UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

BRUNO SHINJI ITO

ANÁLISE DE ACIDENTES DE TRÂNSITO EM BAURU-SP UTILIZANDO CIÊNCIA DE DADOS

BAURU
Novembro/2025

BRUNO SHINJI ITO

ANÁLISE DE ACIDENTES DE TRÂNSITO EM BAURU-SP UTILIZANDO CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Câmpus Bauru.
Orientador: Prof. Dr. Higor Amario de Souza

BAURU
Novembro/2025

Ito, Bruno Shinji
I89a Análise de acidentes de trânsito em Bauru-SP utilizando Ciência de Dados / Bruno Shinji Ito. -- Bauru, 2025
47 p. : il., tabs., fotos, mapas

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru
Orientador: Higor Amario de Souza

1. Acidentes de trânsito. 2. Aprendizado de Máquina. 3. Ciência de Dados. 4. XGBoost. 5. Bauru. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Bruno Shinji Ito

Análise de acidentes de trânsito em Bauru-SP utilizando Ciência de Dados

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Câmpus Bauru.

Banca Examinadora

Prof. Dr. Higor Amario de Souza
Orientador
Universidade de São Paulo
Escola Politécnica
Departamento de Engenharia de Computação e Sistemas Digitais

Prof. Dra. Simone das Graças Domingues Prado
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Profa. Dra. Juliana da Costa Feitosa
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Bauru, 12 de novembro de 2025.

Resumo

Os acidentes de trânsito representam um grande problema social e econômico no ambiente urbano, não sendo diferente no município de Bauru. Compreender os fatores e perfis de vítimas que são mais suscetíveis às fatalidades pode auxiliar no direcionamento de recursos a fim de mitigar a ocorrência de mais óbitos. O seguinte trabalho busca realizar a análise da ocorrência dos sinistros em Bauru utilizando as metodologias e técnicas de Ciência de Dados. Para o estudo foram usados dados públicos de sinistros e vítimas de Infosiga do período de 2019–2025. Para previsão foram testados os algoritmos de Árvore de Decisão e XGBoost, realizando comparação no potencial de previsão de cada modelo, além da realização da Análise Exploratória de Dados e utilização de DBSCAN. Os modelos apresentaram capacidade de predição satisfatórias, em especial o modelo de XGBoost com a técnica de *Grid Search*. Técnicas de DBSCAN apontam *hotspots* as principais vias da cidade e as rodovias ao redor da cidade como maior causadoras de vítimas não fatais e fatais, respectivamente.

Palavras-chave: Acidentes de trânsito; aprendizado de máquina; ciência de dados; XGBoost; DBSCAN; Bauru.

Abstract

Traffic accidents represent a significant social and economic problem in the urban environment, and the municipality of Bauru is no exception. Understanding the factors and victim profiles most susceptible to fatalities can help direct resources to mitigate further deaths. This work seeks to analyze the occurrence of accidents in Bauru using Data Science methodologies and techniques. Public data on accidents and victims from Infosiga, covering the period from 2019 to 2025, were used for the study. For prediction, Decision Tree and XGBoost algorithms were tested, comparing the predictive potential of each model, in addition to performing Exploratory Data Analysis (EDA) and using DBSCAN clustering. The models demonstrated satisfactory predictive performance, particularly the XGBoost model optimized with the Grid Search technique. DBSCAN techniques identified hotspots on the city's main roads as major contributors to non-fatal victims, while the highways surrounding the city were identified as the main contributors to fatal victims, respectively.

Keywords: Traffic Accidents. Data Science. Machine Learning. XGBoost. Bauru.

Lista de figuras

Lista de tabelas

Lista de quadros

Sumário

1 Introdução

A fatalidade no trânsito é um problema de segurança pública mundial. Apesar de existir uma ligeira diminuição de mortes no trânsito entre 2010 e 2021, com redução de 1,25 milhões de vítimas para 1,19 milhões (Organização Mundial de Saúde, 2023) seu número ainda se mantém expressivo, o que gera altos impactos sociais e econômicos. No cenário brasileiro, apenas em 2023 cerca de 35.938 pessoas foram vítimas de sinistros fatais no país (Instituto de Pesquisa Econômica Aplicada, 2023).

Esse cenário alarmante se reflete no estado de São Paulo, que registrou cerca de 6.124 vítimas fatais em 2024. O município de Bauru, foco deste estudo, contribuiu para essa estatística com 41 óbitos no mesmo período (Departamento Estadual de Trânsito de São Paulo, 2025).

E também com as técnicas presentes como em Ciência de Dados é possível utilizar abordagens inteligentes como *Data-Driven Decision Making* (DDDM) para direcionar políticas públicas de prevenção e fiscalização a fim de diminuir a ocorrência de sinistros (PROVOST; FAWCETT, 2013).

Nesse contexto, o presente trabalho visa analisar e entender sobre a natureza dos acidentes e fatalidades de trânsito da cidade de Bauru, utilizando as técnicas e metodologias de Ciência de Dados.

1.1 Problemática

Como mencionado, o cenário de sinistros fatais é um problema social impactante em todas as cidades, principalmente de médio e grande porte, inclusive Bauru. O acesso a essas informações e dados é possibilitado graças à crescente digitalização da informação, e a pesquisa de previsão de acidentes em pesquisas como de Kričković (2024) demonstra que esse campo de estudo há espaço para crescer.

Apesar do Detran desenvolver um artigo técnico realizando a análise da sinistralidade de acidentes (Detran, 2025), o cenário nacional carece de pesquisas e estudos metodológicos que una segurança viária, técnicas de Inteligência Artificial e estudo de fatores de risco com foco em Bauru, e que busque compreender quais são os fatores de risco associados aos sinistros e fatalidade, e em qual medida os modelos são capazes de prever.

1.2 Justificativa

Com isso o desenvolvimento do trabalho se justifica buscando diminuir o número de fatalidades no trânsito, em consonância ao Plano Nacional de Redução de Mortes e Lesões no Trânsito (PNATRANS) e aos seus seis pilares de plano de ação, como gestão de segurança no trânsito (Brasil, 2018).

O foco também é mitigar o impacto econômico que os sinistros causam no país, como por exemplo aponta a notícia sobre Sistema Único de Saúde (SUS), que gastou cerca R$ 449 milhões em 2024 com internações de vítimas de sinistros de trânsito no Brasil, sendo R$ 100 milhões apenas no estado de SP (G1, 2025).

E por último o desenvolvimento do trabalho com essa iniciativa atende à Estratégia Federal de Governo Digital, que visa oferecer políticas públicas e serviços de melhor qualidade, mais simples e acessíveis ao cidadão através da tecnologia. A disponibilidade de dados públicos de sinistros são cada vez de mais fácil acesso para que estimule o uso de análise e ciência de dados na tomada de decisões públicas (Brasil, 2021).

