UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Lucas Verdade Godoy

Análise de padrões em casos de risco de suicídio utilizando aprendizado de máquina

Florianópolis
[2022]

Lucas Verdade Godoy

Análise de padrões em casos de risco de suicídio utilizando aprendizado de máquina

Trabalho de Conclusão de Curso do Curso de Graduação em Ciências da Computação do Campus Florianópolis da Universidade Federal de Santa Catarina para a obtenção do título de bacharel em Ciências da Computação.
Orientador: Prof. Dr. Mateus Grellert
Coorientador: Prof. Dr. Jônata Tyska Carvalho

Florianópolis
[2022]

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Godoy, Lucas Verdade
Análise de padrões em casos de risco de suicídio utilizando aprendizado de máquina / Lucas Verdade Godoy ; orientador, Mateus Grellert da Silva, coorientador, Jônata Tyska Carvalho, 2022.
112 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Ciências da Computação, Florianópolis, 2022.

Inclui referências.

1. Ciências da Computação. 2. aprendizado de máquina. 3. análise de dados. 4. clustering. 5. suicídio. I. Grellert da Silva, Mateus. II. Tyska Carvalho, Jônata. III. Universidade Federal de Santa Catarina. Graduação em Ciências da Computação. IV. Título.

Lucas Verdade Godoy

Análise de padrões em casos de risco de suicídio utilizando aprendizado de máquina

Florianópolis, 07 de março de 2022.

Coordenador do Curso

Banca Examinadora:

Prof. Dr. Mateus Grellert
Orientador

Prof. Dr. Jônata Tyska Carvalho
Coorientador

Prof[a]. Manuella Kaster
Membro da banca

Prof. Rafael de Santiago
Membro da banca

Dedico este trabalho a todos que acreditaram em mim, seja de longe ou de perto.

AGRADECIMENTOS

É difícil botar em palavras a gratidão que sinto por todos que fizeram parte da minha caminhada até aqui. Cada um contribuiu com esse trabalho e comigo de formas e momentos diferentes. E, apesar das contribuições serem tão diferentes, a essência foi a mesma: vocês sempre me seguravam quando eu estava prestes a cair. Quando eu perdia as forças e deixava de acreditar em mim mesmo vocês estavam lá pra me ajudar, me acolher e lembrar do que eu sou capaz. Vocês também me lembravam de viver, relaxar e aproveitar a vida quando eu mesmo já tinha esquecido. Agradeço a cada palavra e ação de incentivo que minha família me deu, acreditando que eu poderia ir além. Cada caixa de papelão da Natura que guardavam para mim na infância, porque eu queria muito transformar aquilo num robô, carro ou nave. Cada LEGO que me davam, porque viam que eu amava brincar com aquilo e ficar inventando coisas novas. Cada palavra de carinho que me motivou a ser uma pessoa melhor. Cada vez que vocês viam brilho em meus olhos, e faziam de tudo para que eu continuasse sendo eu mesmo, me levou a esse momento. Aos meus amigos, que considero minha família também, agradeço profundamente pelo apoio, a parceria, as refeições no RU, as risadas(que foram várias) e os puxões de orelha(que foram vários também). Cada vez que passei madrugadas fazendo trabalhos, estudando para provas, conversando besteiras e escrevendo o TCC com a companhia de vocês foi preciosa. Vou carregar esses momentos para sempre como parte de mim.

Enquanto escrevo esses agradecimentos, meu coração transborda de emoção lembrando desses momentos e até paro de escrever pra apreciar essas memórias. Obrigado a todos me ajudaram a construir o que sou hoje.

"Entre razões e emoções, a saída é fazer valer a pena"
(NX Zero, 2006)

RESUMO

Mortes por suicídio são um fenômeno global que tira a vida de cerca de 800 mil pessoas por ano. Atualmente, a avaliação do risco de suicídio é feita clinicamente e existem múltiplos fatores, externos e internos, que resultam em um risco alto. A literatura aponta um crescimento no uso de técnicas de Inteligência Artificial (IA) para apoiar profissionais da saúde que atuam na área. Levando-se isso em consideração, este trabalho propõe uma solução com mineração de dados, algoritmos de aprendizado de máquina para agrupamento de dados e algoritmo de árvore de decisão para encontrar padrões em perfis de indivíduos que cometeram suicídio. O processo de mineração de dados envolve etapas de agregação de dados de diferentes fontes, como do Sistema de Informação sobre Mortalidade (SIM) do DATASUS, da Classificação Brasileira de Ocupações (CBO-2002), da Classificação Internacional de Doenças (CID10) e de metadados dos municípios. Em seguida, o processo passa pela etapa de limpeza, transformação e de engenharia de atributos. Então, os dados são utilizados nos algoritmos de aprendizado de máquina, como o *K-Prototypes*, com o propósito de criar grupos similares de indivíduos. Os resultados dos agrupamentos são analisados visualmente por meio de técnicas de visualização de dados e de um modelo de árvore de decisão gerado a partir dos agrupamentos. O resultado das análises feitas levou a uma melhor compreensão dos perfis dos indivíduos. Para fins de validação, os resultados dos agrupamentos são comparados com a literatura. Com isso, chegou-se a dois principais perfis: o agrupamento de indivíduos do sexo feminino e o agrupamento dos indivíduos com ocupações relacionadas à agricultura. Isso indica que os padrões obtidos pelos agrupamentos estão alinhados com aqueles encontrados na literatura. Espera-se que este trabalho contribua com os estudos na área da saúde que utilizam de técnicas de mineração de dados e aprendizado de máquina para se obter mais informações do fenômeno de suicídio no Brasil visando a criar políticas públicas de prevenção e visibilidade para esse problema.

Palavras-chave: aprendizado de máquina, análise de dados, clustering, suicídio.

ABSTRACT

Deaths by suicide are a global phenomenon that takes the lives of nearly 800,000 people a year. Nowadays, suicide risk assessment is done clinically, and there are multiple factors, external and internal, that result in a high risk. The literature in the area shows an increasing use of Artificial Intelligence (AI) techniques to support health professionals working in the area. Taking that into consideration, this work proposes a solution using data mining, machine learning algorithms for data clustering and decision tree algorithm to find patterns in profiles of individuals who committed suicide. The data mining process involves data aggregation steps from different sources, such as the DATASUS Mortality Information System (SIM), the Brazilian Classification of Occupations (CBO-2002), the International Classification of Diseases (ICD10) and municipal metadata. After that, the process goes through the stage of cleaning, transformation and attribute engineering. Then the data is used in machine learning algorithms such as K-Prototypes for creating similar groups of individuals. The cluster results are visually analyzed using data visualization techniques and a decision tree model generated from the clusters. The result of the analysis carried out led to a better understanding of the profiles of individuals. For validation purposes, the clustering results are compared with the literature in the area. With that, two main profiles were found: the grouping of female individuals and the grouping of individuals with occupations related to agriculture. This indicates that the patterns obtained by the clusters are in line with those found in the literature. It is expected that this work contributes to studies in the health area that use data mining and machine learning techniques to obtain more information on the phenomenon of suicide in Brazil, in order to create public policies to prevent and give visibility to this problem. **Palavras-chave**: aprendizado de máquina, análise de dados, clustering, suicídio. **Keywords**: Machine Learning. Data Analysis. Clustering. Suicide

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Segundo a Organização Mundial da Saúde (WHO, 2017), cerca de 800 mil pessoas cometem suicídio anualmente pelo mundo. Isso coloca o suicídio como a 18ª maior causa de morte global, sendo que entre o jovens é a 2ª maior. Este problema também é bastante grave no Brasil. Segundo o Ministério da Saúde (SEHNEM; PALOSQUI, 2014), o estado de Santa Catarina concentra cerca de 21% dos casos de suicídios do país.

Esse fenômeno é o resultado de uma interação complexa entre estressores ambientais, como adversidades e eventos da vida, e traços de susceptibilidade da vítima, independente de transtornos psiquiátricos, segundo Van Heeringen e Mann (2014).

Atualmente, a avaliação de risco de suicídio é feita clinicamente (ASSOCIATION; ASSOCIATION *et al.*, 2013). No entanto, a compreensão limitada do aspecto epidemiológico aliada com os diversos outros diagnósticos associados dificulta a detecção de possíveis vítimas e, consequentemente, o estabelecimento de ações de prevenção e de intervenção (INSEL *et al.*, 2010). Isso é agravado pelo caráter multifatorial das causas que aumentam o risco de suicídio. Portanto, é necessário buscar soluções capazes de considerar diferentes fontes de dados para auxiliar na estimativa de risco de suicídio.

No campo da psiquiatria, o uso de técnicas de aprendizado de máquina (*Machine learning - ML*) vem crescendo e tem se mostrado ser um forte aliado. Essas técnicas permitem que seja feita uma análise de grandes quantidade de dados, geração de perfis e modelos entre grupos de pacientes (CABITZA; BANFI, 2018; GRAHAM *et al.*, 2019). No entanto, ainda não existe um conjunto de dados para análise de fatores do suicídio no Brasil, o que impede que as soluções propostas sejam aplicadas no contexto nacional.

O governo brasileiro tem buscado formas de facilitar o acesso automatizado a diversos dados públicos a fim de subsidiar análises objetivas da situação sanitária. O Departamento de Informática do Sistema Único de Saúde (DATASUS) fornece diferentes dados digitais sobre a saúde através de *Application Programming Interface* (API) (SOUZA; AUTRAN *et al.*, 2019). Em 2020, o DATASUS coordenou a criação da Rede Nacional de Dados em Saúde (RNDS), com uma API que disponibiliza acesso a vários dados de saúde (informações sobre pacientes, dados de exame, do examinador etc). Infelizmente, essa API só é acessível a estabelecimentos de saúde.

No entanto, estes dados são usualmente coletados e analisados de forma isolada, o que impede o desenvolvimento de soluções que contemplem os diversos fatores associados ao suicídio. Portanto, a integração destes dados em uma base única não é trivial. Para isso, técnicas de mineração de dados (*Data Mining* - DM) são necessárias para realizar essa integração de maneira correta, assim como realizar a limpeza e as devidas transformações para extrair informação desses dados de forma eficiente.

Este trabalho busca trazer contribuições nessa linha, propondo o uso de técnicas de aprendizado de máquina para auxiliar na busca por soluções relacionadas à saúde mental.

1.1 OBJETIVOS

O objetivo geral deste trabalho é criar um fluxo de mineração de dados para encontrar grupos similares de indivíduos que cometem suicídio. Como objetivos específicos, os seguintes foram identificados:

1. Realizar uma revisão do estado da arte relacionado a epidemiologia do suicídio, assim como a técnicas de aprendizado de máquina e análise de dados, com foco em saúde mental;
2. Criar novos conjuntos de dados através de agregações de dados de diferentes fontes. Utilizando técnicas de *Data Linkage* para efetuar essas integrações de informações diferentes em novos conjuntos únicos;
3. Realizar o pré-processamento dos dados coletados e verificar a presença de dados faltantes, incorretos e discrepantes, para então aplicar as respectivas técnicas de correção para cada problema;
4. Treinar modelos de *Machine Learning* de agrupamento de dados com base nos dados processados;
5. Avaliar o desempenho dos modelos treinados;
6. Comparar os padrões encontrados com as características do suicídio encontradas na literatura;
7. Divulgar e publicar este trabalho em conferência e plataformas de código aberto;

1.2 ESTRUTURA DO TRABALHO

O conteúdo dos capítulos deste trabalho foi estruturado da seguinte maneira:

- Capítulo 2: referencial teórico, que inclui os conceitos básicos e as técnicas usadas; avaliação do estado da arte seguida de uma comparação com o objetivo deste trabalho;
- Capítulo 3: idealização da solução proposta e explicação da metodologia de cada etapa;
- Capítulo 4: discorre sobre a etapa de coleta dos dados das diferentes fontes, sobre a preparação para agrupar alguns desses dados, sobre o significado dos dados coletados e a transformação deles em uma forma mais clara;
- Capítulo 5: apresenta o enriquecimento de dados e engenharia de atributos, em que novos dados são adicionados com base em outros já existentes;

- Capítulo 6: apresenta a aplicação dos algoritmos de aprendizado de máquina usando os dados gerados pelas etapas anteriores, bem como aborda o treinamento e uso dos modelos de aprendizado de máquina, de agrupamento de dados e de árvore de decisão;;
- Capítulo 7: propõe a visualização, comparação e avaliação dos resultados, obtidos no capítulo anterior, buscando interpretar os perfis de indivíduos encontrados pela solução aplicada.
- Capítulo 8: consolidação das conclusões tiradas deste trabalho e melhorias possíveis para trabalhos futuros;

2 REFERENCIAL TEÓRICO

Este capítulo é dividido em duas partes. Primeiro, são apresentados os conceitos básicos deste trabalho na seção 2.1. Então, são listados os trabalhos relacionados e encontrados na literatura 2.2. Ao final dessa segunda parte, há um sumário dos principais achados.

2.1 CONCEITOS BÁSICOS

2.1.1 Aspectos do fenômeno do suicídio

Uma das grandes preocupações de saúde pública no Brasil e no mundo é o fenômeno do suicídio. De fato, trata-se da 18ª maior causa de morte de toda a população mundial. Segundo dados da Organização Mundial da Saúde (WHO, 2017), cerca de 800 mil pessoas cometem suicídio por ano no mundo todo, número que equivale a cerca de 1,5% do total de mortes globais, sendo a maioria dessas mortes de homens. O suicídio não é apenas uma consequência de características do indivíduo; trata-se, na verdade, de um fenômeno que envolve a sociedade e o ambiente em que o indivíduo está condicionado a viver (VAN HEERINGEN; MANN, 2014), ou seja, trata-se de um fenômeno multifatorial, que abarca elementos intrínsecos e extrínsecos ao indivíduo. Os dados da OMS também demonstram como esse fenômeno atinge de maneira preocupante pessoas entre 15 e 29 anos, sendo a segunda principal causa global de morte nessa faixa etária, e é ainda particularmente preocupante em países de baixa e média renda, como o Brasil (SILVA, 2017)Alguns estudos, como os elaborados em "Determinantes espaciais e socioeconômicos do suicídio no brasil: Uma abordagem regional (2011)" (GONÇALVES *et al.*, 2011), indicam ainda que existe uma relação espacial de tentativas de suicídio. Nessa relação, os estudiosos apontam que regiões com uma alta taxa de tentativas provavelmente terão em sua vizinhança regiões que também apresentam taxas elevadas.

Devido à variedade de fatores que influenciam esse fenômeno social, a avaliação do risco de tentativa de suicídio se revela uma tarefa complexa. Numa tentativa de contribuir para os estudos da área, alguns estudos, como o de "(GONÇALVES *et al.*, 2011)", fazem uso de análise de dados e algoritmos de aprendizado de máquina para buscar identificar fatores preditores e associados ao risco da tentativa de suicídio. Nesse mesmo estudo, a localidade espacial foi identificada como um possível fator associado à taxa de tentativas nas microrregiões do Brasil.

2.1.2 Plataformas de dados de saúde

Levando-se em consideração o contexto brasileiro, buscou-se nas plataformas de dados de saúde informações e dados que pudessem enriquecer esta pesquisa. Nesta seção, propõe-se uma apresentação dessas plataformas.

Em 1991, surgiu o Departamento de Informática do Sistema Único de Saúde (DATASUS) e durante seu tempo de atuação foram desenvolvidos mais de 200 sistemas que auxiliam a construção e fortalecimento do Sistema Único de Saúde (SUS). O departamento também fornece informações e dados de diversos aspectos da saúde da população brasileira para que decisões sejam tomadas com base em evidências e indicadores de saúde. Há muitos anos essa informação está disponibilizada para acesso público, mas até recentemente ela dependia de softwares específicos para seu acesso e eles não possuíam nenhum tipo de *Application Programming Interface* (API) para acesso automatizado.

Em 2016, foi criada uma biblioteca em Python, conhecida como *PySUS*, com o objetivo de resolver esse problema (COELHO *et al.*, 2021). A biblioteca PySUS é um pacote para linguagem Python que reúne diversos utilitários para lidar com bancos de dados públicos publicados pelo DATASUS (COELHO, s.d.). Ela foi criada inicialmente para auxiliar nas pesquisas envolvendo os casos de dengue no Brasil através de análises dos dados do banco Sistema de Informações de Agravos de Notificação (SINAN). A versão atual da biblioteca PySUS possui acesso a diversos outros bancos, permitindo que ela seja utilizada em pesquisas as mais variadas voltadas à saúde.

A figura 1 representa a estrutura do DATASUS e módulos, de diferentes setores da saúde, que estão ligados a ele. Através da biblioteca PySUS é possível acessar os bancos de dados dos módulos do SIM, SINASC e SIA [1].

Dentre esses bancos de dados, está incluso o banco do Sistema de Informação sobre Mortalidade (SIM), um dos principais instrumentos para o apoio e criação de políticas de prevenção e cuidado em saúde, que foi desenvolvido em 1975 e coleta dados de mortalidade no país ao longo dos anos e é a fonte oficial que possibilita a recuperação de informações de óbitos por diferentes naturezas. Esse sistema também possui informações por indivíduo a respeito da causa de sua morte, incluindo dados socioeconômicos, local de residência e de ocorrência, causa do óbito, entre outras informações mais específicas, como a faixa etária, o estado civil e o sexo biológico.

2.1.3 Algoritmos de Aprendizado de Máquina

Um agente aprende quando melhora sua performance em tarefas futuras após a realização de observações a respeito do mundo, é o que apontam (RUSSEL; NORVIG) em Artificial Intelligence A Modern Approach Third Edition (s.d.), obra em que também definem que há três tipos principais de aprendizado, que são: o aprendizado supervisionado, o não-supervisionado e o aprendizado por reforço. Neste trabalho são abordados apenas modelos de aprendizado de máquina supervisionado e não-supervisionado.

A etapa de treinamento desses modelos é o momento em que acontece o aprendizado de máquina, pois os dados de treinamento estão sendo usados para que o modelo tente aprender os seus padrões. Modelos de aprendizado supervisionado recebem pares de dados

[1] Disponível em (COELHO *et al.*, 2021)

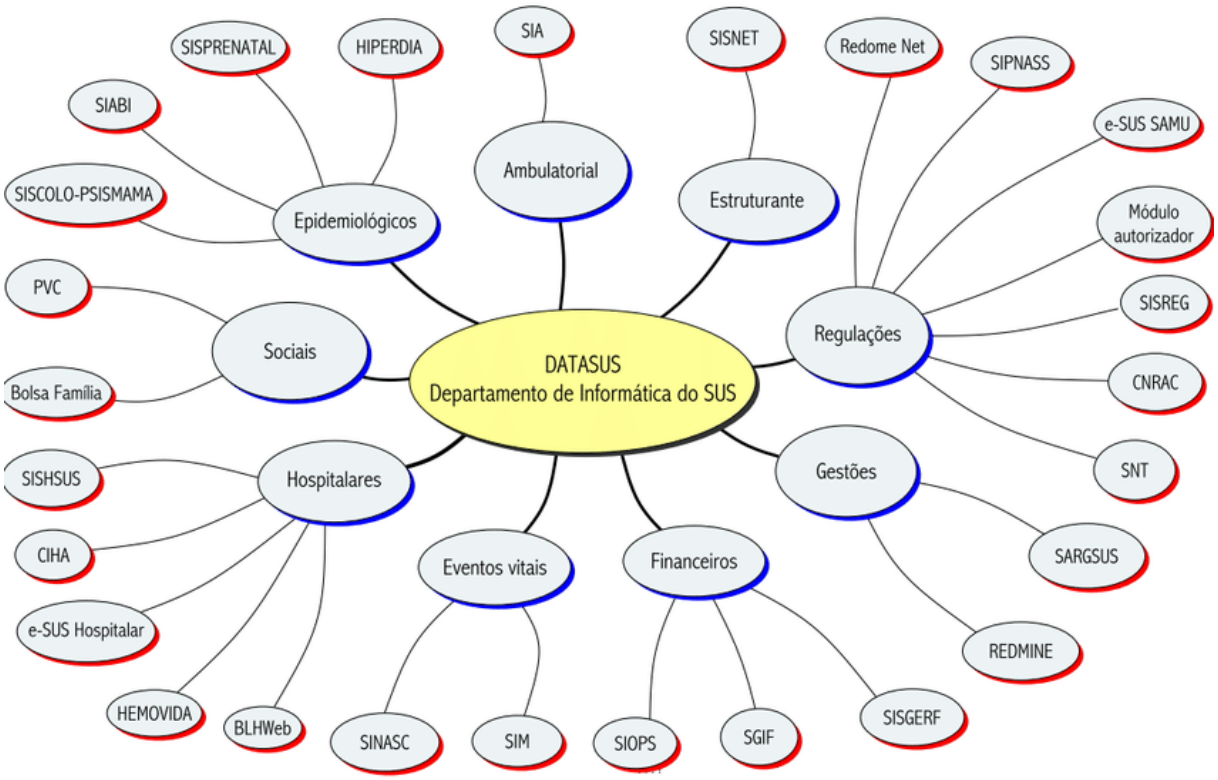

Figura 1 – Mapa Conceitual - Alimentação do DataSUS (Fonte: (SOUZA; AUTRAN *et al.*, 2019))

de entrada e respostas e, a partir disso, o modelo faz observações para tentar encontrar o padrão que mapeia a entrada para o resultado corretamente. Como esse modelo tem as respostas corretas de cada par de entrada, é possível calcular um *feedback* de quão próximo o resultado do modelo, que foi gerado através do padrão aprendido até o momento, chegou ao resultado real, que era a saída para aquele par de entradas no mundo do modelo. A supervisão usada nesse tipo de modelo, então, é o acesso às respostas reais para se calcular o *feedback* e fazer ajustes, caso necessário.

No caso deste trabalho, os agentes são os modelos de aprendizado de máquina e o seu mundo é o conjunto de dados que é usado no treinamento dos próprios agentes. Já o aprendizado não-supervisionado, este não possui os dados de saída e, portanto, não pode calcular um *feedback* explícito baseado na proximidade de seu resultado com o resultado real. Nesse tipo de aprendizado, o modelo tenta aprender os padrões nos dados de entrada apenas. A definição e explicação dos algoritmos de aprendizado de máquina usados neste trabalho se encontram abaixo.

Algoritmo: K-means

É um dos algoritmos de aprendizado de máquina não-supervisionado mais populares, apesar de suas limitações (AHMED; SERAJ; ISLAM, 2020). Ele é usado para fazer

agrupamento, ou clustering, de dados baseado em suas características.

Ou seja, o algoritmo k-means é usado para dados numéricos. Esse algoritmo depende do valor de K, que deve ser sempre especificado para performar uma análise de *clustering* (AHMED; SERAJ; ISLAM, 2020).

Dado um conjunto X = [X0, X1, X2, .., Xn] de dados e um número inteiro, maior que zero, K será o número de agrupamentos a serem encontrados. O funcionamento desse algoritmo segue os passos abaixo:

1: Escolhe-se aleatoriamente K elementos, diferentes entre si, de X para serem os centros de cada agrupamento.

2: Para cada dado em X, calcula-se a dissimilaridade entre o dado atual e os centros dos agrupamentos;

3: A partir das dissimilaridades calculadas, atribuem-se os dados aos agrupamentos que tem a menor dissimilaridade;

4: Recalcula-se o centro dos agrupamentos reatribuindo seus valores com a média dos valores dos dados pertencentes àquele grupo.

5: Repetir os passos 2, 3 e 4 até encontrar a posição ideal do centro dos agrupamentos.

Esse algoritmo é específico para dados com atributos do tipo numéricos e seu cálculo de dissimilaridade é feito usando a distância euclidiana entre o dado e o centro, considerando os atributos numéricos do dado como as coordenadas.

Algoritmo: K-modes

Como aponta K-modes Clustering (2001), o algoritmo K-means não é apropriado para dados categóricos, para tanto, faz-se necessário o algoritmo K-modes. Trata-se de uma extensão do algoritmo K-means, portanto, assim como aquele, é um algoritmo de aprendizado de máquina não-supervisionado usado para fazer agrupamento, ou clusterização, de dados com base em suas características. Porém, essa versão é feita para dados categóricos e, por isso, não é possível usar o cálculo da distância euclidiana entre dois pontos. Sua função de cálculo de dissimilaridade calcula uma pontuação de dissimilaridade baseada na hipótese dos valores dos atributos de dois dados serem ou não diferentes.

Dado um conjunto X = [X0, X1, X2, .., Xn] de dados e um número inteiro, maior que zero, K será o número de agrupamentos a serem encontrados. Em que Xi, para 0 <= i <= n, é composto por (A0, A1, A2,..., Am) que representam tuplas de atributos categóricos, em que Aj, para 0 <= j <= m, é um atributo categórico. O cálculo de diferença de atributos é feito de forma condicional, como explicado a seguir:

```
Sejam Ai0 e Aj0 atributos de diferentes dados:
 - Caso Ai0 = Aj0 então a diferença dos atributos é 0,
 - Caso Ai0 != Aj0 então a diferença dos atributos é 1.
```

Então, o cálculo da dissimilaridade de dois dados com atributos categóricos é feito com o somatório dos cálculos de diferenças entre seus atributos.

Algoritmo: K-prototypes

Trata-se de uma extensão dos dois algoritmos de agrupamentos citados anteriormente. Essa versão, entretanto, foi feita para ser usada em conjuntos de dados mistos, ou seja, que têm atributos numéricos e atributos categóricos. O cálculo de dissimilaridade dessa versão é a soma do cálculo das distâncias euclidianas dos atributos numéricos com a pontuação de dissimilaridade dos atributos categóricos.

Por fim, a partir de um conjunto de protótipos de cluster iniciais, esse processo atribui cada objeto a um cluster e atualiza o protótipo de cluster de acordo após cada atribuição.

Algoritmo: Árvores de decisão

Árvore de decisão é uma das formas de aprendizado de máquina mais bem sucedida e simples, como apontam (RUSSEL; NORVIG). Esse algoritmo recebe um vetor de atributos de entrada e retorna um único valor de saída, que representa a "decisão". Nesse caso, a decisão da árvore é a classificação de a qual cluster os registros pertencem baseado em seus atributos de entrada. O algoritmo faz uma sequência de testes nos atributos de entrada para chegar em suas decisões. Esses testes são representados como nós internos da árvore, ou seja, que não são nós do tipo folha, que, como definem (RUSSEL; NORVIG), são aqueles que não têm filhos na árvore. Os testes consistem em uma comparação de um atributo de entrada com um dos possíveis valores dele e, para os dados em que a comparação é verdadeira, é criada uma ramificação que vai para outro nó teste ou nó folha. Para os dados em que a comparação não é satisfeita, é criada outra ramificação que também pode ir para um nó teste ou nó folha. Dessa forma, a cada teste feito os dados vão se separando até chegarem em seus nós folhas com suas decisões. O algoritmo da árvore de decisão pode ser visto com mais detalhes na figura 2.

Nesse exemplo, (RUSSEL; NORVIG) mostram uma árvore de decisão feita para decidir se vale a pena esperar por uma mesa em um restaurante ou não. Nesse caso, o vetor de atributos de entrada possui variáveis que indicam aspectos da situação de entrada como, por exemplo, se há ou não outra alternativa de restaurante por perto, se o dia da decisão é sexta ou sábado, se os indivíduos da decisão estão com fome ou não, se o restaurante é barato ou caro, se está cheio, vazio ou com algumas pessoas, entre outros aspectos. Os testes então são feitos com os valores possíveis desses atributos. Seu valor de saída é a decisão *VaiEsperar* e seus valores possíveis são verdadeiro e falso, portanto trata-se de uma decisão booleana. A Figura 3 apresenta uma visualização da árvore de decisão gerada para esse exemplo.

function DECISION-TREE-LEARNING($examples, attributes, parent\_examples$) **returns** a tree

if $examples$ is empty **then return** PLURALITY-VALUE($parent\_examples$)
else if all $examples$ have the same classification **then return** the classification
else if $attributes$ is empty **then return** PLURALITY-VALUE($examples$)
else
  $A \leftarrow \text{argmax}_{a \in attributes}$ IMPORTANCE($a, examples$)
  $tree \leftarrow$ a new decision tree with root test A
  for each value v_k of A **do**
    $exs \leftarrow \{e : e \in examples$ **and** $e.A = v_k\}$
    $subtree \leftarrow$ DECISION-TREE-LEARNING($exs, attributes - A, examples$)
    add a branch to $tree$ with label $(A = v_k)$ and subtree $subtree$
  return $tree$

Figura 2 – Fonte: (RUSSEL; NORVIG, s.d.)

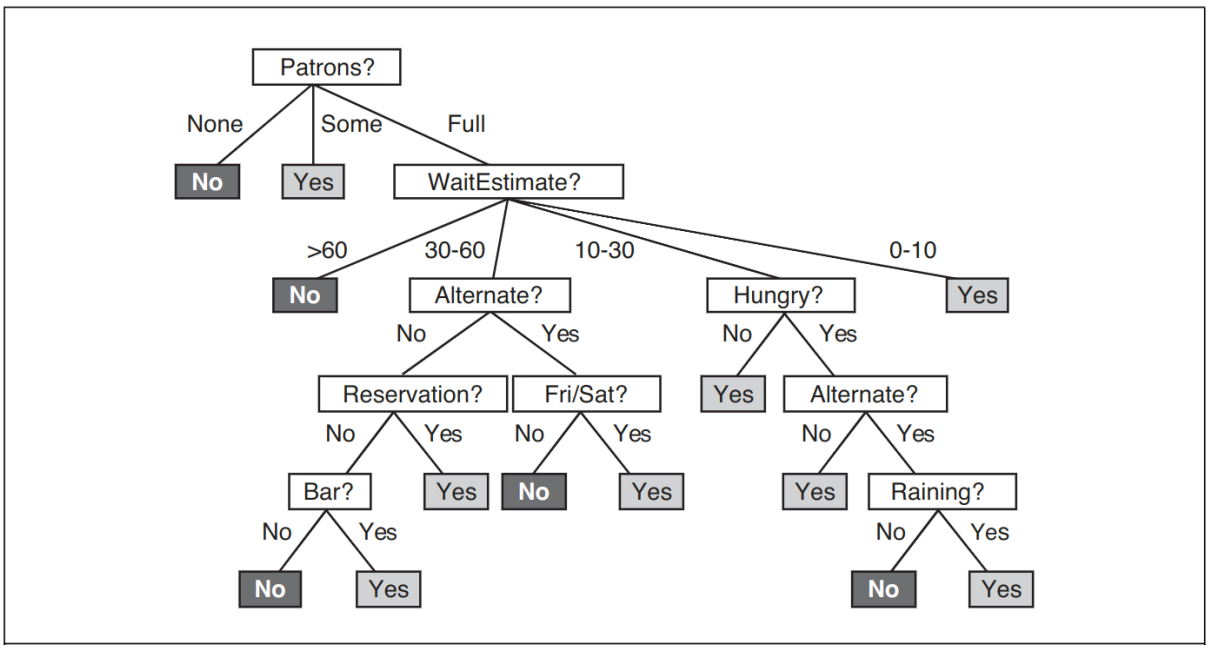

Figura 3 – Fonte: (RUSSEL; NORVIG, s.d.)

De acordo com (SINGH; GIRI, 2014) em *Comparative Study Id3, Cart And C4.5 Decision Tree Algorithm: A Survey (2014)*, três dos principais algoritmos de árvores de decisão são: *Classification and Regression Trees* (CART), C4.5 e *Iterative Dichotomizer* (ID3). As maiores diferenças entre esses algoritmos dizem respeito ao método usado para definir o melhor atributo para se fazer o teste, ou o atributo de maior importância, e o critério de parada do algoritmo. Neste trabalho, a implementação da árvore de decisão utilizada se baseia em uma versão otimizada do algoritmo CART[2] em que uma das diferenças se trata, por exemplo, da escolha do melhor atributo que utiliza a taxa de impureza de Gini. No algoritmo ID3, por sua vez, essa escolha é feita com o Ganho de Informação.

2.1.4 Fluxo de mineração de dados com foco em aprendizado de máquina

A Mineração de Dados (Data Mining - DM) estuda técnicas e ferramentas para coleta, tratamento, análise e extração de informação presente em dados de diversos tipos (textos, valores formatados, imagens etc) (AGGARWAL, 2015). Ou seja, o termo mineração de dados implica uma série de diferentes aspectos do processamento de dados (AGGARWAL, 2015). A Figura 4 apresenta os principais passos desse processo.

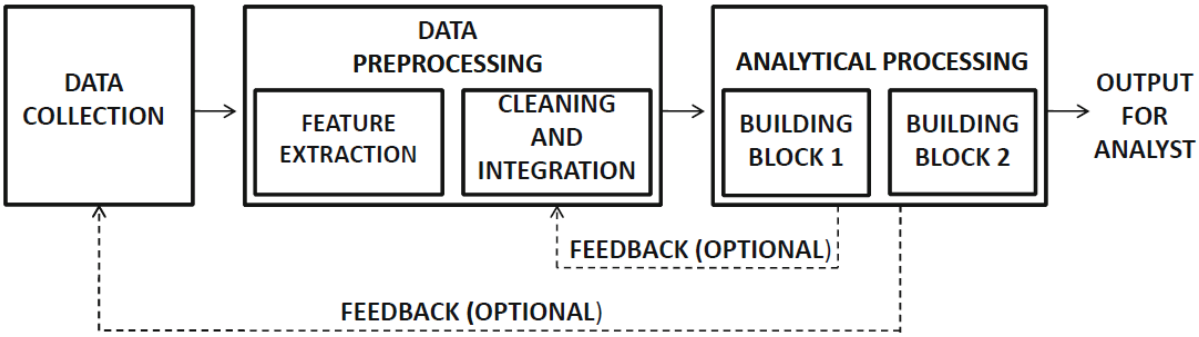

Figura 4 – *Pipeline* de processamento de dados (Fonte: (AGGARWAL, 2015))

Para que o processo inicie é preciso definir quais serão os dados usados na mineração e quais suas fontes. A etapa de coleta trata de buscar e trazer esses dados e, geralmente, armazená-los em algum lugar para uso posterior. A coleta de dados pode ser feita de diversas maneiras, como com questionários respondidos, a partir de bancos de dados já existentes, com *scraping* (garimpagem) de dados de páginas da web, com dados coletados de sensores e com dispositivos que medem algum fenômeno. Nessa etapa é importante que as decisões de como a coleta será feita sejam alinhadas para favorecer o processo de mineração de dados.

