

**UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE SISTEMAS DE INFORMAÇÃO**

Cássio Castro Araujo

**FERRAMENTA WEB PARA ANÁLISES DOS ACIDENTES DE TRÂNSITO NAS
RODOVIAS DO RS COM BASE EM CAMPANHAS DE CONSCIENTIZAÇÃO**

**Santa Maria, RS
2019**

Cássio Castro Araujo

**FERRAMENTA WEB PARA ANÁLISES DOS ACIDENTES DE TRÂNSITO NAS RODOVIAS
DO RS COM BASE EM CAMPANHAS DE CONSCIENTIZAÇÃO**

Trabalho de Conclusão de Curso apresentado à
Coordenação do Curso de Sistemas de Infor-
mação da Universidade Federal de Santa Ma-
ria (UFSM, RS), como requisito parcial para a
obtenção do grau de **Bacharel em Sistemas de
Informação**.

Orientador: Prof. Dr. Joaquim Vinícius Carvalho Assunção

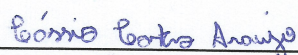
Santa Maria, RS
2019

**Universidade federal de Santa Maria
Centro de Tecnologia
Curso de Sistemas de Informação**

A comissão organizadora, abaixo assinada, aprova o Trabalho de Graduação

**Ferramenta web para análises dos acidentes de trânsito nas rodovias do RS
com base em campanhas de conscientização.**

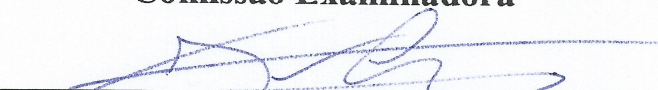
Elaborado por:

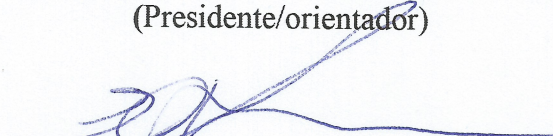


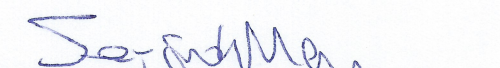
Cássio Castro Araujo

Como requisito parcial para obtenção do grau de
Bacharel em Sistemas de Informação

Comissão Examinadora


Prof. Dr. Joaquim Vinícius Carvalho Assunção
(Presidente/orientador)


Prof. Dr. João Carlos Damasceno Lima


Prof. Dr. Sérgio Luis Sardi Mergen

Santa Maria, 04 de Julho de 2019

DEDICATÓRIA

*[Não posso vê-los, não posso abraçá-los,
mas sei que estão presentes comigo aqui hoje.
Dedico essa conquista a vocês]*

AGRADECIMENTOS

- *Aos meus pais por permitirem que eu tivesse a chance de tentar novamente não criticando uma escolha errada, mas sim apoiando a cada nova decisão que eu tomava.*
- *À minha noiva Andressa por estar presente aguentando meu mal humor por cada código que não dava certo. Também por acreditar e incentivar minha capacidade para chegar até o fim.*
- *Ao meu irmão Rodolfo pela psicologia reversa. Mesmo que essa realmente não fosse a intenção, serviu de combustível para chegar além de onde eu acreditei que um dia chegaria.*
- *Ao orientador Joaquim, pelas sugestões, aprendizado e ajuda durante todo o desenvolvimento deste trabalho.*
- *Aos amigos e colegas de trabalho do Sistema Irriga pelas dicas e apoio.*

“Nossos sonhos são os mesmos há muito tempo, mas não há mais muito tempo para sonhar.”

Humberto Gessinger

RESUMO

FERRAMENTA WEB PARA ANÁLISES DOS ACIDENTES DE TRÂNSITO NAS RODOVIAS DO RS COM BASE EM CAMPANHAS DE CONSCIENTIZAÇÃO

AUTOR : Cássio Castro Araujo

ORIENTADOR : Prof. Dr. Joaquim Vinícius Carvalho Assunção

Os acidentes de trânsito estão presentes em todo nosso território nacional, diante disso, o presente trabalho tem por objetivo não só analisar os acidentes ocorridos nas estradas do Rio Grande do Sul nos anos de 2017 e 2018 como também demonstrar essas informações de maneira visual e interativa para os interessados. Informações mais concretas e uma melhor visualização de dados, importantes para os condutores, tornam-se essenciais nos dias de hoje. Tendo em vista os avanços tecnológicos, alinhados com um crescimento na área de mineração de dados, mais precisamente em KDD (Knowledge Discovery in Databases), é possível apresentar árvores de decisão com resultados que utilizem critérios de idade, dia de semana e mês para definir previsões de probabilidade das pessoas saírem ilesas em acidentes de trânsito, além de apresentar outras informações utilizando como base campanhas de conscientização conhecidas, de forma que se forneçam informações mais concretas e compreensíveis além de apresentar resultados que não sejam visíveis através de relatórios quantitativos de acidentes de trânsito auxiliando na conscientização e servindo como guia para os condutores.

Palavras-chave: Visualização de dados, mineração de dados, KDD, Acidentes de trânsito, Árvore de Decisão.

ABSTRACT

WEB TOOL FOR ANALYSIS OF ROAD TRAFFIC ACCIDENTS IN RS BASED ON AWARENESS CAMPAIGNS

AUTHOR : Cássio Castro Araujo

ADVISOR: Prof. Dr. Joaquim Vinícius Carvalho Assunção

The purpose of this paper is not only to analyze the accidents occurred on the roads of Rio Grande do Sul in the years 2017 and 2018, but also to demonstrate this information in a visual and interactive way for those interested. More accurate information and better data visualization, important to drivers, become essential these days. In light of the technological advances, in line with data mining, more precisely in Knowledge Discovery in Databases, it is possible to present decision trees with results that use criteria of age, weekday and month to define predictions of people's likelihood of being left uninjured in traffic accidents, and to present other information using known awareness campaigns to provide more concrete and understandable information and to provide results that are not visible through quantitative accident reports helping to raise awareness and serving as a guide for drivers.

Keywords: Data visualization, data mining, KDD, Traffic Accidents, Decision Tree.

LISTA DE FIGURAS

FIGURA 1	MODELO KDD	15
FIGURA 2	EXEMPLO DE DATAWAREHOUSE	16
FIGURA 3	EXEMPLO DE ÁRVORE DE DECISÃO	21
FIGURA 4	FLORESTA ALEATÓRIA - EXEMPLIFICAÇÃO	22
FIGURA 5	MAPA SEM PROVIDER	24
FIGURA 6	PROVIDER DE RODOVIAS	24
FIGURA 7	EXEMPLO DE GRÁFICO DE BARRAS	25
FIGURA 8	MAPA DA CRIMINALIDADE	26
FIGURA 9	MAPA - CRIMES EM BAGÉ	27
FIGURA 10	ACIDENTES - TIPO DE VEÍCULO	27
FIGURA 11	LOGO DA CAMPANHA - SE BEBER NÃO DIRIJA	29
FIGURA 12	LOGO DA CAMPANHA - MAIO AMARELO	30
FIGURA 13	LOGO DA CAMPANHA - VERÃO SEGURO	31
FIGURA 14	ANTES DA SELEÇÃO DOS DADOS - 2018	32
FIGURA 15	DEPOIS DA SELEÇÃO DOS DADOS - 2018	32
FIGURA 16	COMPARATIVO - ADENDO LEI SECA	36
FIGURA 17	ÁRVORE DE DECISÃO - PROBABILIDADE DE ILESOS - SE BEBER NÃO DIRIJA	38
FIGURA 18	ESTRUTURA INICIAL	40
FIGURA 19	TEXTOS INFORMATIVOS - ACIDENTES COM CORES DIFERENTES POR SEXO DO CONDUTOR	41
FIGURA 20	CAUSAS DOS ACIDENTES - MAIO AMARELO	41
FIGURA 21	CULPADOS DOS ACIDENTES - INGESTÃO DE ÁLCOOL - SE BEBER NÃO DIRIJA	42
FIGURA 22	COMPARAÇÃO - ACIDENTES POR CAUSA	42
FIGURA 23	COMPARAÇÃO - ACIDENTES POR DIA SEMANAL - PARTE 1	43
FIGURA 24	COMPARAÇÃO - ACIDENTES POR DIA SEMANAL - PARTE 2	43
FIGURA 25	COMPARAÇÃO - ACIDENTES POR DIA SEMANAL - PARTE 3	43

LISTA DE QUADROS

QUADRO 1	CAMPANHAS DE TRÂNSITO E SEUS RESPECTIVOS FILTROS	33
QUADRO 2	RESULTADOS DAS ANÁLISES DE PREDIÇÃO - SAIR ILESO DE ACI- DENTES	39

LISTA DE ABREVIATURAS E SIGLAS

CRAN	Comprehensive R Archive Network
CRBM	Comando Rodoviário da Brigada Militar
DAER	Departamento Autônomo de Estradas de Rodagem
DETRAN	Departamento de Trânsito
DPVAT	Seguro de Danos Pessoais Causados por Veículos Automotores de Vias Terrestres
EGR	Empresa Gaúcha de Rodovias
ERP	Enterprise Resource Planning
KDD	Knowledge Discovery in Databases
MIT	Massachusetts Institute of Technology
OMS	Organização Mundial de Saúde
ONU	Organização das Nações Unidas
PDF	Portable Document Format
PRF	Polícia Rodoviária Federal
RPMON	Regimento de Polícia Montada
RS	Rio Grande do Sul
SQL	Structured Query Language

SUMÁRIO

1	INTRODUÇÃO	12
2	FUNDAMENTAÇÃO TEÓRICA	15
2.1	SELEÇÃO	15
2.1.1	DADOS TABULARES	15
2.1.2	BANCO DE DADOS	16
2.1.3	DATA WAREHOUSES	16
2.2	PRÉ-PROCESSAMENTO E LIMPEZA	16
2.2.1	DADOS AUSENTES	17
2.2.2	DADOS INCONSISTENTES	17
2.3	TRANSFORMAÇÃO	17
2.4	MINERAÇÃO DE DADOS	18
2.4.1	MACHINE LEARNING	18
2.4.1.1	Naive Bayes	20
2.4.1.2	Árvore de Decisão	20
2.4.1.3	Random Forest	21
2.5	INTERPRETAÇÃO	22
2.6	FERRAMENTAS	22
2.6.1	LINGUAGEM R	23
2.6.2	SHINY	23
2.6.2.1	Mapas	23
2.6.2.2	Gráficos	25
3	TRABALHOS CORRELATOS	26
4	DESENVOLVIMENTO	29
4.1	ESCOLHA DE CAMPANHAS DE TRÂNSITO ALVO	29
4.1.1	SE BEBER NÃO DIRIJA	29
4.1.2	MAIO AMARELO	30
4.1.3	ÉPOCAS DE VERANEIO	30
4.2	TÉCNICA DE KDD	31
4.2.1	SELEÇÃO	31
4.2.2	PRÉ-PROCESSAMENTO E LIMPEZA	32
4.2.3	TRANSFORMAÇÃO	33
4.2.4	MINERAÇÃO DE DADOS	33
4.2.5	INTERPRETAÇÃO	34
5	RESULTADOS	35
5.1	ESCOLHA DO ALGORITMO	36
5.1.1	ÁRVORE DE DECISÃO	36
6	CONCLUSÃO	40
6.1	MAPAS	40
6.2	GRÁFICOS	41
6.3	TRABALHOS FUTUROS	44

1 INTRODUÇÃO

Um dos problemas que ocorrem não somente em nosso país, como de âmbito global, são os acidentes de trânsito. Diariamente, surgem notícias trágicas que instigam o ser humano a querer saber mais sobre como ocorreu a tragédia. Seja por questões de imprudência, falha mecânica ou entre outros acontecimentos possíveis; essas fatalidades estão, e provavelmente estarão, presentes no nosso cotidiano.

Analisando o contexto por completo, em 2018 o Brasil estava na quinta colocação do ranking de trânsito entre os mais violentos, com base nas informações contidas no texto do [Senado Federal \(2019\)](#). Paralelo a isso, segundo [Luis Alberto Melchert \(2019\)](#), no ano de 2017 nosso país possuía a décima sexta maior frota de veículos por habitante. Diante dessas informações, várias campanhas de conscientização ao trânsito são lançadas tentando minimizar o problema de estarmos entre os trânsitos mais violentos do mundo.

Diante desse fato, são apresentadas algumas campanhas de conscientização ao motorista, que procuram reduzir esses números. Entretanto, não são suficientes para uma drástica redução desse problema, devido a abordagem ser bastante seletiva. Campanhas do tipo “se beber não dirija” ou também as que aparecem nas épocas de veraneio, do tipo, “se for viajar, faça a revisão do seu veículo”, apesar de serem extremamente focadas em reduzir o número de acidentes de trânsito; ou ocorrem nas épocas de veraneio, apenas, ou com foco aos motoristas que costumam ingerir bebidas alcoólicas. O ideal seria ter uma abrangência maior, se possível que atingisse a todos os condutores e não apenas a uma parcela deles.

Como existe uma quantidade significativa de acidentes e um alto volume de informações, inclusive restringindo apenas às rodovias federais do estado do Rio Grande do Sul, é útil obter dados mais visuais sobre essas tragédias. Conforme os dados vão se acumulando e por consequência, tendem a ter um aumento de informações implícitas, algumas destas podem ser interessantes para análise e utilização. Como característica de grandes conjuntos de dados, cada vez mais comuns, é uma tarefa relativamente difícil a visualização e obtenção de informação útil e não explícita.

