encontrados na base de dados, selecionando assim apenas os dados relevantes para a próxima etapa. De acordo com Tenfen (2003), nessa etapa de limpeza o processo acaba eliminando as consultas desnecessárias que seriam feitas pelo algoritmo na fase de mineração de dados, por exemplo, a atribuição de limites em um banco de dados para que os valores desnecessários sejam desconsiderados.

Uma das técnicas que pode ser definida como análise prévia a ser utilizada nessa etapa é a análise de *outliers* ou detecção de desvios. A técnica tem como objetivo encontrar conjuntos de dados que necessitam de tratamento ou serem excluídos do banco de dados, isto se deve ao comportamento anormal do dado que pode acabar colocando em risco o objetivo inicial de estudo (CORTÊS; PORCARO; LIFSCHITZ, 2002).

A etapa 3 é a fase de transformação de dados, também chamada por alguns autores como redução e projeção de dados. Segundo Fayyad et al. (1996), nesta etapa encontram-se as características úteis para representar os dados, como podem também encontrar representações invariantes, levando em consideração o objetivo final do processo. Panwar e Raiwani (2014) comentam que as vantagens deste processo é de que os resultados são mostrados de forma compacta e fácil de compreender, onde os padrões gerais podem ser observados. Já a desvantagem dessa transformação de dados está na perda dos dados originais, sendo este um processo irreversível.

Ainda de acordo com Panwar e Raiwani (2014), a redução de dados pode reduzir custos e aumentar a eficiência de armazenamento. Existem três tipos de estratégias que podem ser utilizados para redução de dados:

Redução da dimensão: objetiva reduzir a quantidade de variáveis aleatórias em consideração ao conjunto de dados, de modo que a conversão possa transmitir mais informação em menor quantidade de dados.

Clustering: o processo de clusterização seria o agrupamento de dados, a organização de objetos em grupos cujos membros são semelhantes entre si de alguma forma.

Amostragem: a criação de um conjunto menor de dados a partir de outro maior é definido como a estratégia de amostragem. O objetivo da amostragem é selecionar um subconjunto para representar todo o conjunto de dados, onde para esse processo existem diversos tipos para se utilizar, amostragem aleatória simples, adaptativos, estratificado, entre outros.

Segundo Castro e Ferrari (2016), a etapa 4 é a mineração de dados, etapa do processo que engloba a fase de exploração da base de dados pré-processada. A mineração, como o próprio termo pode exemplificar, é responsável por extrair as informações da base utilizando algoritmos e técnicas adequadas específicas para o objetivo da análise. Dentre as técnicas que podem ser utilizadas estão aquelas relacionadas à análise descritiva, agrupamento, predição, associação e detecção de anomalias, as quais serão discutidas posteriormente.

Collier et. al. (1998), diz que, embora os algoritmos de mineração de dados tenham o potencial de produzir um número ilimitado de padrões ocultos nos dados, muitos deles podem não ser significativos ou úteis. A última etapa do processo de KDD, a etapa de interpretação e avaliação de dados é a fase que visa selecionar os

modelos que são válidos e úteis para a tomada de decisões de negócios futuras. Para isso, existem diferentes métricas de avaliação que dependem do tipo de tarefa de mineração de dados que está aplicando.

2.1.1 Aprendizado supervisionado e não supervisionado

Para realizar o processo de mineração de dados é importante saber o objetivo do processo pois, dependendo da sua razão de uso, pode-se trabalhar com os dados de formas diferentes, moldando-os, transformando de várias formas.

Em mineração de dados o termo aprendizagem de máquina refere-se à capacidade em se adaptar ao ambiente, de forma que mesmo com regras preexistentes, ele conduz a uma melhoria de desempenho (CASTRO; FERRARI, 2016).

O aprendizado de máquina (AM), é uma das técnicas de Inteligência Artificial e tem sido muito utilizada em análises de problemas reais. De acordo com Mitchell (1997), o aprendizado de máquina está vinculado diretamente à mineração de dados, já que com um volume grande de dados é possível realizar o treinamento da máquina para aumentar a eficiência em tomadas de decisão.

Segundo Facelli (2011), um dos requisitos importantes para o algoritmo de AM é que seja possível de trabalhar com dados imperfeitos, com presença de ruídos, dados ausentes ou redundantes. As tarefas de aprendizado podem ser divididas em dois subgrupos, preditivas e descritivas.

Os modelos preditivos são utilizados para avaliar a classe de um objeto não rotulado, o que também é chamado de aprendizado supervisionado. De acordo com Castro e Ferrari (2016), o aprendizado supervisionado vem da presença de um supervisor externo que conheça a saída desejada para cada exemplo avaliado ou alguma informação que represente o comportamento que deve ser apresentado pelo sistema.

Os modelos descritivos são de aprendizado não supervisionado, pois a meta é a exploração ou descrição dos conjuntos de dados. Neste tipo de aprendizado não há uma informação de saída, ao contrário do aprendizado supervisionado, o processo busca a identificação de alguma similaridade entre os dados. As semelhanças entre elas, ou regras de associação buscam criar consequências associativas entre as variáveis analisadas em um determinado conjunto (GOLDSCHMIDT *et al*. 2015).

Dentro dos tipos de aprendizados, sendo eles supervisionados e não supervisionados existem tarefas distintas para aplicar a mineração de dados.

2.1.2 Tarefas de mineração de dados

Fayyad (1996) cita como principais tarefas utilizadas para a mineração de dados, os cinco seguintes:

a. Classificação: A tarefa de classificação utiliza a base de dados para aprender uma função para mapear e classificar classes, um gráfico pode ser construído a fim de definir traçando uma reta linear simples por exemplo, onde a localização do dado defina a sua classe. Exemplo: Em uma classificação dos tipos de reclamações no SAC a saída pode ser classificada em duas ou mais classes, sendo assim um atributo de saída do tipo categórico.

b. Regressão: Pode ser definido basicamente como uma previsão de valor real. Um exemplo que pode ser utilizado para essa tarefa é a previsão de número de casos de um vírus em um país utilizando o banco de dados de um país semelhante. Exemplos: Previsão de vendas de acordo com os investimentos no setor de marketing para o produto ou o coeficiente de rendimento de um aluno

c. Clusterização/agrupamento: O agrupamento é uma tarefa descritiva que busca identificar um conjunto finito de categorias ou grupos comuns para poder descrever os dados, agrupando as instâncias de forma natural a partir dos valores de seus atributos. Os grupos podem ser mutuamente exclusivos ou consistir por uma representação mais rica, como categorias hierárquicas ou sobrepostas. Exemplo: Agrupar áreas que possuem similaridades quanto ao clima e problemas de estiagem.

d. Associação: A associação envolve o estudo de dependências/independências significativas entre os dados da base, buscando relações funcionais entre as variáveis. Exemplo: Base de dados de mercado, identificando os produtos que estão associados entre eles.

e. Detecção de anomalias: A detecção de anomalias foca em descobrir mudanças significativas nos dados de valores já analisados, identificando se as atividades das observações se mantêm estável ou se rotulam como um problema.