Os dados coletados muitas vezes não vêm num formato amigável para o processamento analítico e o uso de algoritmos. Na etapa de limpeza, deve ser feita a transformação

[2] Cf. https://scikit-learn.org/stable/modules/tree.html#tree-algorithms

dos dados para um padrão mais adequado para a próxima etapa, de processamento. Às vezes o dado pode vir faltante ou errôneo e precisará ser corrigido ou estimado. É na etapa de limpeza e transformação que se aplicam as técnicas de enriquecimento, em que novas propriedades, ou *features*, são extraídas a partir dos dados existentes com engenharia de atributos. Um exemplo simples da necessidade dessa etapa é a seguinte situação: suponha que você coletou dados de indivíduos e a idade do indivíduo é relevante para sua análise. Porém, nos dados há apenas a data de nascimento, que também está em um padrão diferente do esperado. Portanto, é necessário transformar esse dado, para que fique adequado, e também extrair a nova propriedade 'idade' a partir dele. O resultado será um novo conjunto de dados limpo, possivelmente com novas features e pronto para ser usado na próxima etapa.

Em seguida, a etapa de análise permite que novos conhecimentos, novas relações, novas regras sobre os dados iniciais sejam descobertos. Essa análise pode envolver desde testes estatísticos, como ferramentas de visualização para gerar dados mais interpretáveis aos pesquisadores. Alguns algoritmos, e/ou suas implementações, são elaborados para aceitar apenas um tipo de dado de entrada, como categórico ou numérico. Em situações onde o dado de entrada não é compatível com o tipo exigido pelo algoritmo, pode ser possível fazer uma adaptação desse dado para o tipo adequado. Um exemplo disso é quando há um dado de entrada do tipo categórico, mas o algoritmo que usará eles aceita apenas numéricos. Nesse caso, uma das adaptações possíveis de se fazer é aplicar a técnica chamada *one hot encoding*. Essa técnica foi inicialmente proposta por Huffman (1954) e, a partir de então, foi amplamente aplicada em áreas diversas devido a sua simplicidade, o que a torna também umas das técnicas mais populares quando se trata de estratégia de projetar um algoritmo (YU *et al.*, 2020). No caso específico do aprendizado de máquina, essa técnica é utilizada para processar recursos discretos (YU *et al.*, 2020) e uma de suas vantagens é o fato de ela tornar possível a binarização de inputs categóricos para que então sejam considerados como vetores do espaço euclidiano, o que é amplamente usado para calcular distâncias e/ou similaridades entre atributos em muitos algoritmos para classificação (YU *et al.*, 2020). Colocado simplesmente, quando falamos de one hot encoding, deixamos implícito que todos os valores do mesmo atributo categórico estão igualmente distantes uns dos outros. Ou seja, com a técnica de one hot encoding, cria-se uma coluna nova para cada categoria possível do dado original, em que os valores dessas colunas são 0 ou 1, sendo 0 o valor que indica que o dado não é dessa categoria e 1, que o dado é dessa categoria. Dessa forma o algoritmo consegue usar esses dados em seu treinamento.

Uma das técnicas usadas para agregar mais informações aos dados coletados se chama Data Linkage, que é um método de mineração de dados. Esse método permite a combinação dos dados com fins de resultar em maior abundância de conhecimento relevante (ZHENG; CAI; LI, 2018). Como apontam (ZHENG; CAI; LI, 2018), conteúdos

relacionáveis podem ser utilizados para se chegar a observações abrangentes e confiáveis.

As técnicas discutidas acima são particularmente eficientes quando combinadas com algoritmos de Aprendizado de Máquina. De acordo com (MITCHELL, 1997), soluções baseadas em aprendizado de máquina aprendem a desenvolver uma tarefa através de dados, utilizando uma medida objetiva para medir seu desempenho.

2.2 TRABALHOS RELACIONADOS

Nesta seção, serão apresentados resumos e pontos relevantes de alguns artigos relacionados selecionados através do levantamento bibliográfico feito.

Suicídio : investigando as causas por meio da análise de dados? (SILVA, 2017)

O objetivo dessa obra é mostrar como técnicas de mineração de dados, aprendizado de máquina e análise de padrões podem auxiliar na identificação de risco de suicídio e diagnóstico de depressão no Brasil. O banco de dados usado para a análise foi do IPEA (Instituto de Pesquisa Econômica Aplicada) e nele haviam dados do período de 2001 a 2010.

Os dados presentes nesse banco eram: quantidade de suicídios cometidos separados pelo sexo biológico, quantidade de suicídios cometidos por jovens (19 a 25 anos), tamanho da população de cada município e estado, taxa de suicídio a cada 100 mil habitantes, taxa de suicídio de jovens, homens e mulheres a cada 100 mil habitantes, renda per capita de cada estado e IDHM de Educação, Longevidade e Renda.

É feita uma análise da relação entre a proporção de ocorrências pela população e o IDH de cada região do Brasil. E é possível perceber que existe uma relação inversamente proporcional desses dados.

Em outra análise é mostrado a renda per capita de cada região e o autor menciona que uma possível conclusão dessas análises seria "que a Região que mais precisa de atenção do Governo Federal e Ministério da Saúde para casos de suicídios é o Nordeste, e não a região Sudeste como a maioria das pessoas pensaria por relacionar o cotidiano das grandes metrópoles à depressão e, consequentemente, suicídios."Nas considerações finais conclui-se que a análise dos dados coletados permitiu encontrar possíveis fatores, como a relação inversa do número de suicídios e o índice de desenvolvimento humano da região do Brasil que residem os indivíduos, que influenciam e direcionam pessoas a se suicidar. Também é adicionado que com mais parâmetros nos dados seria possível encontrar mais desses fatores.

Predictors of suicide attempt in patients with obsessive-compulsive disorder: an exploratory study with machine learning analysis **(AGNE *et al.*, 2020)**

Esse artigo busca fatores de risco de tentativa de suicídio em pessoas com Transtorno obsessivo-compulsivo e esclarecer se os fatores encontrados estão mais relacionados com o transtorno ou com fatores extrínsecos, como variáveis sociodemográficas ligadas a pessoa e outras comorbidades.

O estudo incluiu dados de 959 pacientes com TOC, incluindo dados clínicos e sociodemográficos. Seus resultados concluíram que 10.8% da amostra apresentava risco de tentativa de suicídio. Foram encontrados os seguintes fatores relevantes para predizer o risco de tentativa de suicídio: precedente de planejamento de suicídio e de pensamentos suicidas, episódios depressivos durante a vida e desordem explosiva intermitente.

Os autores ressaltam que esse foi o primeiro estudo a avaliar fatores de risco para tentativa de suicídio entre pacientes com TOC, usando algoritmos de aprendizado de máquina. A conclusão desse estudo menciona que os resultados obtidos mostram que é possível criar um algoritmo acurado para predizer risco de suicídio usando dados clínicos e sociodemográficos. Também exalta que comorbidades com sintomas depressivas devem receber atenção extra no momento do diagnóstico clínico e, além disso, pacientes com TOC precisam de um diagnóstico clínico cuidadoso sobre todos os aspectos do fenômeno do suicídio.

Prediction of attempted suicide in men and women with crack-cocaine use disorder in Brazil **(ROGLIO *et al.*, 2020)**

Nesse artigo, é reconhecida a falta de estudos em busca de fatores preditores de risco de suicídio em indivíduos com transtorno por uso de substâncias. O objetivo do artigo é investigar esses fatores preditores nesses indivíduos com transtorno por uso de cocaína e/ou crack usando duas abordagens analíticas diferentes, uma descritiva e outra preditiva. A abordagem descritiva usou regressão de Poisson com variância robusta e a abordagem preditiva usou o algoritmo de aprendizado de máquina chamado Floresta Aleatória. Ambas abordagens foram usadas de forma estratificada por gênero. O banco de dados usado foi mesclado de bancos de dados secundários de duas instituições especializadas no tratamento da dependência química de Porto Alegre/RS.

Os resultados desse estudos indicam que a tentativa de suicídio está associada com depressão, alucinações e internações anteriores por motivos de questões mentais, tanto para homens como para mulheres.

QUEM SÃO OS ESTUDANTES DE MEDICINA QUE TENTAM SUICÍDIO? (MARCON, 2019)

Essa dissertação apresenta o problema que estudantes de medicina tem um maior risco de tentativa de suicídio comparados com a população geral. O objetivo desse estudo é encontrar fatores associados entre essa população e a tentativa de suicídio para que possam acontecer identificações e intervenções mais precoce em alunos de risco.

Para identificar estes fatores foi usado regressão de Poisson multivariada e um algoritmo, de aprendizado de máquina, *Elastic Net Regularization* para reconhecer os padrões dos alunos que tentam suicídio. Esse estudo teve a participação de 4840 estudantes de medicina. E foram coletados dados relacionados à saúde mental e à universidade, estilo de vida e dados demográficos desses indivíduos. Nessa amostra teve uma prevalência de tentativa de suicídio de 8,94%. O estudo conclui que seria possível implementar intervenções personalizadas ao identificar sujeitos sob maior risco de tentativa de suicídio através de algoritmos de risco.

Determinantes espaciais e socioeconômicos do suicídio no Brasil: uma abordagem regional (GONÇALVES *et al.*, 2011)

Esse artigo tem como objetivo avaliar a relação das taxas de suicídio das microrregiões do Brasil com os seus aspectos socioeconômicos, considerando também o aspecto espacial. A base da dados usada nesse estudo, que foi fornecida pelo IPEA (Instituto de Pesquisa Econômica Aplicada) cuja fonte principal foi o Sistema Único de Saúde (SUS), contém dados sobre suicídios por microrregiões brasileiras. Os dados são referentes ao período de 1998 até 2002.

Este estudo propôs uma hipótese de "efeito contágio"espacial que pode ser confirmada nos resultados e também na própria análise exploratória. O estudo mostrou que existe autocorrelação espacial positiva. Também foram analisados os fatores de pobreza e grau de ruralização das microrregiões. Foi encontrada uma relação inversamente proporcional entre o grau de pobreza e as taxas de suicídio. Já o grau de ruralização teve uma relação diretamente proporcional as taxas de suicídio.

O estudo ressalta, em sua conclusão, que o suicídio não envolve apenas o indivíduo, mas também a sociedade como um todo. Por esse motivo é importante sensibilizar a sociedade sobre o assunto e investir em políticas que garantam o cuidado da saúde mental e a promoção da qualidade de vida.

2.2.1 Sumário dos trabalhos relacionados

A tabela abaixo apresenta um resumo comparativo dos trabalho relacionados listados anteriormente.

Artigo	Objetivo	Algoritmos	Dados	Resultados
(GONÇALVES *et al.*, 2011)	Relação espacial da taxa de suicídio.	Índice de Moran; Econometria espacial	Dados de microrregiões brasileiras de 1998 - 2002 do IPEA	Comprovou o ëfeito contágio" a partir da amostra.
(SILVA, 2017)	Fatores que influenciam o suicídio.	Árvore de decisão	Dados de estados e regiões do IPEA 2001-2010	Demonstrou relação entre tamanho da população, Índice de Desenvolvimento Humano, renda e a taxa de suicídio.
(MARCON, 2019)	Fatores de risco de suicídio em alunos de medicina	Regressão de Poisson; Elastic Net Regularization	Estudo próprio com 4840 estudantes.	Modelo Elastic Net teve uma área abaixo da curva de 0,83.
(AGNE *et al.*, 2020)	Fatores de risco de suicídio em indivíduos com TOC	Elastic Net Regularization	Estudo próprio com 959 pacientes com TOC.	Os fatores de risco não estão diretamente ligados com o TOC.
(ROGLIO *et al.*, 2020)	Fatores de risco de suicídio em indivíduos viciados.	Regressão de Poisson; Floresta Aleatória	Dados individuais de instituições de tratamento de dependência química	Fatores: depressão, alucinações e internações por questões mentais.

Tabela 1 – Resumo comparativo dos trabalhos relacionados

Os principais achados nos trabalhos listados acima foram: (i) a relação espacial do suicídio entre microrregiões (GONÇALVES *et al.*, 2011), reforçando a necessidade de incluir dados sobre a residência do indivíduo; (ii) a relação com comorbidades mentais pré-existentes, como depressão, pode influenciar (ROGLIO *et al.*, 2020); (iii) e por fim a relação da taxa de suicídio com a renda do local de residência (SILVA, 2017).

Observou-se ainda que os trabalhos listados usam ou dados de indivíduos, coletados de estudos, ou dados agrupados por regiões, coletados de entidades como a IPEA.

O presente trabalho de conclusão de curso visa utilizar dados de indivíduos de um conjunto do DATASUS e enriquecê-los com atributos do ambiente que residem e que influenciam o risco de suicídio. Assim, com a grande quantidade de dados no DATASUS e com os diversos atributos criados ao enriquecer os dados, o modelo aqui proposto tem a possibilidade de compreender melhor os vários fatores influenciadores e atingir uma acurácia maior, contribuindo, desse modo, com os estudos que vêm sendo elaborados no campo da saúde pública brasileira.

3 SOLUÇÃO PROPOSTA

O fluxo idealizado da solução proposta por este trabalho é apresentado na Figura 5, a qual contém ainda informações dos dados utilizados para fazer *Data Linkage* e estabelece o que são as etapas de preparação e modelagem.

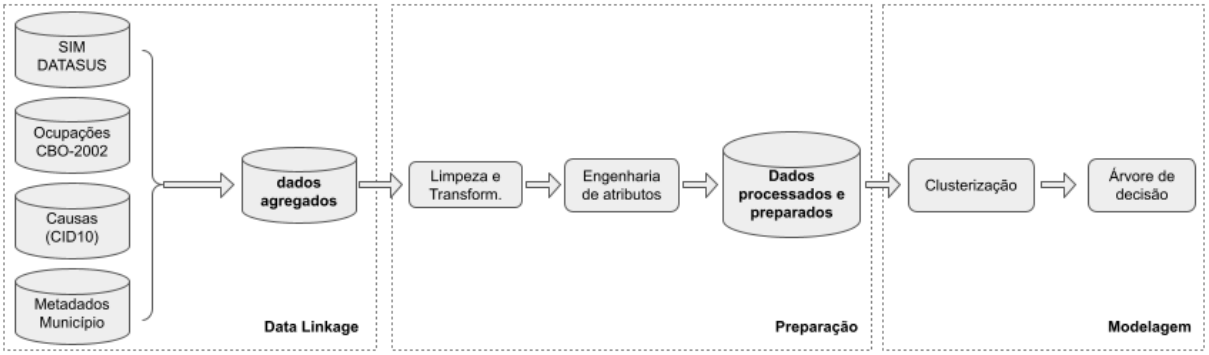

Figura 5 – Fluxo idealizado da solução. (Fonte: o autor)

A partir do que pode ser observado na Figura 5, o modelo idealizado por este trabalho é composto por três etapas: 1) Data Linkage com a agregação de diferentes dados; 2) preparação, com limpeza e transformação de dados, e engenharia de atributos; 3) modelagem, usando modelos de agrupamento e de árvores de decisão.

3.1 *DATA LINKAGE*

Data Linkage é um método de mineração de dados que permite a combinação deles com fins de resultar em maior abundância de conhecimento relevante (ZHENG; CAI; LI, 2018, p. 55). Levando-se em consideração, que conteúdos relacionáveis podem ser utilizados para se chegar a observações abrangentes e confiáveis (ZHENG; CAI; LI, 2018), na primeira etapa do modelo proposto é feito o Data linkage com a coleta de dados de diferentes fontes, no caso em particular desta pesquisa, eles se referem aos dados de mortalidade obtidos pelo SIM (DATASUS), os dados de ocupações da Classificação Brasileira de Ocupações Versão 2002 (CBO-2002) (CLASSIFICAÇÃO..., s.d.), os dados da 10ª revisão da Classificação Internacional de Doenças (CID10) (ICD-10..., s.d.), com suas categorias e subcategorias das enfermidades, e os metadados dos municípios, como nome, se se trata de uma capital, a latitude e a longitude. Tais dados são, então, agregados e servem de entrada para algoritmos de aprendizado de máquina cujo objetivo é gerar modelos que agrupam os dados em grupos homogêneos. No caso em particular, tratam-se de agrupamentos de indivíduos que cometeram suicídio. Em relação aos dados do SIM (DATASUS), eles foram coletados na biblioteca PySUS, que é um pacote para linguagem Python que possui diversas APIs. Essas APIs disponibilizam o acesso público a diferentes dados da área de saúde, como informações sobre os pacientes, dados dos exames e do

examinador, entre outras. Os metadados dos municípios também foram coletados por meio da biblioteca PySUS. Já a coleta dos dados das ocupações e das enfermidades CID10 foi feita manualmente pelo autor. A coleta desses dados será descrita com maior detalhamento mais adiante neste trabalho, na seção 3.4.1. Os dados coletados são então conectados para gerar um novo conjunto, denominado conjunto de dados agregados. Essa agregação faz com que um número maior de informações disponibilizadas a serem analisadas em conjunto, permitindo, com isso, uma complexificação das análises que serão feitas.

3.2 PREPARAÇÃO

A partir da agregação dos dados, entra-se na segunda etapa, denominada preparação. Nessa parte, faz-se necessária uma limpeza dos dados agregados, ou seja, a eliminação de atributos errôneos, faltantes e/ou com valores discrepantes. Essa primeira parte é realizada visando maior eficácia da análise, que então poderá ser feita a partir desses novos dados filtrados. Um exemplo de como a análise pode ser prejudicada nesse sentido, sem que a limpeza fosse feita, é a digitação de dados errados nos sistemas, como um valor de hora do óbito com caracteres inválidos, por exemplo "07/00", ou um horário inválido, por exemplo "9999". Junto da etapa de limpeza são feitas transformações dos dados para deixar seus significados mais claros, como a tradução de um atributo que está codificado para sua versão decodificada ou a transformação do campo da hora do óbito do seu formato de texto, representando horas e minutos, para um formato de variável do tipo horário. Após isso, alguns dados que não foram transformados com sucesso precisam passar pelo processo de limpeza novamente. Depois da limpeza e da transformação, passa-se para a parte em que é feita a avaliação dos atributos existentes pela engenharia de atributos, que consiste na transformação de dados existentes em novos atributos – por exemplo, criar um novo atributo informando o turno do dia baseado na hora de óbito – que sejam relevantes para o desfecho e o treinamento do modelo de aprendizado de máquina aqui proposto.

3.3 MODELAGEM

Após essas duas primeiras etapas, chega-se enfim à etapa de modelagem. Ela consiste na aplicação dos modelos de aprendizado de máquina no conjunto de dados, gerado nas etapas anteriores. A primeira parte é a aplicação dos dados em um algoritmo de agrupamento, cujo objetivo é agrupar casos semelhantes de suicídios em grupos homogêneos. Os resultados desse agrupamento foram utilizados para identificar grupos com características similares, os quais podem servir como base para intervenções direcionadas.

Além deste agrupamento, uma segunda etapa de treinamento supervisionado utilizando árvores de decisão foi realizada. O treinamento considerou como desfecho os próprios grupos gerados na etapa anterior. O objetivo deste experimento foi o de agregar semântica aos grupos com auxílio do modelos treinados, visto que árvores de decisão possuem uma

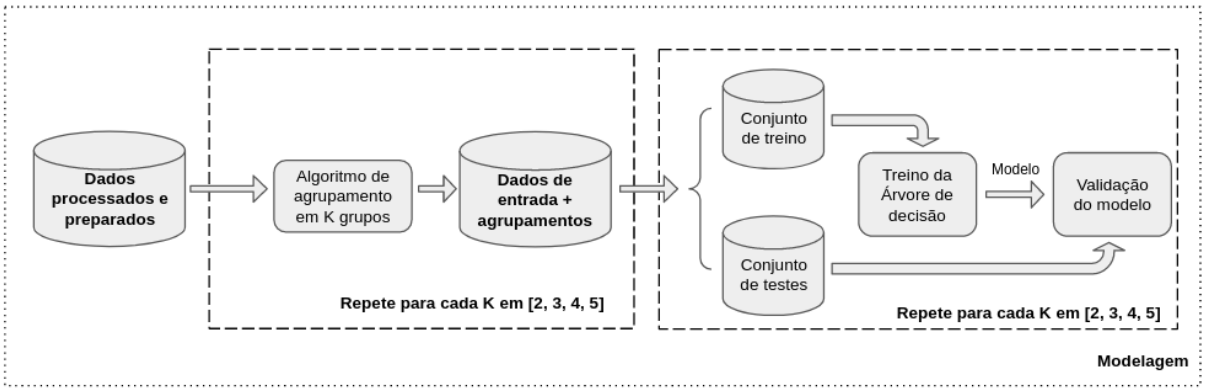

Figura 6 – Detalhes da etapa de modelagem. (Fonte: o autor)

estrutura bastante intuitiva e de fácil explicação. Para comparar a acurácia da árvore de decisão, os dados de entrada são separados em conjunto de treino e conjunto de teste, sendo que cada conjunto é separado entre atributos de entrada e resultado ou desfecho. Ao final dessa etapa, os resultados dos agrupamentos e a árvore de decisão gerada são comparados entre si, bem como com grupos de risco encontrados na literatura.

3.4 METODOLOGIA

A partir da revisão bibliográfica e da análise de dados coletados do DATASUS, o presente trabalho seguiu as etapas elencadas e discutidas abaixo.

3.4.1 Coleta, preparação e transformação dos dados

Utilizou-se o banco de dados do Sistema de Informação sobre Mortalidade (SIM), que é fornecido pelo DATASUS. A biblioteca PySUS disponibiliza uma funcionalidade que facilita o acesso a bancos de dados do DATASUS, incluindo o do SIM. Essa funcionalidade foi usada para agilizar o download de todos os bancos de dados necessários. Os conjuntos de dados são separados por estado e ano, então foi necessário baixar cada um separadamente e salvar em disco. Para facilitar o uso desses dados, criou-se um novo conjunto, contendo todos os dados baixados anteriormente. Nesse processo, foi identificado que a quantidade de atributos variava muito entre estados e anos, entretanto, nos conjuntos dos anos de 2010 em diante a variação era bem menor. Portanto, o novo conjunto de dados foi filtrado para os dados de 2010 até 2019. Além de ter sido limitado para a região sul do Brasil, pois o tamanho do arquivo dificultava a manipulação e exploração dele. O conjunto de dados gerado nessa etapa foi usado como base para as seguintes.

3.4.2 Enriquecimento e tradução dos dados

Partindo do conjunto de dados da etapa anterior, passou-se para o enriquecimento e tradução dos dados desse conjunto. Seus atributos categóricos estavam codificados, com

números inteiros representando seus valores. Para traduzi-los para o valor em texto foi necessário, para cada atributo categórico, usar um dicionário que mapeasse o valor numérico para seu respectivo valor em texto. Nesse caso, os dicionários usados são disponibilizados pela Fiocruz na página de documentação do SIM para a Plataforma de Ciência de Dados aplicada à Saúde (PCDaS). No conjunto há um atributo para o código IBGE do município. Esse foi o valor usado para enriquecer os dados do município de cada registro.

Uma das formas usadas para enriquecer os atributos no conjunto foi a técnica de engenharia de atributos, ou seja, a extração de novas informações a partir de atributos já presentes no conjunto original. Segundo (NARGESIAN *et al.*), a engenharia de atributos é a tarefa de melhorar o desempenho da modelagem preditiva em um conjunto de dados, transformando seu espaço de recursos. Para o presente trabalho, a engenharia de atributos foi aplicada para gerar dois novos atributos. Um deles foi o que indica em qual momento do dia ocorreu o óbito. Para criá-lo, utilizou-se a hora do óbito como informação original. O segundo atributo criado foi a indicação do local em que ocorreu o óbito (residência, trabalho, escola, estrada etc.). Essa informação foi extraída a partir do código CID10 da causa básica do óbito. Uma das subcategorias da categoria de lesões autoprovocadas indica o local do óbito, essa é a informação que se utilizou.

3.4.3 Aplicação dos algoritmos de aprendizado de máquina

Com os dados enriquecidos e traduzidos e os novos atributos criados, foi possível usar o conjunto em algoritmos de aprendizado de máquina. Os algoritmos de aprendizado de máquina foram usados para tentarmos encontrar grupos homogêneos no nosso conjunto de dados, ou algoritmo clustering, em aprendizado não supervisionado. Tal tarefa teve como objetivo deixar que o algoritmo por si só encontrasse agrupamentos que nos levassem a estabelecer grupos de risco, ou de atenção. Apenas os atributos que fizeram mais sentido foram usados no treino. No passo seguinte, usamos o algoritmo treinado para identificar o cluster de cada registro dos dados. Um dos parâmetros que é passado para o algoritmo em seu treinamento é o número de conglomerados a ser encontrado. O algoritmo encontra K agrupamentos de dados em cada execução sua, porém neste trabalho ele foi executado diversas vezes para diferentes números de agrupamentos. Os resultados dos labels de a qual cluster cada registro pertence são salvos em novas colunas. Os valores definidos arbitrariamente foram de 2, 3, 4 e 5 grupos, pois, conforme o número de grupos aumenta, a dificuldade de entender e explicar cada grupo aumenta também. Devido a essa complexidade, definimos o número máximo de agrupamentos como 5. Antes de poder usar esses resultados em um próximo algoritmo de aprendizado de máquina, foi necessário fazer um one hot encoding (get_dummies) das variáveis categóricas do conjunto, pois a implementação do algoritmo usado foi feita para lidar apenas com dados numéricos. Porém, nessa etapa, o algoritmo usa os resultados da clusterização para tentar explicar o que foi feito para decidir a qual cluster pertence cada registro. O algoritmo que faz isso é o de árvore

de decisão. Neste trabalho usamos esse algoritmo implementado na biblioteca scikitlearn[1]. Passamos para esse algoritmo os dados resultantes da etapa anterior (enriquecimento e engenharia de atributos) e pedimos para que ele gerasse a árvore para cada número de clusters.

3.4.4 Avaliação dos resultados

Avaliamos os resultados usando diferentes formas de visualização. Na primeira parte os resultados são avaliados através da visualização dos agrupamentos em um gráfico do mapa da região sul e da comparação entre eles para tentar compreender se há relação espacial ou não. Na segunda parte foram feitas análises dos atributos do conjunto de entrada, sempre comparando entre os agrupamentos, por meio de gráficos da contagem de observações, para atributos categóricos, e de gráficos de diagramas de caixa da distribuição dos valores, para atributos numéricos. A terceira parte foi feita baseada na sugestão do estudo de (HUANG, 1997), intitulado "Clustering large data sets with mixed numeric and categorical values (1997)", em que se combinam algoritmos de clusterização e árvores de decisão para auxiliar na interpretação dos agrupamentos. Nessa parte também analisamos os resultados da árvore de decisão, que então são comparados às análises dos atributos, feitos na etapa precedente. Por último, fizemos uma comparação das análises dos resultados com alguns agrupamentos encontrados na literatura para tentar validar as interpretações das análises.

[1] Disponível em: https://scikit-learn.org/stable/index.html

4 COLETA, PREPARAÇÃO E TRANSFORMAÇÃO DOS DADOS

O objetivo inicial dessa etapa foi o de obter os dados utilizados durante o desenvolvimento deste trabalho, bem como prepará-los para seu uso nas próximas etapas. Após a coleta e preparação dos dados, foi necessária uma pesquisa para entendermos quais dados estavam disponíveis no conjunto, o que eles representavam e de que maneira o faziam. As informações obtidas nessa pesquisa, disponibilizadas por documentos do DATASUS e da Fiocruz, foram de grande importância para entendermos quais informações estavam disponíveis no conjunto coletado. Com um entendimento melhor do conjunto de dados, foram feitas transformações em alguns atributos do conjunto. Essas transformações se tratavam de uma tradução da codificação feita nos valores do conjunto original para o texto que descreve o seu significado. Apenas as colunas que foram julgadas interessantes de serem trabalhadas passaram por essas transformações. Um dos objetivos dessa etapa foi precisamente o de gerar um novo conjunto único que contivesse todos os dados, alguns transformados, e salvá-lo em um arquivo separado para facilitar seu uso nas etapas seguintes. Com o arquivo do novo conjunto, fizemos uma exploração inicial para entender o comportamento e distribuição de seus atributos, bem como fazer a mesma exploração nos dados de óbitos por suicídio.

4.1 COLETA DOS DADOS

Acessamos os arquivos de banco de dados do SIM por meio da funcionalidade da biblioteca PySUS chamada de `pysus.online_data.SIM.download()`, que permite facilmente aceder ao banco de dados de um estado e ano, e salvar em uma variável do tipo DataFrame. A partir dela é possível salvar o banco de dados como um arquivo em disco. Para baixar todos os dados desejados foi necessário criar uma lista de todos os estados do Brasil e suas siglas, bem como definir um ano inicial e final. Com essas variáveis definidas já seria possível percorrer cada ano do intervalo e baixar os bancos de dados de todos os estados, que então poderiam ser usados. Porém, fazer o download de cada arquivo toda vez que fôssemos lidar com os dados demoraria, e com isso atrasaria o desenvolvimento desta pesquisa, portanto criamos um script para automatizar o download e salvamento de todos os dados desejados. Nele, criamos uma pasta para armazenar esses arquivos e, dentro dela, os arquivos foram organizados por pastas de cada ano do intervalo. Cada arquivo é salvo com o padrão de nome 'SIM-sigla do estado-ano.csv'. A função de download dispara um erro caso o arquivo solicitado não exista, portanto o script foi adaptado para capturar esses erros, avisar qual estado e ano não foi possível baixar e, então, prosseguir com os outros downloads. O período foi definido baseado em todos os anos disponíveis no momento da coleta, que, neste caso, foram os anos de 1979 até 2019. O resultado da etapa de coleta dos dados do SIM foi de 1097 arquivos baixados, totalizando pouco mais de 6 Gigabytes, armazenados em disco.

Os dados dos municípios foram obtidos utilizando uma funcionalidade da biblioteca PySUS, que retorna um *DataFrame* com os dados solicitados. Já os conjuntos de dados de ocupação e de categorias e subcategorias CID10 foram coletados de forma manual, ou seja, por meio da busca pelos dados e download dos arquivos feitos pelo autor deste trabalho. Os dados das ocupações coletados estavam disponíveis no formato CSV[1] em um repositório público que transformou a tabela de códigos CBO-2002 disponível em PDF para formato de tabela.

Já os dados de categorias e subcategorias CID10 foram coletados do repositório público que o PCDaS usa como base para enriquecer seu conjunto. Ambos encontram-se no formato CSV[2].

4.2 PREPARAÇÃO DOS DADOS

Com todos os dados baixados, realizamos uma verificação simples de quantas colunas cada arquivo tinha para entender as divergências entre anos e estados. A verificação percorria cada arquivo CSV na pasta de downloads, lia a primeira linha para pegar e imprimir no output a quantidade de colunas do arquivo. Notamos uma variação grande no número de variáveis entre anos e estados dos arquivos de antes de 2010. O DATASUS informa que houve uma padronização dos formulários de dados do SIM a partir do ano de 2011. Em Consolidação da base de dados de 2011 do SIM, lê-se: "a maior parte das UF empenhou-se em alcançar a meta de utilização preferencial dos formulários novos a partir de janeiro de 2011, utilizando os formulários enviados na 2ª remessa de formulários, distribuída no início do 2º semestre de 2010, junto de uma orientação trabalhada em reunião nacional dos sistemas, de recolhimento gradual dos formulários antigos". Devido à grande quantidade de arquivos separados, foi necessário criar um script para unir todos eles em um conjunto único de dados já filtrados pelo período de 2010 a 2019.

O resultado dessa etapa foi um banco de dados com 1.926.737 linhas e 97 colunas, em que as linhas representam cada indivíduo que faleceu e as colunas são atributos desses indivíduos, que também foi salvo em um arquivo separado para facilidade de usos futuros. As próximas etapas usaram esse novo banco dos dados de mortalidade como base, ou ponto de partida. Na preparação também foram coletados os bancos de dados dos códigos de ocupações (CBO-2002) e dos dados dos municípios por meio das funções do PySUS. Além dos conjuntos de categorias e subcategorias CID10, que foram baixados manualmente do repositório público usado no PCDaS, que são referenciados na Documentação fornecida pela Fiocruz.

[1] Disponível em: https://github.com/datasets-br/cbo/blob/master/data/lista_canonicos.csv

[2] Disponíveis em: https://github.com/bigdata-icict/ETL-Dataiku-DSS/blob/master/SIM/CID-10-CATEGORIAS.CSV.utf8 e https://github.com/bigdata-icict/ETL-Dataiku-DSS/blob/master/SIM/CID-10-SUBCATEGORIAS.CSV.utf8

4.3 ENTENDIMENTO DOS DADOS

Para entender os atributos dos dados coletados, usamos duas fontes de informação: o Dicionário de Dados do SIM, fornecido pelo DATASUS (CONSOLIDAÇÃO. . . , s.d.), e a Documentação e Dicionário de variáveis do PCDaS, fornecidos pela Fiocruz. Essas fontes usaram os dados do SIM como base. Nesses documentos há, entre outras informações, o nome de cada variável e sua descrição. A partir da leitura desses detalhes, foi possível compreender o significado das variáveis do conjunto. Com isso, foi feita uma seleção de apenas algumas variáveis, as quais foram julgadas pelo autor como relevantes para a análise ou para o treinamento dos modelos de aprendizado de máquina, como: o sexo biológico ("SEXO"), o estado civil ("ESTCIV"), a cor informada pelo responsável pelas informações do falecido ("RACACOR") e a escolaridade em anos ("ESC"). A informação entre parênteses é o nome da variável. Julgamos relevantes por se tratarem de dados que ajudam a traçar um perfil inicial dos indivíduos, o que poderia contribuir para análises sociais futuras. No caso específico de "IDADE", o dado obtido do SIM vem com o valor codificado em um campo contendo dois subcampos. O primeiro é de um único dígito e indica a unidade da idade (se 1, trata-se de minuto; se 2, hora; se 3, mês; se 4, ano; se 5, idade maior que 100 anos). Já o segundo, de dois dígitos, indica a quantidade de unidades. Para idade menor de 1 hora, o subcampo varia entre 01 e 59 (minutos). Para idade entre 1 e 23 horas, o subcampo varia entre 01 e 23 (horas). Para idade entre 24 horas e 29 dias, o subcampo varia entre 01 e 29 (dias). Para idade entre 1 e menos de 12 meses completos, o subcampo varia entre 01 e 11 (meses). Para idades maiores que 12 meses completos, o subcampo varia entre 00 e 99 (anos). Quando não se sabe a idade, trata-se do código 9.