Esse trabalho visa descobrir informações interessantes e previamente desconhecidas sobre causas e correlações entre acidentes de trânsito para três das campanhas de conscientização mais conhecidas fazendo uso dos dados abertos fornecidos pela PRF e utilizando apenas as informações sobre os acidentes do estado do Rio Grande do Sul. Além da comparação com o conjunto de dados dos acidentes nas rodovias do estado do Rio Grande do Sul como um todo.

Pretende-se ainda comprovar se as campanhas de conscientização apresentam resultados significativos e apresentar predições sobre a chance de sair ileso de um acidente com base nas variáveis disponíveis pelo conjunto de dados. Além de trazer todas essas informações, juntamente com o resultado de análises dos acidentes ocorridos nas rodovias do RS em um software web que foi desenvolvido com uma interface simples e intuitiva, deseja-se traçar um perfil dos acidentes para tipos de condutores, conforme resultados fornecidos por árvores de

decisão.

A ideia de desenvolver um software tornou-se necessária devido a esses dados serem fornecidos em forma de planilha(.csv), isto é, com valores separados por "," ou ";", diante disso, com o auxílio de um sistema web, é possível ter-se uma melhor visualização dessas informações. Salienta-se ainda que o volume de dados impede que um usuário tenha uma visão clara a respeito de potenciais perigos e condições correlacionadas com acidentes.

De acordo com dados da OMS(Organização Mundial de Saúde), o Brasil está em quinto lugar entre os países recordistas em acidentes de trânsito. Além disso, o DPVAT(Seguro de Danos Pessoais Causados por Veículos Automotores de Via Terrestre), pagou em 2015, 42.501 indenizações por morte no país, segundo boletim informativo do [Vias Seguras \(2019\)](#), um dado alarmante que demonstra a necessidade de mudanças visando uma redução desse problema. Salienta-se ainda que, de acordo com dados de 2016, a OMS informou, segundo [Lara Montezuma \(2019\)](#), duas situações importantes e preocupantes, as quais são:

- Cerca de 1,25 milhão de pessoas morrem a cada ano em decorrência de acidentes no trânsito;
- Caso medidas não sejam aplicadas, os acidentes de trânsito devem se tornar a 7ª principal causa de morte em 2030.

Para desenvolvimento desse trabalho, foi necessário a utilização dos dados abertos, fornecidos pela PRF, e a manipulação desses dados. Através de informações contidas nos dados, foi possível implicar a qual(is) campanha(s) o acidente específico se encaixa e também analisar o perfil do condutor de acordo com algumas características contidas no conjunto de dados; como, por exemplo, sexo, idade e tipo de veículo.

O presente trabalho estrutura-se em cinco capítulos, apresentando-se no primeiro informações gerais sobre as fundamentações teóricas, isto é, informações sobre o que foi seguido como base para o trabalho, como por exemplo: Explicações sobre o processo de KDD, proposto por Usama Fayyad em 1996, onde apresenta-se cada passo do processo individualmente, incluindo tipos de aprendizado de *machine learning*, recurso utilizado na etapa 4 do processo de KDD, além das ferramentas utilizadas para o desenvolvimento do trabalho.

No segundo capítulo, apresentam-se informações sobre os trabalhos correlatos. Lá são apresentados trabalhos com ideias e desenvolvimentos próximas a esse.

No terceiro capítulo são citadas as etapas de desenvolvimento. Explicando cada passo que foi seguido até o produto final gerado além de definir as campanhas de trânsito que serão analisadas no contexto dos acidentes do estado do Rio Grande do Sul e .

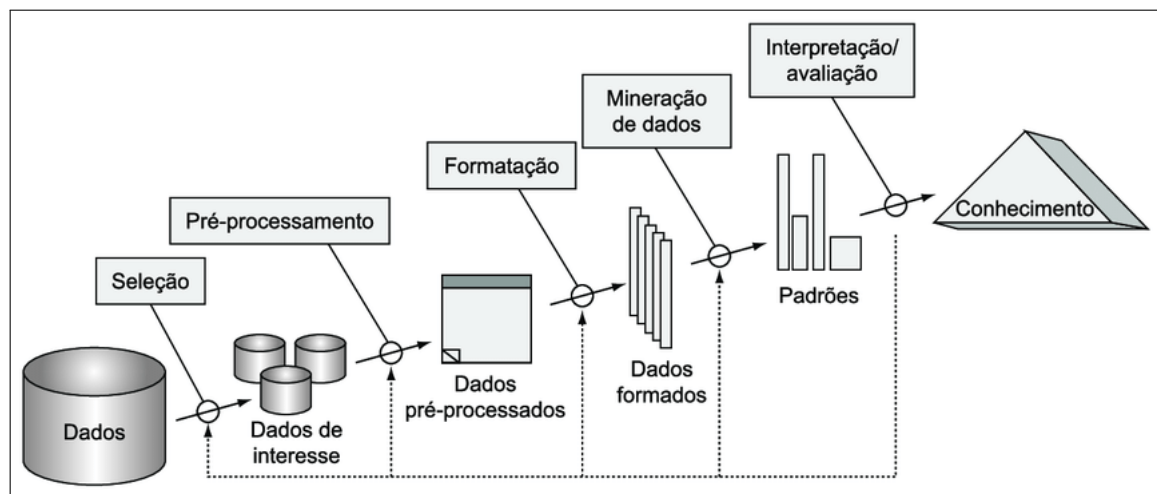
No quarto capítulo são apresentados os resultados do trabalho, tais como o que obtivemos com a utilização da ferramenta, isto é, o que a análise dos dados nos mostra sobre os acidentes ocorridos nas rodovias do RS. Além de informações sobre causadores de acidentes de acordo com critérios utilizados na análise.

No quinto e último capítulo é apresentada a conclusão geral sobre este trabalho além de informações baseadas em dados gerados através da análise feita e sugestões de trabalhos futuros utilizando como base a ferramenta *web*.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será relatada a forma como foi elaborada a base deste trabalho. O ponto inicial foi baseado na sequência dos processos de KDD(Knowledge Discovery in Databases) definidos assim por Usama Fayyad, conforme Figura 1, a qual segue os seguintes passos: seleção, pré-processamento e limpeza, transformação ou formatação, mineração de dados e interpretação ou avaliação.

Figura 1: Modelo KDD



Fonte: [Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth \(1996\)](#) (Versão traduzida)

2.1 SELEÇÃO

Na etapa de seleção, como o próprio nome já informa, é feita a escolha dos dados. Nessa etapa, a busca dos dados é feita e, normalmente, várias fontes que possuam os dados necessários são pesquisadas e analisadas para a escolha. Essa, normalmente, é baseada no formato dos dados, os quais podem se encontrar em fontes diversas, alguns exemplos são:

2.1.1 Dados Tabulares

Esse tipo de dado, o qual possibilitou ser automatizado em 1978 por Dan Bricklin, criador das planilhas eletrônicas, segundo [Redação, do Na Prática \(2017\)](#), é uma das formas mais comuns de se encontrar ou guardar informações computacionalmente. Devido a essa simplicidade, ser fácil de manipular e por ser um dos tipos de dados com formato aceito na lei dos dados abertos, foi o tipo escolhido para ser o fornecedor de informações desse trabalho.

2.1.2 Banco de Dados

Citamos duas formas de armazenamento desse tipo, a primeira delas, a qual chamamos de banco de dados relacional foi idealizada no início da década de 70, através de um artigo publicado por [Codd, E. F. \(1970\)](#). Esse tipo de banco é baseado na linguagem SQL e permite relacionamentos entre tabelas.

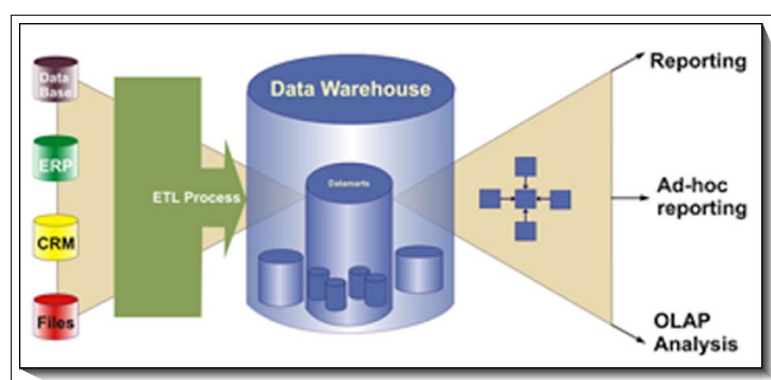
Já a segunda forma de armazenamento, conhecida como *NoSQL* é definida como banco de dados não-relacional e surgiu com a necessidade de se ter uma performance melhor que os bancos relacionais. Atualmente subdivide-se baseado em 4 tipos: documento, coluna, grafos e chave-valor.

2.1.3 Data Warehouses

Em [Turban, E. and Sharda, R. and Aronson, J.E. and King, D. \(2009\)](#), os autores definem esse tipo de armazenamento como um conjunto de dados produzido para oferecer suporte a tomada de decisão. Por se tratar de um enorme banco de dados, esse recurso possui várias fontes de informações da empresa. Comumente está presente nos sistemas ERP, o qual engloba informações de todos os setores e fábricas existentes na organização.

Em outras palavras, o data warehouse seria um grande depósito de dados que contem informações vindas de diversas fontes, inclusive de dados tabulares e/ou banco de dados conforme figura 2, a qual exemplifica fontes possíveis de dados e a estrutura de um *data warehouse*.

Figura 2: Exemplo de DataWarehouse



Fonte: [Wayne Yaddow \(2013\)](#)

2.2 PRÉ-PROCESSAMENTO E LIMPEZA

Nessa etapa de pré-processamento é feito o tratamento dos dados existentes nas colunas, isto é, pode ocorrer a existência de dados redundantes, faltantes ou inconsistentes. Dessa forma,

torna-se necessária uma limpeza desses dados.

2.2.1 Dados Ausentes

Devido ao conjunto de dados ser enorme, esse pode apresentar dados faltantes que poderiam ser significativos para a análise. Se esses dados faltantes forem uma parcela insignificante no conjunto de dados, podem simplesmente serem eliminados da análise juntamente com o resto da linha onde aparecem.

Caso o conjunto de dados não seja muito significativo e caso exista a possibilidade de prever esses dados ausentes, é possível completá-los manualmente e prosseguir com a análise.

Além das duas técnicas já citadas, existe a possibilidade de utilização da média aritmética da variável. Essa técnica é muito utilizada quando lidamos com temperaturas, onde, normalmente, a oscilação dos dados tende a ser baixa.

2.2.2 Dados Inconsistentes

Se ao analisar o todo e notar que existe um dado inconsistente, isto é, dados que são extremamente diferentes dos demais do conjunto. Esse tipo de dado acaba sendo percebido facilmente quando utiliza-se uma análise gráfica. É importante verificar o conjunto todo e estudar se esse dado, apesar de atípico não é verídico. Caso seja realmente um dado inconsistente, deve-se, na maioria dos casos, eliminar o registro onde se encontra essa informação avaliada como errônea.

Existem outras opções que possam ser utilizadas, tais como calcular a mediana ou média dos demais dados e substituir por esse dado que está errado. Além do mais, também pode-se, em determinados casos, optar pela transcrição do valor do vizinho mais próximo, isto é, supondo que esse valor seja semelhante e copiarmos para o local do dado inconsistente. De forma geral, a substituição de dados inconsistentes segue as mesmas técnicas de preenchimento dos dados ausentes.

2.3 TRANSFORMAÇÃO

Após finalizar as etapas anteriores, deve-se armazenar e formatar esses dados de forma que possam ser manipulados pelos algoritmos. Esses podem ser de classificação, predição, associação, entre outros.

2.4 MINERAÇÃO DE DADOS

Nessa etapa, a qual é definida como ponto chave do processo KDD, é onde procura-se padrões e regras existentes no conjunto de dados. É o momento onde tenta-se agrupar informações e apresentar correlações não visíveis sem uma análise mais sucinta.

Através do conjunto de dados como um todo, são procurados padrões existentes. De forma mais clara, nesse momento o foco da análise é definido. Por exemplo, se o interesse é de elaborar um relatório administrativo para a gerência, o foco deve ser os pontos específicos que mostrem dados significantes para o pessoal do administrativo, ou seja, pontos onde temos gastos significativos e que estejam causando problemas ou trazendo lucros. Não é necessário trazer informações da concorrência nesse caso, embora em alguns outros casos seja interessante, como questões de marketing, por exemplo.

Se analisar duas das maiores companhias de refrigerante do mundo, é importante para a segunda colocada saber informações da líder e vice-versa. Mas para empresas pequenas, não é tão interessante assim. Dentro do processo de mineração de dados é possível utilizar algoritmos de Machine Learning.

2.4.1 Machine Learning

"Machine Learning: Campo de estudo que dá aos computadores a capacidade de aprender sem estarem programados explicitamente." [Arthur L. Samuel, 1959]

Essa foi a definição apresentada por Arthur Samuel, engenheiro da MIT (*Massachusetts Institute of Technology*) e pai do que viria a ser o primeiro programa de auto-aprendizagem do mundo, baseado no popular jogo de damas. O termo *machine learning* está presente em vários projetos, seja de reconhecimento de fala ou imagem, dentre outros. Alguns são citados por [Mybridge \(2018\)](#).