Exemplo: Detecção de movimentações bancárias anormais do cliente para identificar roubo.

2.1.3 Partição do conjunto de dados

Na etapa 2 sobre preparação dos dados, parte do processo de KDD onde é realizado a preparação do conjunto de dados, basicamente o objetivo está na transformação dos dados brutos de modo a facilitar a obtenção de conhecimento ao final do processo. Uma vez que o banco de dados está construído, a qualidade do modelo precisa ser avaliada.

Para a realização desta avaliação, separam-se os dados em conjuntos de treinamento e teste, sendo que uma parte dos dados é utilizada na geração do modelo preditivo, para processo de treinamento e outra parte utilizada para avaliar a qualidade do modelo, chamado de conjunto de teste (CASTRO; FERRARI, 2016).

A operação de partição dos dados em dois conjuntos de dados acaba por assumir grande importância, uma vez que interfere diretamente no resultado do processo (GOLDSCHMIDT *et al*. 2015). Existem diversos métodos utilizados na partição do conjunto de dados, onde as diferenciações se dão pelo tipo de aleatoriedade, as quantidades de elementos, formas de distribuição e objetivo do processo.

Dentre os métodos mais conhecidos estão a Validação cruzada com K conjuntos (*K-fold Cross Validation*), Validação Cruzada com K Conjuntos Estratificada (*Stratifield K-fold Cross Validation*), *Leave-One-Out*, *Bootstrap* e *Holdout***.**

Para esta pesquisa, será utilizada o método de Validação Cruzada com K Conjuntos (*k-fold cross validation*, ou simplesmente, *k-fold*) para dividir o conjunto de dados. O método consiste em dividir de forma aleatória o conjunto de dados com *N* elementos em *K* subconjuntos (*folds*), de forma que o número de elementos *N* em cada subcojuntos sejam aproximados. Cada subconjunto *K* é utilizado como conjunto de teste exatamente uma vez e o restante dos subconjuntos utilizados como treinamento, dessa forma o processo se repete *K* vezes, sendo gerados todas as vezes *K* resultados de treinamento e validação (GOLDSCHMIDT *et al*. 2015).

2.1.4 Técnicas de classificação

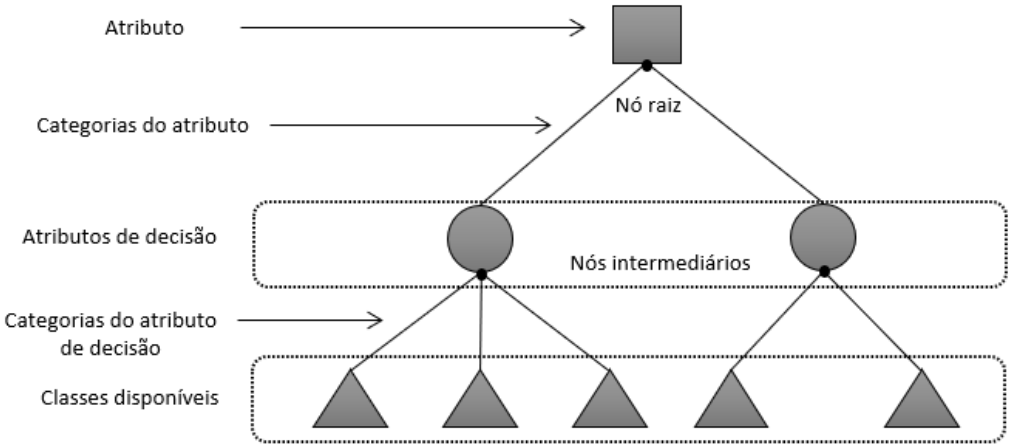
Como citado anteriormente, a classificação é uma das tarefas mais utilizadas no processo de mineração de dados e é uma tarefa preditiva no qual o objetivo está em identificar a classe à qual um objeto ainda não visto pertence. Para realizar esta ação é necessário que um modelo de classificação seja construído, feito com base em um conjunto previamente rotulado (CASTRO; FERRARI, 2016). Dentre os algoritmos de classificação que se destacam, de acordo com Castro e Ferrari (2016), são os algoritmos de k*-vizinhos mais próximos* (K-NN, do inglês k*-nearest neighbors*), as árvores de decisão, regras de classificação 1R e o *Naïve Bayes*.

2.1.4.1 Árvore de decisão

A árvore de decisão é uma técnica utilizada para dividir o problema em porções menores e mais simples, aos quais recursivamente são aplicadas a mesma estratégia (FACELLI, 2011). Os modelos de árvores de decisão são utilizados para problemas de classificação, enquanto que para problemas de regressão existem as árvores de regressão.

De acordo com Collier et. al. (1998), o algoritmo utilizado na árvore de decisão determina divisões naturais nos dados com base em uma variável alvo. As primeiras divisões ocorrem com as variáveis mais significativas do banco de dados. O método se comporta de forma que um galho em uma árvore de decisão pode ser visto como o lado condicional de uma regra. A Figura 2 ilustra uma árvore de decisão.

Figura 2 – Exemplo de estrutura de árvore de decisão



Fonte: Autoria própria (2021)

O algoritmo de árvore de decisão segundo Castro (2016), se desenvolve da seguinte maneira:

- O início da árvore começa com um único nó, sendo ele representado pelos dados da base;
- Se nesse conjunto de dados todos pertencem a uma mesma classe, o nó torna-se uma folha (chamado neste caso de nó puro) e rotula-se com a classe;
- Senão, encontra-se o melhor atributo para separar as classes em classes individuais e torna-o em atributo teste ou atributo de decisão do nó;
- Para cada valor conhecido do atributo teste, a base é particionada seguindo as regras criadas pela árvore de decisão na etapa de treinamento;
- A partir daí os algoritmos continuam com a sua lógica, quando o atributo aparece no nó ele não é mais considerado nos seus descendentes;
- O particionamento é cessado quando uma das condições forem satisfeitas:
    - Se todos os dados para um dado nó pertencem à mesma classe (nó puro);
    - Não há mais classes para se definir o conjunto de dados (máxima profundidade);
    - Quando não há objetos para o atributo teste, nesse caso cria-se uma folha com a classe predominante.

Usualmente os algoritmos utilizados para a construção de uma árvore de decisão exploram heurísticas que executam uma pesquisa sem olhar para trás, isto é, uma vez que a decisão é tomada, nunca mais é reconsiderada.