O significado e valores possíveis de cada uma dessas variáveis se encontra na 2.

Com a documentação do PCDaS foi possível reutilizar os dicionários python disponibilizados na próxima etapa, de transformação dos dados.

Variável	Significado	Valores possíveis (codificados/decodififcados)
SEXO	Sexo biológico	1: masculino, 2: feminino, 9: ignorado,
ESTCIV	Estado civil	1: Solteiro, 2: Casado, 3: Viúvo, 4: Separado Judic./Divorciado, 5: União consensual/estável, 9: Ignorado
RACACOR	Cor informada	1: Branca, 2: Preta, 3: Amarela, 4: Parda, 5: Indígena, 9: Ignorado
ESC	Escolaridade em anos	1: Nenhuma, 2: 1 a 3 anos, 3: 4 a 7 anos, 4: 8 a 11 anos, 5: 12 e mais, 9: Ignorado
OCUP	Ocupação	Código no padrão CB0-2002
CAUSABAS	Causa básica da morte	Código CID10
IDADE	Idade do falecido em minutos, horas, dias, meses ou anos	Campo codificado pelo SIM

Tabela 2 – Tabela da explicação de variáveis do SIM

4.4 TRADUÇÃO/TRANSFORMAÇÃO DOS DADOS

Após entender o significado das variáveis na etapa anterior, selecionamos algumas delas para que fossem transformadas de sua codificação numérica para seu significado em texto. Para isso, reutilizamos os dicionários da etapa de decodificação de variáveis do PCDaS. Utilizamos ainda a função translate_variables_SIM da biblioteca PySUS. Ela traduz os valores de algumas colunas, incluindo “SEXO” e “RACACOR”, que já haviam sido traduzidas. Principalmente, ela traduz a idade do indivíduo para anos de idade. O banco de dados com algumas colunas traduzidas foi o produto dessa etapa. Ele foi salvo em outro arquivo para ser usado nas próximas etapas.

5 ENRIQUECIMENTOS E ENGENHARIA DE ATRIBUTOS

Nesta seção vamos ver mais detalhadamente como se dá o enriquecimento de dados e como é feita a engenharia de atributos nesse processo.

5.1 ENRIQUECIMENTO DOS DADOS

Na etapa de enriquecimento de dados foram agregados novos dados baseados em outros já existentes. Esses dados já existentes se referem às colunas 'MUNCOD', que é o código de identificação do município no padrão do IBGE, e 'OCUP', que é o código da ocupação do indivíduo no padrão da Classificação Brasileira de Ocupações (CBO-2002). Fizemos manualmente a coleta dos conjuntos usados na agregação. Para isso, procuramos uma fonte e baixamos os arquivos desejados nos dados de ocupação e nos dados de códigos CID10, inclusas as categorias e subcategorias dessa fonte, exceto pelos dados dos municípios que coletamos por meio da funcinalidade do PySUS, a já citada get_municipios().

5.1.1 Dados dos munícipios

A partir do conjunto de dados gerado nas etapas anteriores e do conjunto de dados dos municípios coletado através da função get_municipios() da classe SIM da biblioteca PySUS, é possível agregá-los. Dos dados coletados por essa função, que somam 28 colunas, apenas o nome do município ("MUNNOME"), a latitude ("LATITUDE"), a longitude("LONGITUDE"), a altitude ("ALTITUDE") da indicação se é uma capital ou não ("CAPITAL") são usadas nesta pesquisa. As informações entre parênteses e aspas representam o nome das colunas. Para poder agregar os dados por meio da junção dos dois conjuntos de dados foi necessário transformar as colunas que continham o código do município. Como o conjunto com os dados do município possuía o código com tamanho de 6 dígitos, foi necessário transformar o código do conjunto de mortalidade e também limitá-lo a 6 dígitos. Após isso, ambas as colunas foram convertidas para o tipo inteiro e a junção foi feita. Escolhemos o tipo junção à esquerda, conhecida como left join. No caso em específico, o conjunto do lado esquerdo é de óbitos por suicídio e, do lado direito, de dados do município. A junção foi feita no atributo "MUNCOD", que representa o código IBGE do município. Com isso, foram agregados aos dados já existente no conjunto de mortalidade que permitem visualizar mais detalhes do município, como as variáveis de latitude e altitude podem ser usadas para gráficos com mapas geográficos, a variável de capital pode ser usada no treinamento ampliar as dados de entrada disponíveis, a variável de altitude pode ser usada para possíveis visualizações de dados e análises e a variável de nome deixa explícito de qual município está sendo referido.

5.1.2 Dados das ocupações

Para agregar as descrições das ocupações, usamos o conjunto de dados de ocupações fornecido pela biblioteca PySUS em formato `pandas.DataFrame`. O conjunto possui uma coluna chamada 'codigo' que, justamente, contém o código CBO da ocupação em texto, e outra, 'termo', com a descrição da ocupação, também em texto. Foi necessário fazer uma pequena manipulação na coluna 'codigo' para transformar de texto para número inteiro. Pois seus valores possuem um hífen separando os primeiros 4 dígitos dos outros 2 restantes. Já o código da ocupação do falecido está no formato de número inteiro de 6 dígitos, portanto a manipulação feita foi apenas para remover o hífen de 'codigo' e converter para número inteiro. A coluna 'termo' foi renomeada para 'ocup_name'. Com essas adaptações, foi possível fazer a junção dos dados de óbito com os dados de ocupação. A junção à esquerda apresenta o conjunto de óbitos à esquerda e o conjunto de ocupações à direita, como já citado. As variáveis usadas para fazer a junção foram a 'OCUP', no conjunto da esquerda, e a 'codigo', no da direita. Desse modo, foi agregado o dado com a descrição de cada código de ocupação, se o código estava presente nos dois conjuntos, então podemos usar esse novo conjunto em dados, em visualizações e análises de maneira mais evidente que com o código.

5.1.3 Categoria CID10

Usou-se o arquivo coletado com as categorias como uma variável e fez-se a junção numa nova coluna, derivada da categoria 'CAUSABAS', o valor dessa coluna são códigos CID10 da causa básica da morte da vítima. Isso foi necessário porque, em 'CAUSABAS', o código tem 4 caracteres, visto que ele inclui a subcategoria, e no conjunto de categorias, o código tem apenas 3 caracteres. Então, a nova coluna é basicamente uma cópia dos 3 primeiros caracteres da coluna 'CAUSABAS'. O conjunto de categorias foi filtrado para ter apenas as colunas com o código e a descrição dos itens. A partir disso foi possível fazer a junção à esquerda dos dados dos óbitos com as categorias.

5.1.4 Subcategoria CID10

O arquivo coletado com as subcategorias foi utilizado como uma variável e, assim como o conjunto de categorias, teve suas colunas filtradas para conter apenas o código e a descrição da subcategoria. O produto dessa etapa foi agregar ao conjunto de dados mais detalhes sobre os óbitos, mais informações sobre os municípios, bem como as descrições das ocupações. O novo conjunto contendo esses dados foi salvo em um outro arquivo que é usado nas etapas seguintes.

5.2 ENGENHARIA DE ATRIBUTOS

5.2.1 Novo atributo: turno do dia

A criação desse novo atributo usou como base um já existente, intitulado 'HORA-OBITO', que indica a hora em que o óbito ocorreu. Os turnos do dia foram definidos arbitrariamente seguindo os seguintes intervalos: de 00:00 até 11:59 é a parte da manhã do dia; de 12:00 até 17:59, a tarde; e das 18:00 até 23:59, e da noite. Foi definido então uma função para mapear um valor de hora do dia, que precisa ser um número inteiro, para um texto contendo qual turno do dia aquele horário se refere. Os valores dos turnos são "Manhã", "Tarde"e "Noite"e eles seguem os intervalos de horários já citados, que estão representados pelas variáveis "MIDNIGHT", "MORNING_END"e "AFTERNOON_END". Para extrair o turno do dia da hora do óbito de todos os registros do banco de dados foi usado o seguinte trecho:

Ao final da etapa, o novo atributo “turno_dia” possuía 10.335 registros na parte da manhã, 6461 na parte da tarde e 4.973 na parte da noite, como é mostrado abaixo

5.2.2 Novo atributo: Local da ocorrência

O atributo do local da ocorrência foi extraído de uma coluna que já possuía essa informação, a já citada “CAUSABAS”. De acordo com o CID10, o grupo do código X60 até o X84 são denominados 'Lesões auto provocadas intencionalmente', nele existem categorias que definem qual tipo de lesão foi provocada. Cada categoria tem sua subcategoria que indica em que área aconteceu o óbito. Os valores possíveis dessa subcategoria são: “residência”, “habitação coletiva”, “escolas, outras instituições e áreas de administração pública”, “área para a prática de esportes e atletismo”, “rua e estrada”, “áreas de comércio e de serviços”, “áreas industriais e em construção”, “fazenda”, “outros locais especificados” e “local não especificado”. Para extrair essa informação e criar o novo atributo, usamos como base a variável da descrição da subcategoria da causa básica, que foi criada na etapa de enriquecimento de dados. Os valores da variável base têm o seguinte formato: “categoria da lesão-subcategoria”, em que os valores entre chaves dão o significado dos textos dessa posição. Então, para extrair a informação desejada, fizemos uma função para pegar o texto da descrição, separá-lo em duas partes, baseado no divisor “ - ”, e devolver a última parte, que contém a subcategoria. Caso não houvesse esse divisor, o texto assumiria o valor “Indefinido”. Essa função foi executada com os valores da descrição da subcategoria de cada registro e seu resultado foi armazenado no novo atributo chamado “local_ocorr”, que representa o local da ocorrência do óbito. Com esse novo atributo disponível é possível agregar mais informações para o treinamento do modelo e testar hipóteses, como o local da ocorrência ser diferente entre os agrupamentos que serão encontrados. A contagem dos valores do novo atributo podem ser visualizadas na tabela 3.

Contagem dos valores de local_ocorr	
Valor	Contagem
residência	9926
local não especificado	1769
outros locais não especificados	738
rua e estrada	651
áreas de comércio e de serviço	233
Indefinido	193
habitação coletiva	192
escolas, outras instituições e áreas de administração pública	154
fazenda	147
áreas industriais e em construção	34
área para a prática de esportes e atletismo	21

Tabela 3 – Contagem dos valores da novo atributo 'local_ocorr'

5.2.3 Novo atributo: Ocorrência em feriado ou fim de semana

Esse novo atributo foi criado para identificar se o óbito ocorreu em feriado ou final de semana. A motivação para a criação desse novo atributo foi a hipótese de que poderia haver alguma relação entre um grupo de indivíduos que comete suicídio e o dia escolhido não ser um dia útil. Para criar esse novo atributo, usamos como valor base o atributo "DTOBITO". Para extrair a informação desejada, utilizamos as bibliotecas holidays e datetime da linguagem Python. Antes de aplicar as funções a seguir, foi necessário fazer uma transformação da coluna "DTOBITO" do formato texto para o formato de data. Para descobrir se a data era um feriado, criamos uma função que lê as datas dos óbitos e verifica se elas estão presentes na lista de feriados do Brasil, inclusos os feriados estaduais, fornecida pela biblioteca holidays. Criamos ainda outra função para descobrir se o dia do óbito era em fim de semana. Nela, criamos uma nova coluna temporária, contendo o dia da semana, em números inteiros, sendo 0 a representação de segunda-feira e 6, a de domingo.

A partir disso, foi possível criar uma função que, com base no dia da semana, retornava se era sábado ou domingo. Usamos essas duas funções na elaboração de uma terceira, que retorna se as datas dos óbitos foram em feriados ou final de semana. A terceira função foi aplicada ao conjunto de dados e os resultados foram salvos no novo atributo "fim_semana_ou_feriado".

6 APLICAÇÃO DOS ALGORITMOS DE APRENDIZADO DE MÁQUINA

6.1 ALGORITMO DE CLUSTERING

O algoritmo de clustering foi usado para tentar identificar perfis nos grupos dentro do conjunto de dados de mortes por suicídio. Neste conjunto, há dezenas de variáveis a respeito do óbito, porém apenas algumas delas foram selecionadas para serem processadas pelo algoritmo de clustering. A escolha dessas colunas foi baseada em atributos que o autor julga interessantes de se avaliar e explicar sua distribuição entre diferentes clusters. Dentre os atributos selecionados havia um do tipo numérico — a idade do indivíduo em anos ("IDADE_ANOS") — e os outros do tipo categórico — o sexo biológico ("SEXO"), o estado civil ("ESTCIV"), a cor informada pelo responsável pelas informações do falecido ("RACACOR"), a escolaridade em anos ("ESC"), a ocupação ("OCUP"), se houve assistência médica ou não durante a enfermidade que ocasionou o óbito ("ASSISTMED"), se a cidade da ocorrência era uma capital ("CAPITAL"), local da ocorrência ("local_ocorr"), turno do dia da ocorrência ("turno_dia"), se o dia da ocorrência foi em feriado ou final de semana ("fim_semana_ou_feriado"). Devido à natureza mista dos dados, o algoritmo de clustering usado foi o K-prototypes que, como já comentado, é o mais eficiente quando se tratam de dados mistos, isto é, numéricos e categóricos. A proposta desse algoritmo é fazer a clusterização de dados de natureza mista combinando os algoritmos k-means, para dados numéricos, e k-modes, para dados categóricos. As colunas de tipo numérico foram normalizadas usando a técnica de normalização min-max (HAN; KAMBER; PEI, 2012), que trata do processo de redimensionar os dados da coluna para o intervalo fechado de 0 até 1. O parâmetro de iterações máximas foi definido em 40, baseado no artigo de proposta do K-prototypes (HUANG, 1997), que mostra um intervalo ideal de número de iterações. Os dados usados no algoritmo foram os do conjunto de dados de óbitos por suicídio contendo apenas as variáveis escolhidas mencionadas anteriormente. O algoritmo foi executado para encontrar 2, 3 ,4 e 5 clusters. Para cada agrupamento, foram anotadas e suas taxas de custo, que é como a biblioteca chama a taxa de erro, foram salvas num arquivo separado junto de outras informações sobre a execução. A tabela 4 apresenta os valores das taxas de custo do algoritmo K-Prototypes para cada número de agrupamentos.

Número de agrupamentos	Taxa de custo
2	5420,06
3	4968,80
4	4743,96
5	4613,78

Tabela 4 – Taxa de custo do algoritmo k-prototypes para cada execução

Observando a tabela é observado que a menor taxa de custo entre os experimentos foi para encontrar 5 agrupamentos. É possível notar também que a taxa de custo diminui

de acordo com o número de clusters, o que poderia indicar que o aumento do número de clusters fosse interessante. Porém, como já mencionado, explicar tantos clusters se tornaria tarefa demasiado complexa. Em razão disso, optou-se por manter o número máximo de clusters em 5. Com o modelo treinado para detectar um número de clusters, foi possível predizer a qual cluster cada registro foi classificado. Esses resultados foram salvos em uma nova coluna, "CLUSTER_K{num. de clusters}". O arquivo final com o banco de dados usado agregado às colunas dos resultados foi salvo em um arquivo separado, para evitar a sobrescrita dos dados de input.

6.2 ALGORITMO DE ÁRVORE DE DECISÃO

O conjunto de dados produto da etapa anterior foi usado nessa etapa, pois o desfecho da classificação dos dados é o número do cluster a que ele pertence. A implementação do algoritmo de árvore de decisão que foi usada no desenvolvimento foi a `tree.DecisionTreeClassifier` da biblioteca Scikitlearn[1]. Essa implementação não suporta variáveis categóricas, portanto é necessário fazer a transformação dos dados. Nesse caso, trata-se da transformação chamada *one hot encoding* (YU *et al.*, 2020). Ao aplicá-la em uma coluna categórica, uma nova coluna é criada para cada valor único na coluna original. Essas novas colunas só possuem dois valores possíveis, 0 e 1, em que 0 indica que não pertence a categoria e 1, que pertence. Fizemos uma lista das colunas usadas na clusterização. A partir dessa lista, criamos um dicionário com um valor booleano para cada coluna, indicando se se trata ou não de uma coluna categórica. Assim, foi possível iterar sobre a lista de colunas categóricas e aplicar a cada uma a transformação para *dummy variables*. Com as transformações aplicadas, o resultado é um DataFrame com colunas numéricas e colunas categóricas no formato one-hot. Antes de rodar esse algoritmo, fizemos uma separação dos dados nos diferentes conjuntos de treino e de teste. A separação dos dados foi feita na proporção de 80% para o conjunto de treino e 20% para o conjunto de testes. A execução do algoritmo foi feita para 2, 3, 4 e 5 clusters, e cada visualização de árvore foi salva em uma imagem separada. Para 2 clusters, as taxas de acurácia no conjunto de treino e também no conjunto de teste foram de 84%. Para 3 clusters, as taxas de acurácia no conjunto de treino e no conjunto de teste foram de 78%. Para 4 clusters, a taxa de acurácia foi de 67% tanto no conjunto de treino como no conjunto de teste. Para 5 clusters, a acurácia no conjunto de treino foi de 62% enquanto no conjunto de teste foi de 61%.

[1] Disponível em: https://scikit-learn.org/stable/modules/tree.html

7 ANÁLISE DOS RESULTADOS

Esse capítulo visa compreender de diversas formas os produtos dos experimentos anteriores e essa compreensão é feita através de análises visuais dos gráficos dos atributos de cada agrupamento e de gráficos da distribuição dos agrupamentos pelo mapa da região sul do Brasil. As análises dos atributos visam compreender melhor e tentar interpretar o perfil dos indivíduos de cada agrupamento. Já a análise que utiliza o mapa da região sul procura visualizar se há uma relação espacial entre os agrupamentos. Ao final, para validar os resultados dos experimentos, as análises feitas serão comparadas entre si, com o objetivo de descobrir mais alguma informação dos perfis dos agrupamentos, e também serão comparadas com grupos de risco/atenção encontrados na literatura.

7.1 VISUALIZAÇÃO USANDO MAPA

Para essa visualização foi necessário coletar um arquivo do tipo GEOJSON que contém coordenadas e pontos dos distritos brasileiros da região desejada. Os dados das vítimas foram cruzados com os dados geográficos e então foi possível gerar o mapa da região com as informações de cada *cluster*. As figuras nessa seção representam a estimativa da densidade da distribuição dos dados pelo mapa do sul do Brasil e para gerar esses gráficos foi usada a função *kdeplot*[1] da biblioteca *geoplot*. Nessas figuras as cores representam a quantidade de observações na região do mapa, em que quanto mais escura a cor mais perto de 0 observações e quanto mais clara/amarelo mais observações. Essa parte visa analisar visualmente se há ou não relações espaciais entre os agrupamentos.

Mapa para 2 agrupamentos

Ao visualizar a Figura 7 é possível observar que no cluster 1 as áreas do interior dos estados do Paraná e de Santa Catarina e também a área ao redor da cidade de Itajaí, em Santa Catarina, estão indicando uma maior concentração de indivíduos quando comparadas a essas mesmas regiões no cluster 0. No cluster 0 a região em torno de Porto Alegre, no Rio Grande do Sul, tem uma área de concentração maior de indivíduos do que a mesma região no cluster 0. Apesar dessas diferenças ambos os clusters possuem concentrações de indivíduos na região ao redor de Porto Alegre e Curitiba, que são as capitais do Rio Grande do Sul e Paraná respectivamente, porém não há uma concentração tão densa em Florianópolis, cidade de Santa Catarina, mesmo ela sendo uma cidade capital assim como as citadas anteriormente. Isso pode indicar que não há uma relação direta com o indivíduo morar em uma capital.

[1] Disponível em: https://residentmario.github.io/geoplot/plot_references/plot_reference.html#kdeplot

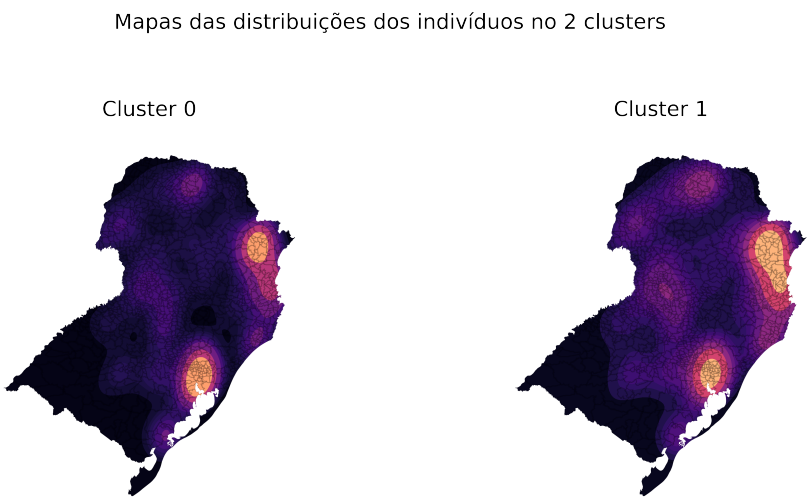

Figura 7 – Mapas da distribuição dos indivíduos entre os 2 cluster (Fonte: o autor)

Mapa para 3 agrupamentos

Ao visualizar a figura 8 é possível observar uma concentração maior de indivíduos no interior da região sul nos cluster 1 e 2 enquanto no cluster 0 a concentração é maior ao redor de Porto Alegre, Curitiba e um pouco ao norte do Paraná. Embora os clusters 1 e 2 possuam semelhança na concentração na região do interior eles diferem nos pontos de maior densidade, já que o cluster 1 tem uma concentração maior na região de Curitiba até Balneário Camboriú enquanto o cluster 2 possui uma concentração maior de indivíduos na região de Porto Alegre. O cluster 1 apresenta uma coloração mais clara distribuída de maneira mais uniforme que os outros mapas e isso pode indicar que o perfil dos indivíduos desse cluster está um pouco mais espalhado que os outros ou seja, não depende tanto da localização e pode ser encontrado em mais lugares do ponto de foco. No mapa do cluster 0 é possível ver mais regiões escuras do que nos outros e isso pode indicar que os indivíduos desse agrupamento tem um perfil mais compatível com cidadãos dessas localizações nas áreas mais densas, que nesse caso são a região de Porto Alegre, de Curitiba e um pouco no norte do Paraná.

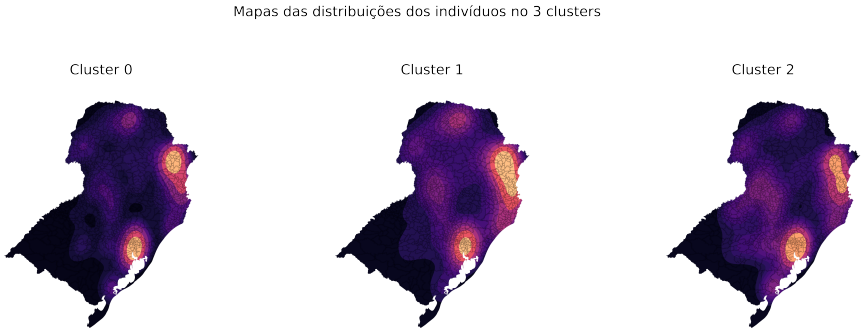

Figura 8 – Mapas da distribuição dos indivíduos entre os 3 cluster (Fonte: o autor)

Mapa para 4 agrupamentos

Ao visualizar a figura 9 é possível notar um destaque da distribuição do cluster 2, pois enquanto os outros clusters têm uma baixa concentração de indivíduos no interior dos estados o cluster 2 possuí uma concentração significativamente maior nessa mesma região e isso poderia indicar que o perfil dos indivíduos desse cluster é compatível não só com cidadães dos pontos de maior densidade, de Porto Alegre e Curitiba, mas também são compatíveis com uma boa parcela de cidadães de interior. Os outros clusters possuem uma distribuição semelhante com os pontos de maior concentração de indivíduos na região de Porto Alegre e Curitiba, embora há algumas diferenças como: no cluster 0 a terceira maior concentração se encontra no interior de Santa Catarina, mais precisamente perto da fronteira oeste do estado, nos clusters 1 e 3 a terceira região com mais concetração seria no norte do Paraná, porém o cluster 3 possui muito mais regiões escuras que o 1.

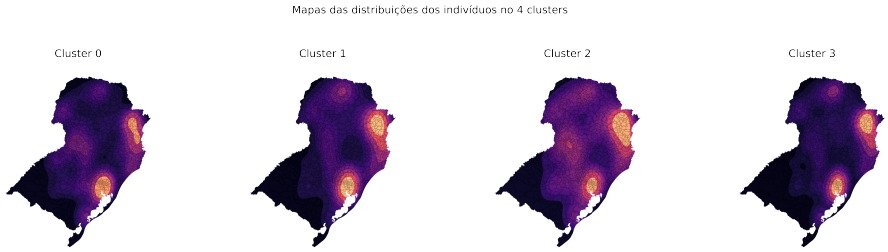

Figura 9 – Mapas da distribuição dos indivíduos entre os 4 cluster (Fonte: o autor)

Mapa para 5 agrupamentos

Ao visualizar a figura 10 é possível notar que as distribuições dos clusters 0, 1, 2 e 3 são muito semelhantes aos cluster 0, 1, 2 e 3 da figura 9, respectivamente. Isso pode indicar que esses clusters semelhantes possuem seus perfis de indivíduos também muito

próximos. As poucos diferenças que pode-se observar são: no cluster 0 do experimento da figura 10 há menos concentração de indivíduos na região ao redor de Balneário Camboriú e também nas regiões do interior dos estados do que no cluster 0 da figura 9; No cluster 1 desse experimento teve um pequeno aumento da concentração de indivíduos nas regiões dos arredores de Balneário Camboriú e do interior de Santa Catarina e Paraná quando comparado ao cluster 1 do experimento anterior. Além disso nesse agrupamento também teve uma aumento pequeno nas áreas que ficam entre as maiores concentrações e isso pode indicar que o perfil desse agrupamento nesse experimento está menos relacionado com a proximidade das regiões de grande concentração. No cluster 2 a diferença foi o aumento da concentração de indivíduos nas regiões do interior de cada estado. A concentração nas regiões do interior é significativamente maior que no cluster 1 então isso pode indicar que nesse cluster há uma relação do perfil de seus indivíduos com o perfil dos cidadães dessas regiões. No cluster 3 houve um pequeno aumento na concentração da região ao redor de Curitiba e uma pequena diminuição da concentração da região de Porto Alegre. E por fim no cluster 4, que não existia no experimento anterior, é possível observar que sua distribuição é semelhante ao cluster 0, porém possui uma concentração um pouco maior na região oeste de Santa Catarina e uma concentração significativamente maior na região de Balneário Camboriú.

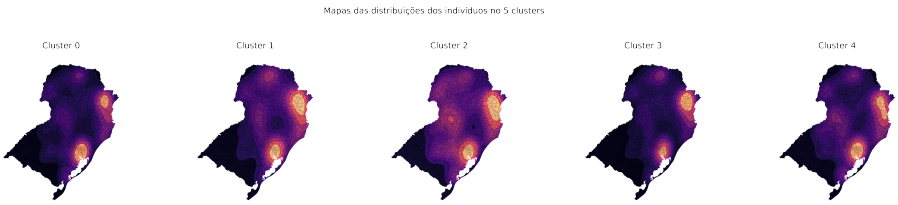

Figura 10 – Mapas da distribuição dos indivíduos entre os 5 cluster (Fonte: o autor)

7.2 COMPARAÇÃO DOS ATRIBUTOS DE CADA CLUSTER

Para poder entender as diferenças dos *clusters*, as distribuições e variações de seus atributos foram analisadas visualmente através do uso de técnicas de visualização de dados. Para número de agrupamentos que foi executado o algoritmo de clusterização foi feita uma análise dos gráficos gerados, em que apenas os mais relevantes são apresentados nessa seção. Para cada agrupamento há uma explicação mais detalhada das demais análises na seção A.1 do apêndice A .

Atributos para 2 clusters

A distribuição da população dos clusters teve uma diferença relevante, como pode ser visto na figura 11.

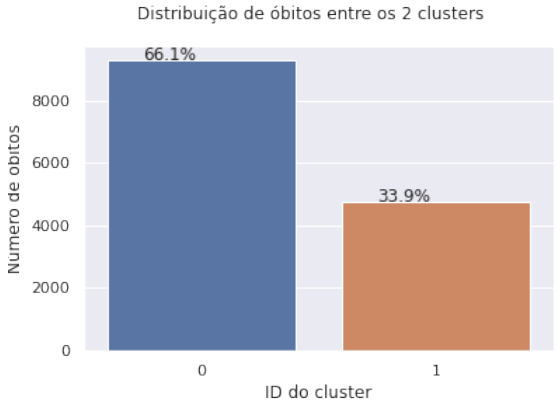

Figura 11 – Distribuição de população nos 2 agrupamentos

- Turno do dia
  Na figura 12, podemos ver que o cluster 0 tem uma predominância de 60% de ocorrências pela manhã enquanto o cluster 1 tem sua maioria na parte da tarde.

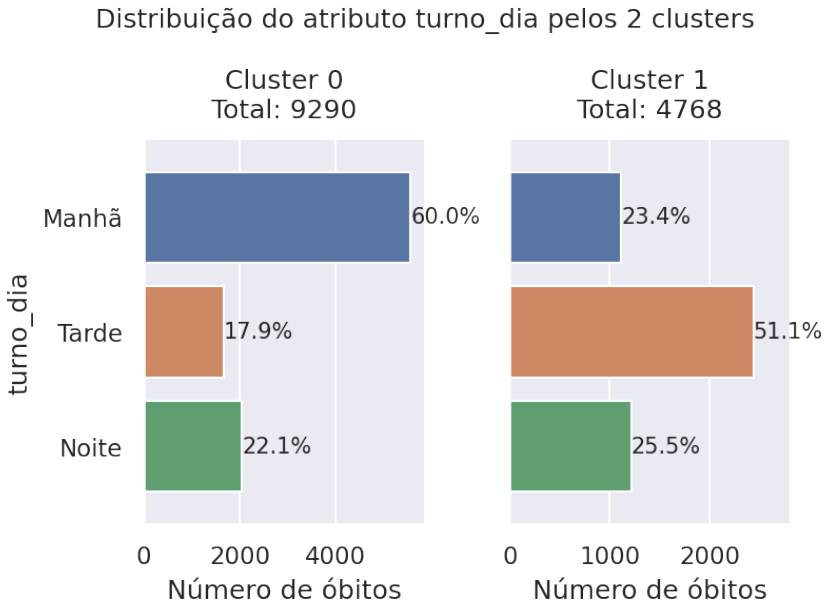

Figura 12 – Distribuição de turno do dia da ocorrência para 2 agrupamentos

- Estado civil
  A ordem decrescente dos estados civis com mais ocorrências dos clusters difere apenas nas duas primeiras posições. Porém, seu valor mais frequente indica uma forte relação com o cluster pertencente. Enquanto o cluster 0 possui cerca de 58% de sua população solteira o cluster 1 possui cerca de 59% de sua população casada. Os dois valores mais frequentes dos clusters apenas trocam de posição e não variam tanto em proporção.

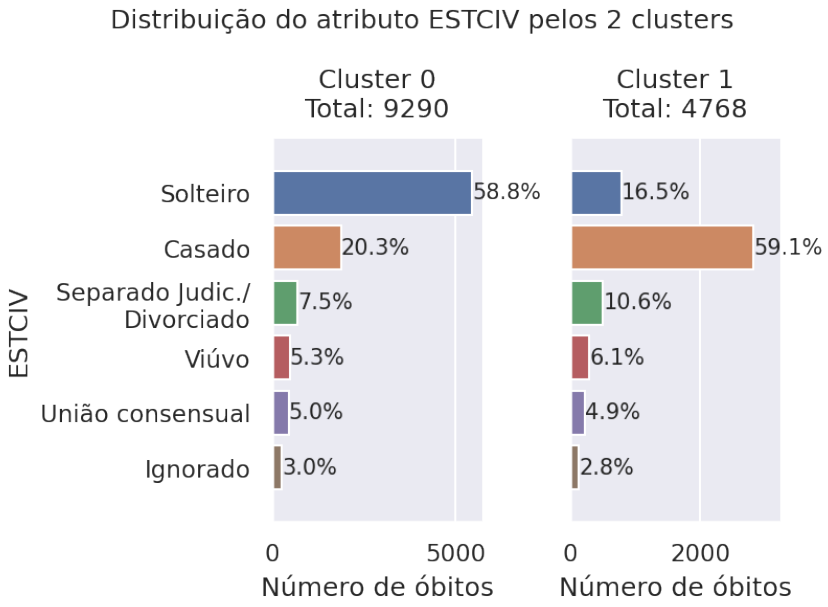

Figura 13 – Distribuição de anos de escolaridade para 2 agrupamentos

- Ocupação

  Na tabela 7.2 pode se ver uma diferença entre as três ocupações com mais ocorrências em cada cluster. No cluster 1 existem duas ocupações relacionadas a agricultura enquanto no cluster 0 há apenas uma. Isso indica que a população do cluster 1 tem uma proporção de trabalhadores da agricultura maior que o cluster 0.