Tipos de Aprendizado

No campo de *machine learning*, são conhecidos, pelo menos, quatro métodos diferentes de aprendizagem. Esses apresentam formas de utilização variadas e cada um tem a sua melhor utilidade. Os mais comuns são: aprendizado não supervisionado, aprendizado por reforço, aprendizado semi supervisionado e aprendizado supervisionado,.

Aprendizado Não Supervisionado

Esse tipo de aprendizado é mais utilizado quando não se tem informações sobre entradas e saídas específicas. Por exemplo, o proprietário de um comércio gostaria de saber o perfil de seus clientes. Não há dados de entrada conhecidos para se ter um padrão. Um exemplo clássico desse tipo de aprendizado é citado por [Helio Gurovitz \(2011\)](#), onde foi percebida uma relação entre a compra de fraldas descartáveis e cerveja. Diante disso, os produtos foram colocados lado a lado e a venda de ambos ficou elevada.

Aprendizado Por Reforço

Segundo [Haykin \(2007\)](#), "O aprendizado de um mapeamento de entrada-saída é realizado através da interação contínua com o ambiente. E a aprendizagem por reforço atrasado é difícil de ser realizada. Dentre uma das causas é o fato de que não existe um professor para fornecer uma resposta desejada em cada passo do processo de aprendizagem."

Esse tipo de aprendizado também foi testado no ramo da robótica conforme exemplificado por [Gedson Faria, Roseli Francelin Romero \(1999\)](#).

Aprendizado Semi Supervisionado

O aprendizado semi supervisionado fica entre o aprendizado supervisionado e o não supervisionado. Normalmente, possui-se dois conjuntos de dados, um com dados de entrada e saída conhecidos, chamados rotulados e outro com dados de saída desconhecidos, ditos não rotulados.

Segundo [Hui LI \(2017\)](#), "Com o aprendizado semisupervisionado, você usa exemplos não rotulados com uma pequena quantidade de dados rotulados para melhorar a precisão do aprendizado."

"Um bom exemplo é um software capaz de identificar o rosto de uma pessoa em uma webcam.", segundo [MRG Tecnologia \(2019\)](#).

Aprendizado Supervisionado

No método de aprendizado supervisionado, a máquina aprende pelo exemplo, isto é, através dos dados fornecidos, de entradas e saídas já conhecidas, a máquina deve identificar padrões nos dados, prever algumas informações e seguirá nesse loop até que seja encontrado um bom método e que a predição está de acordo com o esperado ou pelo menos tenha algum resultado significativo para ser apresentado.

Devido ao presente trabalho possuir um conjunto de dados com entradas e saídas conhecidas, esse foi o tipo de aprendizado selecionado. Nesse tipo de aprendizado, alguns algoritmos são bastante conhecidos e apresentam bons resultados em diversos aspectos.

2.4.1.1 Naive Bayes

O algoritmo de *Naive Bayes* foi criado baseado no teorema de Thomas Bayes, o qual foi pioneiro na utilização de probabilidade indutivamente além de inferir situações, com base em eventos já ocorridos, a probabilidade de ocorrência em eventos futuros.

Estudos indicam que os algoritmos simples de classificação bayesiana, conhecidos como *Naive Bayes*, possuem um bom desempenho, se comparados a redes neurais artificiais e a árvores de decisão para determinados problemas, como por exemplo, para diagnósticos médicos, conforme artigo de Šter, Branko and Dobnikar, Andrej (1996).

Diferentemente do algoritmo árvore de decisão, citado em 2.4.1.2, o Naive Bayes não costuma gerar imagens ilustrando caminhos ou mostrando cada passo que foi gerado. Ele apresenta informações baseada na matriz de confusão, onde é mostrado valores reais na primeira coluna e os valores previstos pelo algoritmo na segunda coluna.

Para a utilização do algoritmo de Naive Bayes, é indicado a separação da amostra total de dados, a ser analisada, em dois conjuntos, normalmente chamados de treino e teste. Os conjuntos de treino e teste são divididos nas proporções que o usuário preferir, entretanto a medida depende da quantidade de dados e da regra de negócio do usuário. Comumente é utilizada a proporção de 70% para 30%, respectivamente.

Além da predição de diagnósticos médicos, outra possibilidade de aplicação desse algoritmo, citada por Pedro César Tebaldi Gomes (2019), é para classificação de SPAM e identificação de assuntos em *emails* com base em seu conteúdo.

2.4.1.2 Árvore de Decisão

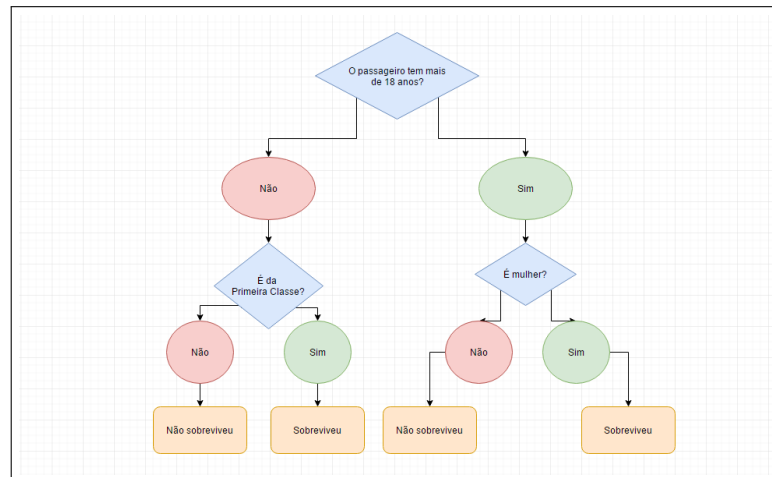
Uma árvore de decisão se apresenta como uma estrutura em formato de árvore, onde seus nós internos correspondem a um teste de atributo, cada ramo dessa árvore seria um resultado do teste e os nós folhas representam as classes utilizadas.

Essas árvores são de fácil entendimento, pois basta seguir os caminhos gerados que tem-se uma facilidade para explicar as classificações propostas, bastando apenas seguir a rota gerada pela árvore. Assim chega-se a classificação que ela gerou para aquele objeto.

Pode-se utilizar como exemplo de fácil entendimento e compreensão, o reino animal. Supondo que uma nova espécie tenha sido encontrada e deseja-se classificá-la, a árvore de decisão é, se não a melhor, uma das melhores opções de classificação. É extremamente fácil para um ser humano analisar que tipo de animal seria, dependendo de suas características, mas do ponto de vista de uma máquina, torna-se necessário um tipo de treinamento. Com isso, são fornecidos dados de amostra para o estudo da máquina. E, posteriormente, efetuam-se testes para que a máquina descubra a espécie, ou, no caso, a espécie mais próxima que o animal analisado se enquadra.

Um exemplo de árvore de decisão foi apresentado em uma competição elaborada pelo *Kaggle*, uma plataforma de competição de aprendizado de máquina. A plataforma forneceu os dados de treino e teste de sobreviventes no navio Titanic e através desses dados era necessário estruturar uma árvore de decisão e prever quais passageiros sobreviveriam à tragédia.

Figura 3: Exemplo de Árvore de Decisão



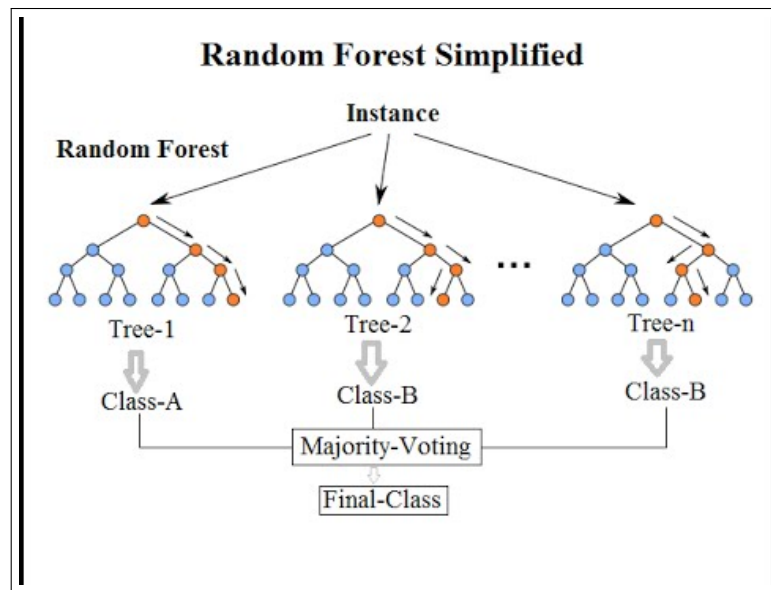
Fonte: [Paulo Vasconcellos \(2017\)](#)

Na ilustração, é possível observar que afirmações se a pessoa sobreviverá ou não conforme os questionamentos. O nó raiz, o qual foi questionado se a pessoa tem mais ou menos que 18 anos, possui dois nós filhos. De um lado se a resposta afirmativa para a pergunta do nó raiz e do outro a resposta negativa. Através dessas informações, são gerados outros nós com perguntas até se chegar a conclusão sobre o que ocorreu com a classe da pessoa, apresentada nos nós folha e que era o objetivo da competição, definir se a pessoa sobreviveria ou não e em quais condições.

2.4.1.3 Random Forest

Conhecida como um conjunto de árvores de decisão, esse algoritmo apesar de ser mais demorado que o citado anteriormente, tende a trazer resultados melhores, na maioria das vezes. O *random forest*, segundo [Tan, P.N. and Steinbach, M. and Kumar, V. and FERNANDES, A. \(2009\)](#) combina previsões feitas por múltiplas árvores de decisão e cada uma dessas utiliza um vetor aleatório que é gerado a partir de alguma distribuição de probabilidade fixa.

Figura 4: Floresta Aleatória - Exemplificação



Fonte: [Cristiano Furquim \(2018\)](#)

Na imagem acima, é explanado o funcionamento do algoritmo *random forest*. São analisadas instâncias diferentes e geradas "n" árvores de decisão. Após isso, é feita uma contagem de ocorrências, exemplificadas na imagem pelas *Class A* e *B*. A classe que possuir a maior contagem de ocorrências será a classe aceita para a definição. Isso reduz a chance de erro pois analisa diversos caminhos e instâncias.

2.5 INTERPRETAÇÃO

Aqui é verificado se o algoritmo trouxe algum resultado significativo. Normalmente essa etapa é feita em conjunto com especialistas na área para qual foi feita a análise. Dependente da resposta do expert, pode-se tentar a troca do algoritmo ou fazer a utilização de outras variáveis significativas para trazer um melhor resultado naquilo que foi solicitado.

2.6 FERRAMENTAS

Nesse tópico, serão abordadas algumas ferramentas utilizadas durante processo de desenvolvimento.

2.6.1 Linguagem R

Uma linguagem focada mais para a estatística mas que também apresenta um crescimento na utilização da tecnologia *machine learning* foi criada na Nova Zelândia por Ross Ihaka e Robert Gentleman.

A escolha pela ferramenta R foi caracterizada pela possibilidade de utilização dos algoritmos de mineração de dados e por ser simples e intuitiva. Além disso, ela possui uma IDE chamada de RStudio, a qual é muito interativa e auxilia nas fórmulas e algoritmos existentes. Salienta-se ainda que essa ferramenta possui um menu com uma aba *helper*, que fornece exemplos de utilização das fórmulas, dessa forma as dúvidas são sanadas mais rapidamente e torna-se dispensável dedicar tanto tempo procurando informações através das ferramentas de pesquisa na internet.

A linguagem R possui também uma biblioteca própria, chamada de *CRAN*, esse recurso possibilita a instalação de pacotes existentes no repositório da linguagem. Assim, as chances de ocorrerem problemas na busca das bibliotecas que continham os algoritmos de classificação utilizados nesse trabalho são reduzidas.

2.6.2 Shiny

Um pacote disponível na linguagem R que pode-se definir como um servidor que faz a interação da linguagem R com ferramentas web, sejam folhas de estilo ou apenas códigos em html.

Através da criação de um servidor local, é feita a integração do código e possibilita um melhor visual para análise dos resultados. Quando cria-se o servidor, a estrutura apresentada é de três arquivos, sendo eles um de interface ao usuário, normalmente chamado de UI.R, o arquivo de controller, server.R, o arquivo para chamarmos todas as bibliotecas e funções que serão executadas chamado de global.R. Salienta-se ainda que é possível gerar um único arquivo chamado de app.R, entretanto, por questões de semântica e separação de código, a primeira estrutura é a mais indicada.

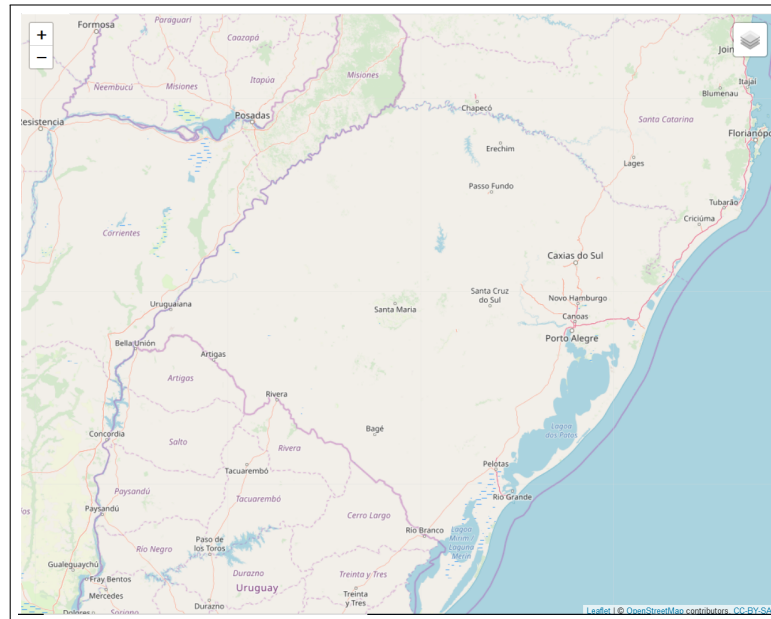
2.6.2.1 Mapas

A ferramenta *shiny*, dentre os vários recursos que possui, permite também a utilização de mapas. Possibilitando o uso de maneira interna do leaflet, uma biblioteca de mapas javascript.