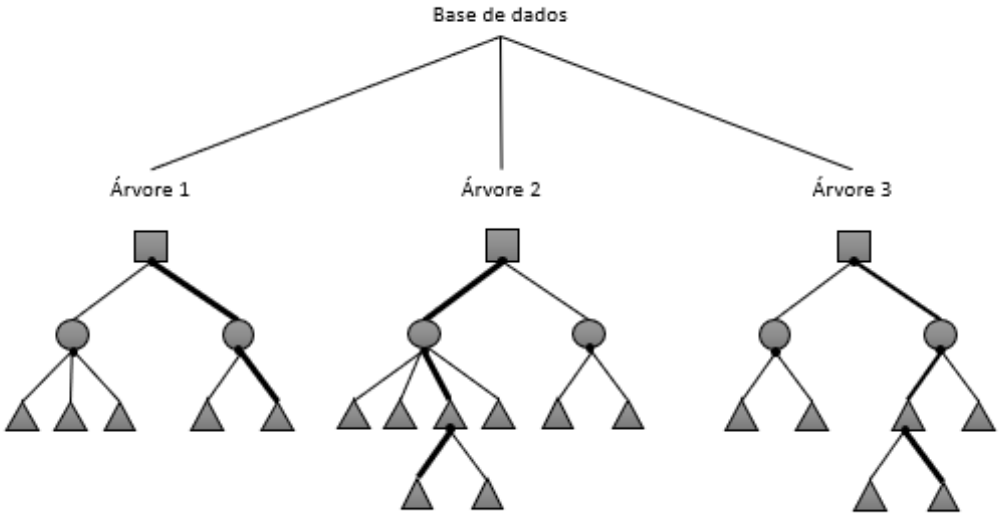
2.1.4.2 Floresta Aleatória (*Random Forest*)

O algoritmo Floresta Aleatória foi desenvolvido por Léo Breiman (2001), o qual consiste em criar combinações de preditores de árvores de tal forma que cada árvore depende dos valores de amostragem aleatória. O autor definiu como um classificador que consiste em uma coleção de árvores estruturadas {h (x, θ_K), k= 1, ...} onde θ_K são vetores independentes identicamente distribuídos e cada árvore lista uma unidade de voto para a classe mais popular de entrada *x* (BREIMAN, 2001).

O método de Floresta Aleatória engloba a técnica de árvore de decisão, cada um dos classificadores é um conjunto de classificador do tipo árvore de decisão (Figura 3), e as árvores de decisão individuais são geradas por seleção aleatória de atributos para cada nó, determinando assim uma divisão (HAN *et al*. 2012).

A amostragem é independente e a mesma distribuição é percorrida durante todas as árvores da floresta (BREIMAN, 2001).

Figura 3 – Exemplo de estrutura de Floresta Aleatória (*random forest*)



Fonte: Autoria própria (2021)

Segundo Wang *et al*. (2020), durante o treinamento, cada modelo básico aprende a partir de diferentes subamostras aleatórias dos dados. As amostras são retiradas pelo processo chamado de "*bagging*" ou "*bootstrapping"*, que significa que algumas amostras podem ser usadas muitas vezes em uma única árvore de decisão.

Breiman (2001) utiliza "*bagging"* em conjunto com a seleção aleatória de recursos, cada novo conjunto de treinamento é desenhado, com substituição do conjunto original e em seguida uma árvore é cultivada no novo conjunto de treinamento usando seleção aleatória de recursos.

De acordo com Wang *et al*. (2020), a floresta aleatória supera os modelos lineares porque pode capturar dados não lineares que possuem a relação entre o objeto e os recursos. A falha do método é que ele não pode funcionar com recursos esparsos porque as árvores de decisão são blocos de construção. Portanto, precisa-se pré-processar as entradas para se adequar ao modelo.

2.1.2 Métricas de avaliação

As métricas de avaliação são utilizadas para averiguar o modelo de dados utilizados para o aprendizado de máquina, verificando se o algoritmo proporcionou bons resultados na classificação dos dados.

A matriz de confusão é composta por todos os outputs possíveis de um algoritmo. Segue abaixo na Tabela x a ilustração dos possíveis resultados de um modelo de dados binário, de forma a auxiliar na exemplificação dos tipos de métricas.

Tabela 1: Matriz de confusão

		Classe predita	
		Positivo	Negativo
Classe original	Positivo	VP (verdadeiro positivo)	FN (falso negativo)
	Negativo	FP (falso positivo)	VN (verdadeiro negativo)

Fonte: Autoria própria (2021)

VP: Verdadeiro Positivo, refere-se à predição correta do modelo, uma previsão exata da classe original;

FN: Falso Negativo, quando o algoritmo prevê uma saída inversa da classe original positiva;

FP: Falso Positivo, o algoritmo prevê como verdadeiro a classe original negativa;

VN: Verdadeiro Negativo, previsão negativa de uma classe original negativa, no caso, predição correta.

- Acurácia

A métrica de acurácia é uma forma de avaliar o desempenho de um classificador, pois calcula-se o percentual da classificação correta. A acurácia não considera o custo de uma decisão incorreta por parte do classificador, ela assume que todas as classes possuem o mesmo grau de importância. Indica uma performance geral do modelo, o quão bom é o algoritmo, somando todas as classificações corretas e dividindo-a pela quantidade total de instâncias de teste.

Os valores geralmente apresentados em percentuais são calculados usando a Equação 1:

$$Acurácia = \frac{VP + VN}{VP + FP + FN + VN} \quad (1)$$

- *Recall* ou sensibilidade

A métrica de *Recall* avalia a eficiência do modelo para classificar a classe de interesse (positivo), seguindo o racional de acordo com a equação 2:

$$Recall = \frac{VP}{VP + FN} \quad (2)$$

- Especificidade

A especificidade avalia o quão bom o modelo é para prever a saída negativa (equação 3):

$$Especificidade = \frac{VN}{VN + FP} \quad (3)$$

- Precisão

A métrica de precisão mede quanto o modelo conseguiu prever corretamente dentro do conjunto classificado como positivo (equação 4):

$$Precisão = \frac{VP}{VP + FP} \tag{4}$$

- Score F

A métrica de *Score F* ou medida F calcula a eficiência do algoritmo equilibrando as métricas de precisão e sensibilidade, realizando uma média harmônica entre as duas métricas (equação 5):

$$ScoreF = 2x\frac{precisãoxsensibilidade}{precisão + sensibilidade} \tag{5}$$

2.2 Ideação suicida

Segundo Nock et. al. (2008), a ideação suicida em si aparece como uma forma de comportamento suicida não fatal, quando o indivíduo passa a ter pensamentos que alimentam a vontade de acabar com a sua própria existência e se agrava com o acompanhamento de um plano suicida. No caso da tentativa de suicídio envolve as condutas para realizar a morte, podendo ou não resultar na mesma e por fim o suicídio é o ato intencional para acabar com a vida.

Nos últimos anos o assunto relacionado a depressão, tentativa de suicídio têm se tornado recorrente, em diversas faixas etárias, desde jovens a idosos o assunto tem sido repercutido. Em 2011, nos Estados Unidos da América (EUA), foi desenvolvido um relatório sobre comportamento suicida, onde 105 mil estudantes universitários foram envolvidos no estudo. Os resultados estão apresentados na Tabela 2.