1.3 Objetivos

1.3.1 Objetivo Geral

Analisar as ocorrências de acidentes de trânsito na cidade de Bauru através de técnicas de Ciência de Dados a fim de identificar possíveis padrões espaciais, demográficos e temporais.

1.3.2 Objetivos Específicos

- Caracterizar o perfil mais suscetível e vulnerável a acidentes e fatalidades;
- Identificar os locais de maior incidência de acidentes através do mapeamento;
- Analisar a correlação entre as características dos acidentes e suas ocorrências; e
- Predizer a fatalidade de um acidente através de algoritmos de Aprendizado de Máquina;

2 Fundamentação Teórica

O trabalho aborda diferentes conceitos e assuntos, como a segurança viária. Segundo a legislação brasileira, a definição oficial de sinistro de trânsito é "evento que resulta em dano ao veículo ou à sua carga e/ou em lesões a pessoas ou animais e que pode trazer dano material ou prejuízo ao trânsito, à via ou ao meio ambiente, em que pelo menos uma das partes está em movimento" (BRASIL, 2023). Por conta da complexidade do assunto, Ciência de Dados é aliada para resolução de problemas dessa escala, pois engloba uma variedade de conceitos, princípios, algoritmos e processos de extrair dados não óbvio e padrões relevantes de grande conjunto de dados, e esses padrões extraídos são só úteis se eles dão a visão do problema que seja possível resolucioná-lo (KELLEHER; TIERNEY, 2018).

O processo começa na Análise Exploratória de Dados, que pode ser definida como o conjunto de técnicas que envolvem coleta, exploração, descrição e interpretação dos dados, em que estes métodos permitem a visualização dos padrões identificados pela exploração do conjunto de dados (LOPES et al, 2019). Para dar continuidade ao trabalho e realizar a modelagem e previsão, a implementação de técnicas de Aprendizado de Máquina são essenciais. O aprendizado e as técnicas do Aprendizado de Máquina são orientados e geram hipóteses a partir de dados (LUDERMIR, 2021). O método é o indutivo, uma forma de obter conclusões de um conjunto de exemplos pela inferência lógica (MONARD; BARANAUSKAS, 2003). Existem diferentes tipos de aprendizado para essas técnicas, e para o seguinte trabalho são utilizados dois: supervisionado, em que o algoritmo gera uma função que mapeia as entrada para as saídas desejadas, como nos problemas de classificação, que é utilizada no trabalho; e o não-supervisionado, em que os exemplos de entrada não são fornecidos, como o DBSCAN (AYODELE, 2010).

A técnica supervisionada utilizada é Árvore de Decisão, cujo método de aprendizado é aproximar a função alvo do seu valor discreto. Pode ser representado como conjunto de "se-então" (MITCHELL, 1997). Em outras palavras, a unidade é um classificador que determina sua saída de acordo com seus dados, mas que muitas vezes é possível aumentar a acurácia ao combinar o seu conjunto (KINGSFORD; SALZBERG, 2008). A Árvore é construída através de um conjunto de instâncias como o modelo de Divisão e Conquista, onde cada teste reparte as instâncias até que sua folha seja constituída por uma única classe (QUINLAN, 1996).

Também é utilizada XGBoost (*eXtreme Gradient Boosting*), que é um modelo aprimorado de *Gradient Boost*, e que possui como diferencial o potencial de escalabilidade e aproveitamento de recursos computacionais de forma otimizada em comparação a modelos antigos, além de alta taxa de escolha e sucesso em competições de Ciência de Dados (CHEN; GUESTRIN, 2016).

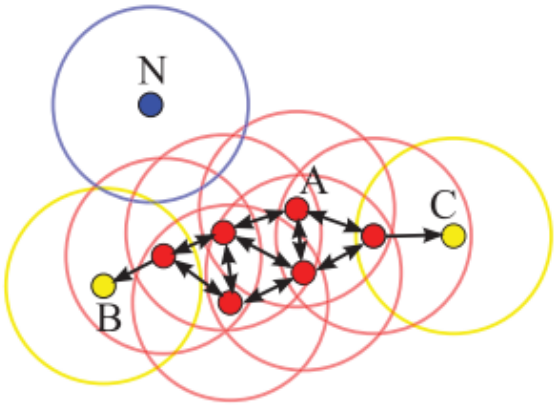
E de método não supervisionado é utilizado DBSCAN (*Density Based Spatial Clustering of Applications with Noise*), que é baseado em formação de *clusters* de tamanho arbitrário em conjunto de dados e de grande tamanho, que também se diferencia pelo conhecimento mínimo de requisitos de entrada, que é um problema a se determinar num grande conjunto de dados (KHAN et al, 2014)

Esses requisitos de entrada são: *minPts*, que indica a quantia limite de vizinhos; e ε que indica o raio de um ponto. Schubert et al. (2017) explica que para a construção dos *clusters* são feitas as seguintes etapas:

- para nomear Ponto Central um ponto arbitrário deve possuir *minPts* pontos dentro do seu raio ε para se formar um *cluster*, e todos esses pontos irão pertencer ao grupo;
- é feita a verificação de Ponto Central aos pontos inseridos ao *cluster*: se for considerado o processo de expansão se repete. Entretanto, se dentro de seu raio não atingir o limite ele é chamado de Ponto de Fronteira e a expansão não continua nesse ponto; e
- os pontos que são inalcançáveis ou insuficientes para formar um *cluster* são chamados de Ruído.

A Figura 1 indica o processo da classificação de alguns desses pontos. O *MinPts* é igual a 4, e os pontos B a C pertecem ao mesmo *cluster*, entretanto ambos são considerados Pontos de Fronteira por não atingirem a quantia mínima de densidade. O ponto A e outros pontos vermelhos são os Pontos Centrais. O ponto N é definido como ruído.

Figura 1 – Representação de formação de um *cluster*



Fonte: SCHUBERT et al. (2017)

3 Metodologia

Neste capítulo são descritas as etapas e os métodos utilizados para o desenvolvimento do trabalho, a fonte da base de dados, as tecnologias utilizadas e como foram usadas.

O trabalho consiste em uma pesquisa aplicada, que possui fim de gerar resultados e soluções de um problema de relevância social, como a segurança no trânsito.

Os dados utilizados no trabalho são provenientes de Infosiga, IBGE e INMET, e compreendem tabelas de formato CSV — que significa *Comma Separated Values*, um tipo de arquivo de texto que armazena dados tabulares — referentes aos acidentes, perfil das vítimas e dados meteorológicos, e arquivos de formato *shapefile*.