Proporção das ocorrências das 3 ocupações mais frequentes em cada cluster			
Cluster 0		Cluster 1	
Ocupação	Proporção	Ocupação	Proporção
Pedreiro	6,09%	Trabalhador volante da agricultura	6,12%
Trabalhador volante da agricultura	5,77%	Pedreiro	4,76%
Empregado doméstico nos serviços gerais	4,41%	Produtor agrícola polivalente	4,11%

Tabela 5 – Distribuição do turno do dia das ocorrências entre os 2 agrupamentos

- Idade em anos

  O atributo idade indica ter sido relevante para determinar a que cluster o indivíduo pertence. O cluster 0 possui uma maior concentração de pessoas mais jovens em relação ao cluster 1, como é possível observar na figura 14. Isso adiciona uma nova informação sobre a concentração da população do cluster 0, que é possível interpretar como um grupo com a maiorira de pessoas mais jovens, sendo a maioria solteiras,

com 4 a 7 anos de escolaridade e que cometeram suicídio na parte da manhã de dias úteis.

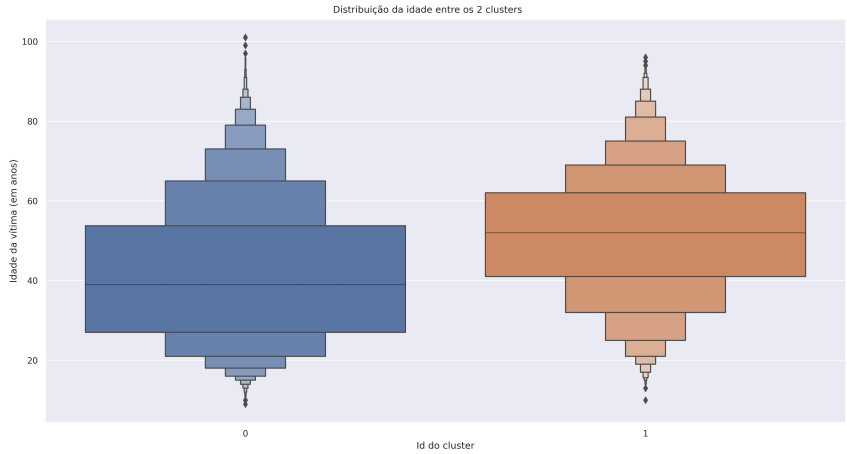

Figura 14 – Distribuição da idade em anos entre 2 agrupamentos

Tabela 6 – Tabela de diferenças mais relevantes entre 2 agrupamentos

Diferenças entre valores de maioria dos clusters 0 e 1			
Atributo	Significado	Cluster 0	Cluster 1
IDADE_ANOS	Idade em anos	25 a 55 anos	40 a 60 anos
ESTCIV	Estado civil	Solteiro	Casado
OCUP	3 ocupações com mais ocorrências	1º Pedreiro, 2º Trabalhador volante da agricultura, 3º Empregado doméstico	1º Trabalhador volante da agricultura, 2º Pedreiro, 3º Produtor agrícola polivante
fim_semana_ ou_feriado	Ocorreu em dia útil ou não?	80% em dia útil	Quase 50/50
turno_dia	Turno do dia	60% pela manhã	60% pela tarde
ESTADO	Estado de residência	1º RS, 2º PR, 3º SC	1º PR, 2º RS, 3º SC

A tabela 6 resume as características dos atributos em suas maiorias em cada cluster. Os atributos que não tiveram uma diferença considerável na distribuição de valores entre os clusters possuem o texto 'Proporção semelhante' indicando essa característica. Portanto, desses clusters os atributos que divergiram mais entre os clusters foram a escolaridade e estado civil do indivíduo e turno do dia da ocorrência. Pode-se dizer que o perfil dos indivíduos do cluster 0 é de pessoas solteiras com 4 a 7 anos de escolaridade que cometeram suicídio pela manhã. Reiterando que o cluster 0 é 11% maior que o cluster 1.

Atributos para 3 clusters

A distribuição da população dos clusters teve uma diferença relevante, como pode ser visto na figura 15. A distribuição dos clusters é diferente entre os 3 agrupamentos, sendo o cluster 0 o maior cluster com 47.7% da população do conjunto inteiro, seguido do cluster 2 com 35.6% e por último o cluster 1 com 16.7%.

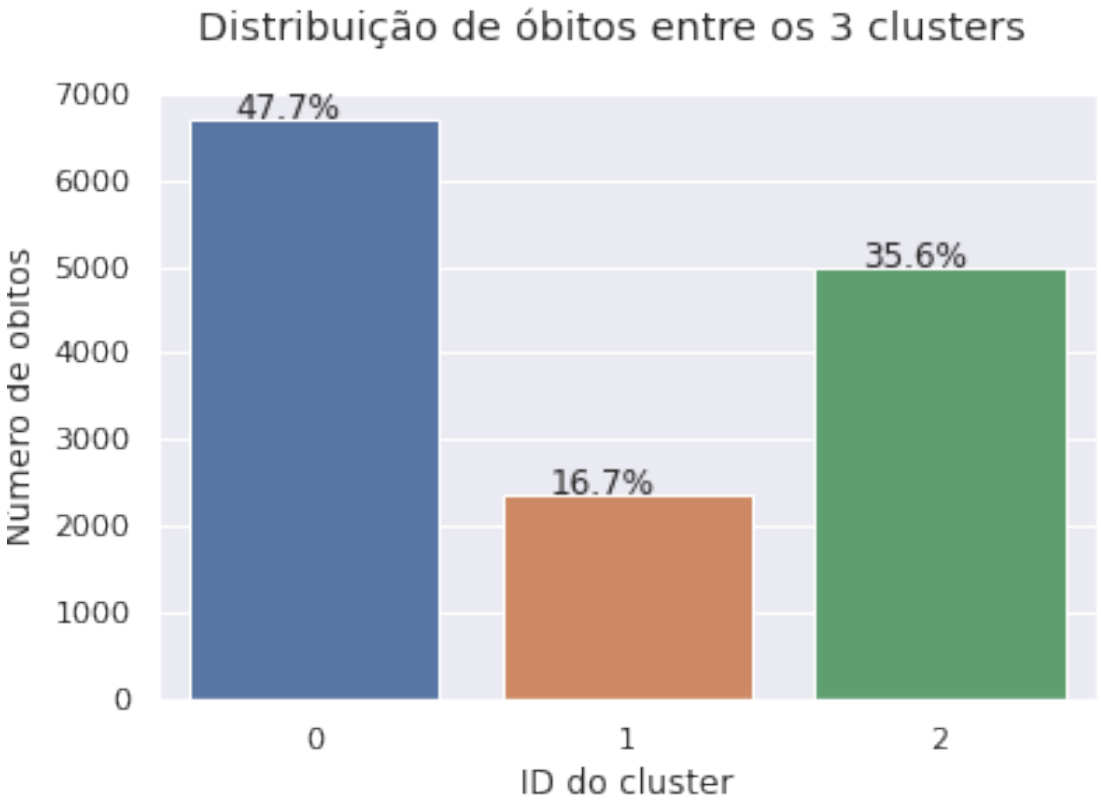

Figura 15 – Distribuição de população nos 3 agrupamentos

- Idade em anos
  Na figura 16 é possível ver uma diferença na faixa etária das concentrações dos indivíduos e isso indica que a idade foi um atributo muito relevante para determinar qual cluster cada dado pertencia. Seguindo na ordem de clusters, que inclusive também é a ordem crescente das faixas etárias mais concentrada, temos o cluster 0

com os indivíduos mais jovens do conjunto inteiro, seguido do cluster 1 em que a concentração começa um pouco antes do fim da concentração do cluster 0 e termina na idade de cerca de 55 anos de idade. Isso nos mostra que cada cluster capturou uma faixa etária bem definida com pouca intersecção entre suas maiores concentrações.

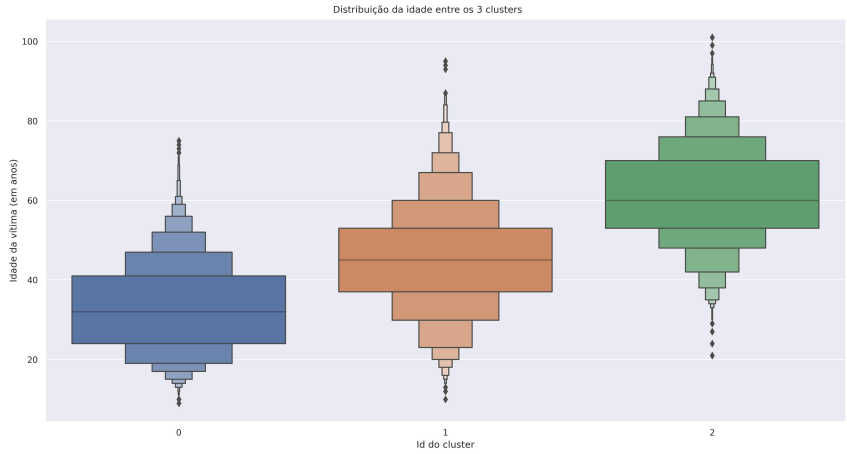

Figura 16 – Distribuição Idade em anos de entre os 3 agrupamentos

- Turno do dia
  Na figura 17 é possível ver que os cluster 0 e 2 são bem semelhantes entre si considerando que ambos tem a predominância de cerca de 50% de ocorrências na parte da manhã enquanto os outros turnos tem cerca de 20% cada. Uma pequena diferença no segundo e terceiro turno mais frequente entre os clusters é que no cluster 0 a parte da noite é a segunda mais frequente, com uma diferença de 1.2% a mais que a parte da tarde. Já no cluster 2, o segundo turno mais frequente é a parte da tarde com uma diferença de 4.8% a mais que a parte da noite. Uma informação bem relevante é que, embora os cluster 0 e 2 sejam bastante semelhantes, o cluster 1 não segue a mesma ordem dos outros. No cluster 1 temos cerca de 50% de ocorrências na parte da tarde diferentemente dos cluste 0 e 2 onde a parte da manhã era mais frequente.

- Assistência médica
  Na figura 18 é possível notar novamente uma diferença grande entre o cluster 1 e os demais. Reforçando que nos resultados do agrupamento para 2 clusters essa coluna de assistência médica não teve diferença relevante entre os clusters. Os clusters 0 e 2 possuem um formato muito semelhante com a distribuição na análise para 2 clusters e esse formato seria cerca de 80% dos indivíduos não receberam atendimento médico durante a ocorrência, cerca de 15% receberam e cerca de 6% foram valores ignorados. Já no cluster 1 a diferença é muito relevante considerando que o valor de mais frequência são indivíduos que receberam assistência com cerca de 53%,o segundo mais relevante são indivíduos que não receberam assistência com cerca de

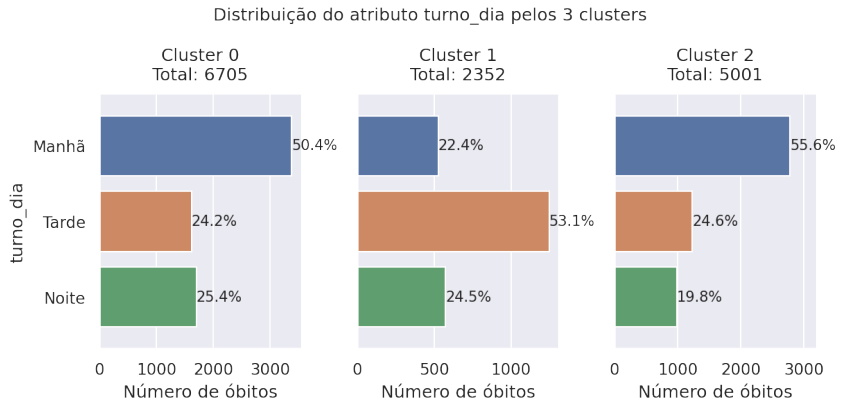

Figura 17 – Distribuição do turno do dia da ocorrência entre os 3 agrupamentos

38% da população e por fim cerca de 8% de valores ignorados. É necessário exaltar a diferença do primeiro valor mais frequente para o segundo entre os clusters já que no cluster 0 temos uma diferença de 63.5% que é muito próxima da diferença de 68.6% do cluster 2 e muito relevante perceber que no cluster 1 temos uma diferença de 15%, que é bem distante da diferença dos outros clusters.

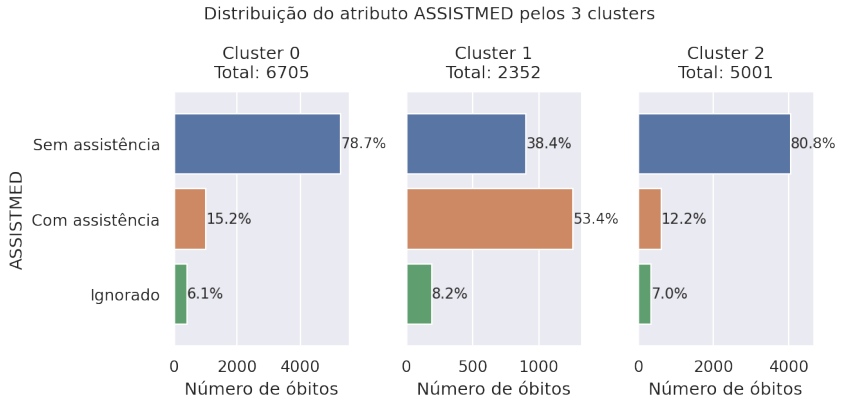

Figura 18 – Distribuição Assistência médica de entre os 3 agrupamentos

- Estado civil
  Podemos perceber que o cluster 0 é formado por cerca de 80% de solteiros enquanto os clusters 1 e 2 tem a predominância de pessoas casadas. Mas há uma diferença entre as duas maiorias dos cluster 1 e 2, que é o tamanho da população do segundo valor frequente. Em que no cluster 1 a segunda população mais frequente é de solteiros que representam cerca de 23% da amostra do cluster no cluster 2 temos a segunda maior população de indivíduos viúvos com 11.6% da população do cluster.

- Sexo biológico
  Assim como nas colunas de assistência médica e de turno da ocorrência os clusters 0 e 2 são muito semelhante e são compostos por mais de 85% de indivíduos do sexo

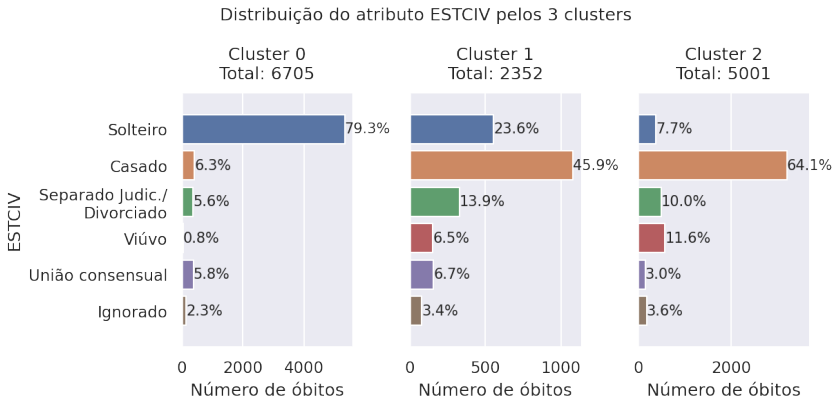

Figura 19 – Distribuição do estado civil entre os 3 agrupamentos

masculino e menos de 15% de feminino. Enquanto o cluster 1 possui cerca de 70% de sua população feminina e cerca de 30% masculina. Isso é uma distribuição bem diferente em relação aos agrupamentos vistos nesse experimento, com 3 grupos, e no anterior, com 2 grupos.

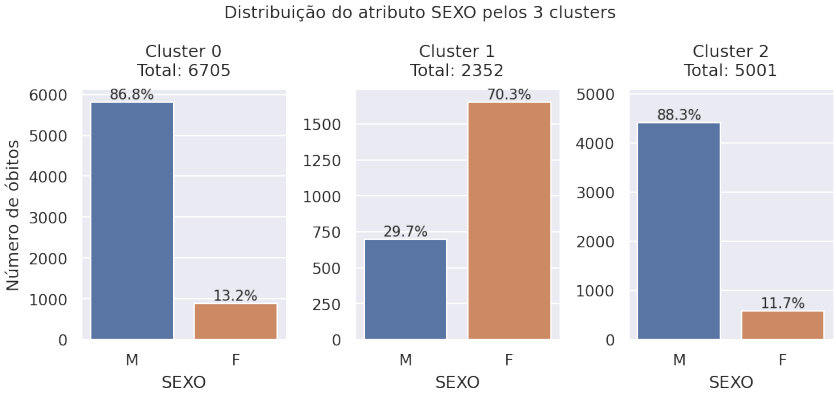

Figura 20 – Distribuição Sexo biológico de entre os 3 agrupamentos

- Causa básica da morte (ou método usado)
  É possível ver na tabela 7 um destaque na diferença dos métodos mais usados no cluster 1 para os demais. Apesar do método mais usado ser o mesmo em todos os agrupamentos ele é cerca de 22% menos utilizado no cluster 1. Enquanto os cluster 0 e 2 possuem o segundo e o terceiro método relacionados a disparo de arma de fogo o cluster 1 tem apenas o quinto método relacionado a uso de arma de fogo. Os dados indicam que o cluster 1 possui mais ocorrências de auto-intoxicação que os demais agrupamentos desse experimento.

Proporção dos métodos mais usados em cada cluster					
Cluster 0		Cluster 1		Cluster 2	
Método	Proporção	Método	Proporção	Método	Proporção
enforcamento, estrangulamento e sufocação	71,11%	enforcamento, estrangulamento e sufocação	48,77%	enforcamento, estrangulamento e sufocação	71,41%
disparo de outra arma de fogo e de arma de fogo não especificada	6,82%	Auto-intoxicação por psicotrópicos não classificados em outra parte	7,87%	disparo de outra arma de fogo e de arma de fogo não especificada	9,28%
precipitação de um lugar elevado	3,21%	Auto-intoxicação por pesticidas	7,19%	disparo de arma de fogo de mão	4,42%
disparo de arma de fogo de mão	3,12%	Auto-intoxicação por substâncias biológicas não especificadas	6,21%	Auto-intoxicação por pesticidas	2,38%
Auto-intoxicação por substâncias biológicas não especificadas	2,27%	disparo de outra arma de fogo e de arma de fogo não especificada	6,21%	precipitação de um lugar elevado	2,12%

Tabela 7 – Distribuição dos 5 métodos mais usados entre os 3 agrupamentos

O cluster 0 tem a maioria de indivíduos do sexo masculino, solteiros, na faixa dos 20 a 40 anos de idade, com 4 a 7 anos de escolaridade, que não receberam assistência médica e cometeram suicídio na parte da manhã. Enquanto no cluster 1 a maioria é de indivíduos do sexo feminino, casados, com escolaridade de 8 a 11 anos, que cometeu suicídio na parte da tarde usando técnicas menos violentas que os outros clusters e recebeu assistência médica. Diferente dos clusters 0 e 2, o cluster 1 não possui a ocupação de pedreiro em suas três ocupações com mais ocorrências. O cluster 2 é composto majoritariamente de indivíduos do sexo masculino, casados, com escolaridade de 1 a 3 anos, em que duas das três ocupações com mais ocorrências estão ligadas a trabalhos agrícolas e a outra é a ocupação de pedreiro, e que cometeram suicídio pela manhã.

Atributos para 4 clusters

A distribuição da população dos clusters teve uma diferença relevante, como pode ser visto na figura 21. Sendo o cluster 0 e cluster 3 os maiores clusters, com cerca de 37% e 34% respectivamente, do conjunto. Os outros dois clusters são cerca de 20% menores do que os dois maiores, porém tem uma proporção semelhante entre si.

Já os valores que as distribuições apresentam diferenças revelantes entre os agrupamentos são:

- Idade em anos

  A concentração da população dos clusters é relativamente bem diferente já que não há nenhum cluster que a concentração da população seja quase igual a de outro. O cluster 0 possui uma população mais velha na faixa dos 50 a 70 anos de idade.

Tabela 8 – Sumário das diferenças entre os 3 agrupamentos

Diferenças entre valores de maioria entre os clusters 0, 1 e 2				
Atributo	Significado	Cluster 0	Cluster 1	Cluster 2
IDADE_ ANOS	Idade em anos	20 e 40 anos	35 e 55 anos	55 e 70 anos
SEXO	Sexo biológico	Masculino	Feminino	Masculino
ESTCIV	Estado civil	Solteiro	Casado	Casado
ESC	Anos de escolaridade da maioria	4 a 7 anos	8 a 11 anos	1 a 3 anos
ocup_ name	3 ocupações com mais ocorrências	1º Pedreiro, 2º Empregado doméstico nos serviços gerais, 3º Trabalhador volante da agricultura	1º Trabalhador volante da agricultura, 2º Comerciante varejista, 3º Empregado doméstico nos serviços gerais,	1º Trabalhador volante da agricultura, 2º Produtor agrícola polivalente, 3º Pedreiro
ASSIST MED	Recebeu assistência médica?	Não	Sim	Não
CAUSA BAS	Categoria CID10 da morte (método usado)	1º sufocação, 2º arma de fogo, 3º precipitação de lugar alto, 4º arma de fogo, 5º intoxicação por substâncias biológicas não especificadas	1º sufocação, 2º intoxicação por remédios pesados, 3º intoxicação por pesticidas, 4º intoxicação por substâncias biológicas não especificadas 5º arma de fogo	1º sufocação, 2º arma de fogo, 3º arma de fogo, 4º intoxicação por pesticidas, 5º precipitação de lugar alto
turno_dia	Turno do dia da ocorrência	Manhã	Tarde	Manhã
ESTADO	Estado de residência	1º Paraná, 2º Rio Gr. do Sul, 3º Santa Cat.	1º Paraná, 2º Rio Gr. do Sul, 3º Santa Cat.	1º Rio Gr. do Sul, 2º Paraná, 3º Santa Cat.

O cluster 1 possui uma população mais concentrada na faixa dos 30 a 50 anos. O cluster 2 possui mais indivíduos na faixa dos 40 a 55 anos. Já o cluster 3 possui a maior concentração de indivíduos mais jovens, considerando a faixa dos 20 a 40 anos de idade.

- Sexo biológico
  Os clusters 0, 2 e 3 tem uma distribuição muito semelhante sendo cerca de 83% masculino e o resto feminino, distribuição também vista em clusters de experimentos anteriores. Enquanto o cluster 1 se destaca devido sua composição de 71,1% de indivíduos do sexo feminino. Reforçando que no experimento de 3 clusters também havia apenas um cluster com a maioria feminina enquanto os outros seguiam o

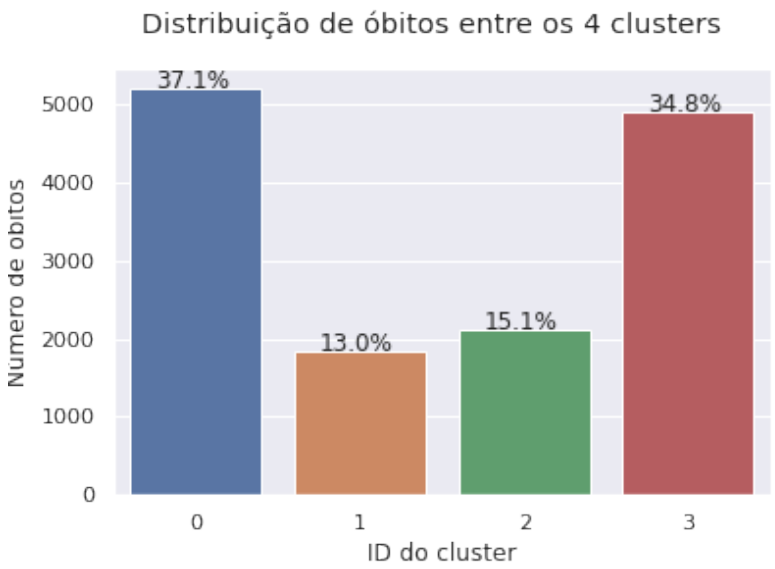

Figura 21 – Distribuição de população nos 4 agrupamentos

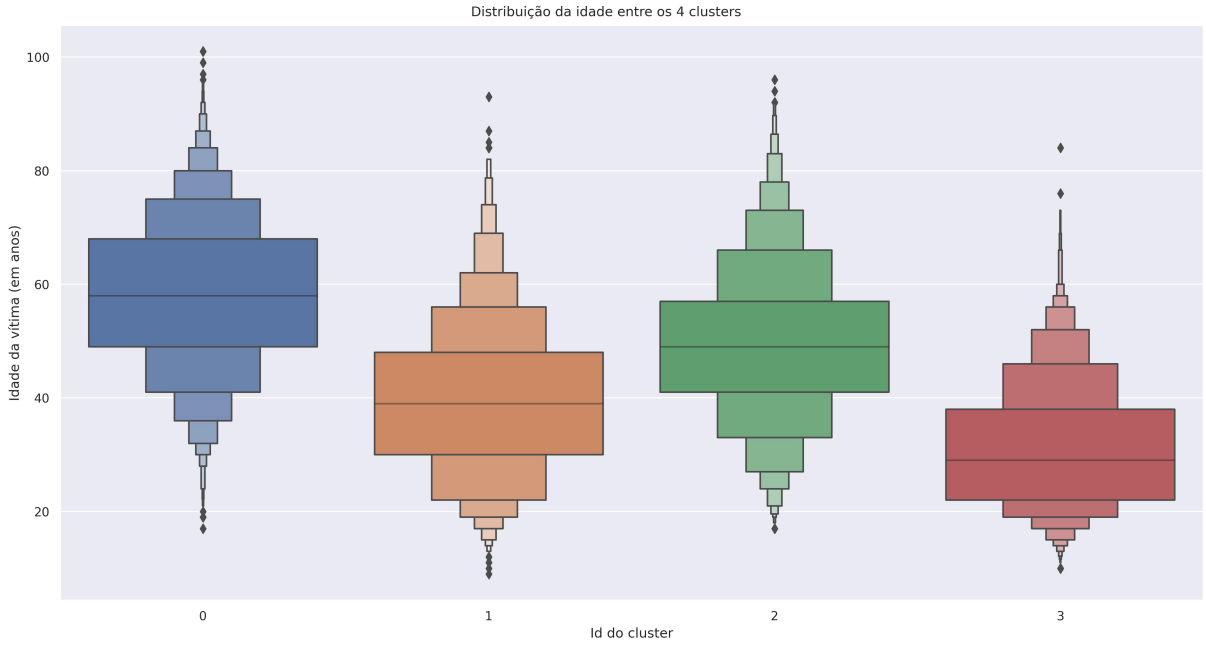

Figura 22 – Distribuição da idade em anos entre os 4 agrupamentos

mesmo padrão de cerca de 80% masculino.

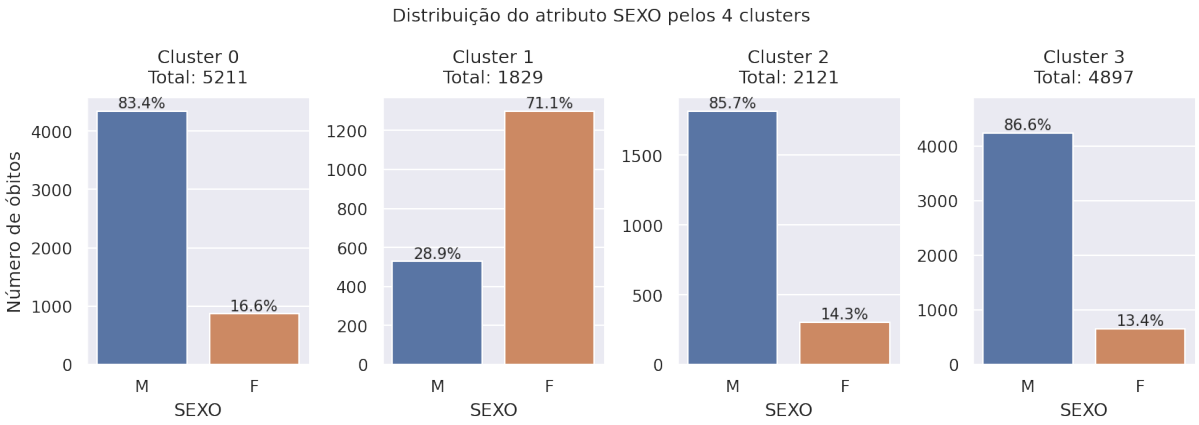

Figura 23 – Distribuição do sexo biológico entre os 4 agrupamentos

- Estado civil
  As maiorias de cada cluster parecem estar bem definidas já que o tamanho da população com o estado civil mais frequente para o segundo mais frequente tem uma diferença relativamente grande. Os clusters 0 e 2 possuem cerca de 57% e 52%, respectivamente, casadas e possuem o mesmo estado civil na posição de segundo mais frequente, porém o cluster 2 possui cerca de 8% mais solteiros que o cluster 0. E já os clusters 1 e 3 possuem uma população majoritariamente solteira, sendo uma parcela de cerca de 60% no cluster 1 e 80% no cluster 3.

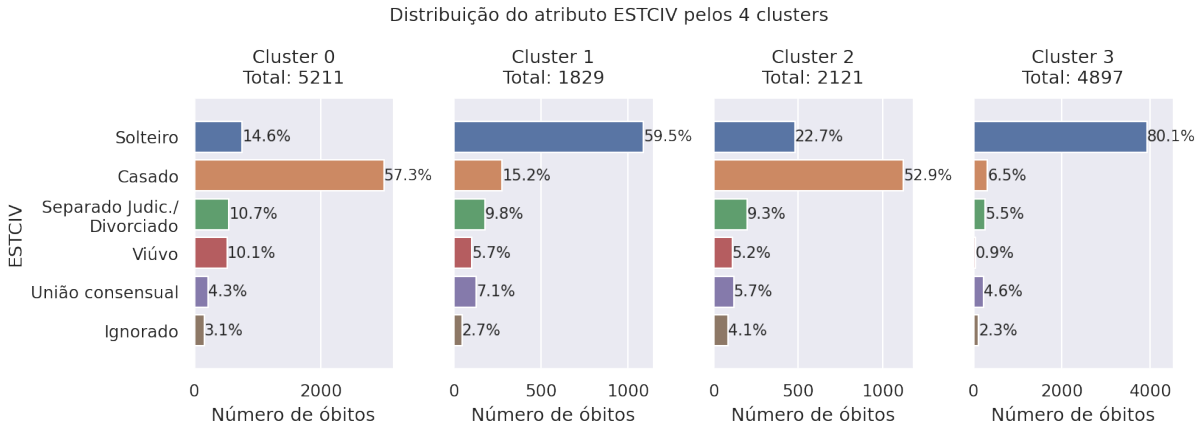

Figura 24 – Distribuição do estado civil entre os 4 agrupamentos

- Recebeu assistência médica?
  Semelhante a distribuição entre os clusters do atributo de sexo, por exemplo, os clusters 0, 2 e 3 possuem o mesmo padrão enquanto o cluster 1 se destaca entre eles. Para esse atributo os cluster 0, 2 e 3 possuem sua grande maioria, de cerca de 77 a

80%, de indivíduos que não receberam assistência médica antes de falecer enquanto no cluster 1 a maioria de 61,4% recebeu assistência médica antes de falecer.

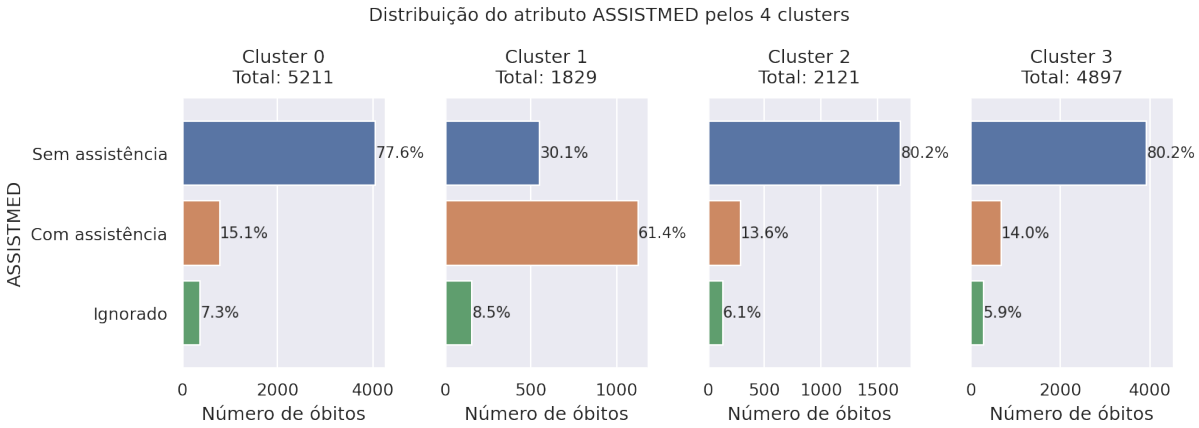

Figura 25 – Distribuição da assistência médica entre os 4 agrupamentos

- Causa básica da morte (ou método usado)
  Apesar de a morte por enforcamento, estrangulamento e sufocação ser a maioria e todos os clusters, o cluster 1 tem cerca de 30% a menos em comparação os outros clusteres. A distribuição dos métodos usados para cometer é muito semelhante nos clusters 0, 2 e 3, sendo a segunda causa o disparo de arma de fogo não especificada. Entre a terceira e quinta causa mais frequentes desses clusters estão presentes também outra causa envolvendo armas de fogo de mão, preciptação de lugares elevados e intoxicação por outras substâncias biológicas não especificadas. Sendo que nesses clusters a segunda causa não chega a 9% e cada uma das outras não chega a 5%. Enquanto no cluster 1, as 5 causas mais frequentes não envolvem armas de fogo, mas as causas entre a segunda e a quarta são todas por auto-intoxicação, seja por medicamentos, substâncias biológicas não especificadas ou pesticidas e a última é a precipitação de lugar elevado. Uma outra diferença do cluster 1 e a proporção das causas estar um pouco mais distribuída em relação aos outros clusters, considerando que as outras causas, exceto a primeira, possuem uma porcentagem de cerca de mais de 5% em contraste com os outros clusters em que essas posições chegavam a muito menos de 5%. Reforçando que a maioria dos indivíduos no cluster 1 recebu assistência médica antes de falecer.