Tabela 2 – Resultados da pesquisa com universitários

	Consideraram seriamente suicídio		
Porcentagem (%)	Homens	Mulheres	Total
Não, nunca	81,9	80,2	80,7
Não, nos últimos 12 meses	11,8	13,4	12,9
Sim, nas últimas 2 semanas	1,4	1,2	1,3
Sim, nos últimos 30 dias	1,0	1,0	1,0
Sim, nos últimos 12 meses	3,9	4,2	4,1
Em qualquer momento nos últimos 12 meses	6,3	6,4	6,4

Fonte: Adaptado de National College Health Assessment I (2011).

Os dados do estudo mostram uma porcentagem alta referente a indivíduos que tiveram ideação suicida nos últimos 12 meses do período analisado, foram em média 6,4% da população de estudo, no qual o sexo feminino/masculino não teve diferença relevante no resultado.

2.2.1 Influência da toxicodependência na ideação suicida

De acordo com Costa (2014), a toxicodependência pode ser considerado um fator de risco para a ideação suicida. Segundo sua análise um número significativo de indivíduos com pensamentos suicidas é toxicodependente entre outras características predominantes na literatura, como: pessoas na faixa etária entre 20 a 29 anos, uma vulnerabilidade maior a ideação e tentativas de suicídio em mulheres toxicodependentes e em indivíduos que passaram por eventos anteriores de depressão, tentativas ou doenças psiquiátricas que podem contribuir para o comportamento suicidário.

Já um estudo realizado por SILVA, V. F. et al. (2006), na cidade de Campinas-SP, analisou um subgrupo de 29 indivíduos de um total de 515 entrevistados. Os 29 selecionados foram os que responderam afirmativo para as duas questões consecutivas: “Alguma vez você já pensou em pôr fim a sua vida?” e “Este pensamento lhe ocorreu alguma vez nos últimos 12 meses?”. Para o grupo de indivíduos que já tiveram/possuem uma ideação suicida, foi aplicado mais um questionário englobando duas atmosferas, características sociodemográficas e

religiosas, de saúde mental e de comportamento. O resultado encontrado pelos autores, utilizando métodos como o *odds-ratio* para comparação foi de que diferentemente da literatura, não se estabeleceu uma diferença entre os grupos quanto ao uso de álcool e outras drogas, a significância estatística se permaneceu apenas nas variáveis relacionadas a falta de energia, humor deprimido, dificuldades emocionais.

3. METODOLOGIA

O presente capítulo traz informações referentes à metodologia aplicada ao estudo, a base de dados considerada, além de discorrer sobre as etapas do processo de KDD.

3.1 Base de dados

A base de dados americana utilizada, realizada pela NSDUH é a principal fonte do país para estimativas anuais de uso de drogas e doenças mentais entre sua população.

Para este estudo foram utilizadas as pesquisas dos anos 2019 e 2020. O tamanho da amostra de nível nacional dessa pesquisa foi de 56.136 indivíduos e de 2.741 variáveis no ano de 2019, com 32.893 indivíduos e 2.892 variáveis em 2020.

A amostra foi alocada aos grupos de idade da seguinte forma: 25% para jovens de 12 a 17 anos, 25% para jovens adultos de 18 a 25, 15% para adultos de 26 a 34 anos, 20% para adultos de 35 a 49 anos e 15% para adultos com 50 anos ou mais. As dimensões abordadas na pesquisa foram perguntas relacionadas a demografia, saúde física e mental, vícios em diversos tipos de substâncias tóxicas, perguntas quanto ao plano de saúde, entre outros.

A maior parcela das perguntas realizadas para a pesquisa foi administrada com o *software* ACASI, tendo como objetivo coletar dados sensíveis, garantindo aos entrevistados uma pesquisa de modo altamente privado, podendo aumentar o nível de sinceridade no relato sobre o uso de drogas ilícitas e outros componentes. Embora haja todo o sistema de confidencialidade, a base de dados pode ser limitada, pois, dependerá da veracidade e da memória do entrevistado sobre o assunto. Todos que terminaram uma entrevista completa, receberam cada um US$ 30 em dinheiro como um sinal de agradecimento por seu tempo.

3.2 Seleção de dados

Os atributos escolhidos para o estudo foram selecionados buscando uma descoberta de conhecimento a respeito da influência da toxicodependência na ideação suicida. Considerando referências de estudo como o de Breet *et al.* (2018)

foram escolhidos também questões relacionadas a outras drogas além de *cannabis*, álcool e cigarro.

Para realizar um estudo com maior amplitude, além dos atributos de toxicodependência, foram considerados também dimensões de aspecto demográfico e saúde mental, onde as instâncias que representavam indivíduos com idades de 12 a 17 anos foram excluídos, pois não participam do questionário relacionado a saúde mental.

Segue na Tabela 3 a listagem dos 13 atributos selecionados:

Tabela 3: Listagem de atributos selecionados para o estudo.

• Categoria de idade • Sexo • Raça
• Já teve problemas com drogas ou álcool • Quantos cigarros em média fumou por dia nos últimos 30 dias • Dias por semana em média que ingeriu álcool nos últimos 12 meses • Dias por semana em média que usou cannabis/raxixe nos últimos 12 meses • Total de dias que usou inalantes nos últimos 12 meses • Total de dias que usou alucinógenos nos últimos 12 meses • Total de dias que usou metanfetamina nos últimos 12 meses
• Já teve problemas com a saúde mental • Tem maior episódio depressivo (levantamento levando-se em consideração outros critérios relacionados a frequência e duração de sentimento depressivo, desanimo, dificuldades de dormir, comer, concentrar por questões emocionais) • Pensou em cometer suicídio quando teve problemas piores

Fonte: Autoria própria (2021).

3.3 Pré-processamento de dados

Como apresentado anteriormente, nas etapas do processo de KDD, o pré-processamento de dados é a etapa no qual realiza-se a limpeza de dados. Para a realização do pré-processamento do banco de dados, foi utilizado a ferramenta *Microsoft Excel* (2019).

O banco de dados do estudo continha respostas em códigos, para o pré-processamento de dados foram analisados os 13 atributos iniciais, cada tipo de código e o que seriam considerados para o estudo. Foram excluídas todas as instâncias com qualquer ausência de informação, seja em um ou mais atributos.

Inicialmente, o conjunto de dados (anos 2019 e 2020) continha 89.012 instâncias (desconsiderados 17 indivíduos que fizeram parte da pesquisa nos dois

anos). Após a etapa de seleção dos 13 atributos, na etapa de pré processamento realizando a exclusão das instâncias com respostas vazias ou inválidas, o conjunto de dados passou para 5.055 instâncias (2.019 com ideação suicida e 3.036 sem ideação suicida).

3.4 Transformação de dados

Na etapa de transformação de dados, foram analisados os atributos e a necessidade de alteração de cada cenário. A informações na Tabela 4 mostram os atributos e alternativas pós processo de limpeza de dados e as alterações realizadas na etapa de transformação de dados.

Tabela 4: Listagem de atributos transformados.