O trabalho foi desenvolvido no Google Colaboratory utilizando a linguagem Python com auxílio de Excel. Diversas bibliotecas foram utilizadas a fim de desenvolver cada etapa do projeto, desde a manipulação e visualização de dados até criação de modelos de Aprendizado de Máquina.

3.1 Fonte de Dados

Esta seção apresenta a origem dos dados adquiridos, para qual finalidade foram utilizados durante o desenvolvimento do trabalho e quais as informações relevantes que estão contidas.

3.1.1 Acidentes

Os dados referentes aos acidentes ocorridos em Bauru são originários da página Infosiga, plataforma de estatística e dados referentes aos acidentes do DETRAN. Disponível[1] para baixar e atualizado todo mês, nela está presente a tabela no formato CSV e estão presentes informações referentes aos sinistros, como data, localização e caráter do acidente.

3.1.2 Mapa

Para visualizar e mapear corretamente os acidentes dentro dos limites administrativos do município foram utilizados o conjunto de arquivos de extensão *shapefile*, que estão disponibilizados no site do IBGE. Alguns dos arquivos que estão disponíveis permitem o mapeamento e visualização de algumas diferentes malhas, como distritos, setores censitários e bairros.

[1] Disponível em: <https://infosiga.detran.sp.gov.br/#referencia>

3.1.3 Perfil Demográfico

O perfil das vítimas envolvidas nos acidentes encontra-se no mesmo arquivo compactado do Infosiga nos mesmos recortes temporais em formato CSV, e possui informações como idade, sexo e ocupação, além das informações relacionadas ao sinistro.

3.1.4 Condições Meteorológicas

A fim de enriquecer as *features* para que o modelo de Aprendizado de Máquina tenha dados que possam ser relevantes e auxiliar no processo de predição foram adicionados os dados meteorológicos como precipitação, umidade e temperatura através de engenharia de *features*. A consulta pode ser feita online e corresponde a dados no período de seis meses. Para o objeto do estudo, foram geradas 14 tabelas do período 01/01/2019 a 30/09/2025.

3.2 Ambiente de Desenvolvimento e Ferramentas

Nesta seção serão abordados as ferramentas, linguagens e ambiente utilizados para o desenvolvimento do trabalho.

3.2.1 Google Colaboratory

O ambiente em que foram feitas as análise de dados, desenvolvimento dos modelos de Aprendizado de Máquina e geração de dados visuais é o Google Colaboratory, que possui facilidade no uso e capacidade de processamento aumentada. Permite a criação e edição de arquivos de formato ipynb, que vem de Interactive Python Notebook, e possibilita a execução dos códigos de Python, geração de gráficos e textos de forma organizada, rápida e simples.

3.2.2 Excel

Para auxiliar na manipulação e visualização de alguns arquivos CSV foi utilizado a ferramenta Excel da Microsoft, que permite manipular, filtrar e visualizar os dados de forma mais simples e rápida em comparação ao ambiente do Google Colaboratory.

3.2.3 Python

Python é uma linguagem interpretada de alto nível e que suporta as aplicações da pesquisa, como análise de dados e implementação de Aprendizado de Máquina, além do vasto suporte de bibliotecas para Ciência de Dados que o tornam mais acessível e preferível para se desenvolver.

3.2.3.1 Bibliotecas

Para o desenvolvimento são utilizadas diversas bibliotecas que são essenciais para as atividades do projeto, desde a manipulação, limpeza e visualização dos dados, até o desenvolvimento e implementação de modelos de Aprendizado de Máquina.

A fim de importar, manipular, filtrar, tratar e outras operações diretas com os dados a biblioteca Pandas se torna fundamental. Os dados são convertidos em *DataFrames*, que é a estrutura de dados bidimensional fundamental para a análise e ciência de dados.

Para o escopo do projeto, que trata de acidentes geolocalizados, utiliza-se a biblioteca GeoPandas que mantém a funcionalidade e estrutura de *DataFrame* do Pandas, entretanto com o adicional de coluna *geometry*, que armazena informações espaciais como linhas, coordenadas e pontos. E por fim, para utilização de algumas funções matemáticas e operações para auxiliar no desenvolvimento a biblioteca NumPy foi escolhida.

Na etapa de Análise Exploratória de Dados, que será abordada mais a frente, a geração de gráficos é imprescindível pois permite que a visualização dos dados seja mais clara e, consequentemente, análises e interpretações melhores orientadas. E para isso, as bibliotecas Matplotlib e Seaborn, que é derivada da primeira, são utilizadas para a geração de diferentes gráficos, como gráfico de barra, dispersão e boxplot.

Para a geração e visualização do mapa de Bauru e as diferentes subdivisões administrativas, a biblioteca Folium é a que atende essa necessidade, pois gera mapas interativos e de fácil visualização dentro do ambiente Google Colaboratory.

E por último, também usando a técnica de engenharia de *features*, a biblioteca OSMnx foi utilizada para a classificação dos tipos de vias dos acidentes, já que a biblioteca permite baixar, analisar e visualizar as informações dos mapas de OpenStreetMap, um projeto colaborativo para a criação do mapa do mundo, e que contém dados urbanos relevantes como ruas e *tags*.

A biblioteca Scikit-learn é a principal biblioteca do Python para a criação de modelos de Aprendizado de Máquina. Foram utilizadas diversas funções para a implementação, como criação do algoritmo de Árvore de Decisão, métricas de desempenho, validação cruzada e PCA.

A biblioteca imblearn, que é compatível com o Scikit-learn, é crucial para tratatamento de dados desbalanceados com técnicas de *oversample* ou *undersample*. Também fornece pipeline que garante uma aplicação correta das técnicas de reamostragem no seu conjunto de dados.

E por último, a biblioteca XGBoost que importa todas funções referentes a esse modelo de Aprendizado de Máquina, como a instanciação, teste, treinamento e métricas de desempenho.

3.3 Procedimentos Metodológicos

Esta seção abordará os procedimentos adotados para o desenvolvimento do trabalho e o que foi realizado em cada etapa do projeto. Compreende a filtragem e tratamento do conjuntos de dados, aplicação da engenharia de *features*, a análise exploratória de dados e a implementação de modelos de Aprendizado de Máquina.

3.3.1 Pré-Processamento de Dados

Corresponde ao processo de filtrar e tratar os dados para que estes estejam adequados para a etapa de AED e para a implementação das técnicas de Aprendizado de Máquina, que devem possuir apenas valores numéricos e sem valores nulos.