- Escolaridade em anos
  Todos os clusters tem uma maioria bem destacada do restante das categorias. No cluster 1 temos a maioria de cerca de 45% de indivíduos com 4 a 7 anos de escolaridade, no cluster 1 e 3 temos a maioria de 8 a 11 anos de escolaridade com cerca de mais de 41% e no cluster 2 temos a maioria de 39.7% de 1 a 3 anos de escolaridade.

Proporção dos métodos mais usados em cada cluster			
Cluster 0		Cluster 1	
Método	Proporção	Método	Proporção
enforcamento, estrangulamento e sufocação	71,58%	enforcamento, estrangulamento e sufocação	42,43%
disparo de outra arma de fogo e de arma de fogo não especificada	8,23%	Auto-intoxicação por psicotrópicos não classificados em outra parte	10,55%
disparo de arma de fogo de mão	3,80%	Auto-intoxicação por substâncias biológicas não especificadas	8,75%
Auto-intoxicação por pesticidas	2,61%	Auto-intoxicação por pesticidas	6,18%
precipitação de um lugar elevado	2,19%	precipitação de um lugar elevado	4,81%
Cluster 3		Cluster 4	
Método	Proporção	Método	Proporção
enforcamento, estrangulamento e sufocação	70,49%	enforcamento, estrangulamento e sufocação	71,17%
disparo de outra arma de fogo e de arma de fogo não especificada	8,72%	disparo de outra arma de fogo e de arma de fogo não especificada	7,58%
disparo de arma de fogo de mão	4,29%	precipitação de um lugar elevado	3,53%
Auto-intoxicação por pesticidas	3,82%	disparo de arma de fogo de mão	3,29%
precipitação de um lugar elevado	1,74%	Auto-intoxicação por substâncias biológicas não especificadas	1,94%

Tabela 9 – Distribuição dos 5 métodos mais usados entre os 4 agrupamentos (Fonte: o autor)

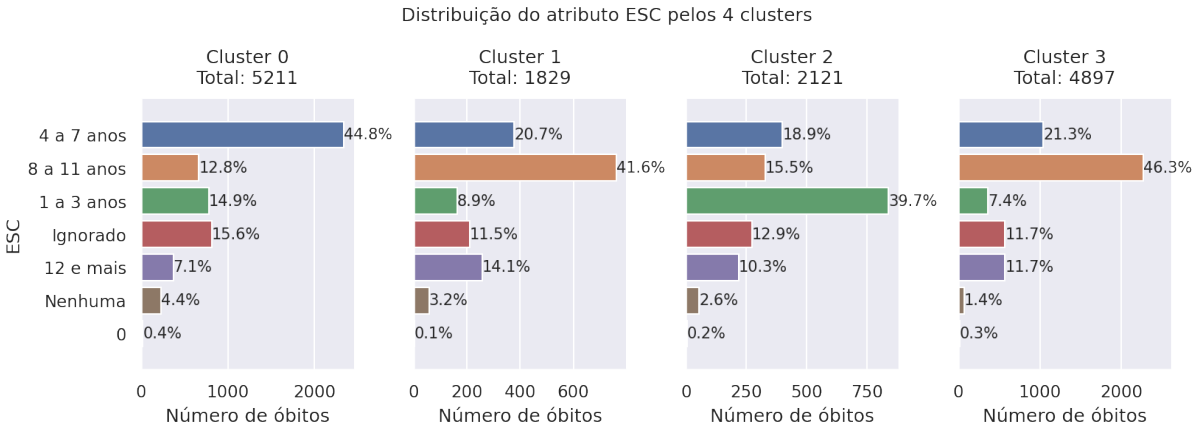

Figura 26 – Distribuição da escolaridade entre os 4 agrupamentos

- Ocupações mais frequentes

  Na tabela 7.2 é possível observar que os cluster 1 e 2 se destacam dos demais por motivos diferentes. O cluster 1 se destaca, pois em suas três ocupações com mais ocorrências não são encontradas as ocupações de pedreiro, trabalhador volante da agricultura ou produtor agrícola polivalente, que são encontradas nos outros agrupamentos. Isso indica um perfil da ocupação da maioria dos indivíduos que não está tão relacionado a ocupações da área de agricultura ou ocupações de pedreiro quanto os demais agrupamentos desse experimento.

Proporção das ocorrências das 3 ocupações mais frequentes em cada cluster			
Cluster 0		Cluster 1	
Ocupação	Proporção	Ocupação	Proporção
Trabalhador volante da agricultura	6,47%	Empregado doméstico nos serviços gerais	4,32%
Pedreiro	5,83%	Comerciante Varejista	2,46%
Produtor agrícola polivalente	4,72%	Representante comercial autônomo	2,02%
Cluster 2		Cluster 3	
Ocupação	Proporção	Ocupação	Proporção
Trabalhador volante da agricultura	15,56%	Trabalhador volante da agricultura	6,21%
Pedreiro	7,12%	Pedreiro	4,68%
Produtor agrícola polivalente	4,38%	Produtor agrícola polivalente	3,68%

Tabela 10 – Distribuição do turno do dia das ocorrências entre os 4 agrupamentos (Fonte: o autor)

Diferenças entre os valores de maioria dos atributos dos 4 clusters					
Atributo	Significado	Cluster 0	Cluster 1	Cluster 2	Cluster 3
IDADE_ANOS	Idade em anos	50 a 70 anos	30 a 50 anos	40 a 58 anos	20 a 40 anos
fim_semana_ ou_feriado	Ocorrreu em fim de semana ou feriado?	Não	Sim	Sim	Não
turno_dia	Turno do dia	Manhã	Manhã	Noite	Tarde, mas manhã fica próximo
SEXO	Sexo biológico	Masculino	Feminino	Masculino	Masculino
ESTCIV	Estado civil	Casado	Solteiro	Casado	Solteiro
OCUP	3 ocupações com mais ocorrências	1º Trabalhador volante da agricultura, 2º Pedreiro, 3º Produtor agrícola polivalente	1º Empregado doméstico nos serviços gerais, 2º Comerciante varejista, 3º Representante comercial autônomo	1º Trabalhador volante da agricultura, 2º Pedreiro, 3º Produtor agrícola polivalente	1º Pedreiro, 2º Empregado doméstico nos serviços gerais, 3º Representante comercial autônomo
ASSISTMED	Teve assistência médica?	Não recebeu	Recebeu	Não recebeu	Não recebeu
CAUSABAS	Categoria CID10 da causa da morte (método usado)	1º sufocação, 2º arma de fogo, 3º arma de fogo, 4º intoxic. pesticida, 5º preciptação de lugar elevado	1º sufocação, 2º intoxic. remédios pesados, 3º intoxic. drogas não especificadas, 4º intoxic. pesticidas, 5º preciptação de lugar elevado	1º sufocação, 2º arma de fogo, 3º arma de fogo, 4º intoxic. pesticida, 5º preciptação de lugar elevado	1º sufocação, 2º arma de fogo, 3º preciptação de lugar elevado, 4º arma de fogo, 5º intoxic. drogas não especificadas
ESC	Anos de escolaridade	4 a 7 anos	8 a 11 anos	1 a 3 anos	8 a 11 anos
ESTADO	Estado de residência	1º Rio Gr. do Sul, 2º Paraná, 3º Santa Catarina	1º Paraná, 2º Rio Gr. do Sul, 3º Santa Catarina	1º Paraná, 2º Rio Gr. do Sul, 3º Santa Catarina	1º Paraná, 2º Rio Gr. do Sul, 3º Santa Catarina

Tabela 11 – Sumário das diferenças entre os valores das maiorias dos 4 clusters (Fonte: o autor)

Um dos clusters mais interessantes de todos os 4 é o cluster 1 que difere principalmente o sexo biológico de sua população, já que é o único em que o sexo feminino predomina. Os dados indicam que esse cluster é composto majoritariamente por indivíduos do sexo feminino, solteiros, na faixa dos 30 a 50 anos de idade, com escolaridade de 8 a 11 anos (até o ensino médio), que cometeram suicídio nas manhãs de feriados ou finais de semana usando principalmente métodos de sufocação ou outros métodos de auto-intoxicação e a maioria recebeu assistência médica.

Outro cluster interessante é o cluster 2 que possui a maioria de indivíduos do sexo masculino, casados, na faixa dos 40 a 58 anos de idade, com escolaridade de 1 a 3 anos (até a 3ª série) que cometeram suicídio na noite de feriados ou finais de semana.

Atributos para 5 clusters

A distribuição da população dos clusters teve uma diferença relevante, como pode ser visto na figura 27. Sendo o cluster 0 o maior cluster, com 28.5% da população do conjunto, seguido do cluster 3, com 23.5%, seguido do cluster 4 com 20.1%, seguido do cluster 2 com 17.7% e por fim o cluster 1 com 10.2%.

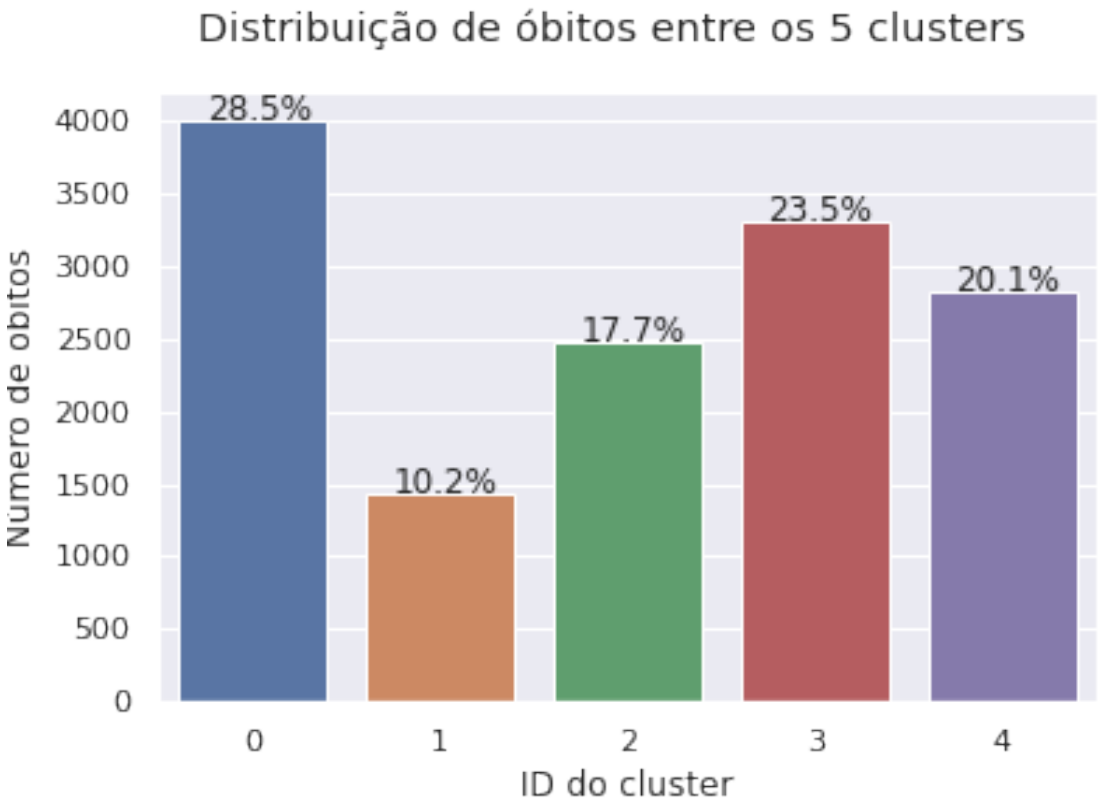

Figura 27 – Distribuição de população nos 5 agrupamentos

Já os valores que as distribuições apresentam diferenças revelantes entre os agrupamentos são:

- Sexo biológico
  Assim como no experimento de 4 clusters, apenas um agrupamento com predominância feminina foi encontrado. O cluster 1 é o cluster com 77.1% de sua população do sexo feminino. Enquanto os outros cluster tem cerca de 85% de suas populações do sexo masculino.

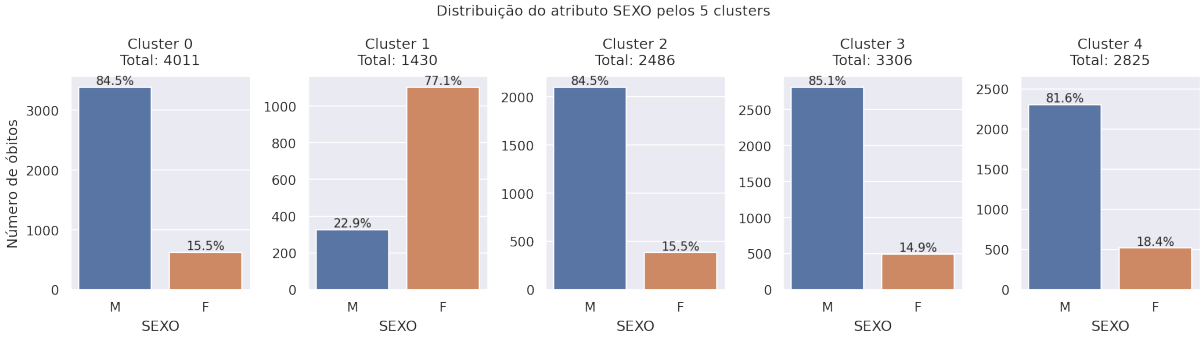

Figura 28 – Distribuição do sexo biológico entre os 5 agrupamentos

- Idade em anos
  Os clusters divergem na distribuição das idades, porém existem pares deles que tem a concentração de sua população semelhante entre si. Como o cluster 0 e 1, que possuem uma população concentrada na faixa dos 30 a 50 anos e 30 a 45 anos de idade, respectivamente. E o cluster 2 e 4 que tem sua concentração de idade entre 50 e 65 anos de idade. Já o cluster 3 não possui nenhuma semelhança com os outros clusters, mas possui a concentração de população mais jovem entre os clusters.

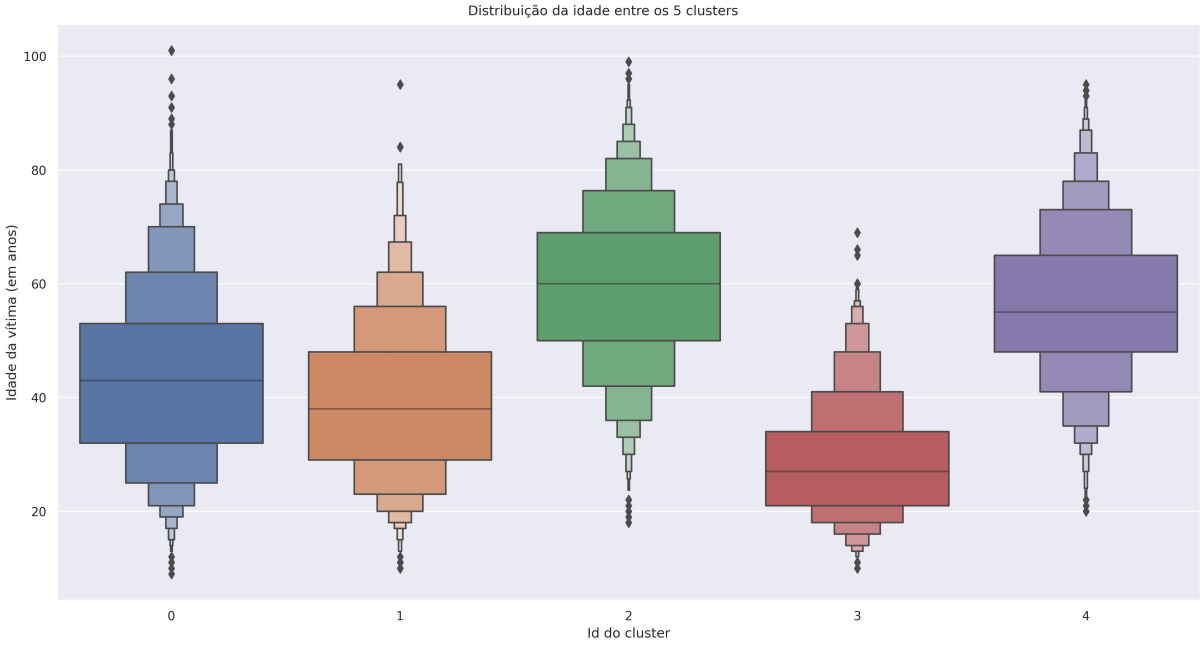

Figura 29 – Distribuição da idade em anos entre os 5 agrupamentos

- Recebeu assistência médica?
  Semelhante aos experimentos para 3 e 4 agrupamentos, o único cluster, nesse caso o cluster 1, com predominância de indivíduos que receberam assistência médica antes de falecer é também o cluster com predominância de indivíduos do sexo feminino. Nos outros clusters, a maioria de suas ocorrências não recebeu assistência médica antes de vir a óbito.

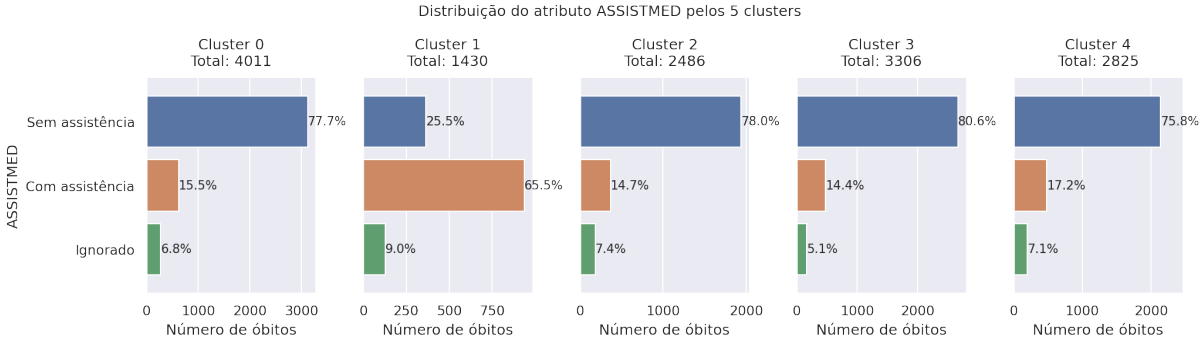

Figura 30 – Distribuição da assistência médica entre os 5 agrupamentos

- Ocupação
  Na tabela 12 são apresentadas as três ocupações com mais ocorrências em cada cluster e sua proporção de ocorrências em relação a todo cluster.

Proporção das ocorrências das 3 ocupações mais frequentes em cada cluster					
Cluster 0		Cluster 1		Cluster 2	
Ocupação	Proporção	Ocupação	Proporção	Ocupação	Proporção
Pedreiro	8.13%	Empregado doméstico nos serviços gerais	4.55%	Trabalhador volante da agricultura	18.1%
Empregado doméstico nos serviços gerais	5.04%	Comerciante varejista	1.96%	Pedreiro	5.59%
Trabalhador volante da agricultura	4.39%	Empregado doméstico diarista	1.89%	Produtor agrícola polivalente	5.59%
Cluster 3		Cluster 4			
Ocupação	Proporção	Ocupação	Proporção		
Pedreiro	5.41%	Pedreiro	4.46%		
Empregado doméstico nos serviços gerais	4.33%	Produtor agrícola polivalente	3.75%		
Representante comercial autônomo	3.3%	Comerciante varejista	3.68%		

Tabela 12 – Distribuição do turno do dia das ocorrências entre os 5 agrupamentos

- Escolaridade
  Nos clusters 1, 3 e 4 a predominância é de indivíduos com escolaridade de 8 a 11 anos, porém nos cluster 1 e 4 suas proporções são um pouco mais distribuídas, já que têm uma parcela de cerca de 20% de sua população com escolaridade de 4 a 7 anos. Já o cluster 0 possui uma maioria de 60.5% de sua população com escolaridade de 4 a 7 anos. E o cluster 2 possui a maioria, cerca de 45%, de seus indivíduos com escolaridade de 1 a 3 anos, que é o menor grau de escolaridade entre as maiorias dos clusters.

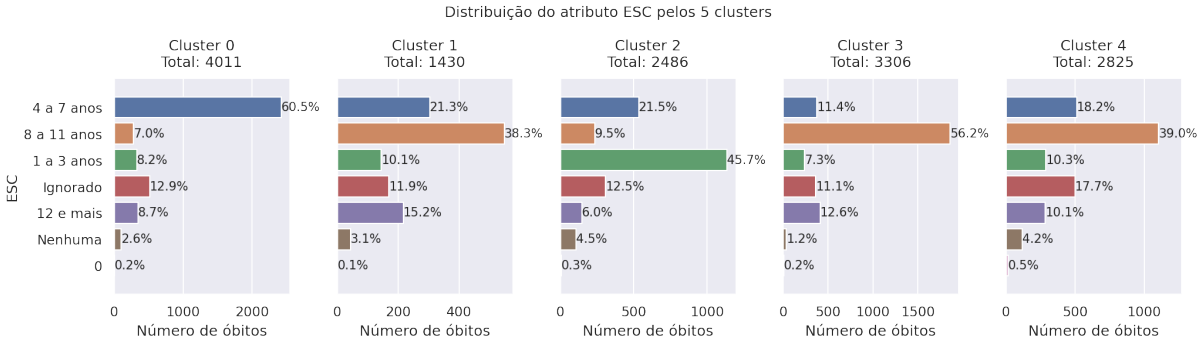

Figura 31 – Distribuição da escolaridade entre os 5 agrupamentos

- Estado civil
  Os clusters 0, 1 e 3 possuem sua maioria de indivíduos solteiros enquanto os clusters 2 e 4 são compostos majoritariamente por casados.

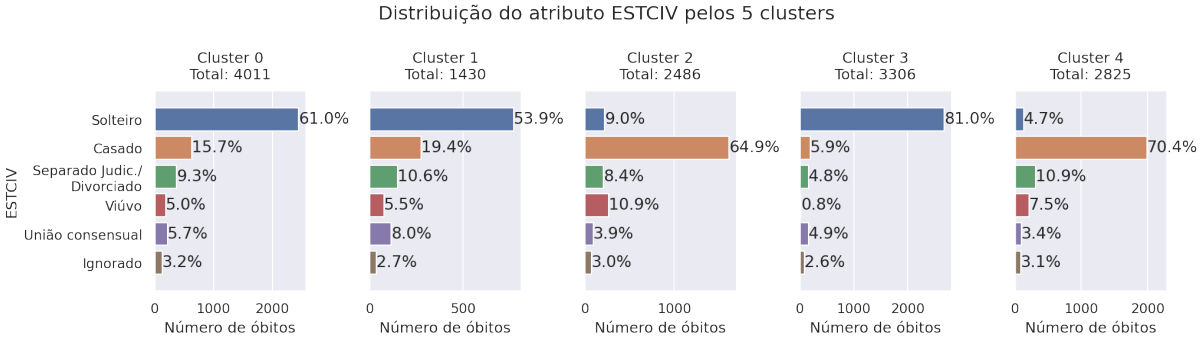

Figura 32 – Distribuição do estado civil entre os 5 agrupamentos

Existem dois clusters semelhantes ao experimento para 4 agrupamentos. Um desse clusteres tem a maioria de indivíduos do sexo feminino, solteiros, na faixa dos 30 a 50 anos com escolaridade de 8 a 11 anos, que cometeram em feriado ou final de semana, que receberam assistência médica e os 5 métodos mais usados não envolveram armas de fogo. Nesse caso esse é o cluster 1 e nele duas das 3 ocupações mais frequentes são de empregado doméstico, sendo um em serviços gerais e outro como diarista. Uma das poucas diferenças,

Diferenças entre os valores de maioria dos atributos dos 5 clusters						
Atributo	Significado	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
SEXO	Sexo biológico	Masculino	Feminino	Masculino	Masculino	Masculino
IDADE_ANOS	Idade em anos	30 a 50 anos	30 a 50 anos	50 e 65 anos	20 a 35 anos	50 a 65 anos
ESTCIV	Estado civil	Solteiro	Solteiro	Casado	Solteiro	Casado
ESC	Anos de escolaridade	4 a 7 anos	8 a 11 anos	1 a 3 anos	8 a 11 anos	8 a 11 anos
OCUP	3 ocupações com mais ocorrências	1º Pedreiro, 2º Empregado doméstico serviços gerais, 3º Trabalhador volante da agricultura	1º Empregado doméstico serviços gerais,, 2º Comerciante varejista, 3º Empregado doméstico diarista	1º Trabalhador volante da agricultura, 2º Pedreiro, 3º Produtor agrícola polivalente	1º Pedreiro, 2º Empregado doméstico de serviços gerais, 3º Representante comercial autônomo	1º Pedreiro, 2º Produtor agrícola polivalente, 3º Comerciante varejista
fim_semana_ ou_feriado	Ocorrreu em fim de semana ou feriado?	não	sim	sim	não	não
turno_dia	Turno do dia	Manhã	Tarde	Manhã	Tarde	Noite
CAUSABAS	Categoria CID10 da causa da morte (método usado)	1º sufocação, 2º arma de fogo, 3º arma de fogo, 4º preciptação de lugar alto, 5º intoxicação com pesticida	1º sufocação, 2º intoxicação remédios pesados, 3º intoxicação drogas não especificadas, 4º intoxicação por pesticidas, 5º precipitação de lugar alto,	1º sufocação, 2º arma de fogo, 3º intoxicação com pesticida, 4º arma de fogo, 5º preciptação de lugar alto,	1º sufocação, 2º arma de fogo, 3º preciptação de lugar alto, 4º arma de fogo, 5º intoxicação drogas não especificadas	1º sufocação, 2º arma de fogo, 3º arma de fogo, 4º intoxicação com pesticida, 5º preciptação de lugar alto
ASSISTMED	Teve assistência médica?	Não	Sim	Não	Não	Não
ESTADO	Estado de residência	1º RS, 2º PR, 3º SC	1º PR, 2º RS, 3º SC	1º RS, 2º PR, 3º SC	1º PR, 2º RS, 3º SC	1º RS, 2º PR, 3º SC

Tabela 13 – Sumário das diferenças entre os valores das maiorias dos 5 agrupamentos

encontradas nas análises, entre esse cluster e o cluster semelhante do experimentod e 4 agrupamentos é o turno do dia que nesse a maioria ocorrou na parte da tarde enquanto no anterior ocorreu na parte da manhã.

O outro cluster semelhante ao experimento anterior é de indivíduos do sexo masculino, mais velhos, casados, com escolaridade de 1 a 3 anos (até a terceira série) e que cometeram suicídio em feriados ou finais de semana e que nesse caso é o cluster 2. Algumas das diferenças desse cluster para o experimento anterior é que nesse o turno mais comum foi a manhã enquanto no anterior foi a noite, a idade nesse cluster é concentrada de 50 a 65 anos enquanto no anterior é de 40 a 58 anos e por final nesse cluster a 2ª e 3ª causa são por arma de fogo e por intoxicação por pesticida, respectivamente, enquanto no cluster anterior a 2ª e 3ª maior causa foi por armas de fogo e a 4ª causa que era por intoxicação por pesticida.

O único dos clusters que predomina ocorrências na parte da noite é o cluster 4 que possui a maioria de indivíduos do sexo masculino, na faixa dos 50 a 65 anos, casados, com escolaridade de 8 a 11 anos (até ensino médio) e que cometeram suicídio no turno da noite de dias úteis.

Já o cluster com maior concentração de pessoas mais jovens é o cluster 3. Esse

cluster possui a maioria da sua população na faixa dos 20 a 35 anos de idade, solteiros, com escolaridade de 8 a 11 anos (até ensino médio) e que cometeram suicídio na tarde de dias úteis.

7.3 RESULTADOS DA ÁRVORE DE DECISÃO

Para os modelos de árvore de decisão gerados para cada número de agrupamentos, foram geradas imagens com suas visualizações, e estas foram salvas em disco. Nessa seção foi feita uma interpretação dos *clusters* através da visualização das decisões da árvore. Cada nó de decisão terá 5 informações dentro de si, que são: o número de identificação do nó, o teste com uma das variáveis, a proporção da amostra do conjunto de dados que chegou nesse nó, uma variável chamada `value`, que representa uma lista com a proporção de amostras por classe/cluster (por exemplo, para 2 clusters, `value = [0.662, 0.338]` representa que naquele nó as amostras estão divididas entre 66.2% do cluster 0 e 33.8% do cluster 1), e por fim o valor classificado, que neste caso é o número do cluster a que os dados desse nó pertencem. As árvores geradas foram reduzidas para no máximo 3 níveis de profundidade devido ao escopo deste trabalho e ao tempo de pesquisa. Com essa quantidade de níveis, a árvore testa menos e produz uma classificação menos detalhada, o que faz com que a interpretabilidade seja menos aprofundada.

Para a divisão dos dados em 2 agrupamentos a árvore obtida é apresentada na figura 33.

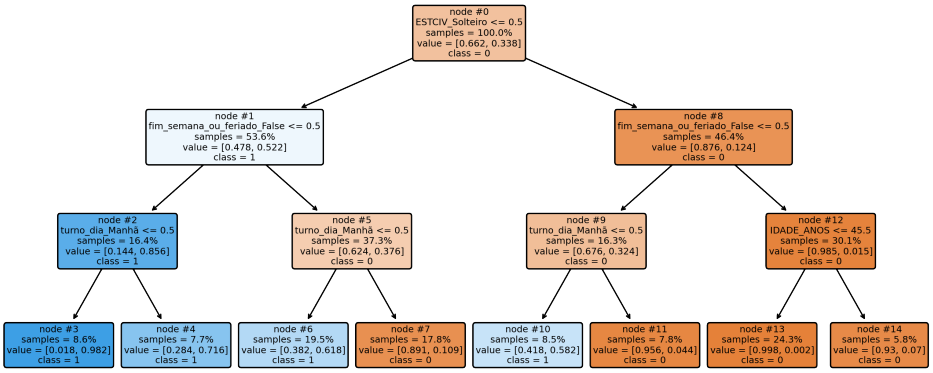

Figura 33 – Visualização da árvore de decisão para 2 *clusters* (Fonte: o autor)

A árvore possui 8 nós folha, sendo 4 deles para cada classificação de cluster, portanto, há 4 caminhos diferentes que os dados podem percorrer pelos testes até chegarem nas classificações. No cluster 0, os caminhos possíveis descrevem os 4 seguintes perfis: indivíduos que não são solteiros (teste do nó #0), cometeram suicídio em dias úteis (teste do nó

#1) e faleceram na parte da manhã (teste do nó #5), o que é equivalente a 17.8% da população do conjunto de dados (nó 7). Outro perfil é o de indivíduos solteiros (teste do nó #0), que cometeram suicídio em feriados ou finais de semana (teste do nó #8) na parte da manhã (teste do nó #9), o que equivale a 7.8% da população do conjunto todo (nó 11). Os outros dois perfis são de indivíduos solteiros (teste do nó #0) que cometeram suicídio em dias úteis (teste do nó #8). Porém, os dois perfis divergem quanto a idade: a decisão com os indivíduos com idade de 45 anos ou menos representa 24.3% (nó 13), enquanto os indivíduos com 46 anos ou mais representam 5.8% (nó 14). Já no cluster 1 os quatro perfis são: indivíduos que não são solteiros, que cometeram suicídio em feriados ou finais de semana na parte da manhã, representando 7.7% (nó 4), ou na parte da tarde, representando 8.6% (nó 3). Outro perfil é de indivíduos que não são solteiros, que cometeram suicídio na parte da tarde ou noite, representando 19.5% (nó 6). Por fim o perfil de indivíduos solteiros, que cometeram suicídio em dias úteis durante a parte da tarde ou noite, que correspondem a 8.5% do conjunto (nó 10).

A árvore obtida pela divisão dos dados em 3 agrupamentos é representada pela Figura 45 do apêndice B.

As decisões da árvore indicam que no cluster 0 há três perfis de indivíduos e todos eles tem 47 anos, ou menos, de idade sendo um dos perfis os indivíduos solteiros do sexo masculino (nó 7), representando 31.3% dos dados, o perfil dos indivíduos solteiros do sexo feminino (nó 6), representando 7.9% dos dados, e por fim o perfil dos indivíduos que não são solteiros e que possuem de 4 a 7 anos de escolaridade (nó 4), representando 6.2% dos dados.

No cluster 1 há apenas 2 perfis identificados pela árvore e eles divergem desde o nó raiz da árvore. Um dos perfis é o de indivíduos com 47 anos, ou menos, de idade, que não são solteiros e não possuem escolaridade de 4 a 7 anos (nó 3), que representa 13.2% dos dados. E o outro perfil é dos indivíduos com mais de 47 anos de idade, do sexo feminino, que receberam assistência médica (nó 14), representando 2.8%.

Já o cluster 2 possui 3 perfis e todos eles são de indivíduos com mais de 47 anos. Um dos perfis é de indivíduos do sexo feminino que não receberam assistência médica (nó 14), representando 6.3% dos dados. E os outros dois são de indivíduos do sexo masculino, porém divergem no estado civil em que um deles é de casados (nó 11), representando 18.3%, e o outro é de indivíduos que não são casados (nó 10), representando 14.2%.

Para a divisão dos dados em 4 agrupamentos a árvore obtida é apresentada na Figura 46 do apêndice B.

A árvore gerada para os 4 agrupamentos encontrou desde clusters com apenas um nó folha para sua classificação até clusters com 3 nós folha de classificação. Os clusters que possuem apenas um nó folha com o perfil dos indivíduos são os clusters 0 e 1.

O cluster 0 é formado por indivíduos com menos de 41 anos de idade, do sexo feminino que cometeram suicídio em feriados ou finais de semana (nó 3), representando 3.4%. E o cluster 1 é composto de indivíduos com 41 anos, ou mais, de idade, que cometaram suicídio em feriados ou finais de semana durante a tarde ou noite, representando 8.5%.