Código	Atributos (questões)	Alternativas (respostas)	Transformação
CATAG3	Categoria de idade	1. 12 a 17 anos 2. 18 a 25 anos 3. 26 a 34 anos 4. 35 a 49 anos 5. 50 anos ou mais	Manteve
IRSEX	Sexo	1. Masculino 2. Feminino	Binarização Feminino 0 – Não 1 – Sim (Todos os 0 considerados como masculinos)
NEWRACE2	Raça	1. Branco 2. Negro/Afro-americano 3. Nativo Americano/Alasca 4. Nativo do Havaí/ilhas do pacífico 5. Asiático 6. Outras raças não hispânicas 7. Hispânico/latino	Binarização por raças Exemplo: Branco 0 – Não 1 – Sim
CASUPROB	Já teve problemas com drogas ou álcool	1. Sim 2. Não	Manteve
CIG30AV	Quantos cigarros em média fumou por dia nos últimos 30 dias	1. Menos de 1 cigarro 2. 1 cigarro 3. 2 a 5 cigarros 4. 6 a 15 cigarros 5. 16 a 25 cigarros 6. 26 a 35 cigarros 7. Mais de 35 cigarros 91. Nunca fumou 93. Não fumou nos últimos 30 dias	0. Nunca fumou/não fumou nos últimos 30 dias 1. Menos de 1 cigarro 2. 1 cigarro 3. 2 a 5 cigarros 4. 6 a 15 cigarros 5. 16 a 25 cigarros 6. 26 a 35

			cigarros 7. Mais de 35 cigarros
ALDAYPWK	Dias por semana em média que ingeriu álcool nos últimos 12 meses	1. 1 dia 2. 2 dias 3. 3 dias 4. 4 dias 5. 5 dias 6. 6 dias 7. 7 dias 91. Nunca ingeriu álcool 93. Não ingeriu álcool nos últimos 12 meses	0. Nunca ingeriu álcool/não ingeriu nos últimos 12 meses 1. 1 dia 2. 2 dias 3. 3 dias 4. 4 dias 5. 5 dias 6. 6 dias 7. 7 dias
MRDAYPYR	Total de dias que usou cannabis/raxixe nos últimos 12 meses	1 a 365 dias 991. Nunca usou inalantes 993. Não usou inalantes nos últimos 12 meses	0. Nunca usou cannabis/raxixe ou não usou nos últimos 12 meses 1. Usou nos últimos 12 meses
INHALYFQ	Total de dias que usou inalantes nos últimos 12 meses	1 a 365 dias 991. Nunca usou inalantes 993. Não usou inalantes nos últimos 12 meses	Binarização: 0 – Não usou nos últimos 12 meses/nunca usou 1 - Usou nos últimos 12 meses
HALLUCYFQ	Total de dias que usou alucinógenos nos últimos 12 meses	1 a 365 dias 991. Nunca usou alucinógenos 993. Não usou alucinógenos nos últimos 12 meses	Binarização: 0 – Não usou nos últimos 12 meses/nunca usou 1 - Usou nos últimos 12 meses
METHAMYFQ	Total de dias que usou metanfetamina nos últimos 12 meses	1 a 365 dias 991. Nunca usou metanfetamina 993. Não usou metanfetamina nos últimos 12 meses	Binarização: 0 – Não usou nos últimos 12 meses/nunca usou 1 - Usou nos últimos 12 meses
CAMHPROB	Já teve problemas com a saúde mental	1. Sim 2. Não	Manteve
AMDELT	Tem maior episódio depressivo	1. Sim 2. Não	Manteve
ADWRSTHK	Pensou em cometer suicídio quando teve problemas piores	1. Sim 2. Não	Manteve

Fonte: Autoria própria (2021).

Após a transformação dos dados, o conjunto de dados chegou na quantidade de 19 atributos, 2.942 instâncias do ano de 2019 e 2.113 instâncias de 2020, totalizando em 5.055 instâncias válidas.

3.5 WEKA

A ferramenta utilizada para o processamento dos dados foi a ferramenta criada na Nova Zelândia chamada *WEKA*. A ferramenta foi desenvolvida na Universidade de *Waikato*, onde o nome são as abreviaturas de *Waikato Environment for Knowledge Analysis*. Fora da universidade, o *WEKA*, pronunciou para rimar com Meca, uma ave que não voa com uma natureza curiosa encontrada apenas nas ilhas de Nova Zelândia (Frank *et. al.* 2016).

O *WEKA* é uma ferramenta com uma grande coleção de algoritmos de aprendizado de máquina, disponibiliza também opções de pré-processamento de dados, transformação de conjuntos de dados, como os algoritmos para discretização e amostragem.

A ferramenta *WEKA* inclui métodos para os principais problemas de mineração de dados: regressão, classificação, armazenamento em cluster, mineração de regras de associação e seleção de atributos. Conhecer os dados é uma parte integrante do trabalho, e muitos recursos de visualização de dados e ferramentas de pré-processamento de dados são fornecidos. Todos os algoritmos recebem sua entrada na forma de uma única tabela relacional que pode ser lida de um arquivo ou gerado por uma consulta de banco de dados (Frank *et. al.* 2016).

A ferramenta foi escolhida para processar os dados pela facilidade de utilização e rapidez na geração de resultados.

4. RESULTADOS E DISCUSSÕES

Para o estudo foram utilizadas duas técnicas de classificação, os algoritmos de Árvore de decisão (J48) e Floresta aleatória (*Random forest*), ambas foram configuradas para utilizar como método de partição de dados o *K-fold Cross Validation* com 10 subconjuntos (*K=10*).

A fim de analisar as dimensões do conjunto de dados, foram processados três grupos de dados diferentes no *WEKA*, o conjunto pré-processado com todos os 19 atributos selecionados, o segundo conjunto de dados sem os atributos relacionados a toxicodependência e o outro sem os atributos relacionados à saúde mental, todos tendo como atributo de saída a ideação suicida.

Como atributo de saída utilizou-se a questão “Pensou em cometer suicídio quando teve problemas piores”, sendo duas as classes nominais disponíveis, sim e não.

4.1 Análise Exploratória

A distribuição das classes de saída da base de dados após o pré-processamento e transformação de dados, estão apresentadas na Tabela 5.

Tabela 5 – Total de instâncias/classe de saída

	Classe de saída - Ideação suicida	
	SIM	NÃO
Quantidade de instâncias	2.019	3.036
Porcentagem	40%	60%

Fonte: Autoria própria (2021)

A quantidade de instâncias com saída positiva no conjunto de dados foi um total de 2019, representando 40% das instâncias, já as instâncias com saída negativa representaram 60% dos dados. Observou-se com essa proporção de que o conjunto de dados, em geral, não estava desbalanceado.

As informações contidas na Tabela 6 mostram a classe de saída por cada um dos 18 atributos de entrada.