O quadro 1 apresenta o nome, principais informações, fonte e formato dos arquivos utilizados:

Quadro 1 – Arquivos utilizados

Nome do arquivo	**Principais informações**	**Fonte**	**Formato**
pessoas_2015-2021 pessoas_2022-2025 sinistros_2015-2021 sinistros_2022-2025	Dados dos acidentes e perfil das vítimas (id_sinistro, data, tipo de acidente, etc.)	Infosiga	CSV
Amostra_Pessoas_35_outras	Informações de menor escala da população de Bauru	IBGE	TXT
BAURU_area_de_ponderacao	Mapa das áreas de ponderação de Bauru	IBGE	DBF, PRJ, SHP, SHX
bauru_meteorologia[2]	Informações meteorológicas por hora e dia de Bauru	INMET	CSV

Fonte: Elaborado pelo autor

3.3.1.1 Filtragem dos Dados

Adquiridos todas os conjuntos de dados dos sites de Infosiga, IBGE e INMET, é necessário realizar a filtragem. Os arquivos referentes aos acidentes e pessoas do Infosiga e do mapa da cidade de Bauru do IBGE correspondem ao estado de São Paulo e os das condições meteorológicas de INMET corresponde ao município de Bauru, este já estando diretamente filtrado. Tendo em vista que Bauru é o campo da pesquisa, a primeira parte foi filtrar dados apenas referentes a cidade nos arquivos do Infosiga.

3.3.1.2 Tratamento de Dados

Com os arquivos filtrados para o município de Bauru, é necessário fazer o tratamento desses dados, a fim de que todas as colunas tenham valores com o seu formato correspondente e que não propague algum tipo de erro no modelo e na análise, como o caso de valores nulos.

Em alguns dos conjuntos de dados da pesquisa foi observado que em algumas células continham valores nulos, o que é incompatível para o treinamento dos modelos de Aprendizado de Máquina. Como exemplo, existe um padrão para os valores de chuva, logo na coluna *precipitacao* foi utilizada a interpolação linear, já que o valor da chuva tende a seguir um padrão de variação por hora, sem aumentar ou diminuir drasticamente. Entretanto em alguns casos a recuperação e tratamento desses dados é inviável e a melhor alternativa é a eliminação, como no caso da ausência das coordenadas geográficas, que não permitem mapear o ponto de um acidente.

Alguns dados possuem conteúdo de tipo *Object*, e que por conta do formato a leitura do dado não é feita corretamente, logo sendo necessário utilizar a função *replace()*. Por exemplo, todas as célula da coluna *hora_sinistro* da tabela pessoas_2022-2025 estavam com a formatação de tempo incorreta, gerando o texto *00/01/1900* em frente do horário. Outro caso recorrente e que é presente nesse trabalho é sobre o separador decimal, já que no Brasil utiliza-se a vírgula entretanto muitas linguagens de programação, que são comumente atribuídas em inglês, utiliza-se o ponto.

Em algumas linhas dos arquivos de sinistro e de pessoas do Infosiga existem incosistências nos seus dados após a realização do *merge()*, como presença de campos nulos, declarados como não disponíveis ou com informações que gera conflito. Há quatro conjuntos de colunas nessas tabelas em que foi aplicada a imputação lógica, que são:

- *qtd_*(tipo de veículo): indica a quantidade de veículos envolvidos num determinado sinistro;
- *tp_sinistro_*(categoria do sinistro): indica o tipo do sinistro, como atropelamento, colisão, capotamento, entre outros;
- *tipo_veiculo_vitima*: indica o tipo de veículo em que a vítima estava presente no momento do acidente; e
- *tipo_de_vitima*: indica a categoria da vítima, como condutora, passageira ou pedestre.

Em uma linha de dado onde indica os dados da vítima e do sinistro consta que *tp_sinistro_atropelamento* é igual a 1, entretanto o *tipo_de_vitima* consta como "NAO DISPONIVEL" ou nulo no lugar de pedestre.

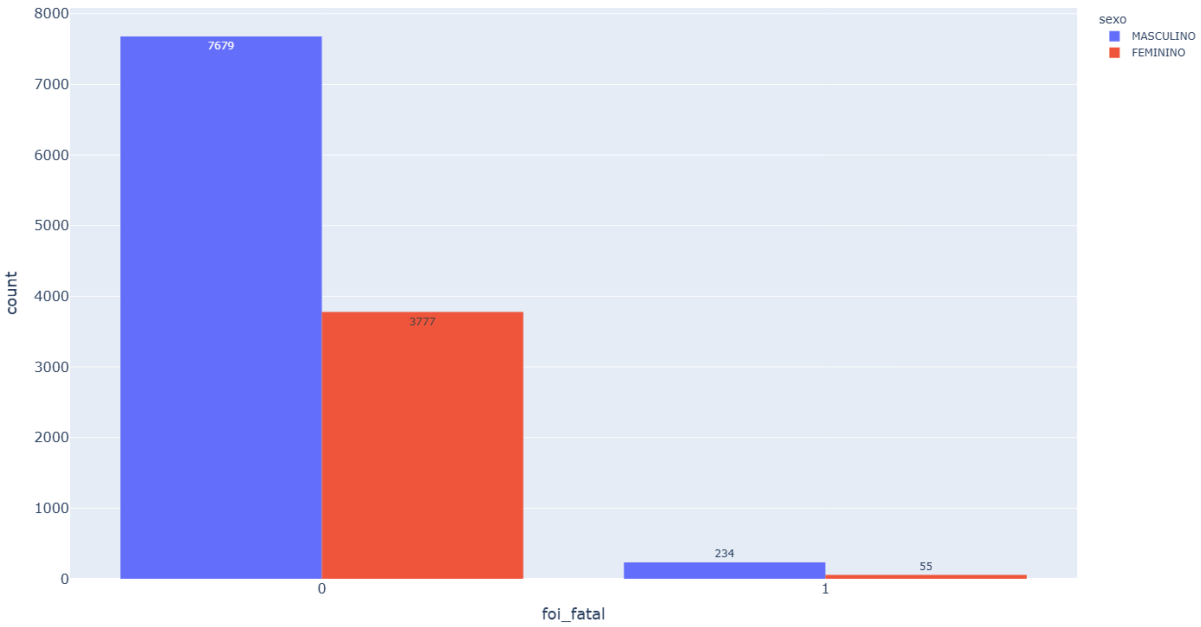
3.3.2 Análise Exploratória de Dados

Esta subseção realiza a primeira análise dos dados adquiridos e tratados, realizando visualizações e análise por meio de gráficos e mapas.

3.3.2.1 Análise na Fatalidade

A primeira análise a ser realizada é justamente a variável alvo, em que buscou-se compreender a relação de número de acidentes (vítimas) e fatalidades. Foi feita a divisão por sexo para ilustrar que existe uma diferença entre o perfil de vítima que tende a sofrer o acidente, como mostra a Figura 2, sendo o principal sendo a população masculina.