Com a utilização dessa biblioteca, pode-se adicionar mapas nos códigos de linguagem R, possibilitando uma maior interatividade com o código. Além disso, esses mapas apresentam como base a figura 5 e podem receber *Provider Tiles* de acordo com o interesse do utilizador da

ferramenta.

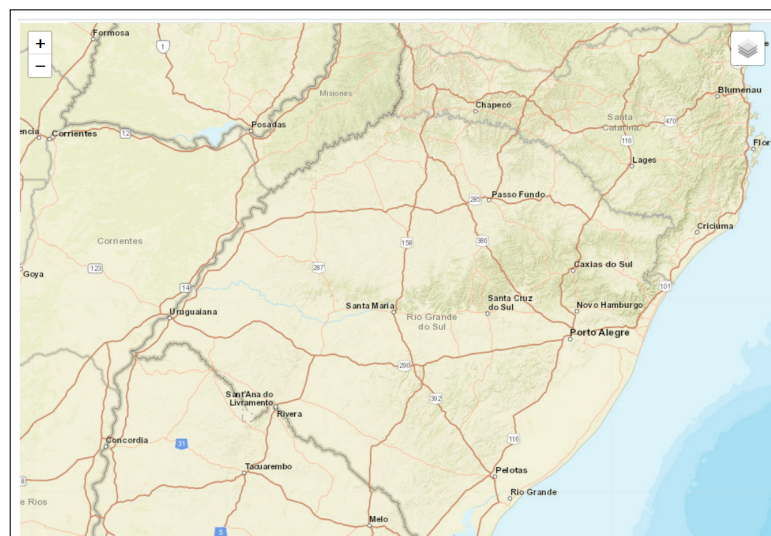
Figura 5: Mapa sem Provider



Fonte: [Fundação OpenStreetMap](#) (2019)

O leaflet permite utilizar outros *Provider Tiles* ou provedores de mapas, o que torna a ferramenta mais interessante, pois pode-se utilizá-la para diversas situações como por exemplo, trânsito, com foco maior dos mapas com rodovias, ilustrado pela figura 6. Existem outras várias possibilidades de uso, tais como relevo, hidrografias, dentre os vários providers existentes.

Figura 6: Provider de Rodovias



Fonte: [Esri](#) (2019)

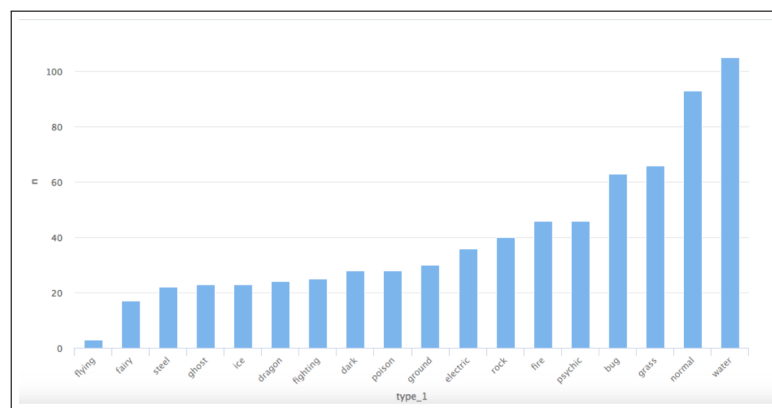
Além dos recursos da escolha de "estrutura do mapa", a adição de pontos no mapa também é possível. Para isso basta que sejam estipuladas as coordenadas geográficas desses pontos. Certamente, essa possibilidade permite uma melhor apresentação ao usuário, definindo

o local exato do ocorrido. Outro recurso disponível é a adição de *popups* nos pontos, isso é, informações relevantes ao usuário que, ao clicar no ponto específico terá informações relevantes sobre o local. No caso do trabalho atual, o foco dos pontos foi, além de marcar o local do acontecimento, informar questões climáticas do momento, qual foi o motivo do acidente, além da quantidade de pessoas envolvidas, dia da semana, dentre outros.

2.6.2.2 Gráficos

Para a geração dos gráficos, existe a biblioteca *Highcharter*, uma biblioteca que utiliza a linguagem R para a geração de diversos tipos gráficos que auxiliam na apresentação dos dados de maneira mais intuitiva e gerada por [Joshua Kunst \(2019\)](#), baseada na biblioteca *Highcharts*. Um dos tipos de gráfico gerado é o de colunas, exemplificado na figura 7, o qual é útil para se exemplificar quantidades, normalmente apresentadas no eixo **y**, de ocorrências, comumente ilustradas no eixo **x**.

Figura 7: Exemplo de Gráfico de Barras



Fonte: [Nishant Singh \(2018\)](#)

Outro tipo possível de se gerar através da biblioteca *Highcharter* é o gráfico *pie*, também conhecido como gráfico de pizza ou de torta. Esse tipo é útil para se analisar, de maneira proporcional, dados que possam ser interessantes de fazer uma verificação por proporção, como por exemplo: vendas individuais por vendedor, frequência de acidentes de avião, dentre outros.

3 TRABALHOS CORRELATOS

Algumas pesquisas foram feitas com relação a trabalhos que tenham alguma proximidade com o presente trabalho. Dentre os quais, cito os trabalhos de [Cassel \(2016\)](#) e de [Robaina \(2017\)](#), que fizeram plotagem de dados e utilização de dados abertos. Além de destacar o trabalho de [Junior \(2014\)](#), o qual apresenta uma análise clara de acidentes ocorridos em uma grande empresa do sul do Brasil.

[Cassel \(2016\)](#) coletou informações de jornais locais sobre os crimes ocorridos por região da cidade de Santa Maria no Rio Grande do Sul e através dessas informações, marcou no mapa da cidade esse pontos e cada cor do marcador representava uma quantidade de crimes ocorridos.

A coleta dos dados era feita utilizando um dicionário de termos criminalísticos, o qual foi elaborado pelo próprio autor. E através desses termos, era feito uma busca edições diárias de dois jornais da região sobre textos que continham essas palavras. Quando obtia sucesso na coleta, o artefato de saída gerado era o endereço de onde havia ocorrido o crime. Diante desses dados, era feita uma contagem de ocorrências no local e, de acordo com a quantidade, o marcador no mapa recebia uma cor determinada pela faixa estipulada pelo autor, conforme figura 8.

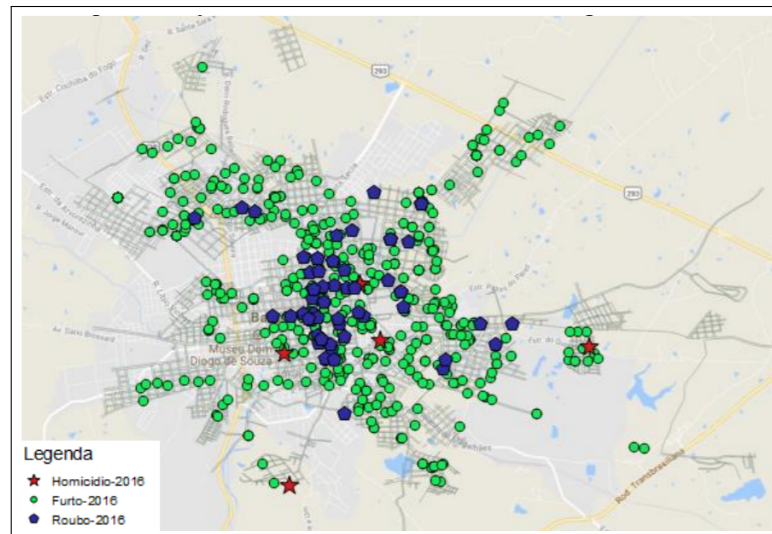
Figura 8: Mapa da Criminalidade



Fonte: [Cassel \(2016\)](#)

Já [Robaina \(2017\)](#) fez a utilização dos dados fornecidos pelo 6ºRPMON e mapeou dados onde ocorreram 3 tipos de crimes diferentes na cidade de Bagé no Rio Grande do Sul. Através da utilização de um mapa base da cidade, [Robaina \(2017\)](#) definiu formas e cores diferentes para os ícones dos crimes de homicídio, furto e roubo na cidade nos anos de 2014 à 2016, este apresentado na figura 9. Além disso, gerou mapas de calor ilustrando o local de maior concentração desses crimes.

Figura 9: Mapa - Crimes em Bagé

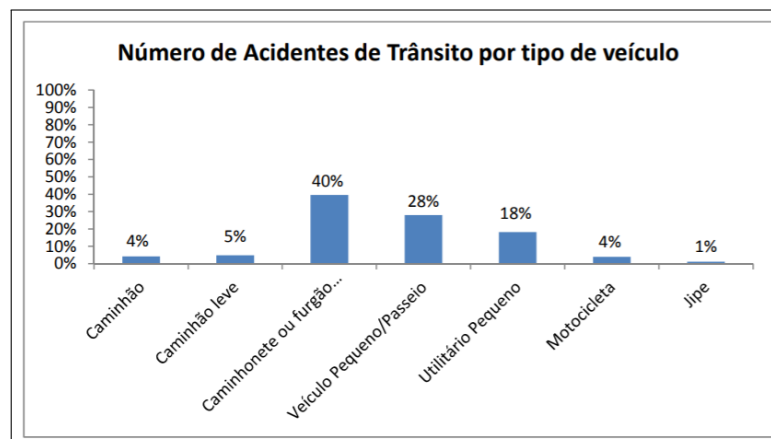


Fonte: [Robaina \(2017\)](#)

O trabalho de [Junior \(2014\)](#), faz uma análise estatística e gráfica sobre acidentes ocorridos em uma grande empresa do sul do Brasil. Além disso, o autor faz uso de variáveis diversas como datas, causas, tipo de veículo, esta última exemplificada na figura 10.

Com utilização de gráficos de barras, ele expõe as diversas situações dos acidentes ocorridos na empresa citada e apresenta análises correlacionando os fatos. Salienta-se ainda, que são apresentadas recomendações para cada tipo de ocorrência.

Figura 10: Acidentes - Tipo de Veículo



Fonte: [Junior \(2014\)](#)

Paralelo aos trabalhos citados, foi encontrado no site do Detran-RS um relatório, contendo diversas informações sobre acidentes no Rio Grande do Sul. Apesar de ser um relatório com um conjunto de dados de tamanho significativo, apresenta o problema de ser "apenas" um relatório técnico, isto é, inviável para leitura de um usuário leigo, por exemplo.

De maneira mais específica, falta algo mais utilizável para os condutores, ou seja, algo mais interativo, que não apresente somente informações, mas que permita ao usuário um acesso facilitado ou que seja um relatório mais focado ao que é desejado pelo utilizador da ferramenta.

Salienta-se ainda a exemplificação de utilização da ferramenta *shiny* com a biblioteca *dygraphs*, elaborada por [Angelo Salton \(2019\)](#).

De forma geral, o trabalho atual coincide com os trabalhos citados. A utilização de mapas e dados abertos é relacionada com os trabalhos de [Cassel \(2016\)](#) e [Robaina \(2017\)](#). Já um estudo sobre os acidentes é feito conforme o trabalho de [Junior \(2014\)](#). Este trabalho diferencia-se entretanto, por fazer uma análise sobre os acidentes gerais em todo o estado do Rio Grande do Sul, além de possuir um mapa interativo para o usuário final poder verificar informações pontuais sobre cada acidente ocorrido nos anos de 2018 e 2019, de acordo com as campanhas de trânsito selecionadas.

Salienta-se ainda a utilização de algoritmos para geração de árvores de decisão, elaboradas para cada campanha de trânsito analisada durante o desenvolvimento do trabalho.

4 DESENVOLVIMENTO

O trabalho foi estruturado seguindo os conceitos de KDD, conforme citado no capítulo 2. Como resultado, foi projetada uma ferramenta *web* que apresenta no mapa os locais dos acidentes para cada campanha de trânsito contida no trabalho e seus respectivos focos. Além de apresentar gráficos para as informações mais importantes e significativas para o usuário final, e também gráficos contendo informações não explícitas nos dados fornecidos pela PRF.

O desenvolvimento ficou subdividido em duas partes: escolha de campanhas de trânsito alvo e técnica de KDD.

4.1 ESCOLHA DE CAMPANHAS DE TRÂNSITO ALVO

As campanhas de conscientização no trânsito estão muito presentes em nosso país e apresentam vários temas importantes a serem abordados com intuito de estimular os condutores a terem mais cuidado. Podem-se citar campanhas como: se beber não dirija, maio amarelo, além de, campanhas de conscientização em épocas de veraneio.