Tabela 6 – Classe de saída por atributo

				Classe de saída - Ideação suicida			
Código original	**Atributos**	**Alternativas**	**Detalhamento**	**SIM**	**NÃO**	**SIM**	**NÃO**
CATAG3	IDADE	2	28 a 25 anos	782	801	49%	51%
		3	26 a 34 anos	401	637	39%	61%
		4	35 a 49 anos	543	930	37%	63%
		5	50 anos ou mais	293	668	30%	70%
IRSEX	Sexo Feminino	0	Masculino	825	1.055	44%	56%
		1	Feminino	1.194	1.981	38%	62%
NEWRACE2	Branco	0	NÃO	535	889	38%	62%
		1	SIM	1.484	2.147	41%	59%
	Negro/Afro-americano	0	NÃO	1902	2804	40%	60%
		1	SIM	117	232	34%	66%
	Nativo Americano/Alasca	0	NÃO	2.000	3.007	40%	60%
		1	SIM	19	29	40%	60%
	Nativo do Havaí/ilhas do pacífico	0	NÃO	2.018	3.025	40%	60%
		1	SIM	1	11	8%	92%
	Asiático	0	NÃO	1.950	2.907	40%	60%
		1	SIM	69	129	35%	65%
	Outras raças não hispânicas	0	NÃO	1.914	2.935	39%	61%
		1	SIM	105	101	51%	49%
	Hispânico/latino	0	NÃO	1.795	2.649	40%	60%
		1	SIM	224	387	37%	63%
CASUPROB	Já teve problemas com drogas	2	NÃO	1.433	2.458	37%	63%
		1	SIM	586	578	50%	50%
CIG30AV	Quantidade de cigarro consumido por dia nos últimos 30 dias	0	Zero/Nunca fumou	1.651	2.511	40%	60%
		1	Menos de 1 cigarro	32	40	44%	56%
		2	1 cigarro	55	50	52%	48%
		3	2 a 5 cigarros	93	140	40%	60%
		4	6 a 15 cigarros	96	160	38%	63%
		5	16 a 25 cigarros	66	96	41%	59%
		6	26 a 35 cigarros	18	26	41%	59%
		7	Mais de 35	8	13	38%	62%
ALDAYPWK	Dias por semana em média que ingeriu álcool nos últimos 12	0	Zero/Nunca ingeriu	1.054	1.536	41%	59%
		1	1 dia	170	283	38%	62%
		2	2 dias	216	371	37%	63%
		3	3 dias	223	354	39%	61%
		4	4 dias	133	162	45%	55%
		5	5 dias	108	166	39%	61%
		6	6 dias	54	87	38%	62%
		7	dias	61	77	44%	56%

MRDAYPYR	Total de dias que usou cannabis/raxixe nos últimos 12 meses	0	Zero/Nunca usou	1.673	2.648	39%	61%
		1	Usou nos últimos 12 meses	346	388	47%	53%
INHALYFQ	Total de dias que usou inalantes nos últimos 12 meses	0	Zero/Nunca usou	1.988	3.012	40%	60%
		1	Usou nos últimos 12 meses	31	24	56%	44%
HALLUCYFQ	Total de dias que usou alucinógenos nos últimos 12 meses	0	Zero/Nunca usou	1.954	2.983	40%	60%
		1	Usou nos últimos 12 meses	65	53	55%	45%
METHAMYFQ	Total de dias que usou metanfetamina nos últimos 12 meses	0	Zero/Nunca usou	1.991	3.018	40%	60%
		1	Usou nos últimos 12 meses	28	18	61%	39%
CAMHPROB	Já teve problemas com a saúde mental	1	SIM	1.728	1.921	47%	53%
		2	NÃO	291	1.115	21%	79%
AMDELT	Tem maior episódio depressivo	1	SIM	1.968	2.619	43%	57%
		2	NÃO	51	417	11%	89%

Fonte: Autoria própria (2021)

De acordo com as informações da Tabela 5, observou-se com maior detalhamento a distribuição do conjunto de dados. Das cinquenta alternativas, apenas duas alternativas apresentaram desbalanceamento com porcentagem acima de 80%, um dos atributos relacionado a questão racial e um atributo relacionado a saúde mental. Dos doze indivíduos Nativos do Havaí/ilhas do pacífico que participaram da pesquisa, apenas um respondeu que teve ideação suicida e onze responderam que não tiveram. O atributo que questionava se o indivíduo tem maior episódio depressivo, 468 responderam que não, desses, 11% quando questionados sobre ideação suicida responderam que sim, já tiveram ideação e os 89% não.

Visto que de forma geral a classe de saída do conjunto de dados positivo era de 40% e entre as 50 alternativas apenas duas alternativas apresentaram desbalanceamento, não foi realizado nenhum processo para balancear o conjunto de dados.

4.2 Experimento 1 - com todas as dimensões do conjunto de dados

Para o primeiro experimento foram consideradas todas as dimensões do conjunto de dados, demográfico, atributos relacionados a toxicodependência e saúde mental, sendo os 19 atributos: CATAG3 (Idade), IRSEX (Sexo), NEWRACE2 (7 raças), CASUPROB (Toxicodependência), CIG30AV (Quantidade de cigarro consumido), ALDAYPWK (Frequência de uso de Álcool), MRDAYPYR (Frequência de uso de Cannabis), INHALYFQ (Frequência de uso de Inalante), HALLUCYFQ (Frequência de uso de Alucinógeno), METHAMYFQ (Frequência de uso de Metanfetamina), CAMHPROB (Questão relacionada a Saúde mental), AMDELT (Questão relacionada a Saúde mental), ADWRSTHK (Questão de Ideação suicida).

Os resultados obtidos após processar os algoritmos de Árvore de decisão (J48) e Floresta aleatória (*Random forest*) estão apresentadas nas Tabelas 7, 8 e 9.

Tabela 7 – Matriz de confusão para a Árvore de decisão (Experimento 1)

		Classe predita	
		1 - SIM	2 - NÃO
Classe original	1 - SIM	992 (verdadeiro SIM)	1027 (falso NÃO)
	2 - NÃO	806 (falso SIM)	2230 (verdadeiro NÃO)

Fonte: Autoria própria (2021)

Tabela 8 – Matriz de confusão para a Floresta Aleatória (Experimento 1)

		Classe predita	
		1 - SIM	2 - NÃO
Classe original	1 - SIM	718 (verdadeiro SIM)	1301 (falso NÃO)
	2 - NÃO	550 (falso SIM)	2486 (verdadeiro NÃO)

Fonte: Autoria própria (2021)

Ao analisar os resultados apresentados na matriz de confusão de ambas as técnicas, pode-se apontar que com esse conjunto de dados o aprendizado foi superior na predição da classe de saída “não”. O motivo dessa diferença de resultado pode estar na própria técnica utilizada ou até mesmo no tamanho do conjunto de dados.