Figura 2 – Número de vítimas não fatais e fatais por sexo



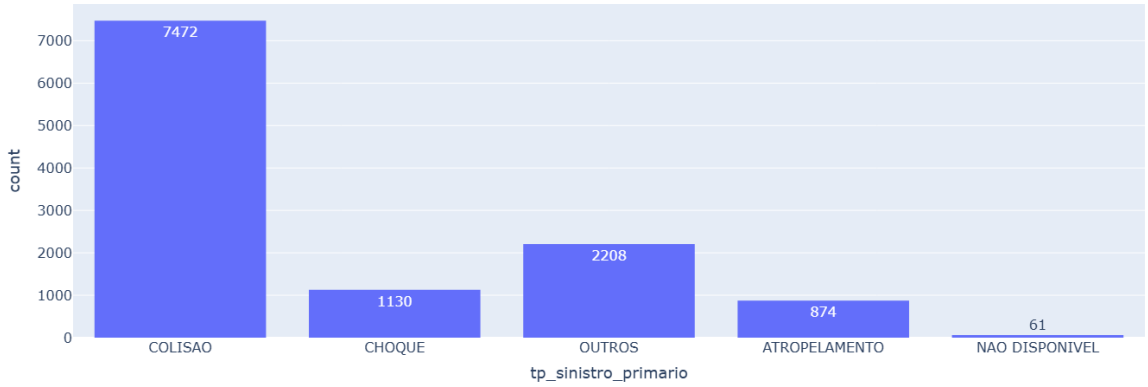
Fonte: Elaborada pelo autor

3.3.2.2 Análise de Acidentes

Nessa subseção será feita a análise nos acidentes. Na Figura 3 é possível perceber que a principal causa de acidente em Bauru é por colisão, que é o sinistro envolvendo dois veículos em movimento, e corresponde a mais de 70% dos casos de sinistros na cidade.

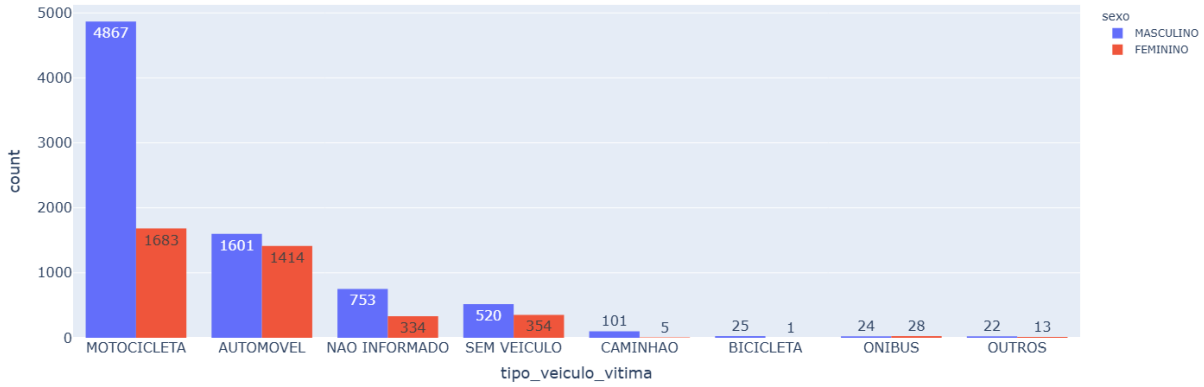
Entre os tipos de veículo, as motocicletas são o veículo com maior número de vítimas, em especial os motociclistas homens que possuem mais que o dobro de vítimas que as mulheres, como é possível observar pela Figura 4.

Figura 3 – Número de vítimas por tipo de sinistro



Fonte: Elaborada pelo autor

Figura 4 – Número de vítimas por tipo de veículo e dividido em sexo



Fonte: Elaborada pelo autor

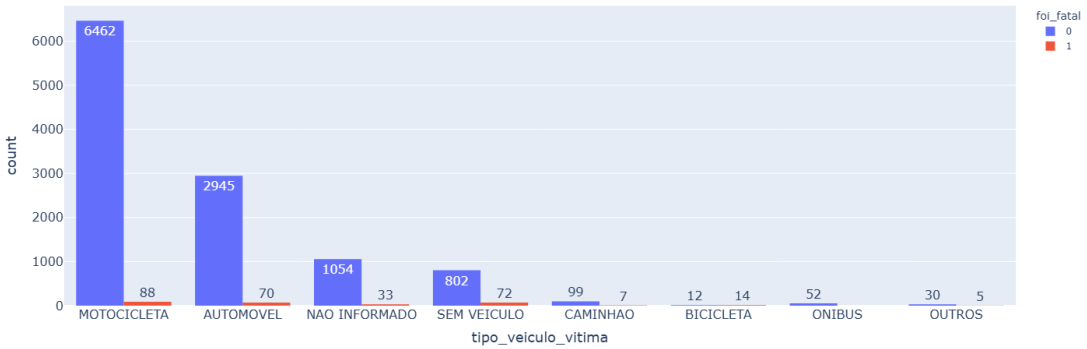
Por fim, há uma relação entre a quantidade de acidentes e de vítimas fatais. Ciclistas é o grupo com maior fatalidade, com mais de 53% de fatalidade em todos os acidentes registrados na cidade de Bauru, como demonstra a Figura 5.

3.3.2.3 Análise Temporal

Na análise temporal irá ser feita a análise sobre o padrão temporal de ocorrência dos acidentes e como é sua distruibuição.

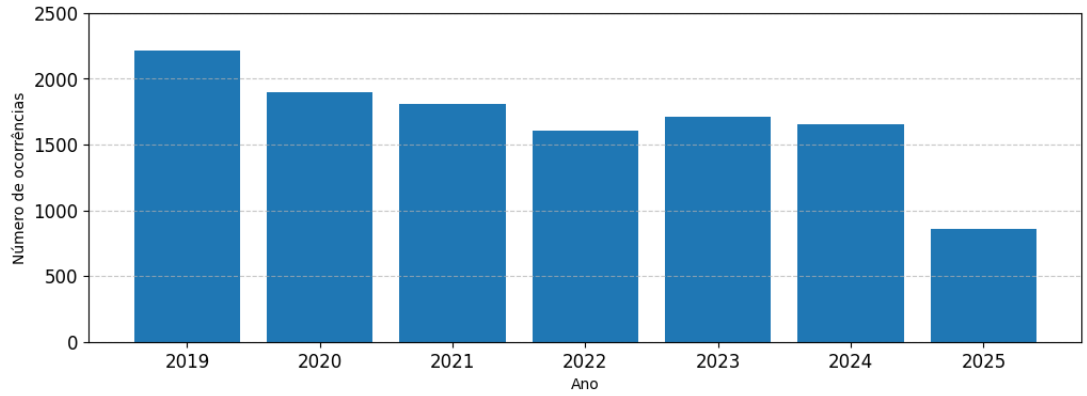
Na Figura 6 é possível observar que existe uma tendência de diminuição dos acidentes em Bauru ao decorrer dos anos, tendo atingido seu ápice em 2019. Vale ressaltar que os dados estão atualizados até 30/09/2025, portanto os dados para o ano de 2025 estão 25% incompletos.