4.1.1 Se beber não dirija

Essa campanha é embasada com base na lei seca, a qual teve entrou em vigor apenas em 2008 e foi sofrendo alterações até a sua versão atual no ano de 2018, segundo [G1 \(2018\)](#). De acordo com a lei atual, não existe tolerância de quantidade alcoólica no organismo, além de aplicar uma multa no valor de quase três mil reais e considerar uma pena de 5 anos se o condutor embriagado causar um acidente que deixe vítimas gravemente feridas ou uma pena de 8 anos caso ocorra óbitos na tragédia causada pelo condutor, estipuladas pela lei número 13.546, conforme o [Governo-Federal \(2017\)](#).

Figura 11: Logo da Campanha - Se beber não dirija



Fonte: [EPTC \(2018\)](#)

4.1.2 Maio Amarelo

Segundo informações de [Organização - Maio Amarelo \(2019\)](#), a campanha tem por objetivo colocar em pauta, para a sociedade, o tema trânsito; o mês de maio foi definido por ser o mesmo mês que a ONU decretou a década de ação para a segurança no trânsito e a cor amarela foi escolhida por simbolizar atenção no trânsito.

Nessa campanha, são elaboradas várias ações de conscientização, colocando em pauta acidentes de trânsito com o propósito de orientar a população dos riscos e das fatalidades que têm ceifado a vida de várias pessoas no trânsito.

Figura 12: Logo da Campanha - Maio Amarelo



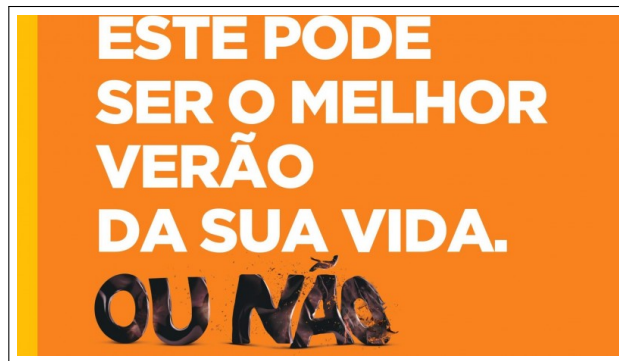
Fonte: [Prefeitura - Rio Grande \(2019\)](#)

4.1.3 Épocas de Veraneio

Segundo informações do [DAER \(2017\)](#), essa campanha existe em outros estados do país com nomes diferentes. Entretanto, no Rio Grande do Sul foi gerada em conjunto com quatro órgãos do estado: DAER, DETRAN, EGR e CRBM.

Apresenta como objetivo conscientizar a população dos riscos durante as viagens nos períodos de veraneio. Um dos focos da campanha é combater o excesso de velocidade dos condutores com a intenção de chegar ao destino mais rapidamente.

Figura 13: Logo da Campanha - Verão Seguro



Fonte: [DAER \(2017\)](#)

4.2 TÉCNICA DE KDD

4.2.1 Seleção

Para a etapa de seleção, relatada em [2.1](#), como eram necessários os dados sobre acidentes, foi feita uma busca pelos órgãos que poderiam possuir essas informações, de acordo com a definição da lei de acesso a informação pública, a qual regula o acesso a dados e informações retidas por órgãos públicos e foi sancionada em 18 de novembro de 2011, segundo informações do [Governo Federal \(2012\)](#).

Atráves de pesquisas sobre fontes possíveis, os dados da **Polícia Estadual do Rio Grande do Sul** ¹ estavam inacessíveis devido a *links* quebrados. Além disso, os dados fornecidos pelo *datapoa* ², esses estavam desatualizados, continham informações mais recentes apenas do ano de 2016. Além de serem referentes apenas à cidade de Porto Alegre, ou seja, não possuindo informações sobre as rodovias, que são ponto chave do trabalho.

Já no caso dos dados fornecidos pelo **DETRAN-RS** ³, o arquivo existente é do tipo PDF e não é um tipo aceito pela lei dos dados abertos, conforme [Governo Federal \(2012\)](#).

Os arquivos desse tipo já se apresentam em forma de relatório e impossibilitam a coleta individual de qualquer informação contida no arquivo.

Já os dados da PRF foram os escolhidos devido a existência de dados dos anos 2017 e 2018 e do mês de janeiro para o ano de 2019, contendo também informações referentes aos acidentes em rodovias no âmbito nacional.

Como a fonte de dados existente era apenas de dados tabulares, esse foi o tipo de dado escolhido. Além de serem os mais atualizados e coerentes dentre as opções possíveis.

¹Data do Acesso: 21/05/2019

²Data do Acesso: 24/05/2019

³Data do Acesso: 03/04/2019

4.2.2 Pré-Processamento e Limpeza

Com relação ao pré-processamento e limpeza, citado na seção 2.2, foi necessária a utilização de um editor de texto que permitisse a busca por um registro inteiro. Devido aos dados conterem informações de todo o território nacional e o trabalho tendo como foco apenas o estado do Rio Grande do Sul, foi preciso uma coleta de todos os registros que fossem da respectiva unidade federativa.

Figura 14: Antes da Seleção dos Dados - 2018

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	pesid	data_invers	dia_semana	horario	uf	br	km	município	causa_princi	causa_acidei	ordem_tipo	tipo_acident	classificacao	fase_dia	sentido_via	condicao_m
2	99973	216524	01/01/2018	segunda-feir	00:20:00	RJ	116	303,5	RESENDE	Sim	Condutor Dc	1	Salida de leit	Com Vitimas	Plena Noite	Decrescente	Nublado
3	99973	216530	01/01/2018	segunda-feir	00:20:00	RJ	116	303,5	RESENDE	Sim	Condutor Dc	1	Salida de leit	Com Vitimas	Plena Noite	Decrescente	Nublado
4	99973	216523	01/01/2018	segunda-feir	00:20:00	RJ	116	303,5	RESENDE	Sim	Condutor Dc	1	Salida de leit	Com Vitimas	Plena Noite	Decrescente	Nublado
5	99973	216527	01/01/2018	segunda-feir	00:20:00	RJ	116	303,5	RESENDE	Sim	Condutor Dc	1	Salida de leit	Com Vitimas	Plena Noite	Decrescente	Nublado
6	99973	216532	01/01/2018	segunda-feir	00:20:00	RJ	116	303,5	RESENDE	Sim	Condutor Dc	1	Salida de leit	Com Vitimas	Plena Noite	Decrescente	Nublado
7	99976	216291	01/01/2018	segunda-feir	00:40:00	SC	282	0,4	FLORIANOPC	Sim	Nao guardar	1	Colisao tras	Com Vitimas	Plena Noite	Crescente	Céu Claro
8	99976	216292	01/01/2018	segunda-feir	00:40:00	SC	282	0,4	FLORIANOPC	Sim	Nao guardar	1	Colisao tras	Com Vitimas	Plena Noite	Crescente	Céu Claro
9	99976	216295	01/01/2018	segunda-feir	00:40:00	SC	282	0,4	FLORIANOPC	Sim	Nao guardar	1	Colisao tras	Com Vitimas	Plena Noite	Crescente	Céu Claro
10	99977	218441	01/01/2018	segunda-feir	00:30:00	RJ	493	1	ITABORAI	Sim	Ultrapassag	1	Colisao front	Com Vitimas	Plena Noite	Crescente	Nublado
11	99977	218442	01/01/2018	segunda-feir	00:30:00	RJ	493	1	ITABORAI	Sim	Ultrapassag	1	Colisao front	Com Vitimas	Plena Noite	Crescente	Nublado
12	99977	218443	01/01/2018	segunda-feir	00:30:00	RJ	493	1	ITABORAI	Sim	Ultrapassag	1	Colisao front	Com Vitimas	Plena Noite	Crescente	Nublado
13	99977	218449	01/01/2018	segunda-feir	00:30:00	RJ	493	1	ITABORAI	Sim	Ultrapassag	1	Colisao front	Com Vitimas	Plena Noite	Crescente	Nublado
14	99977	218440	01/01/2018	segunda-feir	00:30:00	RJ	493	1	ITABORAI	Sim	Ultrapassag	1	Colisao front	Com Vitimas	Plena Noite	Crescente	Nublado
15	99981	216325	01/01/2018	segunda-feir	01:15:00	RS	386	134	SARANDI	Sim	Ingestao de	1	Colisao tran	Sem Vitimas	Pleno dia	Crescente	Céu Claro
16	99981	216327	01/01/2018	segunda-feir	01:15:00	RS	386	134	SARANDI	Sim	Ingestao de	1	Colisao tran	Sem Vitimas	Pleno dia	Crescente	Céu Claro
17	99982	216320	01/01/2018	segunda-feir	00:20:00	RS	293	151,7	CANDIOTA	Sim	Falta de Atei	1	Salida de leit	Com Vitimas	Plena Noite	Decrescente	Nublado

Diante de tal necessidade, foram utilizadas as expressões regulares que segundo Jargas (2016), "são métodos informais de se especificar um padrão de texto". Para tal, era necessária a utilização de alguma expressão que conseguisse coletar informações apenas do estado do Rio Grande do Sul. E isso foi possível, através da expressão: `*;"RS";*`. Com esse recurso, todos os acidentes ocorridos no RS foram coletados e foram gerados três novos *datasets* contendo informações apenas dos acidentes que aconteceram no Rio Grande do Sul.

Figura 15: Depois da Seleção dos Dados - 2018

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	pesid	data_invers	dia_semana	horario	uf	br	km	município	causa_princi	causa_acidei	ordem_tipo	tipo_acident	classificacao	fase_dia	sentido_via	condicao_m
2	99981	216325	01/01/2018	segunda-feir	01:15:00	RS	386	134	SARANDI	Sim	Ingestão de	1	Colisao tran	Sem Vitimas	Pleno dia	Crescente	Céu Claro
3	99981	216327	01/01/2018	segunda-feir	01:15:00	RS	386	134	SARANDI	Sim	Ingestão de	1	Colisao tran	Sem Vitimas	Pleno dia	Crescente	Céu Claro
4	99982	216320	01/01/2018	segunda-feir	00:20:00	RS	293	151,7	CANDIOTA	Sim	Falta de Atei	1	Salida de leit	Com Vitimas	Plena Noite	Decrescente	Nublado
5	99984	216303	01/01/2018	segunda-feir	01:30:00	RS	386	395,8	TRIUNFO	Sim	Animais na f	1	Atropelamei	Sem Vitimas	Plena Noite	Crescente	Céu Claro
6	99990	216321	01/01/2018	segunda-feir	02:10:00	RS	386	116,9	SARANDI	Sim	Ingestão de	1	Colisao front	Sem Vitimas	Plena Noite	Crescente	Céu Claro
7	99990	216322	01/01/2018	segunda-feir	02:10:00	RS	386	116,9	SARANDI	Sim	Ingestão de	1	Colisao front	Sem Vitimas	Plena Noite	Crescente	Céu Claro
8	100017	217406	01/01/2018	segunda-feir	03:30:00	RS	392	638,7	GUARANI D	Sim	Desobediênc	1	Colisao tran	Com Vitimas	Plena Noite	Crescente	Céu Claro
9	100017	217407	01/01/2018	segunda-feir	03:30:00	RS	392	638,7	GUARANI D	Sim	Desobediênc	1	Colisao tran	Com Vitimas	Plena Noite	Crescente	Céu Claro
10	100017	217413	01/01/2018	segunda-feir	03:30:00	RS	392	638,7	GUARANI D	Sim	Desobediênc	1	Colisao tran	Com Vitimas	Plena Noite	Crescente	Céu Claro
11	100017	217412	01/01/2018	segunda-feir	03:30:00	RS	392	638,7	GUARANI D	Sim	Desobediênc	1	Colisao tran	Com Vitimas	Plena Noite	Crescente	Céu Claro
12	100021	216367	01/01/2018	segunda-feir	06:20:00	RS	287	241,5	SANTA MARIA	Sim	Velocidade l	1	Salida de leit	Com Vitimas	Amanhecer	Crescente	Nublado
13	100024	216362	01/01/2018	segunda-feir	07:00:00	RS	116	246,8	SAO LEOPOL	Sim	Velocidade l	1	Capotament	Com Vitimas	Amanhecer	Crescente	Céu Claro
14	100032	216789	01/01/2018	segunda-feir	08:25:00	RS	116	42,6	VACARIA	Sim	Ultrapassag	1	Salida de leit	Sem Vitimas	Pleno dia	Decrescente	Sol
15	100032	216788	01/01/2018	segunda-feir	08:25:00	RS	116	42,6	VACARIA	Sim	Ultrapassag	1	Salida de leit	Sem Vitimas	Pleno dia	Decrescente	Sol
16	100045	216539	01/01/2018	segunda-feir	06:30:00	RS	471	506,5	RIO GRANDE	Sim	Condutor Dc	1	Salida de leit	Com Vitimas	Amanhecer	Crescente	Chuva
17	100063	216492	01/01/2018	segunda-feir	07:40:00	RS	392	5,2	RIO GRANDE	Sim	Ingestão de	1	Salida de leit	Sem Vitimas	Pleno dia	Decrescente	Céu Claro
18	100077	216558	01/01/2018	segunda-feir	14:55:00	RS	116	160,5	CAXIAS DO S	Sim	Falta de Atei	1	Colisao com	Com Vitimas	Pleno dia	Decrescente	Céu Claro
19	100077	216559	01/01/2018	segunda-feir	14:55:00	RS	116	160,5	CAXIAS DO S	Sim	Falta de Atei	1	Colisao com	Com Vitimas	Pleno dia	Decrescente	Céu Claro
20	100077	216560	01/01/2018	segunda-feir	14:55:00	RS	116	160,5	CAXIAS DO S	Sim	Falta de Atei	1	Colisao com	Com Vitimas	Pleno dia	Decrescente	Céu Claro
21	100082	216637	01/01/2018	segunda-feir	14:55:00	RS	101	21	TRES CACHO	Sim	Falta de Atei	1	Colisao tras	Sem Vitimas	Pleno dia	Crescente	Sol
22	100082	216638	01/01/2018	segunda-feir	14:55:00	RS	101	21	TRES CACHO	Sim	Falta de Atei	1	Colisao tras	Sem Vitimas	Pleno dia	Crescente	Sol
23	100102	216599	01/01/2018	segunda-feir	16:45:00	RS	101	70	OSORIO	Não	Falta de Atei	1	Colisao tras	Com Vitimas	Pleno dia	Crescente	Céu Claro
24	100102	216597	01/01/2018	segunda-feir	16:45:00	RS	101	70	OSORIO	Não	Falta de Atei	1	Colisao tras	Com Vitimas	Pleno dia	Crescente	Céu Claro
25	100102	216601	01/01/2018	segunda-feir	16:45:00	RS	101	70	OSORIO	Não	Falta de Atei	1	Colisao tras	Com Vitimas	Pleno dia	Crescente	Céu Claro
26	100103	216928	01/01/2018	segunda-feir	14:45:00	RS	471	636,5	SANTA VITO	Sim	Falta de Atei	1	Colisao front	Com Vitimas	Pleno dia	Decrescente	Céu Claro
27	100103	219408	01/01/2018	segunda-feir	14:45:00	RS	471	636,5	SANTA VITO	Sim	Falta de Atei	1	Colisao front	Com Vitimas	Pleno dia	Decrescente	Céu Claro