Tabela 9 – Resultados Experimento 1

Técnica	Acurácia	Precisão	Recall
Árvore de decisão	63,7%	55,2%	49,1%
Floresta Aleatória	63,3%	56,6%	35,6%

Fonte: Autoria própria (2021)

Percebeu-se nos resultados apresentados na Tabela 9 que embora a diferença seja pequena, a técnica de Árvore de decisão mostrou ser melhor em acertar as classificações quando comparado com a Floresta Aleatória, a porcentagem de acurácia foi um pouco maior. A técnica da Árvore de decisão nesse conjunto de dados obteve um *Recall* também superior ao da Floresta Aleatória, isto é, o "acerto" na classe de interesse (quem possui a ideação suicida), porém, quando avalia-se os resultados pela precisão, percebe-se que a Floresta Aleatória teve melhor desempenho.

4.3 Experimento 2 – Dimensões relacionadas a saúde mental

Para processar o segundo conjunto de dados, selecionou-se os atributos das dimensões demográficas e os atributos relacionadas a saúde mental, foram consideradas 12 atributos: CATAG3 (Idade), IRSEX (Sexo), NEWRACE2 (7 raças), CAMHPROB (Questão relacionada a Saúde mental), AMDELT (Questão relacionada a Saúde mental), ADWRSTHK (Questão de Ideação suicida).

No experimento 2 foram consideradas as mesmas técnicas de classificação, a única diferenciação foi o conjunto de dados de entrada. Na sequência, estão apresentados os resultados do experimento (Tabelas 10, 11 e 12).

Tabela 10 – Matriz de confusão para a Árvore de decisão (Experimento 2)

		Classe predita	
		1 - SIM	2 - NÃO
Classe original	1 - SIM	1046 (verdadeiro SIM)	973 (falso NÃO)
	2 - NÃO	861 (falso SIM)	2175 (verdadeiro NÃO)

Fonte: Autoria própria (2021)

Tabela 11 – Matriz de confusão para a Floresta Aleatória (Experimento 2)

		Classe predita	
		1 - SIM	2 - NÃO
Classe original	1 - SIM	805 (verdadeiro SIM)	1214 (falso NÃO)
	2 - NÃO	678 (falso SIM)	2358 (verdadeiro NÃO)

Fonte: Autoria própria (2021)

Seguindo o mesmo raciocínio do experimento 1, pelo resultado da matriz de confusão das técnicas, nota-se que mesmo com um conjunto de dados diferente, a máquina tem maior facilidade de ler os verdadeiros “Não”.

Tabela 12 – Resultados Experimento 2

Técnica	Acurácia	Precisão	Recall
Árvore de decisão	63,7%	54,9%	51,8%
Floresta Aleatória	62,6%	54,3%	39,9%

Fonte: Autoria própria (2021)

No experimento 2, a técnica de Árvore de decisão também teve um desempenho melhor quanto à acurácia, a predição da técnica foi de 63,7%, enquanto que a Floresta Aleatória classificou corretamente apenas 62,6%. As outras métricas de avaliação, precisão e *recall* também tiveram valores superiores, mostrando que a técnica de classificação da Árvore de decisão prevê melhor dentro do conjunto classificado como positivo.

4.4 Experimento 3 – Dimensões relacionadas a toxicodependência

O experimento 3 foi realizado excluindo dois atributos do conjunto original, o CAMHPROB e AMDELT, relacionados à saúde mental. Novamente, foram aplicadas as duas técnicas de classificação de dados, Árvore de decisão e Floresta Aleatória, da mesma forma como seguiu nos outros dois experimentos anteriores.

A seguir (Tabelas 13, 14 e 15), estão os resultados encontrados para este conjunto de dados.

Tabela 13 – Matriz de confusão para a Árvore de decisão (Experimento 3)

		Classe predita	
		1 - SIM	2 - NÃO
Classe original	1 - SIM	644 (verdadeiro SIM)	1375 (falso NÃO)
	2 - NÃO	620 (falso SIM)	2416 (verdadeiro NÃO)

Fonte: Autoria própria (2021)

Tabela 14 – Matriz de confusão para a Floresta Aleatória (Experimento 3)

		Classe predita	
		1 - SIM	2 - NÃO
Classe original	1 - SIM	227 (verdadeiro SIM)	1792 (falso NÃO)
	2 - NÃO	185 (falso SIM)	2851 (verdadeiro NÃO)

Fonte: Autoria própria (2021)

Tabela 15 – Resultados Experimento 3

Técnica	Acurácia	Precisão	Recall
Árvore de decisão	60,5%	50,9%	31,9%
Floresta Aleatória	60,9%	55,1%	11,2%

Fonte: Autoria própria (2021)

Diferente dos experimentos 1 e 2, o experimento 3 apresentou melhor desempenho com a técnica de Floresta Aleatória, o algoritmo conseguiu prever 60,9% das saídas corretamente. A precisão também foi superior a técnica da Árvore de decisão, mostrando que em 55,1% das instâncias o modelo classificou como “Sim” o que eram realmente “Sim”. O *Recall* mostra que a técnica que melhor avaliou o modelo para classificar a classe de interesse em relação as demais que foram classificadas como “Sim”, foi a árvore de decisão.

4.5 Comparativo dos resultados com as diferentes dimensões de entrada para ideação suicida

A Tabela 16 traz as técnicas que obtiveram melhor resultados em cada experimento realizado.

Tabela 16 – Melhores desempenhos – Experimentos 1, 2 e 3

	Técnica com maior desempenho	Acurácia	Precisão	Recall
Conjunto geral	Árvore de decisão	63,7%	55,2%	49,1%
Conjunto de atributos relacionados a saúde mental	Árvore de decisão	63,7%	54,9%	51,8%
Conjunto de atributos relacionados a toxicodependência	Floresta Aleatória	60,9%	55,1%	11,2%

Fonte: Autoria própria (2021)

Com o conjunto de dados utilizado para o estudo, o algoritmo que apresentou melhor desempenho no geral, foi a árvore de decisão, com exceção do experimento 3 que obteve melhor porcentagem de acurácia com o modelo de Floresta Aleatória.

Quanto aos atributos notou-se que embora o resultado tenha sido baixo, a predição teve um maior desempenho quando utilizado atributos relacionados a saúde mental. Porém, por se tratar de porcentagens baixas de acurácia e uma diferença de menos de 3%, não se pode afirmar sobre a influência da dimensão de toxicodependência.

Analisando os resultados encontrados e o equilíbrio das classes de saída, é possível que, a baixa eficiência do modelo tenha sido causada pela quantidade insuficiente de dados, talvez uma maior quantidade de dados para treino resultasse em melhores conclusões. Outro ponto a ser considerado são alterações em um dos primeiros processos do KDD, outros tipos de pré-processamento poderiam vir a gerar um melhor resultado de aprendizado de máquina para estes conjuntos elaborados. Algumas investigações podem ser feitas no conjunto de dados, nas técnicas e parâmetros utilizados com a tentativa de evitar possíveis *overfitting* ou *underfitting*.

Sabe-se que a área da saúde envolve diretamente os seres humanos e contém diversos dados sensíveis, sendo assim, é de extrema importância que um modelo de classificação utilizando dados dessa área faça uma predição correta. Mesmo o melhor dos resultados encontrados pelo modelo de estudo não alcançou um nível alto de acuracidade e quanto ao *Recall* máximo alcançado pelo modelo foi de 51,8%, o que indica que o modelo não deve ser recomendado para a predição em situações reais.