Figura 5 – Número de vítimas fatais por veículo



Fonte: Elaborada pelo autor

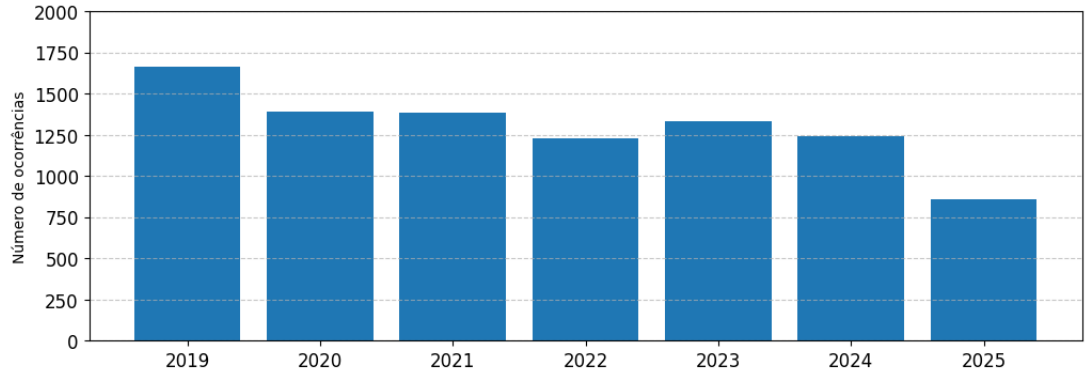
Figura 6 – Número de acidentes por ano



Fonte: Elaborada pelo autor

Entretanto a Figura 7 e a Tabela 1 mostram que as ocorrências de acidentes estão menores no mesmo período em comparação a anos anteriores, o que representa uma redução de 31% de acidentes para o ano de 2025.

Figura 7 – Número de acidentes por ano até Setembro



Fonte: Elaborada pelo autor

[2] Arquivo final gerado pelo autor

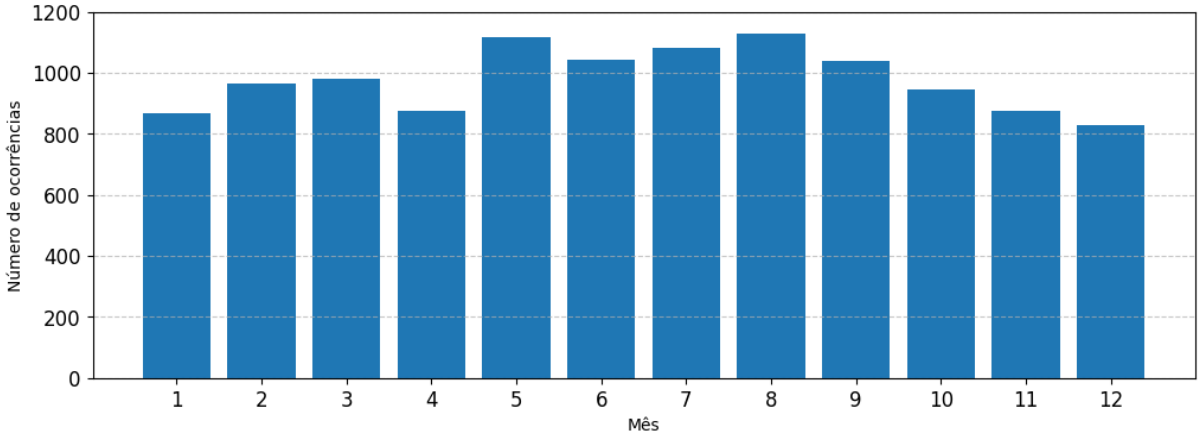
Tabela 1 – Número de acidentes até Setembro de cada ano

2019	2020	2021	2022	2023	2024	2025
1665	1389	1382	1226	1331	1245	858

Fonte: Elaborada pelo autor

Na Figura 8 é possível observar a quantia de acidentes ocorridos a cada mês de forma cumulativa. Apesar de possuir valores maiores no período entre maio a agosto, não é possível apontar uma possível causa ou correlação apenas com essa análise.

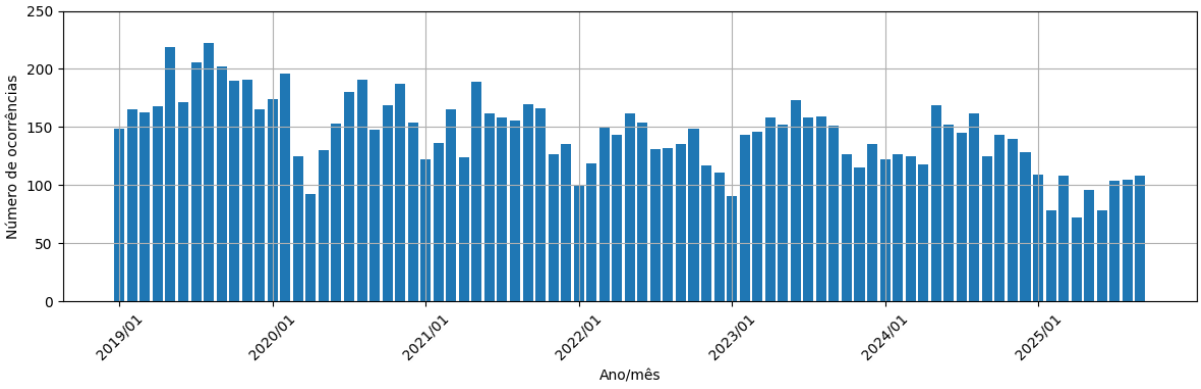
Figura 8 – Número de acidentes cumulativos por mês



Fonte: Elaborada pelo autor

Na distribuição da Figura 9, que está mais granulada em comparação aos outros, é mais difícil realizar a percepção de algum padrão no conjunto de dados. Mas como observado anteriormente, as ocorrências em 2025 estão abaixo da média, além de ser possível perceber que no mês 3/2020 a ocorrência é mais baixa por conta da quarentena causada pela pandemia da Covid-19.