Após a utilização de expressões regulares, já utilizando a linguagem R, foram eliminadas três colunas do *dataset* referente aos anos de 2018 e 2019. Essas colunas continham informações sobre a delegacia, região da mesma e unidade operacional. Diante disso, foi necessária a remoção para equivaler ao *dataset* do ano de 2017 e podermos dar continuidade a etapa

seguinte do processo de KDD.

Dando sequência a limpeza dos dados, foi feita uma procura por dados inconsistentes ou ausentes, e foram encontrados alguns com relação a idade dos envolvidos nos acidentes, a qual, em alguns casos, foi simplesmente preenchida com um valor incorreto, tal como zero ou um valor acima de cem.

Conforme relatado na seção 2.2, nesse caso a escolha de eliminar todo o registro é a mais viável, tendo em vista que não podemos definir idades de pessoas através da média dos demais ou coletando o dado no exemplo anterior.

4.2.3 Transformação

Para a etapa de transformação, seguindo a estrutura KDD, situada na seção 2.3, foram feitas modificações nos dados de acordo com a necessidade dos algoritmos de classificação que seriam testados.

Os dados dos três anos foram agrupados em um único conjunto de dados. Devido a facilidade de manipulação e possibilidade de gerar conjuntos menores, mas de acordo com o propósito do trabalho.

4.2.4 Mineração de Dados

Dando continuidade ao processo de KDD, segue-se para a etapa de mineração de dados, já citada na seção 2.4. Nesse momento, após finalizada a atualização do tipo das variáveis, conforme citado na etapa anterior, foram criadas novas colunas para facilitar nas filtragens futuras, de tal forma foram criadas as colunas ano e mês.

Após alimentar essas novas colunas com os dados pertinentes, era necessária a criação de filtros que agrupassem dados utilizados nas análises das campanhas de trânsito, conforme tabela abaixo:

Tabela 1: Campanhas de trânsito e seus respectivos filtros

Campanha de Trânsito	Filtro
Se beber não dirija	Causa do Acidente = Ingestão de Álcool
Maio Amarelo	Mês do Acidente = Maio
Épocas de Veraneio	Meses do Acidente = Dezembro/Janeiro/Fevereiro

Finalizando as filtragens e criando três *datasets* distintos, parte-se para a alteração do tipo de dado referente a latitude e longitude. Isso deve-se ao fato de que o mapa necessita do tipo *numeric* para conseguir indentificar esses valores e adicioná-los como pontos aceitáveis.

Prosseguindo para a parte visual, além dos mapas com os seus respectivos pontos já apresentáveis, foram gerados gráficos de barras e pizza com informações importantes sobre cada campanha. Partindo do princípio de filtrar as informações para o usuário final.

Paralelamente, foram instaladas e utilizadas as bibliotecas da rede de servidores *CRAN*, *rpart* e *rpart.plot*. Ambas são relacionadas a árvore de decisão, citada na subseção [2.4.1.2](#), entretanto, a primeira gera a árvore em lista e a segunda permite uma visualização gráfica da árvore gerada.

4.2.5 Interpretação

A etapa de interpretação, citada na seção [2.5](#), é apresentada nos capítulos resultados e conclusão. Devido a ser a parte final do trabalho e precisar ter uma análise dos dados gerados pelo conjunto como um todo.

5 RESULTADOS

Os acidentes de trânsito não estão reduzindo se analisarmos o todo ou algumas causas específicas. Se considerarmos as várias campanhas de conscientização existentes, isso gera problemas para o nosso país. Salienta-se que a maior parte dos acidentes ocorrem devido a falta de atenção à condução, conforme figura 22 e recomenda-se que sejam feitas mais campanhas de conscientização com esse propósito, os quais são detalhados por [Circula Seguro \(2019\)](#) e dentre os quais encontram-se: Utilização do aparelho celular ou GPS no volante, fumar, dentre outras. Além do mais, se considerarmos os acidentes por falta de atenção à condução, ultrapassagem indevida e velocidade incompatível, temos ocorrências de mais de 50% da quantidade de acidentes, independente do contexto, se dentro de alguma campanha de trânsito ou não.

Paralelo a isso, recomenda-se que as campanhas com foco na redução de conduzir veículos após ingestão de álcool, deem mais ênfase ao público masculino. Apesar de ser o público que mais causa acidentes em números totais, sendo uma média de 87% a 13%, os homens representam 93% dos acidentes causados por ingestão de álcool, conforme ilustra a figura 21. Salienta-se ainda que, de acordo com [Governo Federal \(2012\)](#), a lei que dispõe sobre o acesso a informação foi criada em 18 de novembro de 2011 e surgiu no decreto em maio de 2012. Diante desse fato, não se garante o acesso aos dados sobre ingestão de álcool de anos anteriores. De tal forma, isso impede que seja feita uma comparação dos dados pós campanha "se beber não dirija" e antes.

Entretanto, um dos pontos positivos é que, possivelmente, com o advento da lei seca, ocorrido em 2018, conforme informações de [G1 \(2018\)](#), tivemos uma redução significativa do ano anterior para o corrente. Informações que são provadas através do gráfico abaixo, onde observa-se uma redução significativa de aproximadamente -48% de ocorrências:

Figura 16: Comparativo - Adendo lei seca



Frisa-se ainda que os acidentes analisados pelo trabalho foram ocorridos apenas no estado do Rio Grande do Sul, podendo gerar resultados diferentes em outros locais analisados. Por meio da ferramenta gerada, é possível concluir que os resultados apresentados são coerentes e podem ser utilizados para um propósito maior além de servir apenas como orientação aos condutores.

5.1 ESCOLHA DO ALGORITMO

Nessa seção, será apresentado o algoritmo utilizado e os resultados obtidos nas análises feitas para cada campanha de conscientização apresentada no trabalho.

5.1.1 Árvore de Decisão

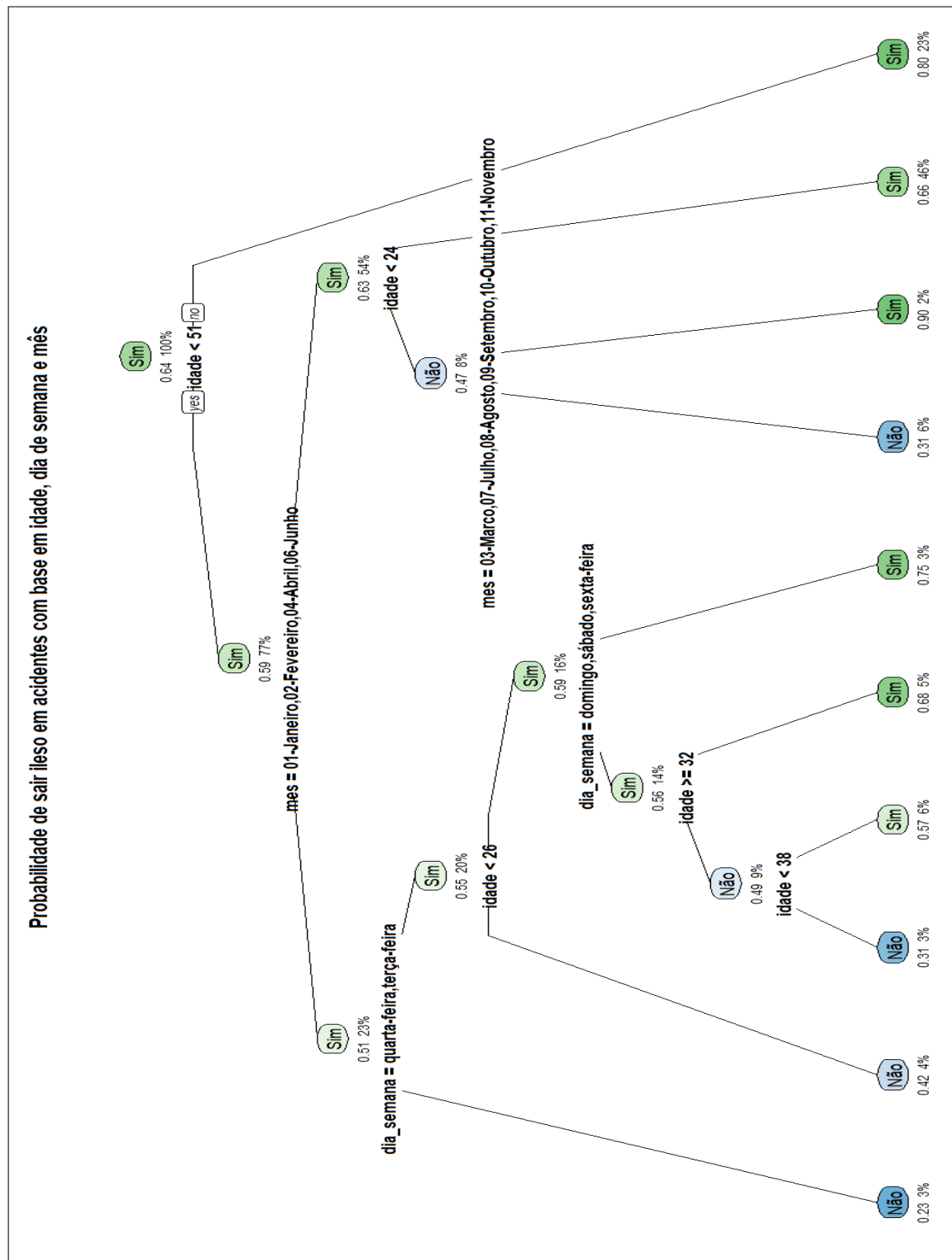
O algoritmo árvore de decisão foi escolhido pois apresenta uma melhor visualização das predições, mostrando também as variáveis utilizadas para cada predição em cada um dos nós criados. Esse algoritmo se apresenta melhor que o Naive Bayes pois este mostra apenas um resultado final, sem considerar subsituações existentes para cada faixa etária, por exemplo. Salienta-se ainda que, a probabilidade da pessoa sair ilesa de um acidente não será 100% tendo em vista que ocorreram acidentes com vítimas feridas e vítimas fatais. Logo o algoritmo árvore

de decisão obteve resultados significativos e coerentes para cada campanha de conscientização.

Cabe ressaltar que o algoritmo *Random Forest* foi preterido pelo de árvore de decisão pois a quantidade de dados não é relativamente alta a ponto de se utilizar um algoritmo que gere várias árvores de decisão, conforme é o *Random Forest*.

Conforme a figura 17, a qual demonstra um dos resultados mais otimista, obtido pela predição feita por árvore de decisão, ilustram-se também as "subpredições" que vão demonstrando variações melhores ou piores dependendo da escolha.

Figura 17: Árvore de Decisão - Probabilidade de Ilesos - Se beber não dirija



O resultado apresentado pela predição da árvore de decisão apresenta o valor de 0.64, que seria a acurácia ou acerto da árvore. Além de apresentar 100%, no caso o total de ocorrências no conjunto de dados e a resposta da probabilidade em um balão acima. Nesse caso "Sim", a árvore informa que a probabilidade de sair ileso em um acidente após a ingestão de álcool, analisando todos os conjuntos possíveis.

Fazendo uma breve descrição das ocorrências, partindo do nó raiz, onde verifica a veracidade da idade do condutor ser menor que 51 anos, para seus nós filhos, a esquerda, premissa verdadeira e a direita, falsa. São feitas análises para cada condição estipulada abaixo do resultado do nó. Por exemplo, supondo que a idade do condutor seja menor que 51 anos, a árvore prediz que o condutor sairá ileso do acidente, entretanto, apresenta uma acurácia de 0.59, ou 59%. E o total de dados que possuem essa situação e foram utilizados para a análise, representam 77% do conjunto de dados.