O cluster 2 possui toda sua população de indivíduos com menos de 41 anos. Detro desse grupo de indivíduos há 3 perfis encontrados pela árvore. Um perfil é de indivíduos do sexo feminino que cometeram suicídio em dias úteis (nó 4), representando 6.9%. Enquanto os outros dois outros perfis são de indivíduos do sexo masculino, porém, divididos em quem faleceu em feriados ou finais de semana (nó 6), representando 13.2%, e quem faleceu em dia útil (nó 7), representando 22.9%.

Por fim o cluster 3 também possui 3 perfis e todos seus indivíduos possuem 41 anos, ou mais, de idade. O primeiro perfil é de indivíduos que cometeram suicídio na manhã de feriados ou finais de semana (nó 14), representando 7.5%. Os outros dois perfis são de indivíduos que faleceram em dias úteis, divergindo apenas no turno do dia em que um dos perfis teve as ocorrências na parte da manhã (nó 11), representando 18.3%, e o outro teve durante a tarde ou noite (nó 10), representando 19.2% dos dados.

Para a divisão dos dados em 5 agrupamentos a árvore obtida é apresentada na

Figura 47 do apêndice B.

Devido o número máximo de nós folhas ser 8, há vários clusters com apenas um nó folha para sua classificação. Apenas o cluster 3 possui mais de um nó folha e, portanto, mais de um perfil de indivíduos. Seus indivíduos são divididos entre indivíduos que não são solteiros, que cometeram suicídio em dias úteis durante a parte da manhã (nó 7), representando 17.8%, e indivíduos solteiros, com escolaridade diferente de 8 a 11 anos que faleceram na parte da manhã (nó 11), representando 13.5%, e por fim indivíduos com escolaridade de 8 a 11 anos com 36 anos, ou mais, (nó 14), representando 4.6%.

O cluster 0 tem um perfil de indivíduos do sexo feminino que não são solteiros, que cometeram suicídio em feriados ou finais de semana (nó 3), representando 4.3%.

O cluster 1 possui uma população de indivíduos do sexo masculino que não são solteiros e faleceram em dias úteis (nó 4), representando 12.1%.

O cluster 2 é composto de indivíduos não solteiros que cometeram suicídio em dias úteis durante a tarde ou noite (nó 6), representando 19.5% dos dados.

E por fim o cluster 4 tem sua popução composta por indivíduos solteiros, com escolaridade de 8 a 11 anos e com menos de 36 anos (nó 13), representando 13.5% de todos os dados.

É possível notar que a limitação da profundidade da árvore dificultou a interpretação dos testes dos modelos e essa dificuldade aumenta junto da quantidade de clusters. Entretanto, os perfis interpretados a partir das árvores validaram alguns aspectos observados na análise dos clusters.

7.4 VALIDAÇÃO DOS CLUSTERS

Seguindo a sugestão de Huang (1997), além do algoritmo K-Prototypes, também foi usado o algoritmo de árvores de decisão para auxiliar na interpretação de cada cluster. As análises das distribuições dos indivíduos no mapa do sul do Brasil, dos atributos e das árvores, foram todas comparadas entre si para se tentar compreender o perfil dos indivíduos de cada grupo e, então, comparadas a perfis de atenção/risco presentes na literatura. Com o objetivo de obter mais informações dos agrupamentos encontrados nos resultados, as análises feitas dos mapas, dos atributos e das árvores de decisão, foram todas comparadas e a síntese desses resultados é apresentada nesta seção.

Comparação das análises dos resultados

Os resultados obtidos no primeiro experimento para encontrar 2 agrupamentos usaram as seguintes informações para diferenciar os indivíduos de cada clusters: escolaridade, estado civil, turno da ocorrência e se ocorreu em feriado ou final de semana. Os atributos que mais se destacaram foram o estado civil e o turno da ocorrência. Já no experimento para encontrar 3 agrupamentos, um dos clusters que mais se destacou foi o 1, que possui

uma maioria de indivíduos do sexo feminino. Ele também possui uma população mais distribuída, em relação ao cluster 0 e ao experimento anterior, nas regiões do interior de Santa Catarina e no norte do Paraná.

Além de ser o único cluster, desse experimento ou do anterior, em que a maioria dos indivíduos recebeu assistência médica, e quatro dos cinco métodos mais usados nesse agrupamento não causam uma morte tão violenta quanto outras, por disparo de fogo. O cluster 1 apresenta indivíduos mais velhos, na faixa de 55 a 70, com até a 3ª série do ensino fundamental, ou seja, de 1 a 3 anos de escolaridade, em que as duas ocupações mais numerosas estão relacionadas à agricultura. Em relação aos outros clusters, esse também apresenta uma concentração um pouco maior no interior de Santa Catarina. Vale ressaltar que apenas o cluster 0 não possui uma concentração alta de indivíduos na região de Balneário Camboriú como possui nas regiões de Porto Alegre e Curitiba. O cluster 0 também possui menos concentração nas regiões entre o interior dos estados e as regiões mais concentradas das capitais de SC e do RS.

No experimento para encontrar 4 agrupamentos, há dois clusters que possuem muitas semelhanças com os clusters 1 e 2 do experimento anterior, mas há algumas pequenas e relevantes diferenças. No cluster 1 desse experimento a população é majoritariamente do sexo feminino, com escolaridade de 8 a 11 anos, na faixa etária de 30 a 50 anos, que recebeu assistência médica e cujos métodos mais usados não eram tão violentos, principalmente porque os cinco métodos mais utilizados não envolvem armas de fogo. As diferenças desse cluster para o seu semelhante do experimento anterior, que era o 1, também com predominância de indivíduos do sexo feminino, são o estado civil, que nesse caso é de maioria solteira, e não casada, como no anterior; o turno da maioria das ocorrências é a manhã, enquanto o outro era a parte da tarde; as três ocupações mais frequentes no grupo não são relacionadas à agricultura; e o uso de armas de fogo não está mais listado nos cinco mais frequentes, já no outro, era o quinto mais usado. Isso pode indicar que o agrupamento possui uma especificidade maior, pois manteve várias das características vistas no experimento anterior, mas não agrupou tantos dados relacionados a trabalhadores da agricultura ou de ocorrências que envolveram disparo de arma de fogo. Entretanto, a mudança do turno do dia e do estado civil precisa ser observada no próximo experimento para se compreender melhor o motivo dessa mudança.

Outro cluster semelhante a de experimentos anteriores é o 2, que é o agrupamento da maioria dos indivíduos de sexo masculino, casados, com baixa escolaridade (1 a 3 anos de estudo), que cometeu suicídio pelos seguintes cinco métodos, em ordem de frequência: sufocação, disparo de armas de fogo não especificadas, disparo de armas de fogo de mão, intoxicação por pesticidas e precipitação de lugar elevado. As características se mantêm semelhantes ao que é encontrado no cluster 2 do experimento anterior, porém as diferenças encontradas dizem respeito à faixa etária, visto que o agrupamento desse experimento concentra indivíduos na faixa de 40 a 58 anos, enquanto, no experimento anterior, concentra

na faixa de 55 a 70. O turno da ocorrência nesse experimento é majoritariamente na parte da noite, enquanto no anterior, na parte da manhã. Além do estado com mais ocorrências ser o Paraná nesse experimento, e no anterior, o Rio Grande do Sul. Apesar dessas diferenças, os dados indicam que o perfil dos indivíduos desse agrupamento está cada vez mais específico, apontando para indivíduos com perfil de trabalhadores relacionados à agricultura, somando cerca de 14% das amostras do cluster anterior, enquanto no experimento atual a primeira ocupação, de trabalhador volante da agricultura, representa cerca de 15,5%, e, quando somada à terceira ocupação mais frequente, que é produtor agrícola polivalente, chega a 19,9%. Os mapas apresentam comportamentos da distribuição que parecem estar relacionados ao que foi observado nas outras análises, já que o cluster 1 desse experimento possui menor concentração de indivíduos no interior da região oeste de Santa Catarina, quando comparado ao experimento anterior. Isso pode estar relacionado à diminuição nos trabalhadores da agricultura nesse agrupamento.

O inverso ocorre no cluster 2, em que há um aumento na proporção dos indivíduos com ocupações da agricultura, e o mapa apresenta uma concentração no interior dos estados significativamente maior quando comparado ao cluster 2 do experimento anterior. Por fim, no experimento que encontra 5 agrupamentos de indivíduos, há novamente dois clusters semelhantes aos encontrados no experimento anterior e no experimento de 3 agrupamentos. São eles: o cluster 1, com a maioria dos indivíduos do sexo feminino, e o 2, com a maioria dos indivíduos do sexo masculino e trabalhadores da agricultura. O mapa do cluster 2 desse experimento indica que a concentração de indivíduos no interior dos estados é maior do que é observado no cluster 2 do experimento anterior. Embora a distribuição do cluster 2 desse experimento seja um pouco maior no interior, quando comparada ao anterior, algumas características permanecem semelhantes, como ambos os clusters serem majoritariamente formados por indivíduos do sexo masculino, casados, com escolaridade baixa (de 1 a 3 anos de estudo), com ocupações relacionadas à agricultura, que cometeram suicídio em feriados ou finais de semana, e cujos métodos mais utilizados foram a sufocação, o disparo de armas de fogo não especificadas e armas de fogo de mão, a intoxicação por pesticidas e a precipitação de lugares elevados. Uma das poucas diferenças entre esses clusters é a faixa etária, que no caso desse último experimento tem concentração entre os 50 e os 65 anos, enquanto no anterior era entre os 40 e os 58. Outra diferença é que o estado do Rio Grande do Sul possui mais ocorrências nesse experimento, mas, no anterior, era o Paraná. Por fim, o turno do dia no agrupamento desse experimento é a manhã, enquanto no anterior era a noite. Uma última diferença, que não parece tão relevante, porém, pode estar relacionada com outras diferenças já mencionadas, é a ordem dos métodos mais utilizados, em que, no cluster do experimento anterior, o segundo e o terceiro métodos envolviam disparo de armas de fogo, o quarto, auto-intoxicação por pesticidas. Nesse experimento, por outro lado, o método que usa pesticidas possui mais ocorrências e se torna o terceiro mais usado, seguido do disparo por arma de fogo.

Por meio da análise dos resultados dos experimentos e da comparação entre eles, é possível perceber que, ao aumentar o número de agrupamentos a serem encontrados, há uma especialização de cada agrupamento em um perfil de indivíduos. Isso pode ser percebido no cluster de indivíduos do sexo feminino e no de indivíduos com ocupações da agricultura, que foram encontrados no experimento de 3 agrupamentos. De todo modo, esses grupos seguiram aparecendo nos experimentos seguintes. Mesmo que haja algumas diferenças, as semelhanças podem ajudar a criar um perfil de indivíduos de grupos de risco. Ao observar as árvores de decisão geradas a partir desses dados, é possível notar que a limitação a 3 níveis de profundidade parece ter dificultado a classificação correta dos agrupamentos. Isso fica mais evidente conforme o nível de agrupamentos aumenta. Veja-se que as taxas de erro dos conjuntos de teste diminuíram de 84%, na árvore de 2 agrupamentos, para 61%, na de 5 agrupamentos. Entretanto, mesmo com a profundidade limitada, a árvore permite a visualização de alguns dos atributos mais relevantes para dividir ao máximo os dados do conjunto, o que parece confirmar as análises feitas.

Comparação com a literatura

As análises feitas ao longo desta pesquisa tiveram como objetivo não apenas reunir uma série de dados que ajudasse a criar um perfil de grupos de risco de pessoas que podem vir a cometer suicídio, mas também usar esses dados reunidos para dialogar, buscando validação, na literatura de estudos que se voltam para o fenômeno do suicídio. Nesta seção, busca-se fazer uma comparação entre as análises feitas e seus resultados e a literatura de suicídio.

Como vimos, o método de disparo por armas de fogo e auto asfixiação é recorrente nos casos de suicídio no Brasil. De acordo com (MENEGHEL *et al.*, 2004), no artigo intitulado "Características epidemiológicas do suicídio no Rio Grande do Sul (2004)", no período estudado de 1980 a 1999, o método mais adotado para cometer suicídio é o enforcamento (62,5%), seguido de disparo de armas de fogo (21,5%) e lesões não especificadas (6,9%). No artigo, os pesquisadores salientam o aumento das mortes por enforcamento e armas de fogo, assim como o decréscimo nos meios não especificados. Ou seja, as análises deste trabalho parecem confirmar essa tendência. De fato, segundo o que se pôde reunir aqui, o método de enforcamento, ou sufocação, permanece como o mais utilizado em todos os experimentos, já os métodos não especificados, não foram listados entre as 5 categorias mais comuns de causa da morte de nenhum cluster dos experimentos.

Vimos também como o número de indivíduos do sexo masculino é mais alto que o de indivíduos do sexo feminino. Além disso, a preferência por métodos mais violentos coincide com o primeiro grupo. Em "Prevalência de suicídio no Sul do Brasil, 2001-2005 (2008)", os autores concluem que há uma predominância de indivíduos do sexo masculino e que eles se matam de maneira mais violenta. De fato, nas análises dos clusters encontrados neste trabalho, havia poucos clusters em que o sexo feminino predominasse e, ainda, a

lista dos cinco métodos mais utilizados incluía duas categorias diferentes de disparo de armas de fogo e uma de precipitação de lugar alto, que se configuram enquanto métodos mais violentos. Os autores do artigo supracitado também concluem que, para todos os sexos, o método mais utilizado era o enforcamento, o que também pode ser percebido nas análises aqui feitas.

Observamos ainda que a ocupação mais frequente nos dados de suicídio estava ligada à agricultura. Em (DREBES *et al.*, 2018), os autores observam que as políticas públicas de modernização da agricultura acarretaram novos elementos para o problema social dos suicídios rurais. Destaca-se, nesse artigo, que, enquanto os países afetados por ele não tomam medidas em relação a essas políticas públicas, esse problema estaria estimulando as mortes de agricultores. Nos resultados deste trabalho foram encontrados agrupamentos compostos por uma grande parcela de trabalhadores da agricultura que, em relação aos demais grupos percebidos nos outros agrupamentos, eles estão mais distribuídos nas regiões do interior de seus estados. Esse perfil foi traçado desde o experimento feito para encontrar 3 agrupamentos até o último experimento, para 5 agrupamentos. O que pode indicar que esses agrupamentos refletem o problema do suicídio em zonas rurais e/ou do suicídio de agricultores que foi destacado no estudo "Legislação, Política Pública e Suicídio: A Influência do Estado sobre Vida e Morte de Agricultores Familiares (2018)".

8 CONCLUSÃO E TRABALHOS FUTUROS

O fenômeno do suicídio é um problema que afeta países de todo o mundo e tem sua natureza multivariada, ou seja, possui diversas variáveis, intrínsecas e extrínsecas ao indivíduo, as quais influenciam a ocorrência desse fenômeno. Observou-se que estudos sobre esse problema, aliados a técnicas de mineração de dados e/ou aprendizado de máquina, têm obtido resultados relevantes e que podem contribuir para a compreensão do fenômeno do suicídio, colaborando com diversas áreas de atuação, principalmente no campo da saúde.

Este trabalho propôs uma solução que aborda, juntamente, a mineração de dados e algoritmos de aprendizado de máquina para agrupamento de dados aliados a técnicas de visualização de dados e algoritmos de árvores de decisão para ampliar a capacidade de compreensão dos agrupamentos encontrados. Foram apresentadas as etapas de metodologias da solução de maneira detalhada, bem como se mostraram análises individuais comparativas dos experimentos realizados.

O objetivo de perceber padrões em casos de suicídio utilizando algoritmos de aprendizado de máquina foi alcançado, considerando que diversos agrupamentos encontrados pelo modelo são semelhantes a grupos de indivíduos observados na literatura. Com isso, o presente trabalho colabora para a pesquisa na área de ciências da computação e sua aplicação, no caso específico, de estudos da área da saúde, da psiquiatria e de políticas públicas em território nacional.

Para trabalhos futuros, uma das sugestões de melhorias é a agregação de novas variáveis relacionadas às vítimas e ao ambiente com os quais elas convivem, trazendo mais informações sobre os indivíduos, isso pode fazer com que o modelo de agrupamento de dados obtenha resultados com mais informações das características de cada grupo, ou seja, que ele se torne cada vez mais especializado. Algumas sugestões de melhorias mais especificamente a respeito de quais variáveis podem ser interessantes de se agregar e a hipótese de motivação se encontram abaixo.

- Variáveis sobre a ocupação
  A agregação de mais informações a respeito dos aspectos financeiros da ocupação da vítima poderia trazer um novo ponto de vista a ser analisado, assim como é feito em relação ao grau de pobreza em (WHO, 2017).

- Variáveis sobre a população dos municípios
  Ao se obterem mais informações das populações dos municípios, seria possível analisar uma hipótese de relação entre o perfil do indivíduo e o do município. Entretanto, seria necessário cuidado para se agregar esses dados não só pelo município, mas também pelo ano que o dado se encontra. Outra melhoria possível com esses dados seria poder calcular uma taxa de número de suicídios a cada 100 mil habitantes para se ter uma visão dos dados em relação ao tamanho de seus municípios.

Outras sugestões de trabalhos futuros são:

- A aplicação desta solução proposta em outras regiões do País
  Os dados do SIM são nacionais e, portanto, não devem ter diferenças nos padrões entre as regiões. Sendo assim, seria possível adaptar a solução proposta para que fosse aplicada em dados de outras regiões do Brasil e até comparar os resultados das regiões, observando-se, com isso, detalhes pertinentes para cada estado e microrregião.

- Análise com foco em criação de políticas públicas
  A partir deste trabalho e de outros similares, abre-se caminho para um trabalho futuro que possivelmente seja feito em parceria com profissionais da saúde mental, por exemplo, e de ciências de dados com foco em analisar e validar as informações obtidas do trabalho com foco de criação de políticas públicas de apoio a vida e prevenção do suicídio.

REFERÊNCIAS

AGGARWAL, Charu C. **Data Mining: The Textbook**. Cham: Springer, 2015. ISBN 978-3-319-14141-1. DOI: 10.1007/978-3-319-14142-8.

AGNE, Neusa Aita *et al.* Predictors of suicide attempt in patients with obsessive-compulsive disorder: an exploratory study with machine learning analysis. **Psychological medicine**, Cambridge University Press, 2020. ISSN 1469-8978. DOI: 10.1017/S0033291720002329. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32669156/.

AHMED, Mohiuddin; SERAJ, Raihan; ISLAM, Syed Mohammed Shamsul. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. **Electronics**, v. 9, n. 8, 2020. ISSN 2079-9292. DOI: 10.3390/electronics9081295. Disponível em: https://www.mdpi.com/2079-9292/9/8/1295.

ASSOCIATION, American Psychiatric; ASSOCIATION, American Psychiatric *et al.* Diagnostic and statistical manual of mental disorders: DSM-5. **United States**, 2013.

CABITZA, Federico; BANFI, Giuseppe. Machine learning in laboratory medicine: waiting for the flood? **Clinical Chemistry and Laboratory Medicine (CCLM)**, De Gruyter, v. 56, n. 4, p. 516–524, 2018.

CHATURVEDI, Anil *et al.* K-modes Clustering. **Journal of Classification**, Springer-Verlag PUB3755 Berlin, Heidelberg, v. 18, p. 35–55, 1 jan. 2001. ISSN 01764268. DOI: 10.1007/S00357-001-0004-3. Disponível em: https://dl.acm.org/doi/10.1007/s00357-001-0004-3.

CLASSIFICAÇÃO Brasileira de Ocupações (CBO) – Portal Emprega Brasil. [*S.l.*: *s.n.*]. Disponível em: https://empregabrasil.mte.gov.br/76/cbo/.

COELHO, Flávio Codeço. **Data Sources — PySUS 0.1.13 documentation**. [*S.l.*: *s.n.*]. Disponível em: https://pysus.readthedocs.io/en/latest/data-sources.html.

COELHO, Flávio Codeço *et al.* **AlertaDengue/PySUS: Vaccine**. [*S.l.*]: Zenodo, mai. 2021. DOI: 10.5281/zenodo.4883502. Disponível em: https://doi.org/10.5281/zenodo.4883502.

CONSOLIDAÇÃO SIM. [*S.l.*: *s.n.*]. Disponível em: http://tabnet.datasus.gov.br/cgi/sim/Consolida_Sim_2011.pdf.

DREBES, Laila Mayara *et al.* Legislação, Política Pública e Suicídio: A Influência do Estado sobre Vida e Morte de Agricultores Familiares. **Desenvolvimento em Questão**, Editora Unijui, v. 16, p. 285–321, 44 ago. 2018. ISSN 2237-6453. DOI:

10.21527/2237-6453.2018.44.285-321. Disponível em: https://revistas.unijui.edu.br/index.php/desenvolvimentoemquestao/article/view/6570.

GONÇALVES, Ludmilla R.C. *et al.* Determinantes espaciais e socioeconômicos do suicídio no brasil: Uma abordagem regional. **Nova Economia**, v. 21, p. 281–316, 2 mai. 2011. ISSN 01036351. DOI: 10.1590/S0103-63512011000200005.

GRAHAM, Sarah *et al.* Artificial intelligence for mental health and mental illnesses: an overview. **Current psychiatry reports**, Springer, v. 21, n. 11, p. 1–18, 2019.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data Preprocessing. **Data Mining**, Elsevier, p. 83–124, 2012. DOI: 10.1016/B978-0-12-381479-1.00003-4.

HUANG, Zhexue. Clustering large data sets with mixed numeric and categorical values. *In:* IN The First Pacific-Asia Conference on Knowledge Discovery and Data Mining. [*S.l.*: *s.n.*], 1997. P. 21–34.

ICD-10 Version:2019. [*S.l.*: *s.n.*]. Disponível em: https://icd.who.int/browse10/2019/en#/.

INSEL, Thomas *et al.* **Research domain criteria (RDoC): toward a new classification framework for research on mental disorders**. [*S.l.*]: Am Psychiatric Assoc, 2010.

MARCON, Grasiela. **QUEM SÃO OS ESTUDANTES DE MEDICINA QUE TENTAM SUICÍDIO?** 2019. Disponível em: https://www.lume.ufrgs.br/handle/10183/199035.

MENEGHEL, Stela Nazareth *et al.* Características epidemiológicas do suicídio no Rio Grande do Sul. **Revista de Saúde Pública**, Faculdade de Saúde Pública da Universidade de São Paulo, v. 38, p. 804–810, 6 2004. ISSN 0034-8910. DOI: 10.1590/S0034-89102004000600008. Disponível em: http://www.scielo.br/j/rsp/a/xpNxhWkXKS7p6bTZRXwMctD/.

MITCHELL, T.M. **Machine Learning**. [*S.l.*]: McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em: https://books.google.com.br/books?id=EoYBngEACAAJ.

NARGESIAN, Fatemeh *et al.* Learning Feature Engineering for Classification. *In:* PROCEEDINGS of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. [*S.l.*: *s.n.*], 2017. P. 2529–2535. DOI: 10.24963/ijcai.2017/352. Disponível em: https://doi.org/10.24963/ijcai.2017/352.

ROGLIO, Vinícius Serafini *et al.* Prediction of attempted suicide in men and women with crack-cocaine use disorder in Brazil. **PLOS ONE**, Public Library of Science, v. 15, e0232242, 5 mai. 2020. ISSN 1932-6203. DOI: 10.1371/JOURNAL.PONE.0232242.

Disponível em:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232242.

RUSSEL, Stuart; NORVIG, Peter. **Artificial Intelligence A Modern Approach Third Edition**. [*S.l.*: *s.n.*].

SEHNEM, Scheila Beatriz; PALOSQUI, Vanusa. Caracteriêsticas epidemiológicas do suiciêdio no estado de Santa Catarina. **Fractal: Revista de Psicologia**, SciELO Brasil, v. 26, p. 365–378, 2014.

SILVA, Rômulo Pereira da. Suiciêdio: investigando as causas por meio da análise de dados? Niterói, 2017.

SINGH, Sonia; GIRI, Manoj. Comparative Study Id3, Cart And C4.5 Decision Tree Algorithm: A Survey. **International Journal of Advanced Information Science and Technology (IJAIST)**, v. 3, jul. 2014. ISSN 2319-2682. DOI: 10.15693/ijaist/2014.v3i7.47-52.

SOUZA, Pollianna Marys de; AUTRAN, Marynice Medeiros Matos de *et al.* Repositório datasus: organização e relevância dos dados abertos em saúde para a vigilância epidemiológica. **P2P E INOVAÇÃO**, v. 6, p. 50–59, 2019.

VAN HEERINGEN, K; MANN, JJ. **The neurobiology of suicide. Lancet Psychiatry 1 (1), 63–72**. [*S.l.*: *s.n.*], 2014.

VIANA, Greta Nazario *et al.* Prevalência de suicídio no Sul do Brasil, 2001-2005. **Jornal Brasileiro de Psiquiatria**, Instituto de Psiquiatria da Universidade Federal do Rio de Janeiro, v. 57, p. 38–43, 1 jan. 2008. ISSN 0047-2085. DOI: 10.1590/S0047-20852008000100008. Disponível em: http://www.scielo.br/j/jbpsiq/a/N3jbcBMYDCG9WVNxHrBFgkR/?lang=pt.

WHO. Other common mental disorders: global health estimates. **Geneva: World Health Organization**, p. 1–24, 2017.

YU, Lean *et al.* Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation? **https://doi.org/10.1080/1540496X.2020.1825935**, Routledge, v. 58, p. 472–482, 2 2020. ISSN 15580938. DOI: 10.1080/1540496X.2020.1825935. Disponível em:
https://www.tandfonline.com/doi/abs/10.1080/1540496X.2020.1825935.

ZHENG, Xu; CAI, Zhipeng; LI, Demin. Data Linkage in Smart Internet of Things Systems: A Consideration from a Privacy Perspective. **IEEE Communications Magazine**, v. 56, p. 55–61, set. 2018. DOI: 10.1109/MCOM.2018.1701245.

APÊNDICE A – ANÁLISES COMPLEMENTARES

Na seção 7 são apresentadas as análises dos atributos mais relevantes para as características dos agrupamentos. Então, neste apêndice são apresentadas as análises dos demais atributos na seção A.1.

A.1 ANÁLISES COMPLEMENTARES

As análises serão apresentadas separadas em subseções, onde cada subseção corresponde a uma quantidade de agrupamentos encontrados.

A.1.1 Para 2 agrupamentos

Os atributos em que a distribuição dos valores não possui uma diferença muito relevante entre os 2 agrupamentos são:

- O sexo biológico
  Sendo a distribuição dos valores no cluster 0 de 80.4% masculino e 19.6% feminino e no cluster 1 há 72.7% masculino e 27.3% feminino.

- O local da ocorrência
  Ocorrências com o local sendo a residência do indivíduo formam aproximadamente 70% da população de cada cluster e ocupam a posição de local mais frequente. Seguidos de local não especificado, com cerca de 12%, de outros locais especificados, com cerca de 5% e de rua e estrada, com cerca de 4.6%, em ambos clusters.

- Cidade da ocorrência é uma capital?
  Ocorrências em cidades que não são capitais são a maioria nos dois clusters, com mais de 86% em ambos os clusters. O possível motivo dessa desproporção é o desquilibrio de cidades capitais e não capitais já que o conjunto de dados continha apenas 3 estados e, portanto, 3 capitais.

- Se recebeu assistência médica
  Em ambos clusters cerca de mais de 71% das amostras não receberam assistência médica durante a ocorrência. E cerca de mais de 20% receberam assistência. Existem também cerca de 6% de valores ignorados.

- Raça e/ou cor declarada
  Indivíduos declarados brancos são a maioria em ambos clusters, com cerca de 85% das amostras do cluster 0 e 90% no cluster 1. Seguidos de indivíduos declarados pardos, com cerca de 10% no cluster 0 e 6.7% no cluster, e indivíduos declarados pretos, com cerca de 3.7% no cluster 0 e 2.3% no cluster 0. Os indivíduos declarados amarelos ou indígenas ocupam cerca de 0.5% de cada agrupamento.

- Causa básica da morte (método usado)
  Nos dois clusters desse experimento as cinco maiores causas de morte por lesões auto-provocadas, ou os cinco métodos mais usados, são: esforcamento, estrangulamento e sufocação; disparo de arma de fogo não especificada; disparo de arma de fogo de mão; precipitação de lugar elevado; auto-intoxicação por pesticidas. Sendo que a única pequena diferença entre os agrupamentos é a ordem da quarta e quinta causa, em que no cluster 0 elas seguem a ordem listada anteriormente e no cluster 1 essas duas causas trocam de posição.

Já os atributos que apresentaram alguma diferença, mas não representam as características destaque dos agrupamentos são:

- Estado de residência
  Os 3 estados estão distribuídos entre os clusters de maneira que não parecem se um dos atributos mais relevantes para classificação. Embora vale mencionar que a ordem dos dois estados com mais ocorrências é diferente entre os clusters. No cluster 0 o estado do Rio Grande do Sul predomina com 39.6% da população seguido do estado do Paraná com 36.2%. Já no outro cluster o estado que predomina é o Paraná com 36.8%, porém o estado do RS fica como segundo mais frequente por 1% a menos que o estado com mais ocorrências.

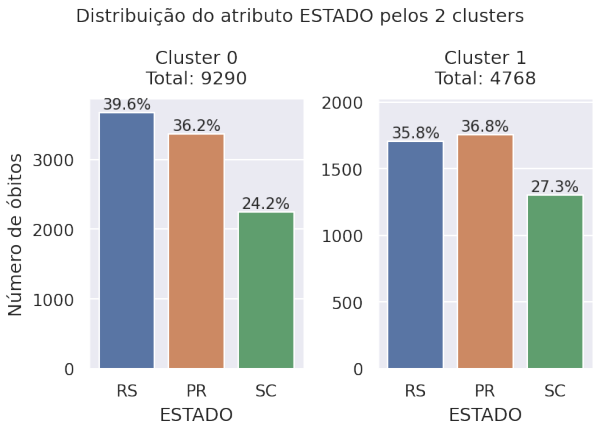

Figura 34 – Distribuição de ocorrências por estado federal para 2 agrupamentos

- Escolaridade em anos
  Uma população com a maioria de indivíduos com 4 a 7 anos de escolaridade é vista no cluster 0 enquanto no cluster 1 a maioria é de pessoas com 8 a 11 anos de estudo. O perfil dos indivíduos do cluster 0 parece ser de indivíduos com 4 a 7 anos de escolaridade, algo como ensino fundamental, em que a ocorrência foi na manhã de um dia útil. Já o cluster 1 tem um perfil com a maioria de indivíduos com 8 a 11 anos de escolaridade, algo como ensino médio, que cometeram na parte da tarde de dias tanto úteis como feriados e fins de semanas.

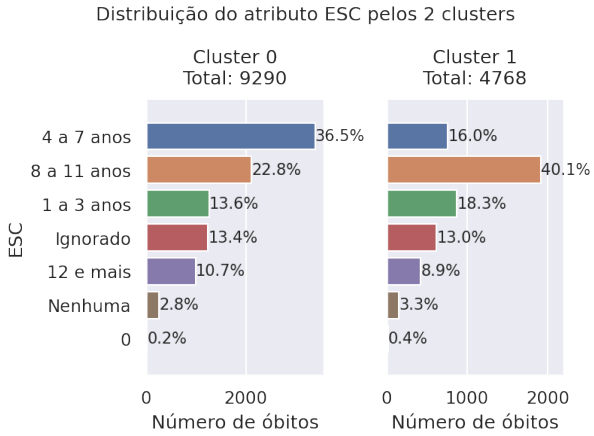

Figura 35 – Distribuição anos de escolaridade para 2 agrupamentos

- Ocorrência em fim de semana ou feriado?
  No cluster 0 é possível ver que cerca de 80% das ocorrências não aconteceram em feriados ou fim de semana e cerca de 20% ocorreram. Entretanto, no cluster 1 as distribuições dos valores são muito mais próximas, sendo que 57.7% ocorreram em dias úteis e 42.3% ocorreram em dias nos dias da hipótese. Isso pode indicar um traço de diferença entre os agrupamentos. Para o algoritmo de clusterização a ocorrência em dia útil parece ser extremamente relevante para classificar um dado como pertencente ao cluster 0 enquanto no cluster 0 essa informação não afeta tanto o agrupamento.

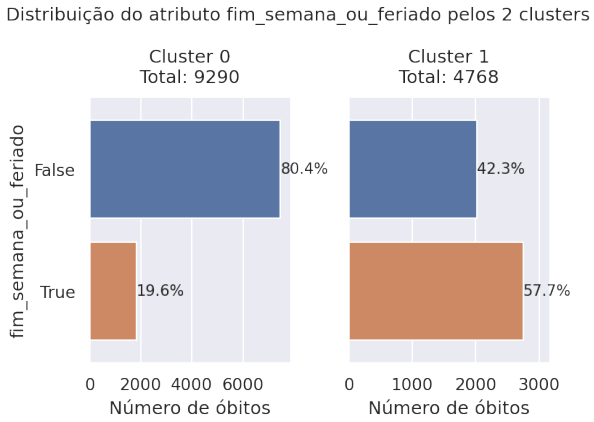

Figura 36 – Distribuição de ocorrências em feriados ou fins de semana para 2 agrupamentos

A.1.2 Para 3 agrupamentos

Os atributos em que a distribuição dos valores não possui tanta diferença entre os 3 agrupamentos são:

- Fim de semana ou feriado?
  Os três agrupamentos possuem cerca de 70% de suas ocorrências em dias úteis e cerca de 30% em feriados ou finais de semana.

- Local da ocorrência
  A predominância de ocorrências em residência é de cerca de 70% em todos os agrupamentos. E são seguidas das ocorrências em locais não especificados, outros locais não especificados e rua e estrada. Os outros locais de ocorrência possíveis após a rua e estrada não chegam a 2% das ocorrências cada.

- Ocorreu em capital?
  Há uma predominância de mais de 85% de ocorrências em cidades que não capitais e cerca de 15% em cidades capitais em todos os agrupamentos.