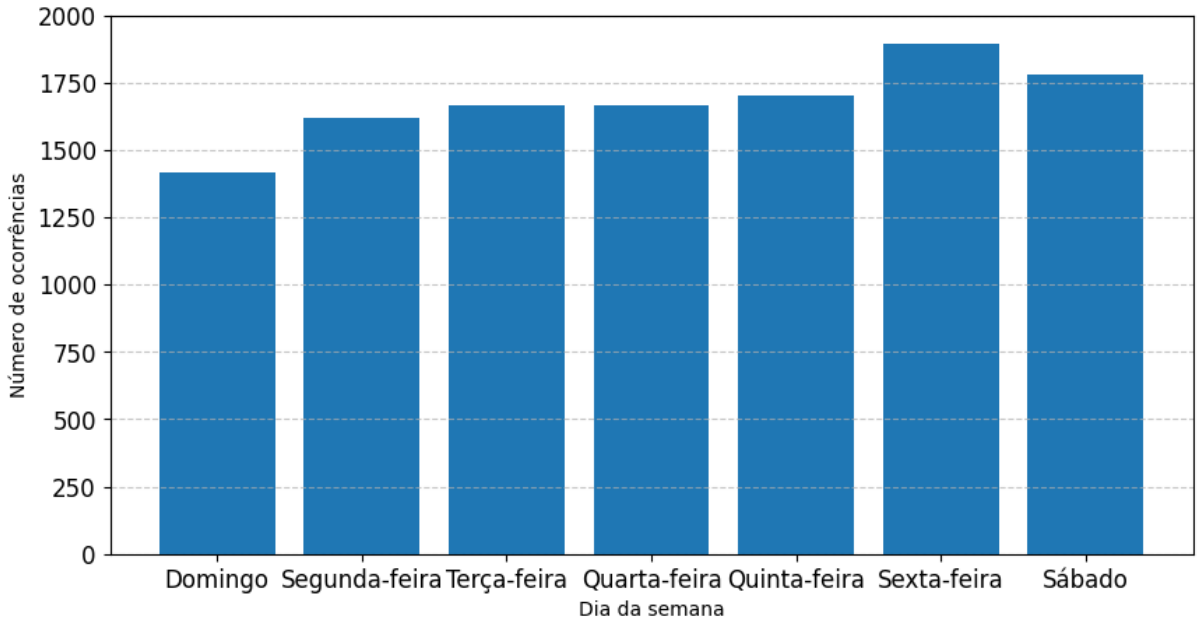
Figura 9 – Número de acidentes por ano/mês



Fonte: Elaborada pelo autor

Na Figura 10 é perceptível um padrão: o número de ocorrências é mais baixo no domingo, dia em que há a menor locomoção de pessoas, e esse número aumenta gradativamente até chegar na sexta-feira, dia da semana em que há maior registro de acidentes, causa que pode ser atribuída pela mobilidade de pessoas por conta do horário comercial e de pessoas se locomovendo a viagens e outros locais durante o período noturno.

Figura 10 – Número de acidentes por dia da semana



Fonte: Elaborada pelo autor

Na Figura 11 pode-se observar que a ocorrência de acidentes é extremamente baixa de madrugada e se mantém distribuída durante a manhã, tarde e noite.

3.3.2.4 Análise Espacial

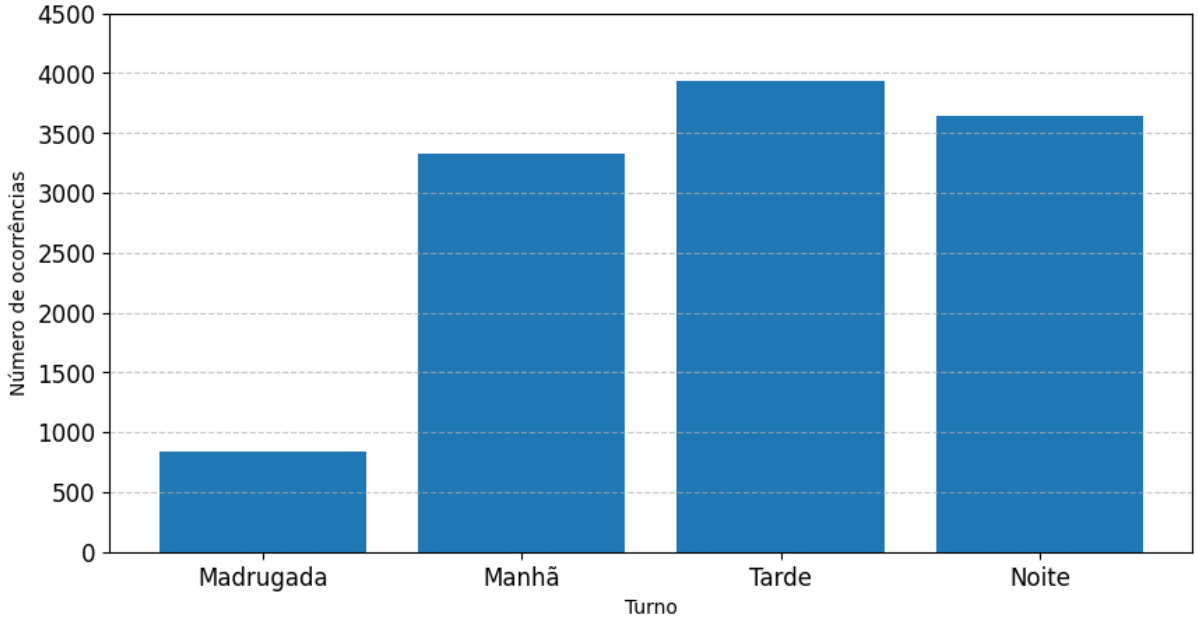
Na Figura 12 cada acidente é representada num ponto e é possível observar sua localização dentro do contorno do limite administrativo da cidade de Bauru.

A Figura 13 mostra o mapa de calor de acidentes no período de 2019 a 2025 em Bauru. Essa primeira análise não permite uma visualização e identificação completa de zonas com o maior risco de acidentes, então no parágrafo 3.3.3.1 será abordado sobre o que fazer para tratar esse problema.