Caso a idade do condutor seja maior que 51 anos, o nó da direita informa que, provavelmente, o condutor sairá ileso do acidente, a taxa de acerto para essa predição é de 0.80 ou 80% e o conjunto de dados que foram utilizados para essa análise representam 23% do conjunto total.

Dessa forma, a árvore vai fazendo as verificações para cada um de seus nós e como análise geral traz que provavelmente o condutor sairá ileso e apresenta a taxa de 64% de acurácia. Porém, se analisarmos os nós folhas, vemos que a taxa de acerto varia de 23% a 90%.

Na tabela abaixo, veremos as predições feitas pelo algoritmo árvore de decisão para cada campanha de conscientização. Cabe resaltar que os valores não teriam como chegar próximos a 100% pois existem acidentes com pessoas feridas e mortas, o que inviabiliza a hipótese de se obtermos uma predição mais próxima do valor citado.

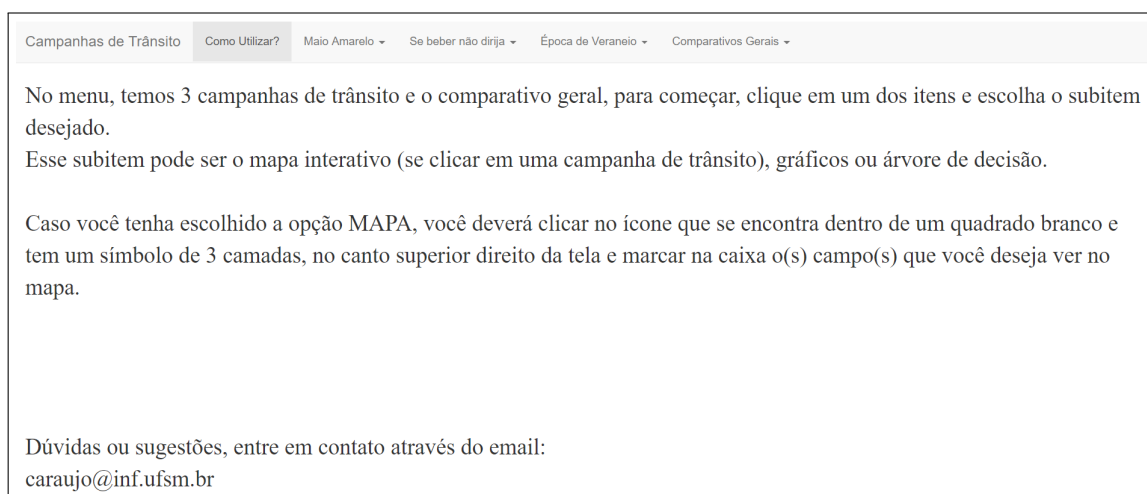
Tabela 2: Resultados das Análises de Predição - Sair Ileso de Acidentes

Campanha de Trânsito	Predição
Maio Amarelo	59%
Se beber não dirija	64%
Época de Veraneio	64%

6 CONCLUSÃO

Como produto final deste trabalho, foi desenvolvida uma ferramenta, que está sendo disponibilizada para acesso dos interessados ¹ e os respectivos códigos encontram-se no repositório ². A ferramenta, inicialmente abre uma breve explicação de utilização e possui um menu contendo itens como o nome de cada campanha de trânsito e também um item chamado comparativos gerais. Através desses, campos, é possível acessar formas de análise visuais individuais, incluindo mapas e gráficos.

Figura 18: Estrutura Inicial



6.1 MAPAS

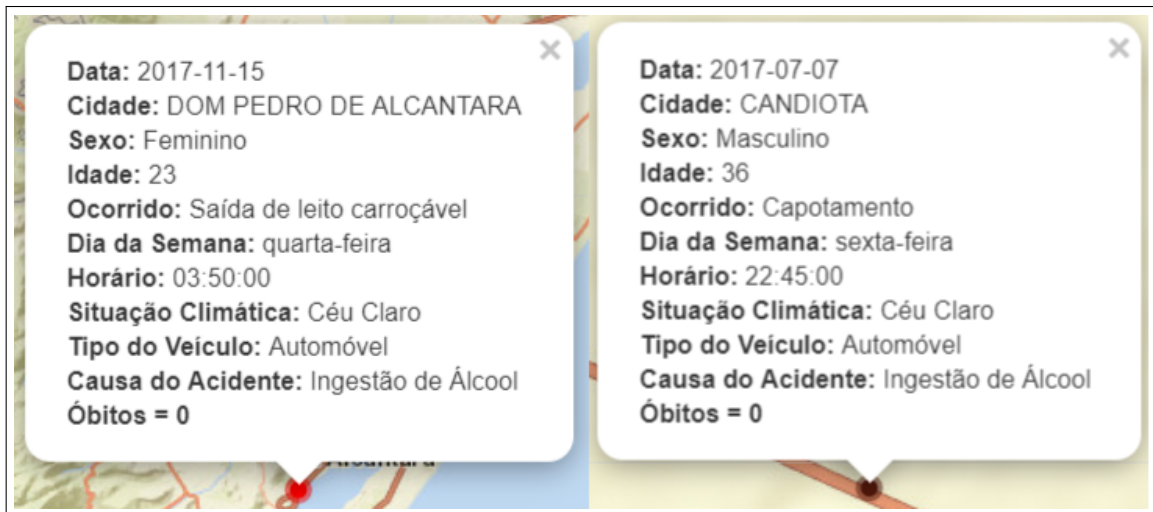
Para cada campanha de trânsito, são disponibilizados mapas individuais contendo, inicialmente, os acidentes separados por causa. Entretanto, para a campanha se beber não dirija, os acidentes foram separados por sexo ou total de óbitos. Ao clicar em qualquer ponto no mapa, são apresentadas informações úteis que possam interessar aos usuários.

Para as campanhas época de veraneio e maio amarelo, ao selecionarmos a(s) causa(s) disponíveis, elas são plotadas no mapa de acordo com a localização dos pontos e para a campanha se beber não dirija, os acidentes podem ser plotados por sexo do causador ou pelo total de óbitos, marcando os pontos de acordo com a seleção do usuário. Pode-se verificar os acidentes plotados no mapa clicando em um dos pontos gerados. Seguindo esse passo, será apresentado um pequeno texto contendo informações sobre a data do acidente, a causa, dentre outras opções, conforme figura 19.

¹Site da Ferramenta: <https://caraujo.shinyapps.io/ProjetoFinal/>

²Códigos: <https://github.com/caraujofsm/FerramentaR>

Figura 19: Textos Informativos - Acidentes com cores diferentes por sexo do condutor



Como o foco do trabalho são acidentes de trânsito em rodovias, o *provider tile* fornecido pelo [Esri \(2019\)](#) foi escolhido devido ao destaque maior para as rodovias.

6.2 GRÁFICOS

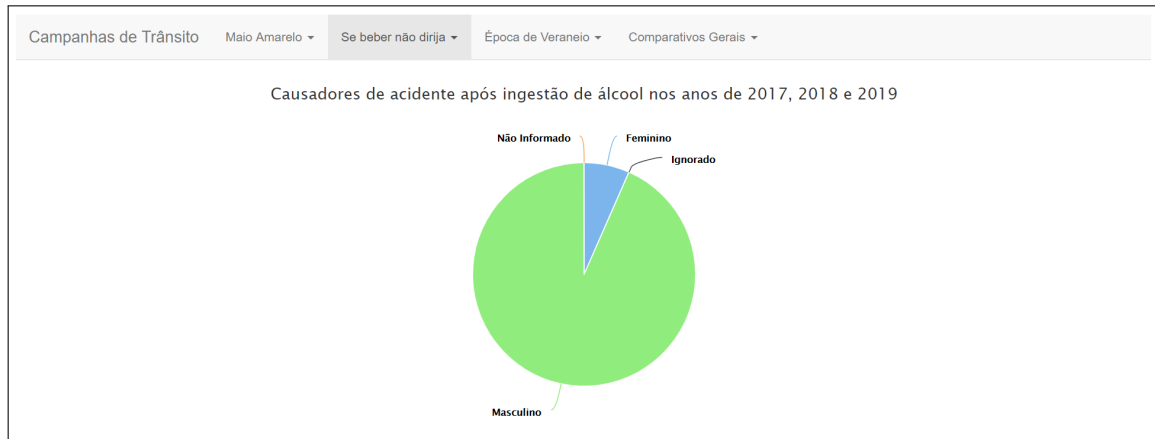
Todas as campanhas de trânsito também possuem gráficos ilustrando para os usuários pontos importantes sobre cada uma delas, como por exemplo, as causas dos acidentes, conforme figura 20 que representa o tipo de acontecimento referente à campanha maio amarelo.

Figura 20: Causas dos Acidentes - Maio Amarelo



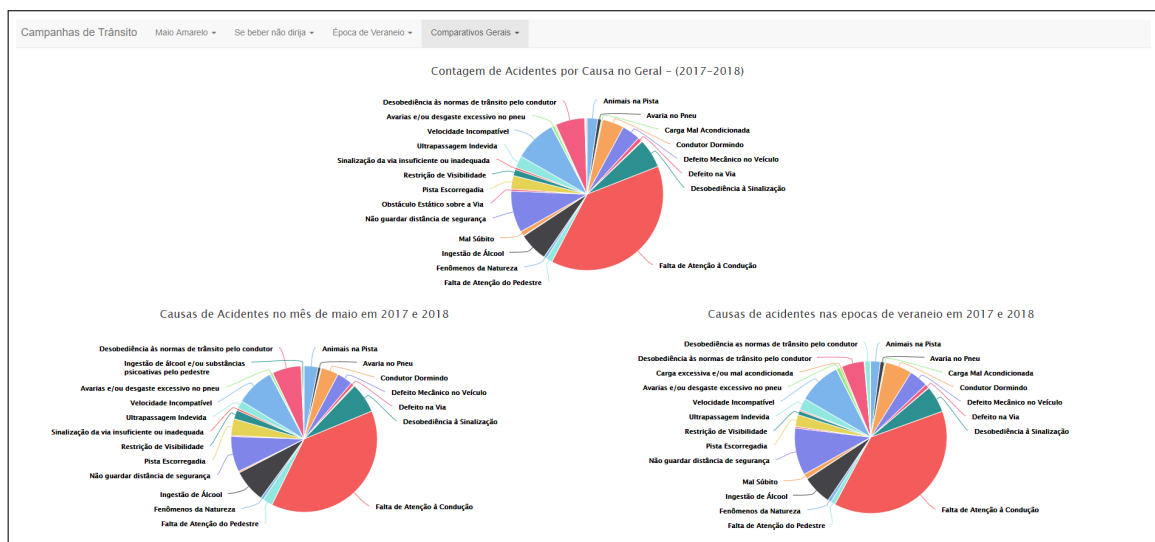
A ferramenta também possui informações sobre os maiores causadores de acidentes que tem como causa principal a ingestão de álcool, conforme ilustração 21. Imagem retirada dos dados da campanha "Se beber não dirija".

Figura 21: Culpados dos Acidentes - Ingestão de Álcool - Se Beber Não Dirija



Além das informações individuais de cada campanha, estão disponibilizadas na aba Comparativos Gerais, informações comparando o geral pontos de campanha de trânsito. Uma das possibilidades de gráfico foi o comparativo sobre as causas de acidente. Obviamente, com a retirada da campanha "Se beber não dirija" pois essa é totalmente causada por ingestão de álcool.

Figura 22: Comparação - Acidentes por Causa



Outra possibilidade é a comparação sobre os dias semanais, essa feita entre a totalidade de acidentes com cada campanha de trânsito, conforme figuras 23, 24 e 25, as quais são uma imagem apenas mas foram divididas em três para melhor visualização.