- Raça/cor declarada
  Os óbitos de indivíduos declarados brancos chega a mais de 80% em todos os agrupamentos. As outras 2 categorias mais frequentes são óbitos de indivíduos declarados pardos e indivíduos declarados pretos, respectivamente. Os indivíduos declarados amarelos e indígenas não somam 1% da população do respectivo agrupamento.

Já os atributos que apresentaram alguma diferença, mas não representam as características destaque dos agrupamentos são:

- Escolaridade
  Ao observar as duas faixas de escolaridade mais frequentes de cada cluster podemos ter uma visão generalizada da maior parte da população dos clusters. No cluster 0 temos grande parte dos indivíduos distribuídos entre 4 a 7 anos, sendo 38.4%, e 8 a 11 anos, sendo 30.5%, de escolaridade, sendo a primeira mais frequente e isso poderia significar uma escolaridade entre ensino fundamental e médio completo/incompleto. No cluster 1 a diferença do tamanho das populações das duas faixas mais frequentes é bem maior do que no cluster 0 chegando em cerca de 30%. Nesse cluster a faixa de 8 a 11 anos de idade equivale a quase 50% do cluster e é seguida da faixa de 4 a 7 anos de idade com cerca de 18% da população do cluster. E por fim o cluster 2 contra a maior parte de sua população na faixa de 1 a 3 anos, com cerca de 30%, e na faixa de 4 a 7 anos, com cerca de 23%, indicando uma tendência de ter indivíduos com anos de escolaridade mais baixa que os outros clusters.

- Ocupação
  Pode-se perceber que embora a ocupação de trabalhador volante da agricultura tenha mais ocorrências nos clusters 1 e 2, o cluster 2 possui a segunda ocupação com mais ocorrências ainda na área de agrícola, enquanto no cluster 1 é a ocupação de comerciante varejista. Vale também ressaltar que a ocupação de pedreiro se encontra

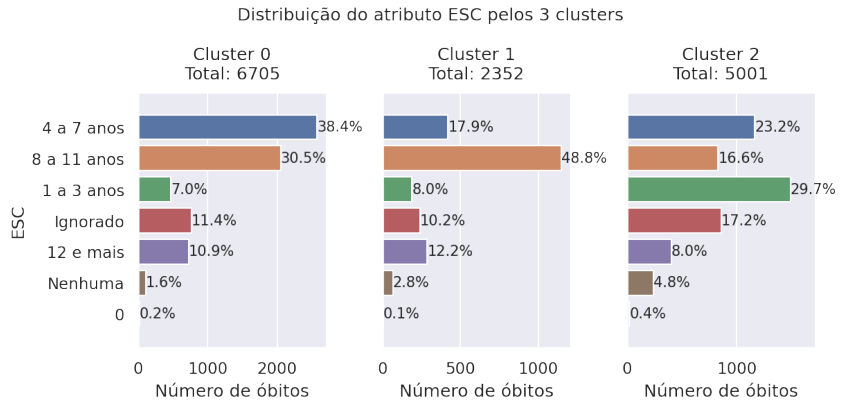

Figura 37 – Distribuição Escolaridade de entre os 3 agrupamentos

na lista das 5 ocupações com mais ocorrências dos cluster 0 e 2, mas do cluster 1 não. Isso indica uma outra diferença entre o perfil dos indivíduos dos clusters. No cluster 1 temos a ocupação de pedreiro como mais frequente, seguida de empregado doméstico e de trabalhador volante agrícola.

Proporção das ocorrências das 3 ocupações mais frequentes em cada cluster					
Cluster 0		Cluster 1		Cluster 2	
Ocupação	Proporção	Ocupação	Proporção	Ocupação	Proporção
Pedreiro	7,38%	Trabalhador volante da agricultura	4,00%	Trabalhador volante da agricultura	8,92%
Empregado doméstico nos serviços gerais	5,29%	Comerciante varejista	3,27%	Produtor agrícola polivalente	5,26%
Trabalhador volante da agricultura	4,30%	Empregado doméstico nos serviços gerais	2,76%	Pedreiro	5,02%

Tabela 14 – Distribuição do turno do dia das ocorrências entre os 3 agrupamentos

- Estado de residência
  Apesar da população de cada cluster estar relativamente bem distribuída entre os estadoshá uma diferença entre o cluster 2 e os demais. No cluster 0 e 1 o estado com mais ocorrências é o Paraná, seguido do Rio Grande do Sul e por fim Santa Catarina. Porém, no cluster 2 temos o Rio grande do Sul com mais ocorrências, seguido do Paraná e por último Santa Catarina. É possível perceber também que, mesmo que os clusters 0 e 1 tenham a ordem parecida, o estado menos frequente possui uma diferença de 13.2% para o segundo mais frequente no cluster 0 e porém no cluster 1 essa diferença é de apenas 3.3%.

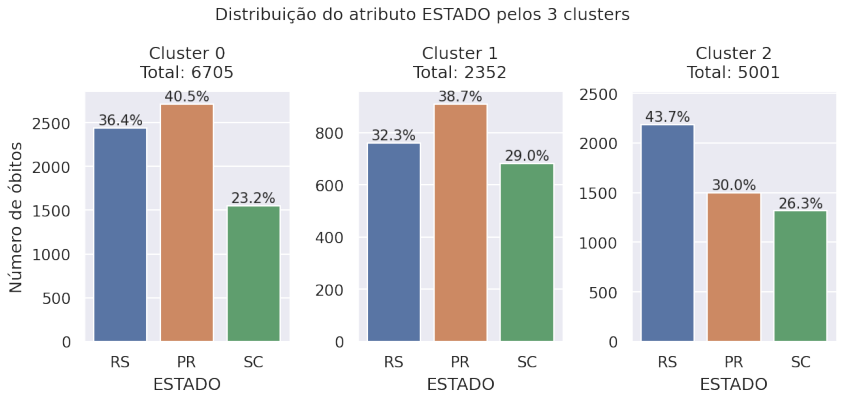

Figura 38 – Distribuição Estado de residência de entre os 3 agrupamentos

A.1.3 Para 4 agrupamentos

Os atributos em que a distribuição dos valores não possui tanta diferença entre os 4 agrupamentos são:

- Raça/cor declarada
  A distribuição da coluna "RACACOR" nesse experimento segue o mesmo padrão dos resultados dos anteriores. Os clusters são compostos por mais de 80% de indivíduos brancos, seguidos por 7 a 12% de indivíduos pardos, seguidos de menos de 5% de indivíduos pretos e por fim os indivíduos amarelos e indígenas compõem menos de 0.5% cada. Exceto o cluster 2 que não possui nenhum indivíduo declarado amarelo.

- Local da ocorrência
  A distribuição dos valores desse atributo mantém as mesmas proporções dos experimentos anteriores. Sendo os 3 lugares que mais ocorreram: na residência, seguidas de local não especificado e outros locais especificados.

- Cidade da ocorrência era capital?
  Seguindo o mesmo padrão dos outros experimentos a maioria das ocorrências foram em cidades não-capitais.

Já os atributos que apresentaram alguma diferença, mas não representam as características destaque dos agrupamentos são:

- Ocorreu em fim de semana ou feriado
  Existem duas distribuições parecidas entre os clusters para os valores desse atributo. Existe a distribuição onde cerca de 80% das ocorrências não foi em feriado ou final de semana enquanto o resto aconteceu sim. Essa distribuição é vista nos cluster 0 e 3. E existe a outra distribuição em que cerca de 60 a 75% das ocorrências aconteceu em feriados ou fim de semana e o resto não. Os clusters 0 e 1 apresentam essa distribuição.

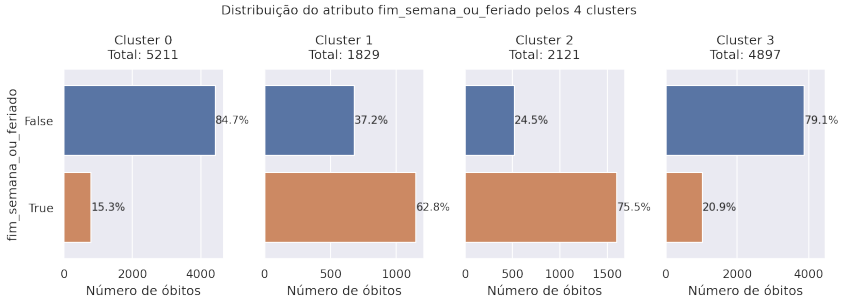

Figura 39 – Distribuição das ocorrências em fim de semana ou feriado entre os 4 agrupamentos

- Turno do dia da ocorrência
  Os clusters 0 e 1 a predominância de ocorrências pela manhã é de um pouco mais de 60%, porém no cluster 0 temos a parte da tarde como segunda mais frequente enquanto no cluster 1 a segunda mais frequente é a parte da noite, ambas com cerca de 20% da população do cluster. No cluster 2 a parte da noite é a maioria enquanto a parte da manhã e tarde são bem próximas a 22% cada. Já o cluster 3 apresenta uma distribuição diferente do resto, pois seus valores estão mais distribuídos já que, mesmo o turno da tarde sendo o mais frequente, o turno mais frequente e o segundo mais frequente têm uma diferença de 12.3%, que é pequena quando comparados a diferença de cerca de mais de 40% que os cluster 0 e 1 possuem entre seus dois valores com mais população.

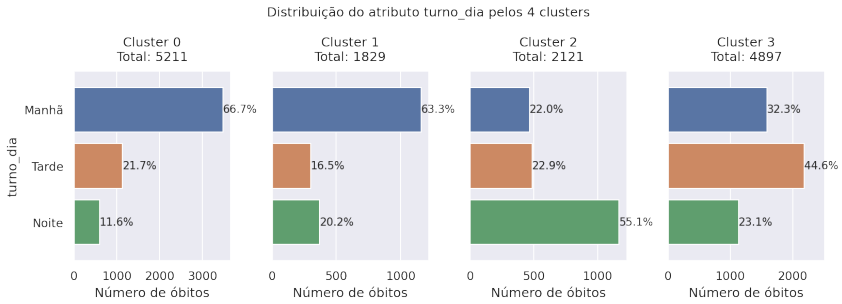

Figura 40 – Distribuição do turno do dia entre os 4 agrupamentos

- Estado de residência
  Os clusters 1, 2 e 3 têm a mesma ordem decrescente de estados que é Paraná, Rio Grande do Sul e Santa Catarina. Já o cluster 0 possui a maioria de seus indivíduos residentes no Rio Grande do Sul, seguidos do Paraná e não muito menos de Santa Catarina.

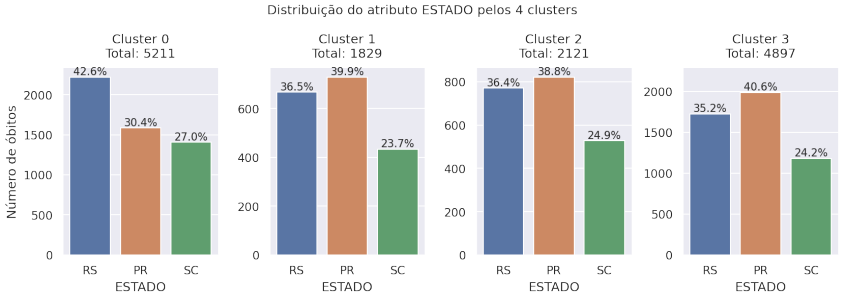

Figura 41 – Distribuição do estado de residência entre os 4 agrupamentos

A.1.4 Para 5 agrupamentos

Os atributos em que a distribuição dos valores não possui tanta diferença entre os 5 agrupamentos são:

- Local da ocorrência
  A proporção e ordem os locais mais frequentes se mantém igual aos experimentos dos grupos anteriores. A maioria de cerca de 70% ocorre na residência, seguidos de locais não especificados e outros locais especificados.

- Cidade da ocorrência é capital?
  Assim como os experimentos anteriores a distribuição de ocorrências em cidade não-capitais predomina por cerca de 80% ou mais.

- Raça/cor declarada
  A maioria é de indivíduos declarados brancos, com mais de 83% da população dos clusters, assim como nos outros experimentos.

Já os atributos que apresentaram alguma diferença, mas não representam as características destaque dos agrupamentos são:

- Fim de semana ou feriado
  Os cluster 0, 3 e 4 são clusters com ocorrências majoritariamente em dias úteis, cerca de 75 a 88% de suas populações. Enquanto no cluster 1 e 2 a maioria das ocorrências foram em feriados ou finais de semana, com cerca de 66 a 69% de suas populações.

- Turno do dia
  Nos clusters 0 e 2 tem uma predominância de ocorrências pela parte da manhã por uma diferença relativamente grande, cerca de 40 a 60% de diferença do segundo turno mais frequente. Os clusters 1 e 3 tem a predominância de ocorrências na parte de tarde, porém o cluster possui a proporção de seus valores muito equilibrados sendo 41.6% parta da tarde, 32.2% da parte da manhã e 26.2% da parte da noite. Enquanto o cluster 3 possui 52.4% de suas ocorrências na parte da tarde, 25.3% na parte da manhã e 22.4% na parte da noite.

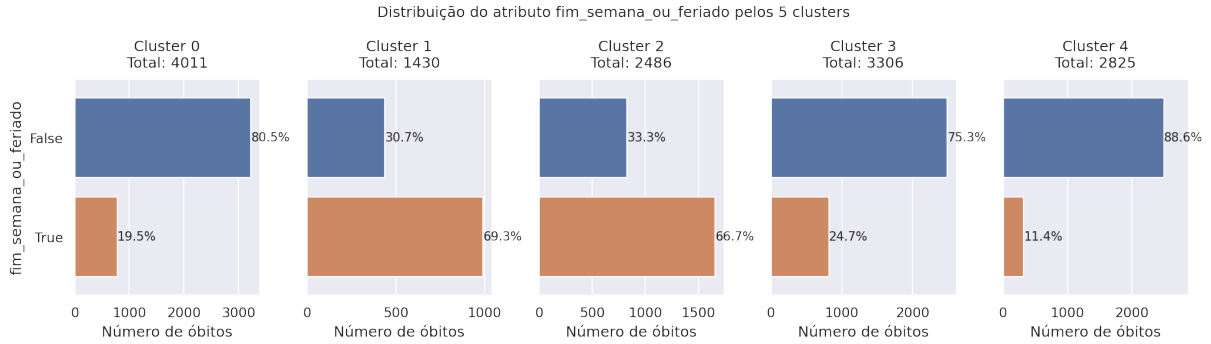

Figura 42 – Distribuição das ocorrências em feriados ou finais de semana entre os 5 agrupamentos

Já o cluster 4 possui a maioria de suas ocorrências na parte da noite, cerca de 50% de sua população, seguidas de 27.8% das ocorrências na parte da tarde e 22.4% na parte da manhã.

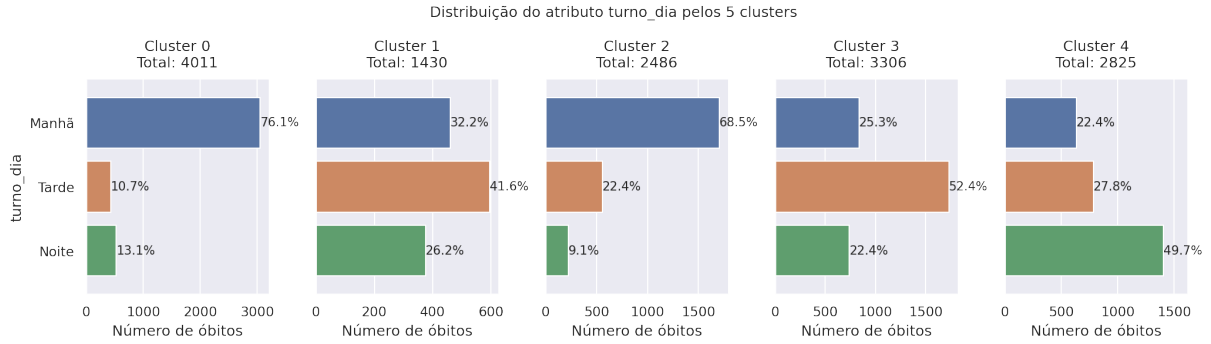

Figura 43 – Distribuição do turno do dia das ocorrências entre os 5 agrupamentos

- Estado de residência
  A distribuição dos estados é relativamente bem distribuída considerando que não há nenhum estado que se destaque com mais de 50% da população do cluster. Temos a predominância do estado do Paraná seguido do Rio Grande do Sul e por fim Santa Catarina nos clusters 1 e 3. E nos clusters 0, 2 e 4 o estado do Rio Grande do Sul é que possui mais ocorrências, seguido do Paraná e por fim Santa Catarina.

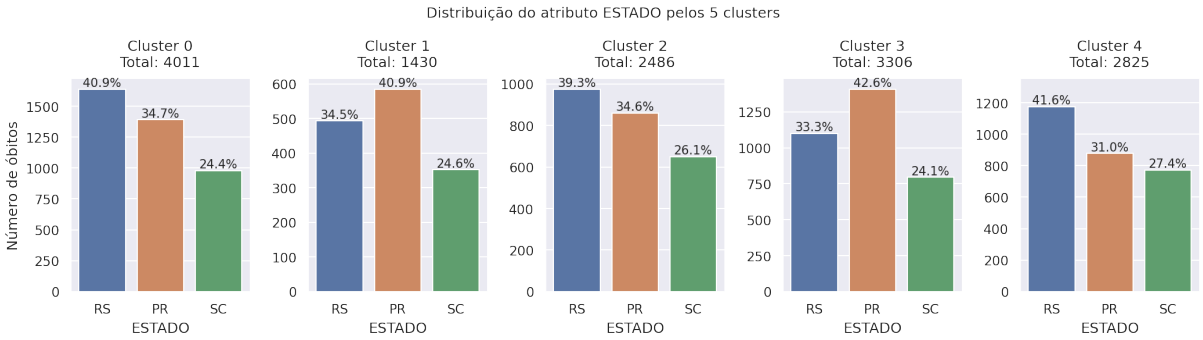

Figura 44 – Distribuição do estado de residência entre os 5 agrupamentos

APÊNDICE B – FIGURAS DAS ÁRVORES DE DECISÃO

As figuras deste anexo estão rotacionadas na para melhor visualização devido seu tamanho. As mesmas figuras estão disponíveis no repositório público do código fonte deste trabalho, para mais detalhes sobre o acesso visite o apêndice C.

A árvore obtida pela divisão dos dados em 3 agrupamentos é representada pela figura 45.

Para a divisão dos dados em 4 agrupamentos a árvore obtida é apresentada na figura 46.

Para a divisão dos dados em 5 agrupamentos a árvore obtida é apresentada na figura 47.

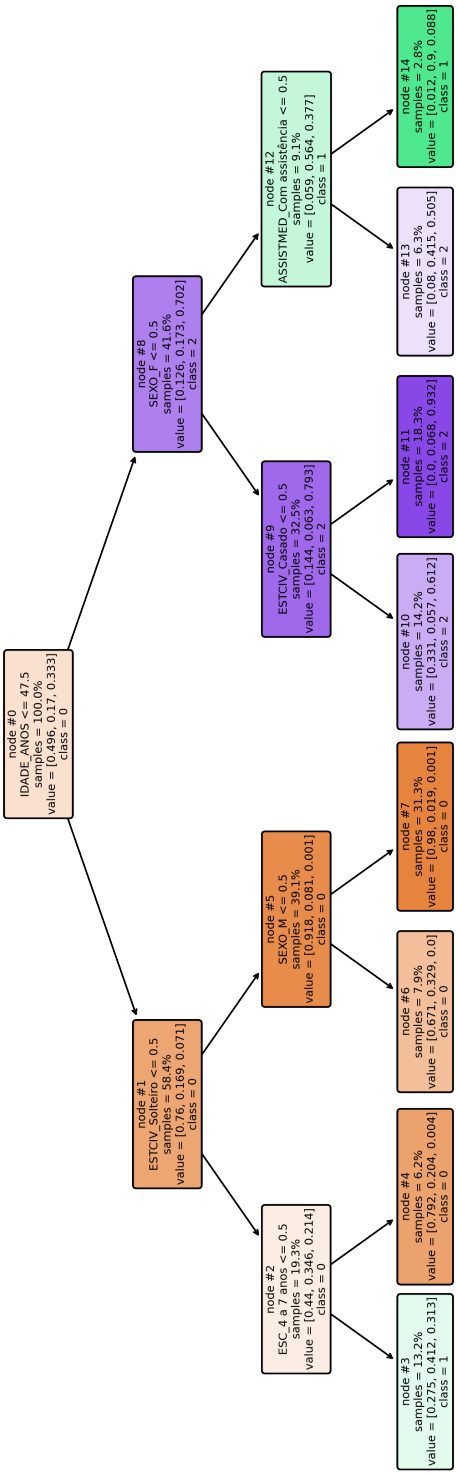

Figura 45 – Visualização da árvore de decisão para 3 *clusters* (Fonte: o autor)

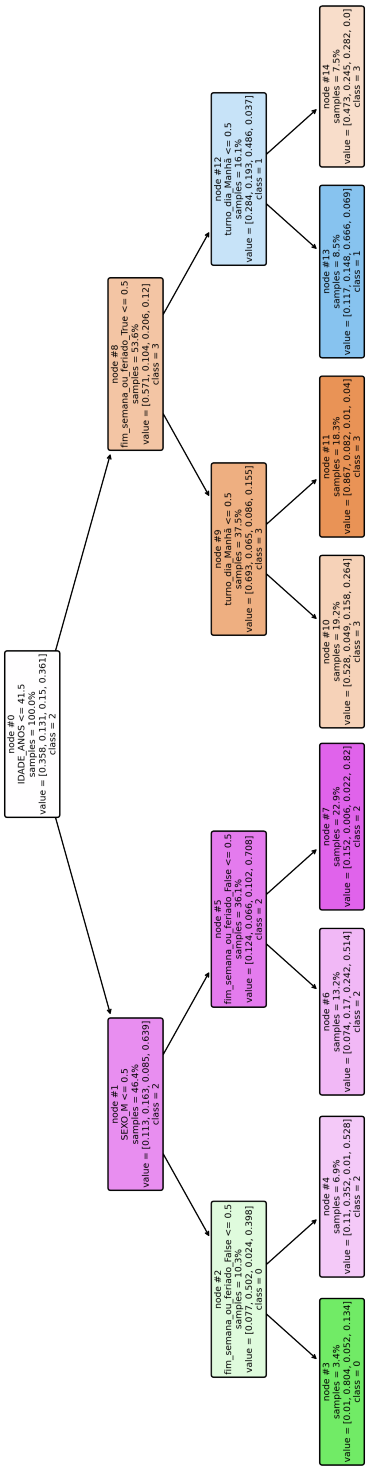

Figura 46 – Visualização da árvore de decisão para 4 *clusters* (Fonte: o autor)

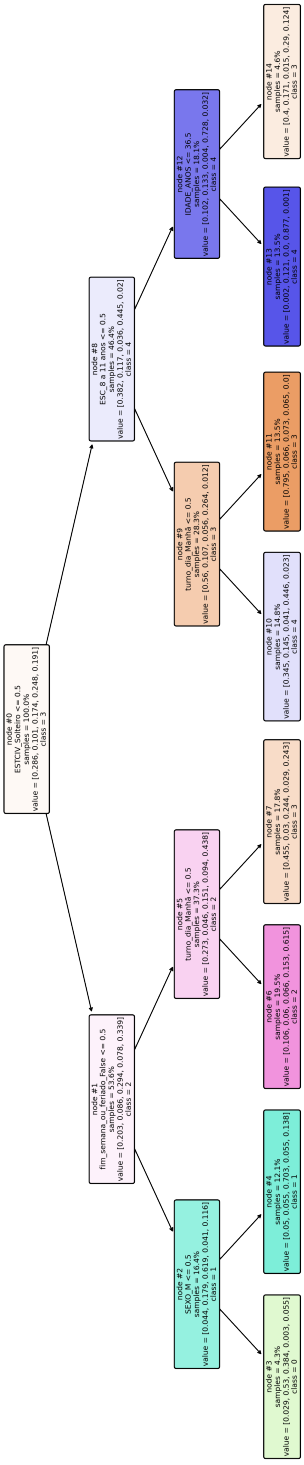

Figura 47 – Visualização da árvore de decisão para 5 *clusters* (Fonte: o autor)

APÊNDICE C – CÓDIGO FONTE

O código fonte desenvolvido pelo autor e utilizado para a geração dos resultados deste trabalho está disponível no repositório público do autor na plataforma Github e pode ser acessado através do link: `https://github.com/LucasVerdade/bachelor-degree-final-project`

Dentro do repositório há um arquivo `README.md` que contém as instruções detalhadas de como executar corretamente o código do trabalho.

APÊNDICE D – ARTIGO

Neste apêndice é apresentado o artigo sobre este trabalho seguindo o padrão da Sociedade Brasileira de Computação.

Análise de padrões em casos de risco de suicídio utilizando aprendizado de máquina

Lucas Verdade Godoy[1], **Mateus Grellert**[1], **Jônata Tyska**[1], **Manuella Kaster**[2], **Rafael de Santiago**[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC) – Florianópolis – SC – Brasil

[2]Departamento de Bioquímica
Universidade Federal de Santa Catarina (UFSC) – Florianópolis – SC – Brasil

Abstract. *Suicide is a worldwide concerning phenomenon that is difficult to understand due to its multifactorial nature. Machine learning techniques (ML) have been used to facilitate the aggregation of large amounts of isolated data, as well as the profiling and modeling among groups of individuals. In Brazil, such studies are still incipient. The Brazilian government has been facilitating access to data related to this phenomenon, however in an isolated manner. This article proposes the aggregation of these data through data mining techniques (DM) and the use of machine learning techniques to assist in the search for solutions related to mental health in Brazil.*

Resumo. *O fenômeno do suicídio é uma preocupação global de difícil compreensão devido o seu caráter multifatorial. Técnicas de aprendizado de máquina (Machine Learning - ML) têm sido utilizadas para facilitar a agregação de grandes quantidades de dados individuais, bem como a geração de perfis e modelos entre grupos de indivíduos. No Brasil, estudos desse tipo ainda são incipientes. O governo tem facilitado o acesso a dados relacionados a esse fenômeno, entretanto de maneira isolada. O presente artigo propõe a agregação desses dados por meio de técnicas de mineração de dados (Data Mining - DM) e o uso de técnicas de aprendizado de máquina para auxiliar na busca por soluções relacionadas à saúde mental no país.*

1. Introdução

O suicídio é a 18ª maior causa de morte global, a 2ª maior entre jovens, e em torno de 800 mil pessoas cometem suicídio anualmente pelo mundo [WHO 2017]. No Brasil, de acordo com o Ministério da Saúde [Sehnem and Palosqui 2014], o estado de Santa Catarina concentra cerca de 21% dos casos de suicídios no país. Trata-se de um fenômeno resultante de uma interação complexa entre estressores ambientais, como adversidades e eventos da vida, e traços de susceptibilidade da vítima, independente de transtornos psiquiátricos, segundo [Van Heeringen and Mann 2014].

Atualmente, a avaliação de risco de suicídio é feita clinicamente [Association et al. 2013]. No entanto, a compreensão limitada do aspecto epidemiológico aliada a outros diversos diagnósticos associados dificulta a detecção de possíveis vítimas e, consequentemente, o estabelecimento de ações de prevenção e de intervenção [Insel et al. 2010]. Isso é agravado pelo caráter multifatorial das causas

que aumentam o risco de suicídio. Portanto, é necessário buscar soluções capazes de considerar diferentes fontes de dados para auxiliar na estimativa de risco de suicídio.

No campo da psiquiatria, o uso de técnicas de aprendizado de máquina (*Machine learning - ML*) vem crescendo e tem se mostrado um forte aliado. Essas técnicas permitem que se faça a análise de grandes quantidade de dados, bem como a geração de perfis e modelos entre grupos de indivíduos de risco [Cabitza and Banfi 2018, Graham et al. 2019]. No entanto, ainda não existe um conjunto de dados para análise dos múltiplos fatores do suicídio no Brasil, o que impede que as soluções propostas em outros trabalhos, internacionais, sejam aplicadas no contexto nacional.

O governo brasileiro tem buscado formas de facilitar o acesso automatizado a diversos dados públicos a fim de subsidiar análises objetivas da situação sanitária. O Departamento de Informática do Sistema Único de Saúde (DATASUS) fornece diferentes dados digitais da saúde por meio de *Application Programming Interface* (API) [de Souza et al. 2019]. Em 2020, o DATASUS coordenou a criação da Rede Nacional de Dados em Saúde (RNDS), com uma API que disponibiliza o acesso a vários dados de saúde (informações sobre pacientes, dados de exame, do examinador etc). Infelizmente, essa API só é acessível a estabelecimentos de saúde. Além disso, esses dados são usualmente coletados e analisados de forma isolada, o que impede o desenvolvimento de soluções que contemplem os diversos fatores associados ao suicídio no Brasil. Portanto, a integração desses dados em uma base única não é trivial. São necessárias técnicas de mineração de dados (*Data Mining* - DM) para realizar essa integração de maneira correta, bem como para realizar a limpeza e as devidas transformações para extrair informação desses dados de forma eficiente.

O presente artigo busca trazer contribuições nessa linha, propondo o uso de técnicas de aprendizado de máquina que auxiliem na busca por soluções relacionadas à saúde mental, particularmente por meio da construção de perfis de grupos de risco.

2. Conceitos básicos e trabalhos relacionados

2.1. Conceitos básicos

2.1.1. Aspectos do fenômeno do suicídio

Uma das grandes preocupações de saúde pública no Brasil e no mundo é o fenômeno do suicídio. De fato, trata-se da 18ª maior causa de morte de toda a população mundial e, segundo dados da Organização Mundial da Saúde [WHO 2017], o número de pessoas que comete suicídio equivale a cerca de 1,5% do total de mortes globais, sendo a maioria dessas mortes de homens. O suicídio não é apenas uma consequência de características do indivíduo; trata-se, na verdade, de um fenômeno que envolve a sociedade e o ambiente em que se está condicionado a viver [Van Heeringen and Mann 2014], ou seja, é um fenômeno multifatorial, que abarca elementos intrínsecos e extrínsecos ao indivíduo. Os dados da OMS mostram que esse fenômeno atinge de maneira preocupante pessoas entre 15 e 29 anos, sendo a segunda principal causa global de morte nessa faixa etária, e é particularmente preocupante em países de baixa e média renda, como o Brasil [Silva 2017]. Existe uma relação espacial de tentativas de suicídio [Gonçalves et al. 2011], em que regiões com uma alta taxa de tentativas provavelmente terão, em sua vizinhança, regiões que também apresentam taxas elevadas. Essa relação foi verificada por meio da análise

de dados e algoritmos de aprendizado de máquina com o intuito de buscar identificar fatores preditores e associados ao risco da tentativa de suicídio [Gonçalves et al. 2011]. A localidade espacial é identificada como um possível fator associado à taxa de tentativas nas microrregiões do Brasil.

2.1.2. Plataformas de dados de saúde

Em 1991, surgiu o DATASUS e, durante seu tempo de atuação, foram desenvolvidos mais de 200 sistemas que auxiliam a construção e fortalecimento do Sistema Único de Saúde (SUS). O departamento também fornece informações e dados de diversos aspectos da saúde da população brasileira para que decisões sejam tomadas com base em evidências e indicadores de saúde. Há muitos anos essa informação está disponibilizada para acesso público, mas até recentemente ela dependia de softwares específicos para seu acesso e eles não possuíam nenhum tipo de API para acesso automatizado. Em 2016, foi criada uma biblioteca em Python, conhecida como *PySUS*, com o objetivo de resolver esse problema [Coelho et al. 2021]. Essa biblioteca é um pacote para linguagem Python que reúne diversos utilitários para lidar com bancos de dados públicos do DATASUS [Coelho]. Sua versão atual possui acesso a diversos bancos, permitindo que ela seja utilizada em pesquisas as mais variadas voltadas à saúde. Dentre eles, está incluso o banco do Sistema de Informação sobre Mortalidade (SIM), um dos principais instrumentos para o apoio e criação de políticas de prevenção e cuidado em saúde. Esse sistema também possui informações por indivíduo a respeito da causa de sua morte, incluindo dados socioeconômicos, local de residência e de ocorrência, causa do óbito, entre outras, mais específicas, como a faixa etária, o estado civil e o sexo biológico.

2.1.3. Algoritmos de Aprendizado de Máquina

Um agente aprende quando melhora sua performance em tarefas futuras após a realização de observações a respeito do mundo, é o que apontam [Russel and Norvig] em [Russel and Norvig], obra em que também definem que há três tipos principais de aprendizado, que são: o aprendizado supervisionado, o não-supervisionado e o aprendizado por reforço. Neste trabalho são abordados apenas modelos de aprendizado de máquina supervisionado e não-supervisionado.

A etapa de treinamento desses modelos é o momento em que acontece o aprendizado de máquina, pois os dados de treinamento estão sendo usados para que o modelo tente aprender os seus padrões. Modelos de aprendizado supervisionado recebem pares de dados de entrada e respostas e, a partir disso, o modelo faz observações para tentar encontrar o padrão que mapeia a entrada para o resultado corretamente. Como esse modelo tem as respostas corretas de cada par de entrada, é possível calcular um *feedback* de quão próximo o resultado do modelo, que foi gerado através do padrão aprendido até o momento, chegou ao resultado real, que era a saída para aquele par de entradas no mundo do modelo. A supervisão usada nesse tipo de modelo, então, é o acesso às respostas reais para se calcular o *feedback* e fazer ajustes, caso necessário.

No caso deste trabalho, os agentes são os modelos de aprendizado de máquina e o seu mundo é o conjunto de dados que é usado no treinamento dos próprios agentes. Já o

aprendizado não-supervisionado, este não possui os dados de saída e, portanto, não pode calcular um *feedback* explícito baseado na proximidade de seu resultado com o resultado real. Nesse tipo de aprendizado, o modelo tenta aprender os padrões nos dados de entrada apenas. A definição e explicação dos algoritmos de aprendizado de máquina usados neste trabalho se encontram abaixo.