5. CONCLUSÃO

O trabalho teve como objetivo analisar os atributos relevantes que podem levar à ideação suicida, pré-processar o conjunto de dados para a etapa de mineração e comparar as técnicas de classificação.

O pré-processamento do conjunto de dados foi realizado, sendo que, das 89.012 instâncias iniciais do conjunto de dados, após o processo de seleção, limpeza, o conjunto teve redução de 94%, resultando em 5.055 instâncias.

Durante o referencial teórico pode-se compreender que muitas variáveis podem influenciar uma pessoa a ter um pensamento suicida, desde questões sociais, faixa etária de idade, utilização de drogas. O objetivo principal do estudo estava na comparação das técnicas de classificação, porém, o trabalho buscou também contribuir de alguma forma com a descoberta de informações relacionados a ideação suicida. Foi possível observar neste conjunto de dados que os atributos relacionados a toxicodependência, com a metodologia aplicada neste trabalho, não mostraram uma forte relação com a ideação suicida.

Apesar dos algoritmos não terem apresentado alta acurácia com as técnicas de Árvore de decisão e Floresta Aleatória, os 63,7% são referentes a predições corretas em um conjunto de dados sensíveis. Os experimentos mostraram que para o conjunto de dados, no geral, a melhor técnica para classificação nesse caso é a Árvore de decisão. Há outras formas de pré-processar a base de dados e organizar os atributos, é possível que essa mesma base de dados, quando exposta a outras técnicas e combinações de dados, possa trazer um maior desempenho quanto ao aprendizado de máquina.

Portanto, a predição da ideação suicida utilizando as técnicas de árvore de decisão e floresta aleatória, considerando o conjunto de dados contendo dimensões de saúde mental, características demográficas e questões relacionadas a toxicodependência, desempenhou uma aprendizagem máxima de 63,7% de acurácia com algoritmo de árvore de decisão. Com o valor máximo de desempenho encontrado, não se recomenda utilizar o modelo, porém, sugere-se que trabalhos futuros realize testes como a inclusão de outros atributos e aplicação de outras técnicas de mineração de dados.

REFERÊNCIAS

BREET, E; GOLDSTONE, D; BANTJES, J. Substance use and suicidal ideation and behaviour in low-and Middle-income countries: a systematic review. BMC Public Health. 2018.

BREIMAN, L. **Random Forests**. Machine Learning. Kluwer Academic Publishers. Manufactured in The Netherlands. 45, 5–32. 2001.

CASTRO, L. N. FERRARI, D. G. Introdução a Mineração de dados: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CENTER FOR BEHAVIORAL HEALTH STATISTICS AND QUALITY. **National Survey on Drug Use and Health.** Public use file codebook. 2019.

CENTER FOR BEHAVIORAL HEALTH STATISTICS AND QUALITY. **National Survey on Drug Use and Health.** Public use file codebook. 2020.

FRANK, E; HALL, M; WITTEN, I. **The WEKA Workbench**. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, 4º ed., 2016.

COLLIER, Kenneth; CAREY, Bernard; GRUSY, Ellen; MARJANIEMI, Curt; SAUTTER, Donald. **A Perspective on Data Mining**. The Center for Data Insight. 1998.

CÔRTES, S. C; PORCARO, R. M; LIFSCHITZ, S. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens**. PUC-Rio. 2002.

COSTA, R. M. P. **A ideação suicida na Toxicodependência**. 2014. Dissertação (Mestrado) – Programa de mestrado em Enfermagem de Saúde Mental e Psiquiatria. Instituto Politécnico de Viseu. Viseu, 2014.

CHACHAMOVICK, E. et al. **Quais são os recentes achados clínicos sobre a associação entre depressão e suicídio?**. Revista Brasileira Psiquiatra. Vol. 31. Supl.1 São Paulo. 2009.

FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowlegde discovery**. American Association for Artificial Intelligence. v.17, n. 3, p. 39. 1996.

GRADIM, J. G. P; SILVA, A. C; PEREIRA, C. C. M; VEDANA, K. G. G. **Análisis de posturas sobre suicidio y comunidad LGBTQ en Twitter**. Salud & Sociedad, 10(3), 286-294. 2020.

GOLDSCHMIDT, R; PASSOS, E; BEZERRA, E. Data mining: conceitos, técnicas, algorítmos, orientações e aplicações. 2 ed. 2015.

GONÇALVES, R. E. M; PONCE, J. C.; LEYTON, V. **Uso de álcool e suicídio**. Saúde, Ética & Justiça. Pag. 09-14. 2015.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data Mining. **Concepts and Techniques**. Morgan Kaufmann Publishers is an imprint of Elsevier. 2012.

NOCK, M. K. et. al. **Suicide and Suicidal Behavior**. National Institutes of Health. NIH Public Access. Pag. 133-154. 2008.

PANWAR, Shailesh; RAIWANI, Y. **Data reduction techniques to analyze NSL-KDD dataset**. Internacional Journal of computer engineering & technology (IJCET). Vol 5. Pag. 21-31. 2014.

SHAPIRO, G. P; MATHEUS, C; SMYTH, P; UTHURUSAMY, R. **KDD-93: Progress and Challenges in Knowledge Discovery in Databases.** Al Magazine Vol. 15, n. 3. 1994.

SILVA, V. F. et al. **Fatores associados à ideação suicida na comunidade: um estudo de caso-controle**. Faculdade de Ciências Médicas, Universidade Estadual de Campinas. Cad. Saúde Pública. Pag. 1835-1843. 2006.

SILVA, A. G. et al. **Saúde mental: porque ainda é importante no meio de uma pandemia.** Revista Brasileira de Psiquiatria. Vol. 42. Ed. 3/2020. Pag. 229-231. 2020.

SOMASUNDARAM, G; SHRIVASTAVA, A. **Armazenamento e Gerenciamento de Informações**. Editora Bookman. Ed. 2011.

TENFEN, Emerson. A técnica de Knowledge Discovery In Databases (KDD) aplicada nas ocorrências atendidas pela polícia militar. 2003. Dissertação (Graduação), Universidade Regional de Blumenau. Blumenau, 2003.

WANG, Yuelin; ZHANG; Yihan; LU; Yan; YU; Xinran. **A Comparative Assessment of Credit Risk Model Based on Machine Learning-a case study of bank loan data**. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019). Procedia Computer Science 174 (2020) 141-149. Elsevier. 2020.

WORLD HEALTH ORGANIZATION. **Suicide in the world**. **Global Health Estimates**. World Health Organization. 2019.

ZHONG, Ning; LIU, Chunnian; KAKEMOTO, Yoshitsugu; OHSUGA, Setsuo. **KDD Process Planning**. Association for the Advancement of Artificial Intelligence. KDD-97. p. 291-294.1997.