Figura 23: Comparação - Acidentes por Dia Semanal - Parte 1

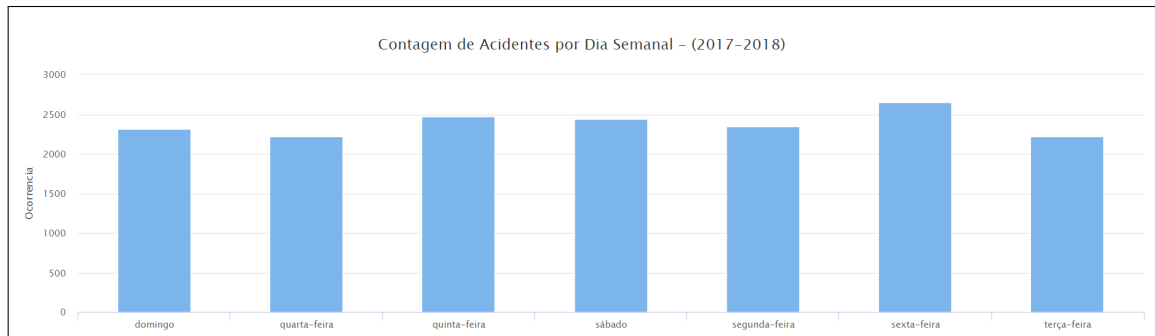


Figura 24: Comparação - Acidentes por Dia Semanal - Parte 2

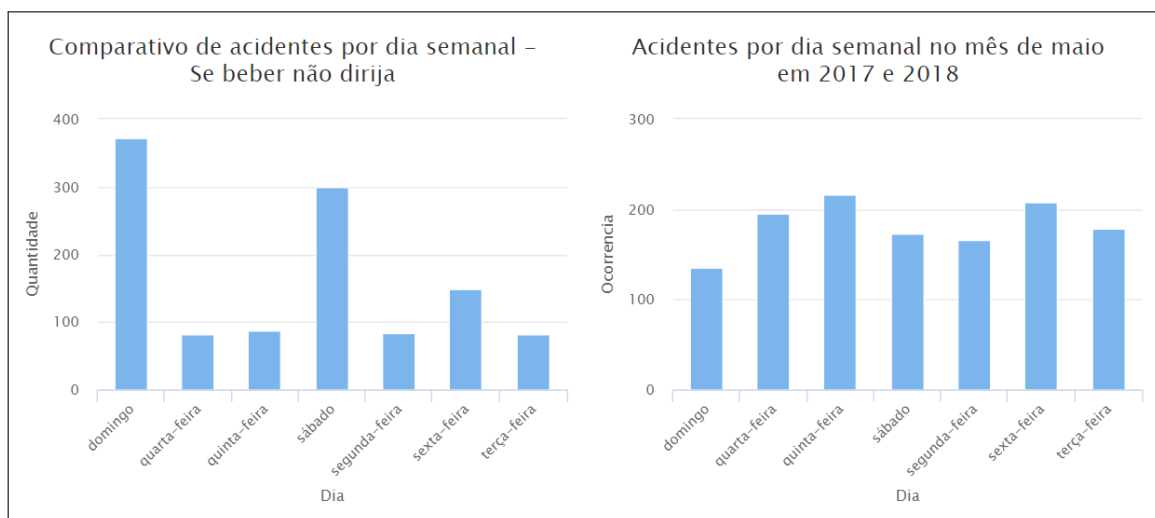
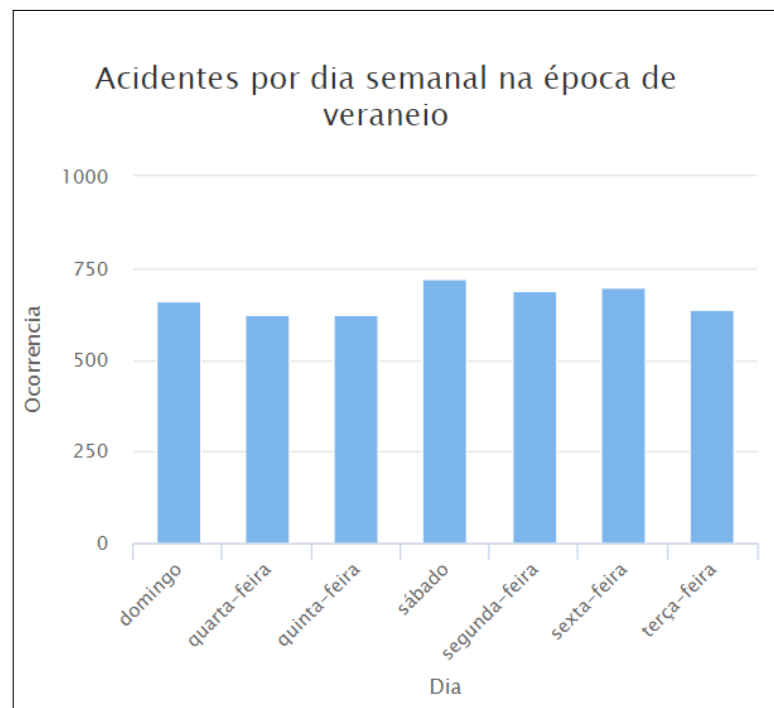


Figura 25: Comparação - Acidentes por Dia Semanal - Parte 3



6.3 TRABALHOS FUTUROS

A ferramenta final elaborada nessa pesquisa poderá servir de base para outras análises. Diante disso, são especificadas algumas possibilidades de dar continuidade a esse projeto, as quais são:

- Adicionar novos dados a pesquisa, conforme adição de novos dados no site da PRF além de ícones nos pontos plotados para diferenciação dos acidentes com morte e sem morte;
- Fazer uma correlação dos acidentes causados por falta de atenção, a maioria dos casos ocorridos, com a utilização de smartphones e aplicativos de mensagens instantâneas e aplicativos de navegação por satélite;
- Analisar o que seria pago pelo DPVAT com base no número de mortos ou feridos graves contidos nos dados fornecidos pela PRF;
- Expandir a pesquisa e as ideias futuras para outros locais: Estados, Regiões e até mesmo o Brasil como um todo.

REFERÊNCIAS

ANGELO SALTON. Gráficos dinâmicos no R: Um exemplo com shiny e dygraphs.

Disponível em: <https://wp.ufpel.edu.br/economia/files/2016/04/cadernos_economicos_10.pdf>.

Acesso em: 10 abr. 2019.

CASSEL, Otávio. Reconhecimento de Entidades Nomeadas em Textos Jornalísticos Criminais. 2016. Bachelor's Thesis – Universidade Federal de Santa Maria - UFSM.

CIRCULA SEGURO. As dez coisas que mais nos distraem ao volante. Disponível em:

<<http://www.circulaseguro.pt/conductor-e-ocupantes/as-dez-coisas-nos-distraem-ao-volante>>. Acesso em: 2 jun. 2019.

CODD, E. F. A Relational Model of Data for Large Shared Data Banks. **Commun. ACM**, ACM, New York, NY, USA, v. 13, n. 6, p. 377–387, jun. 1970. ISSN 0001-0782. DOI:

10.1145/362384.362685. Disponível em:

<<http://doi.acm.org/10.1145/362384.362685>>.

CRISTIANO FURQUIM. 10 Algoritmos de Aprendizagem de Máquinas (Machine Learning) que você precisa saber. 14 fev. 2018. Disponível em:

<https://cdn-images-1.medium.com/max/1200/1*rLYBTFztVF0pAy7m30tElg.jpeg>.

DAER. 2017. Disponível em:

<<https://www.daer.rs.gov.br/transito-seguro-verao-para-todos-2017-2018>>.

Acesso em: 29 mai. 2019.

EPTC. 31 ago. 2018. Disponível em:

<https://twitter.com/EPTC_POA/status/1035646090748612609>. Acesso em: 29 mai. 2019.

ESRI. Streets. Disponível em: <<http://esri.maps.arcgis.com/home/index.html>>.

Acesso em: 20 abr. 2019.

FUNDAÇÃO OPENSTREETMAP. Disponível em: <<https://www.openstreetmap.org>>.

Acesso em: 18 abr. 2019.

G1. Lei Seca ficou mais rígida nos últimos anos; veja o que pode e o que não pode. 18 jun.

2018. Disponível em: <<https://g1.globo.com/carros/noticia/lei-seca-ficou-mais-rigida-nos-ultimos-anos-veja-o-que-pode-e-o-que-nao-pode.ghtml>>.

GEDSON FARIA, ROSELI FRANCELIN ROMERO. Explorando o Potencial de Algoritmos de Aprendizado com Reforço em Robôs Móveis. **Proceedings of the IV Brazilian Conference on Neural Networks**, 1999.

GOVERNO FEDERAL. DECRETO Nº 7.724, DE 16 DE MAIO DE 2012. 16 mai. 2012. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/Decreto/D7724.htm>. Acesso em: 16 abr. 2019.

GOVERNO-FEDERAL. LEI Nº 13.546, DE 19 DE DEZEMBRO DE 2017. 19 dez. 2017. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2017/Lei/L13546.htm#art4>. Acesso em: 26 mai. 2019.

HAYKIN, S. Redes Neurais: Princípios e Prática. [S.l.]: Artmed, 2007. ISBN 9788577800865. Disponível em: <<https://books.google.com.br/books?id=bhMwDwAAQBAJ>>.

HELIO GUROVITZ. O que cerveja tem a ver com fraldas. 18 fev. 2011. Disponível em: <<https://exame.abril.com.br/revista-exame/o-que-cerveja-tem-a-ver-com-fraldas-m0053931/>>. Acesso em: 27 mai. 2019.

HUI LI. 2017. Disponível em: <<https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/>>. Acesso em: 15 jun. 2019.

JARGAS, Aurélio Marinho. Expressões Regulares: Uma Abordagem Divertida. [S.l.]: NOVATEC, 2016. ISBN 9788575221730. Disponível em: <https://books.google.com.br/books?id=fT%5C_CwAEACAAJ>.

JOSHUA KUNST. Highcharter. Disponível em: <<http://jkunst.com/highcharter/>>. Acesso em: 26 mai. 2019.

JUNIOR, João Alberto Kucek. Análise dos Acidentes de Trânsito Ocorridos em uma Grande Empresa do Sul do Brasil. 2014. Universidade Tecnológica Federal do Paraná - UTFPR.

LARA MONTEZUMA. Acidentes de trânsito: um problema de saúde pública. Disponível em: <<https://www.unifor.br/web/osv/acidentes-saude>>. Acesso em: 23 abr. 2019.

LUIS ALBERTO MELCHERT. Carros: como está o Brasil no cenário mundial. Disponível em: <<https://www.http://bestcars.uol.com.br/bc/informe-se/carros-como-esta-o-brasil-no-cenario-mundial/>>. Acesso em: 7 abr. 2019.

MRG TECNOLOGIA. 2019. Disponível em: <<https://blog.mrgtecnologia.com.br/voce-sabe-o-que-e-poc-prova-de-conceito-entenda-a-importancia/>>. Acesso em: 15 jun. 2019.

MYBRIDGE. 30 Amazing Machine Learning Projects for the Past Year (v.2018). 4 jan. 2018. Disponível em: <<https://medium.mybridge.co/30-amazing-machine-learning-projects-for-the-past-year-v-2018-b853b8621ac7>>. Acesso em: 1 jun. 2019.

NISHANT SINGH. Data Visualization with Highcharter in R. 20 ago. 2018. Disponível em: <<https://www.datacamp.com/community/tutorials/data-visualization-highcharter-r>>. Acesso em: 26 mai. 2019.

ORGANIZAÇÃO - MAIO AMARELO. Disponível em: <<https://maioamarelo.com/o-movimento/>>. Acesso em: 29 mai. 2019.

PAULO VASCONCELLOS. Como participar de uma competição de Machine Learning no Kaggle. 31 mai. 2017. Disponível em: <<https://paulovasconcellos.com.br/competicao-kaggle-titanic-tutorial-5b11993774f7>>. Acesso em: 1 jun. 2019.

PEDRO CÉSAR TEBALDI GOMES. Classificação com Naive Bayes. Disponível em: <<https://www.datageeks.com.br/naive-bayes/>>. Acesso em: 26 mai. 2019.

PREFEITURA - RIO GRANDE. 26 abr. 2019. Disponível em: <<http://www.riogrande.rs.gov.br/wp-content/uploads/2019/04/Me0u%C3%A7a-400x240.png>>. Acesso em: 29 mai. 2019.

REDAÇÃO, DO NA PRÁTICA. Foi para resolver um problema de sala de aula que esse estudante inventou as planilhas eletrônicas. 22 fev. 2017. Disponível em: <<https://www.napratica.org.br/o-que-o-inventor-das-planilhas-eletronicas-tem-dizer-sobre-invencoes/>>.

ROBAINA, Rafael. Utilização do Sistema de Informações Geográficas (SIG) para o Mapeamento da Criminalidade da Cidade de Bagé-RS. 2017. Bachelor's Thesis – Universidade Federal do Pampa - UNIPAMPA.

SENADO FEDERAL. Estudo da Organização Mundial da Saúde (OMS) sobre mortes por acidentes de trânsito em 178 países é base para década de ações para segurança. Disponível em: <<http://www.senado.gov.br/noticias/Jornal/emdiscussao/motos/saude/estudo-da-organizacao-mundial-da-saude-oms-sobre-mortes-por-acidentes-de-transito-em-178-paises-e-base-para-decada-de-aco-es-para-seguranca.aspx>>. Acesso em: 7 abr. 2019.

ŠTER, BRANKO AND DOBNIKAR, ANDREJ. Neural networks in medical diagnosis: Comparison with other methods. **Proceedings of the International Conference on Engineering Applications of Neural Networks**, jan. 1996.

TAN, P.N. AND STEINBACH, M. AND KUMAR, V. AND FERNANDES, A. Introdução ao datamining: mineração de dados. [S.l.]: Ciencia Moderna, 2009. ISBN 9788573937619. Disponível em: <<https://books.google.com.br/books?id=69d6PgAACAAJ>>.

TURBAN, E. AND SHARDA, R. AND ARONSON, J.E. AND KING, D. Business Intelligence: Um enfoque gerencial para a inteligência do negócio. [S.l.]: Bookman, 2009. cap. 2. ISBN 9788577804252. Disponível em: <<https://integrada.minhabiblioteca.com.br/#/books/9788577804252/cfi/51!/4/2@100:0.00>>.

USAMA FAYYAD, GREGORY PIATETSKY-SHAPIO, AND PADHRAIC SMYTH. From Data Mining to Knowledge Discovery in Databases. **AI Magazine Volume 17 Number 3 (1996)** (© AAAI), p. 18, mar. 1996.

VIAS SEGURAS. Estatísticas do seguro obrigatório DPVAT. Disponível em: <<http://vias-seguras.com/content/download/6944/41150/file/DPVAT%20Boletim-Estatistico%202015.pdf>>. Acesso em: 23 abr. 2019.

WAYNE YADDOW. Ensuring Data Warehouse Quality: A Business Mandate. 31 mai. 2013. Disponível em: <<https://www.ibmbigdatahub.com/blog/ensuring-data-warehouse-quality-business-mandate>>. Acesso em: 29 mai. 2019.