Algoritmo: K-means

É um dos algoritmos de aprendizado de máquina não-supervisionado mais populares, apesar de suas limitações [Ahmed et al. 2020]. Ele é usado para fazer agrupamento, ou clustering, de dados baseado em suas características.

Ou seja, o algoritmo k-means é usado para dados numéricos. Esse algoritmo depende do valor de K, que deve ser sempre especificado para performar uma análise de *clustering* [Ahmed et al. 2020].

Dado um conjunto X = [X0, X1, X2, .., Xn] de dados e um número inteiro, maior que zero, K será o número de agrupamentos a serem encontrados. O funcionamento desse algoritmo segue os passos abaixo:

1: Escolhe-se aleatoriamente K elementos, diferentes entre si, de X para serem os centros de cada agrupamento.

2: Para cada dado em X, calcula-se a dissimilaridade entre o dado atual e os centros dos agrupamentos;

3: A partir das dissimilaridades calculadas, atribuem-se os dados aos agrupamentos que tem a menor dissimilaridade;

4: Recalcula-se o centro dos agrupamentos reatribuindo seus valores com a média dos valores dos dados pertencentes àquele grupo.

5: Repetir os passos 2, 3 e 4 até encontrar a posição ideal do centro dos agrupamentos.

Esse algoritmo é específico para dados com atributos do tipo numéricos e seu cálculo de dissimilaridade é feito usando a distância euclidiana entre o dado e o centro, considerando os atributos numéricos do dado como as coordenadas.

Algoritmo: K-modes

Como aponta [Chaturvedi et al. 2001], o algoritmo K-means não é apropriado para dados categóricos, para tanto, faz-se necessário o algoritmo K-modes. Trata-se de uma extensão do algoritmo K-means, portanto, assim como aquele, é um algoritmo de aprendizado de máquina não-supervisionado usado para fazer agrupamento, ou clusterização, de dados com base em suas características. Porém, essa versão é feita para dados categóricos e, por isso, não é possível usar o cálculo da distância euclidiana entre dois pontos. Sua função de cálculo de dissimilaridade calcula uma pontuação de dissimilaridade baseada na hipótese dos valores dos atributos de dois dados serem ou não diferentes.

Dado um conjunto X = [X0, X1, X2, .., Xn] de dados e um número inteiro, maior

que zero, K será o número de agrupamentos a serem encontrados. Em que Xi, para $0 \geqslant i \geqslant$ n, é composto por (A0, A1, A2,..., Am) que representam tuplas de atributos categóricos, em que Aj, para $0 \geqslant j \geqslant m$, é um atributo categórico. O cálculo de diferença de atributos é feito de forma condicional, como explicado a seguir:

Sejam Ai0 e Aj0 atributos de diferentes dados:
Caso Ai0 = Aj0 então a diferença dos atributos é 0,
Caso Ai0 != Aj0 então a diferença dos atributos é 1.

Então, o cálculo da dissimilaridade de dois dados com atributos categóricos é feito com o somatório dos cálculos de diferenças entre seus atributos.

Algoritmo: K-prototypes

Trata-se de uma extensão dos dois algoritmos de agrupamentos citados anteriormente. Essa versão, entretanto, foi feita para ser usada em conjuntos de dados mistos, ou seja, que têm atributos numéricos e atributos categóricos. O cálculo de dissimilaridade dessa versão é a soma do cálculo das distâncias euclidianas dos atributos numéricos com a pontuação de dissimilaridade dos atributos categóricos.

Por fim, a partir de um conjunto de protótipos de cluster iniciais, esse processo atribui cada objeto a um cluster e atualiza o protótipo de cluster de acordo após cada atribuição.

Algoritmo: Árvores de decisão

Árvore de decisão é uma das formas de aprendizado de máquina mais bem sucedida e simples, como apontam [Russel and Norvig]. Esse algoritmo recebe um vetor de atributos de entrada e retorna um único valor de saída, que representa a "decisão". Nesse caso, a decisão da árvore é a classificação de a qual cluster os registros pertencem baseado em seus atributos de entrada. O algoritmo faz uma sequência de testes nos atributos de entrada para chegar em suas decisões. Esses testes são representados como nós internos da árvore, ou seja, que não são nós do tipo folha, que, como definem [Russel and Norvig], são aqueles que não têm filhos na árvore. Os testes consistem em uma comparação de um atributo de entrada com um dos possíveis valores dele e, para os dados em que a comparação é verdadeira, é criada uma ramificação que vai para outro nó teste ou nó folha. Para os dados em que a comparação não é satisfeita, é criada outra ramificação que também pode ir para um nó teste ou nó folha. Dessa forma, a cada teste feito os dados vão se separando até chegarem em seus nós folhas com suas decisões. O algoritmo da árvore de decisão pode ser visto com mais detalhes na figura 1.

Nesse exemplo, [Russel and Norvig] mostram uma árvore de decisão feita para decidir se vale a pena esperar por uma mesa em um restaurante ou não. Nesse caso, o vetor de atributos de entrada possui variáveis que indicam aspectos da situação de entrada como, por exemplo, se há ou não outra alternativa de restaurante por perto, se o dia da decisão é sexta ou sábado, se os indivíduos da decisão estão com fome ou não, se o restaurante é barato ou caro, se está cheio, vazio ou com algumas pessoas, entre outros

function DECISION-TREE-LEARNING($examples$, $attributes$, $parent\_examples$) **returns** a tree

if $examples$ is empty **then return** PLURALITY-VALUE($parent\_examples$)
else if all $examples$ have the same classification **then return** the classification
else if $attributes$ is empty **then return** PLURALITY-VALUE($examples$)
else
  $A \leftarrow \text{argmax}_{a \in attributes}$ IMPORTANCE(a, $examples$)
  $tree \leftarrow$ a new decision tree with root test A
  for each value v_k of A **do**
    $exs \leftarrow \{e : e \in examples$ **and** $e.A = v_k\}$
    $subtree \leftarrow$ DECISION-TREE-LEARNING(exs, $attributes - A$, $examples$)
    add a branch to $tree$ with label $(A = v_k)$ and subtree $subtree$
  return $tree$

Figure 1. Fonte: [Russel and Norvig]

aspectos. Os testes então são feitos com os valores possíveis desses atributos. Seu valor de saída é a decisão *VaiEsperar* e seus valores possíveis são verdadeiro e falso, portanto trata-se de uma decisão booleana. A Figura 2 apresenta uma visualização da árvore de decisão gerada para esse exemplo.

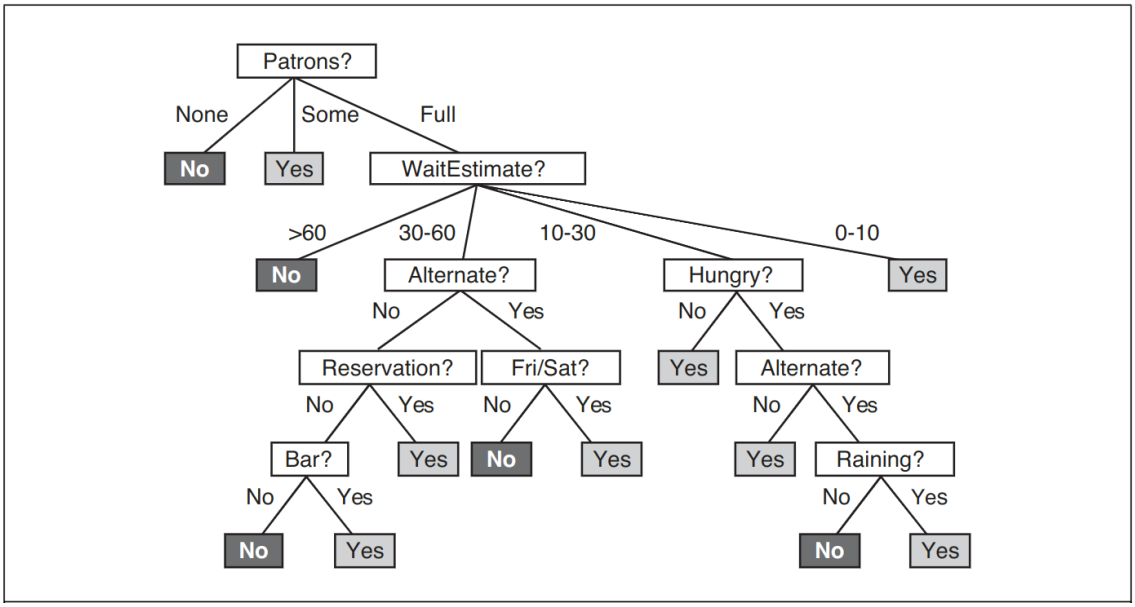

Figure 2. Fonte: [Russel and Norvig]

De acordo com [Singh and Giri 2014] em *[Singh and Giri 2014]*, três dos principais algoritmos de árvores de decisão são: *Classification and Regression Trees* (CART), C4.5 e *Iterative Dichotomizer* (ID3). As maiores diferenças entre esses algoritmos dizem respeito ao método usado para definir o melhor atributo para se fazer o teste, ou o atributo de maior importância, e o critério de parada do algoritmo. Neste trabalho, a implementação da árvore de decisão utilizada se baseia em uma versão otimizada do al-

goritmo CART[1] em que uma das diferenças se trata, por exemplo, da escolha do melhor atributo que utiliza a taxa de impureza de Gini. No algoritmo ID3, por sua vez, essa escolha é feita com o Ganho de Informação.

2.2. Fluxo de mineração de dados com foco em aprendizado de máquina

A Mineração de Dados (Data Mining - DM) estuda técnicas e ferramentas para coleta, tratamento, análise e extração de informação presente em dados de diversos tipos (textos, valores formatados, imagens etc) [Aggarwal 2015]. Ou seja, o termo mineração de dados implica uma série de diferentes aspectos do processamento de dados [Aggarwal 2015]. A Figura 3 apresenta os principais passos desse processo.

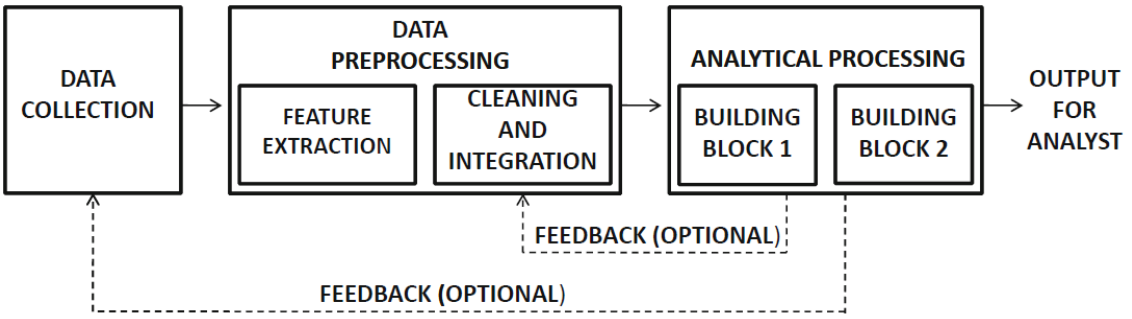

Figure 3. *Pipeline* de processamento de dados (Fonte: [Aggarwal 2015])

Para que o processo inicie é preciso definir quais serão os dados usados na mineração e quais suas fontes. A etapa de coleta trata de buscar e trazer esses dados e, geralmente, armazená-los em algum lugar para uso posterior. A coleta de dados pode ser feita de diversas maneiras, como com questionários respondidos, a partir de bancos de dados já existentes, com *scraping* (garimpagem) de dados de páginas da web, com dados coletados de sensores e com dispositivos que medem algum fenômeno. Nessa etapa é importante que as decisões de como a coleta será feita sejam alinhadas para favorecer o processo de mineração de dados.

Os dados coletados muitas vezes não vêm num formato amigável para o processamento analítico e o uso de algoritmos. Na etapa de limpeza, deve ser feita a transformação dos dados para um padrão mais adequado para a próxima etapa, de processamento. Às vezes o dado pode vir faltante ou errôneo e precisará ser corrigido ou estimado. É na etapa de limpeza e transformação que se aplicam as técnicas de enriquecimento, em que novas propriedades, ou *features*, são extraídas a partir dos dados existentes com engenharia de atributos. Um exemplo simples da necessidade dessa etapa é a seguinte situação: suponha que você coletou dados de indivíduos e a idade do indivíduo é relevante para sua análise. Porém, nos dados há apenas a data de nascimento, que também está em um padrão diferente do esperado. Portanto, é necessário transformar esse dado, para que fique adequado, e também extrair a nova propriedade 'idade' a partir dele. O resultado será um novo conjunto de dados limpo, possivelmente com novas features e pronto para ser usado na próxima etapa.

Em seguida, a etapa de análise permite que novos conhecimentos, novas relações, novas regras sobre os dados iniciais sejam descobertos. Essa análise pode envolver desde

[1]Cf. https://scikit-learn.org/stable/modules/tree.html#tree-algorithms

testes estatísticos, como ferramentas de visualização para gerar dados mais interpretáveis aos pesquisadores. Alguns algoritmos, e/ou suas implementações, são elaborados para aceitar apenas um tipo de dado de entrada, como categórico ou numérico. Em situações onde o dado de entrada não é compatível com o tipo exigido pelo algoritmo, pode ser possível fazer uma adaptação desse dado para o tipo adequado. Um exemplo disso é quando há um dado de entrada do tipo categórico, mas o algoritmo que usará eles aceita apenas numéricos. Nesse caso, uma das adaptações possíveis de se fazer é aplicar a técnica chamada *one hot encoding*. Essa técnica foi inicialmente proposta por Huffman (1954) e, a partir de então, foi amplamente aplicada em áreas diversas devido a sua simplicidade, o que a torna também umas das técnicas mais populares quando se trata de estratégia de projetar um algoritmo [Yu et al. 2020]. No caso específico do aprendizado de máquina, essa técnica é utilizada para processar recursos discretos [Yu et al. 2020] e uma de suas vantagens é o fato de ela tornar possível a binarização de inputs categóricos para que então sejam considerados como vetores do espaço euclidiano, o que é amplamente usado para calcular distâncias e/ou similaridades entre atributos em muitos algoritmos para classificação [Yu et al. 2020]. Colocado simplesmente, quando falamos de one hot encoding, deixamos implícito que todos os valores do mesmo atributo categórico estão igualmente distantes uns dos outros. Ou seja, com a técnica de one hot encoding, cria-se uma coluna nova para cada categoria possível do dado original, em que os valores dessas colunas são 0 ou 1, sendo 0 o valor que indica que o dado não é dessa categoria e 1, que o dado é dessa categoria. Dessa forma o algoritmo consegue usar esses dados em seu treinamento.

Uma das técnicas usadas para agregar mais informações aos dados coletados se chama Data Linkage, que é um método de mineração de dados. Esse método permite a combinação dos dados com fins de resultar em maior abundância de conhecimento relevante [Zheng et al. 2018]. Como apontam [Zheng et al. 2018], conteúdos relacionáveis podem ser utilizados para se chegar a observações abrangentes e confiáveis.

As técnicas discutidas acima são particularmente eficientes quando combinadas com algoritmos de Aprendizado de Máquina. De acordo com [Mitchell 1997], soluções baseadas em aprendizado de máquina aprendem a desenvolver uma tarefa através de dados, utilizando uma medida objetiva para medir seu desempenho.

2.3. Trabalhos relacionados

Suicídio : investigando as causas por meio da análise de dados? [Silva 2017]

Esse artigo mostra como técnicas de mineração de dados, aprendizado de máquina e análise de padrões podem auxiliar na identificação de risco de suicídio e diagnóstico de depressão no Brasil a partir de dados do número de suicídios agrupados por cidade, estado e região do Brasil, além de sexo biológico, população de jovens (de 19 a 25 anos), dados sobre a população, renda per capita e IDHM de Educação, Longevidade e Renda. Os autores analisam a relação entre a proporção de ocorrências pela população e o IDH de cada região do país, concluindo-se que existe uma relação inversamente proporcional desses dados. A análise dos dados coletados permitiu encontrar possíveis fatores que influenciam e direcionam pessoas a se suicidarem. Evidencia-se que com mais parâmetros nos dados seria possível encontrar mais desses fatores.

Predictors of suicide attempt in patients with obsessive-compulsive disorder: an exploratory study with machine learning analysis **[Agne et al. 2020]**

Esse artigo é o primeiro a usar aprendizado de máquina na busca por fatores de risco de tentativa de suicídio em pessoas com Transtorno obsessivo-compulsivo e no esclarecimento a respeito dos fatores encontrados, se estão mais relacionados ao transtorno ou a variáveis sociodemográficas ligadas à pessoa e outras comorbidades. A partir dos dados de 959 pacientes com TOC, conclui-se que 10.8% da amostra apresentava risco de tentativa de suicídio. Foram encontrados os seguintes fatores relevantes para predizer esse risco: precedente de planejamento de suicídio e de pensamentos suicidas, episódios depressivos durante a vida e desordem explosiva intermitente. Chega-se à conclusão de que é possível criar um algoritmo acurado para predizer o risco usando dados clínicos e sociodemográficos.

Prediction of attempted suicide in men and women with crack-cocaine use disorder in Brazil **[Roglio et al. 2020]**

Nesse artigo, reconhece-se a falta de estudos em busca de fatores preditores de risco de suicídio em indivíduos com transtorno por uso de substâncias. Para tanto, buscou-se investigar tais fatores em indivíduos com transtorno por uso de cocaína e/ou crack usando duas abordagens analíticas diferentes: descritiva e preditiva. A primeira usa regressão de Poisson com variância robusta, já a segunda, o algoritmo de aprendizado de máquina chamado Floresta Aleatória. Ambas as abordagens foram usadas de forma estratificada por gênero. Os resultados desse estudo indicam que a tentativa de suicídio está associada a depressão, alucinações e internações anteriores por questões mentais, tanto para homens como para mulheres.

QUEM SÃO OS ESTUDANTES DE MEDICINA QUE TENTAM SUICÍDIO? [Marcon 2019]

Essa dissertação apresenta que estudantes de medicina têm um maior risco de tentativa de suicídio comparados à população em geral. Buscou-se encontrar fatores associados entre essa população e a tentativa de suicídio para que se possam identificar e intervir mais precocemente. Para identificar tais fatores, usou-se a regressão de Poisson multivariada e um algoritmo de aprendizado de máquina, *Elastic Net Regularization*, para reconhecer os padrões dos alunos que tentam suicídio. Com os dados de 4840 estudantes de medicina, foram coletados dados relacionados à saúde mental e à universidade, ao estilo de vida e aos dados demográficos desses indivíduos. Nessa amostra houve uma prevalência de tentativa de suicídio de 8,94%. Conclui-se que seria possível implementar intervenções personalizadas ao se identificarem sujeitos sob maior risco de tentativa de suicídio por meio de algoritmos de risco.

Determinantes espaciais e socioeconômicos do suicídio no Brasil: uma abordagem regional [Gonçalves et al. 2011]

Esse artigo avalia a relação das taxas de suicídio das microrregiões do Brasil com os seus aspectos socioeconômicos, considerando também o aspecto espacial. Os dados coletados são referentes ao período de 1998 até 2002. Propõe-se a hipótese de ”efeito contágio” espacial, que pode ser confirmada nos resultados e na própria análise exploratória. Mostrou-se que existe autocorrelação espacial positiva. Foram analisados ainda os fatores de pobreza e grau de ruralização das microrregiões, encontrando-se uma relação inversamente proporcional entre o grau de pobreza e as taxas de suicídio; já o grau de ruralização mostrou uma relação diretamente proporcional às taxas de suicídio. Conclui-se que o suicídio não envolve apenas o indivíduo, mas toda a sociedade. Por isso, ressalta-se a importância da sensibilização da sociedade sobre o assunto e o investimento em políticas que garantam o cuidado da saúde mental e a promoção da qualidade de vida.

2.3.1. Principais achados

Os principais achados nos trabalhos listados acima foram: (i) a relação espacial do suicídio entre microrregiões [Gonçalves et al. 2011], reforçando a necessidade de se incluir dados sobre a residência do indivíduo; (ii) a relação com comorbidades mentais pré-existentes, como depressão, pode influenciar o risco de suicídio [Roglio et al. 2020]; (iii) por fim, a relação da taxa de suicídio com a renda do local de residência [Silva 2017].

3. Solução proposta

A solução proposta por este trabalho é um fluxo composto por três etapas: 1) Data Linkage; 2) Preparação; 3) Modelagem.

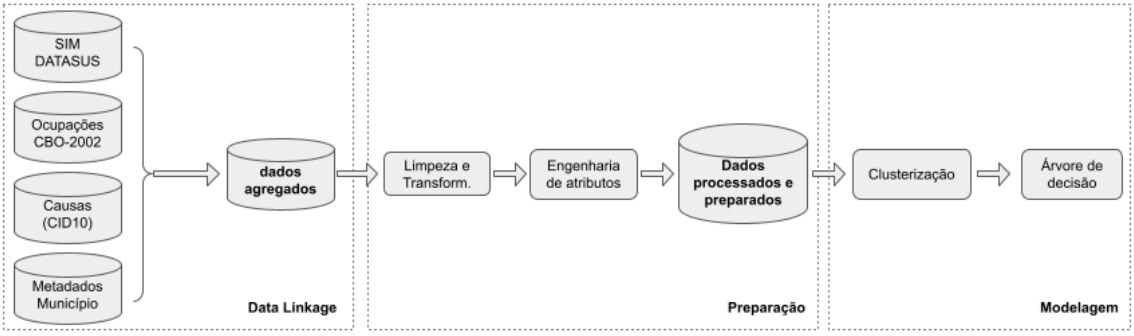

Figure 4. Fluxo da solução proposta. (Fonte: o autor)

Na etapa de Data Linkage, foi feita a coleta, preparação e transformação dos dados. Os dados coletados foram: dos municípios, das enfermidades da décima Classificação Internacional de Doenças (CID10) [2], dos nomes e código de ocupações da Classificação Brasileira de Ocupações (CBO-2002) [3] e principalmente do conjunto do Sistema de Informação sobre Mortalidade (SIM), fornecido pelo DATASUS e coletado

[2]Mais informações disponíveis em: https://icd.who.int/browse10/2019/

[3]Mais informações disponíveis em: https://empregabrasil.mte.gov.br/76/cbo/

por meio da biblioteca PySUS. Esses dados se referem a indivíduos que faleceram no Brasil. Os dados do SIM estão separados em arquivos por estado e ano do conjunto. Para facilitar o uso desses dados, criou-se um novo conjunto, contendo todos os dados baixados anteriormente, com a agregação de dados de mortalidade. O novo conjunto foi filtrado para os dados de 2010 até 2019, além de ter sido limitado para a região sul do Brasil, pois o tamanho do arquivo dificultava sua manipulação e exploração. Esse novo conjunto, com todos os dados de mortalidade filtrados, e os conjuntos dos outros dados coletados nessa etapa foram usados como base para as seguintes. Esses dados coletados são então conectados (linked) para gerar outro conjunto, denominado conjunto de dados agregados, o que faz com que um número maior de informações disponibilizadas sejam analisadas em conjunto, permitindo, com isso, uma complexificação das análises.

A partir disso, os dados precisam ser preparados. Faz-se necessária uma limpeza desses dados agregados, para eliminar atributos errôneos, faltantes e/ou com valores discrepantes, o que possibilita uma maior eficácia da análise. Junto da etapa de limpeza são feitas transformações dos dados para deixar seus significados mais claros, como a tradução de um atributo que está codificado para sua versão decodificada. Após isso, alguns dados que não foram transformados com sucesso precisam passar pelo processo de limpeza novamente. Depois da limpeza e da transformação, passa-se para a parte em que é feita a avaliação dos atributos existentes pela engenharia de atributos, que consiste na transformação de dados existentes em novos atributos que sejam relevantes para o desfecho e o treinamento do modelo de aprendizado de máquina aqui proposto.

Então, chega-se à modelagem, que consiste na aplicação dos modelos de aprendizado de máquina no conjunto de dados gerado nas etapas anteriores. Aplicam-se os dados em um algoritmo de agrupamento, cujo objetivo é agrupar casos semelhantes de suicídios em grupos homogêneos. Os resultados desse agrupamento são utilizados para identificar grupos com características similares, os quais podem servir como base para intervenções direcionadas. Depois, passa-se para um treinamento supervisionado, utilizando árvores de decisão, considerando como desfecho os próprios grupos gerados na etapa anterior. O objetivo desse experimento foi agregar semântica aos grupos, com auxílio dos modelos treinados, visto que árvores de decisão possuem uma estrutura bastante intuitiva e de fácil explicação. Para comparar a acurácia da árvore de decisão, os dados de entrada são separados em conjunto de treino e conjunto de teste, sendo que cada conjunto é separado entre atributos de entrada e resultado ou desfecho. Ao final dessa etapa, os resultados dos agrupamentos e a árvore de decisão gerada são comparados entre si, bem como com grupos de risco encontrados na literatura.

4. Análise dos resultados

Os dados agrupados possibilitam reunir uma série de informações que ajudam na criação de um perfil de grupos de risco. A análise dos gráficos dos atributos visa compreender melhor e tentar interpretar o perfil dos indivíduos de cada agrupamento, bem como a distribuição dos agrupamentos pelo mapa da região sul do Brasil para verificar se há uma relação espacial entre os agrupamentos. Para validar os resultados dos experimentos, é feita uma comparação das análises, com o objetivo de descobrir informações ulteriores dos perfis dos agrupamentos. Além disso, as análises são comparadas com grupos de risco/atenção encontrados na literatura.

As análises deste trabalho são validadas pelo que [Meneghel et al. 2004], no artigo intitulado [Meneghel et al. 2004], sugerem, isto é, que o método de disparo de armas de fogo e autoasfixiação é recorrente nos casos de suicídio no Brasil, sendo o primeiro equivalente a 62,5%, e o segundo, (21,5%). Há também a confirmação do que [Viana et al. 2008] apontam: o número de indivíduos do sexo masculino é mais alto que o de indivíduos do sexo feminino, e a preferência por métodos mais violentos coincide com o primeiro grupo. Os autores do artigo supracitado também concluem que, para todos os sexos, o método mais utilizado era o enforcamento, o que também pode ser percebido nas análises desta pesquisa.

No que diz respeito à ocupação, [Drebes et al. 2018] observam que as políticas públicas de modernização da agricultura acarretaram novos elementos para o problema social dos suicídios rurais. As análises dos agrupamentos evidenciam que a ocupação mais frequente nos dados de suicídio estava ligada à agricultura. Também, foram encontrados agrupamentos compostos por uma grande parcela de trabalhadores da agricultura que estão mais distribuídos nas regiões do interior de seus estados do que indivíduos dos demais grupos percebidos nos outros agrupamentos. Esse perfil foi traçado desde o experimento feito para encontrar 3 agrupamentos até o último experimento, para 5 agrupamentos. O que pode indicar que esses agrupamentos refletem o problema do suicídio em zonas rurais e/ou do suicídio de agricultores que foi destacado no estudo [Drebes et al. 2018].

5. Conclusão e trabalhos futuros

O fenômeno do suicídio é um problema que afeta países de todo o mundo e sua natureza possui diversas variáveis, que podem ser intrínsecas e/ou extrínsecas ao indivíduo. Estudos sobre esse problema, aliados a técnicas de mineração de dados e/ou aprendizado de máquina, têm obtido resultados relevantes e podem contribuir para a compreensão do fenômeno do suicídio, colaborando com diversas áreas de atuação, principalmente no campo da saúde.

Este trabalho propôs uma solução que aborda, juntamente, a mineração de dados e algoritmos de aprendizado de máquina para agrupamento de dados aliados a técnicas de visualização de dados e algoritmos de árvores de decisão para ampliar a capacidade de compreensão dos agrupamentos encontrados. Com isso, tinha-se por objetivo perceber padrões em casos de suicídio utilizando algoritmos de aprendizado de máquina, resultado que foi alcançado, considerando que diversos agrupamentos encontrados pelo modelo são semelhantes a grupos de indivíduos observados na literatura. O presente trabalho colabora para a pesquisa na área de ciências da computação e sua aplicação, no caso específico, de estudos da área da saúde, da psiquiatria e de políticas públicas em território nacional.

Para trabalhos futuros, uma das sugestões de melhorias é a agregação de novas variáveis relacionadas às vítimas e ao ambiente nos quais elas convivem, trazendo mais informações a respeito dos indivíduos, isso pode fazer com que o modelo de agrupamento de dados obtenha resultados com mais informações das características de cada grupo, ou seja, que ele se torne cada vez mais especializado. Algumas sugestões de melhorias mais especificamente a respeito de quais variáveis podem ser interessantes de se agregar e a hipótese de motivação se encontram abaixo.

- Variáveis sobre a ocupação
  A agregação de mais informações a respeito dos aspectos financeiros da ocupação

da vítima poderia trazer um novo ponto de vista a ser analisado, assim como é feito em relação ao grau de pobreza em [WHO 2017].

- Variáveis sobre a população dos municípios
  Ao se obterem mais informações das populações dos municípios, seria possível analisar uma hipótese de relação entre o perfil do indivíduo e o do município. Entretanto, seria necessário cuidado para se agregarem esses dados não só pelo município, mas também pelo ano em que o dado se encontra. Outra melhoria possível com esses dados seria poder calcular uma taxa de número de suicídios a cada 100 mil habitantes para se ter uma visão dos dados em relação ao tamanho de seus municípios.

- A aplicação desta solução proposta em outras regiões do País
  Os dados do SIM são nacionais e, portanto, não devem ter diferenças nos padrões entre as regiões. Sendo assim, seria possível adaptar a solução proposta para que fosse aplicada em dados de outras regiões do Brasil e até comparar os resultados das regiões, observando-se, com isso, detalhes pertinentes para cada estado e microrregião.
- Análise com foco em criação de políticas públicas
  A partir desta pesquisa e de outras similares, faz-se caminho para um trabalho futuro que possivelmente seja feito em parceria com profissionais da saúde mental, por exemplo, e de ciências de dados com foco em analisar e validar as informações obtidas do trabalho com foco de criação de políticas públicas de apoio à vida e prevenção do suicídio.

References

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer, Cham.

Agne, N. A., Tisott, C. G., Ballester, P., Passos, I. C., and Ferrão, Y. A. (2020). Predictors of suicide attempt in patients with obsessive-compulsive disorder: an exploratory study with machine learning analysis. *Psychological medicine*.

Ahmed, M., Seraj, R., and Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8).

Association, A. P., Association, A. P., et al. (2013). Diagnostic and statistical manual of mental disorders: Dsm-5. *United States*.

Cabitza, F. and Banfi, G. (2018). Machine learning in laboratory medicine: waiting for the flood? *Clinical Chemistry and Laboratory Medicine (CCLM)*, 56(4):516–524.

Chaturvedi, A., Foods, K., Green, P. E., and Carroll, J. D. (2001). K-modes clustering. *Journal of Classification*, 18:35–55.

Coelho, F. C. Data sources — pysus 0.1.13 documentation.

Coelho, F. C., Baron, B. C., de Castro Fonseca, G. M., Reck, P., and Palumbo, D. (2021). Alertadengue/pysus: Vaccine.

de Souza, P. M., de Autran, M. M. M., et al. (2019). Repositório datasus: organização e relevância dos dados abertos em saúde para a vigilância epidemiológica. *P2P E INOVAÇÃO*, 6:50–59.

Drebes, L. M., Oliveira, T., Bohner, L., Celestino, V., and Silveira, P. (2018). Legislação, política pública e suicídio: A influência do estado sobre vida e morte de agricultores familiares. *Desenvolvimento em Questão*, 16:285–321.

Gonçalves, L. R., Gonçalves, E., de Oliveira, L. B., de Oliveira, L. B., and de Oliveira, L. B. (2011). Determinantes espaciais e socioeconômicos do suicídio no brasil: Uma abordagem regional. *Nova Economia*, 21:281–316.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., and Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21(11):1–18.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research domain criteria (rdoc): toward a new classification framework for research on mental disorders.

Marcon, G. (2019). Quem sÃo os estudantes de medicina que tentam suicÍdio?

Meneghel, S. N., Victora, C. G., Faria, N. M. X., de Carvalho, L. A., and Falk, J. W. (2004). Características epidemiológicas do suicídio no rio grande do sul. *Revista de Saúde Pública*, 38:804–810.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill.

Roglio, V. S., Borges, E. N., Rabelo-Da-Ponte, F. D., Ornell, F., Scherer, J. N., Schuch, J. B., Passos, I. C., Sanvicente-Vieira, B., Grassi-Oliveira, R., von Diemen, L., Pechansky, F., and Kessler, F. H. P. (2020). Prediction of attempted suicide in men and women with crack-cocaine use disorder in brazil. *PLOS ONE*, 15:e0232242.

Russel, S. and Norvig, P. *Artificial Intelligence A Modern Approach Third Edition*.

Sehnem, S. B. and Palosqui, V. (2014). Características epidemiológicas do suicídio no estado de santa catarina. *Fractal: Revista de Psicologia*, 26:365–378.

Silva, R. P. d. (2017). Suicídio: investigando as causas por meio da análise de dados?

Singh, S. and Giri, M. (2014). Comparative study id3, cart and c4.5 decision tree algorithm: A survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, 3.

Van Heeringen, K. and Mann, J. (2014). The neurobiology of suicide. lancet psychiatry 1 (1), 63–72.

Viana, G. N., Zenkner, F. D. M., Sakae, T. M., and Escobar, B. T. (2008). Prevalência de suicídio no sul do brasil, 2001-2005. *Jornal Brasileiro de Psiquiatria*, 57:38–43.

WHO (2017). Other common mental disorders: global health estimates. *Geneva: World Health Organization*, pages 1–24.

Yu, L., Zhou, R., Chen, R., and Lai, K. K. (2020). Missing data preprocessing in credit classification: One-hot encoding or imputation? *https://doi.org/10.1080/1540496X.2020.1825935*, 58:472–482.

Zheng, X., Cai, Z., and Li, D. (2018). Data linkage in smart internet of things systems: A consideration from a privacy perspective. *IEEE Communications Magazine*, 56:55–61.