3.3.2.5 Análise de Vítimas

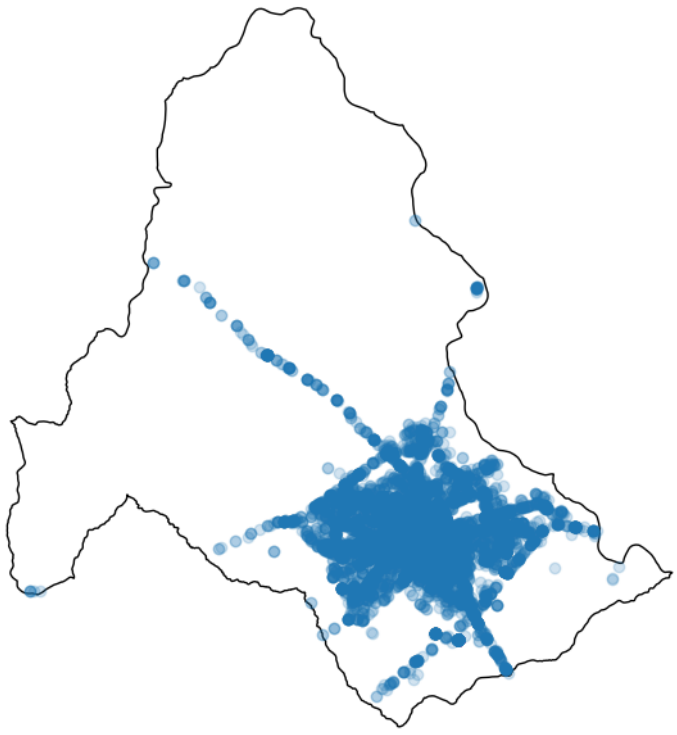
E por fim, serão realizadas duas análises, a primeira sobre o tipo de vítima (condutor, passageiro, pedestre ou não informado) e sobre a distribuição etária dos vitimados.

Figura 11 – Número de acidentes por turno



Fonte: Elaborada pelo autor

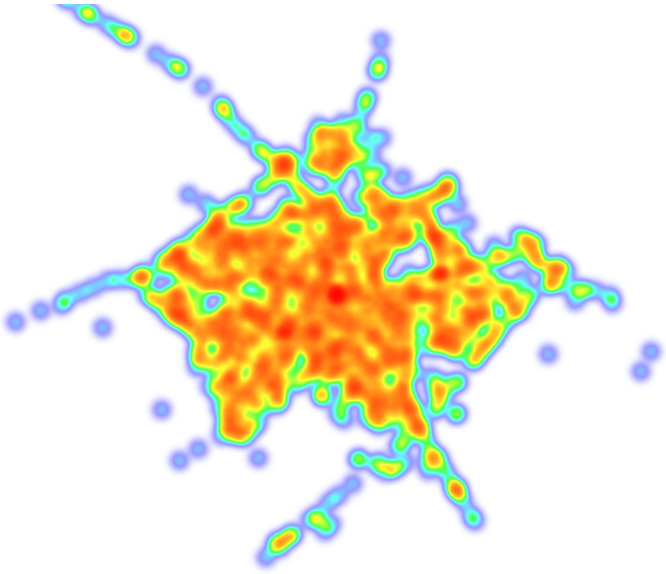
Figura 12 – Mapa de acidentes de Bauru dentro do limite administrativo



Fonte: Elaborada pelo autor

Na Figura 14 observa-se que a maior parte das vítimas são condutores, com grande maioria masculina. Por conta do alto índice de vítimas de motocicletas é possível concluir que

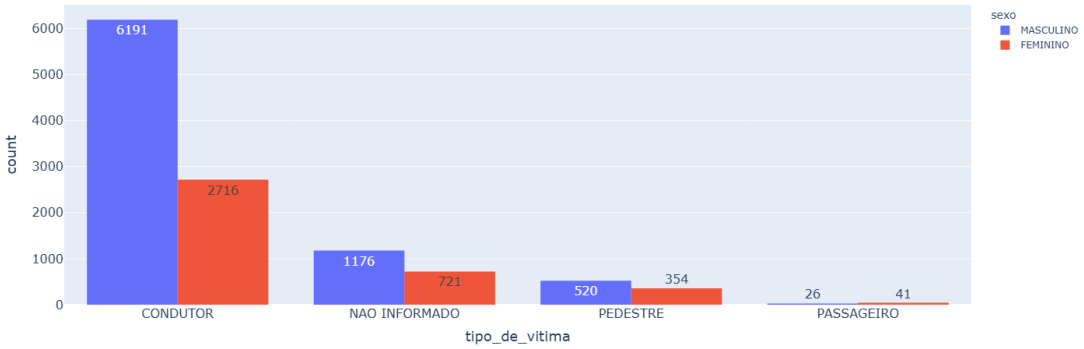
Figura 13 – Mapa de calor dos acidentes na cidade de Bauru



Fonte: Elaborada pelo autor

a maioria dos condutores são motociclistas.

Figura 14 – Número de acidentes por tipo de vítima e sexo



Fonte: Elaborada pelo autor

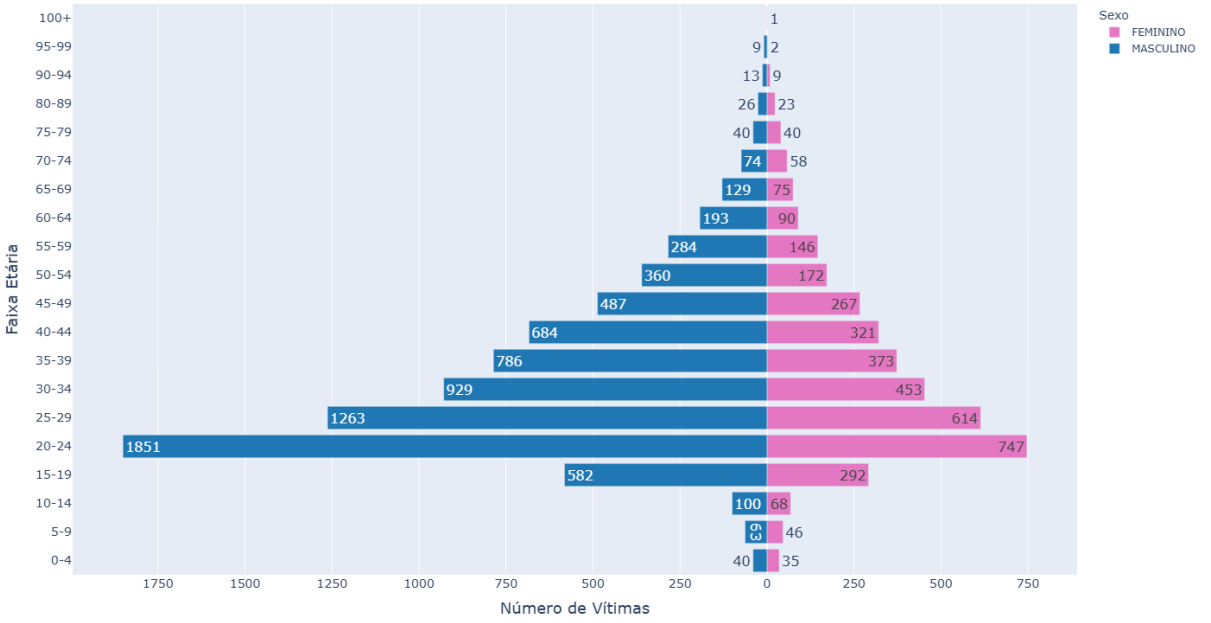
E por fim, a Figura 15 mostra em quais faixas de idade se concentram o maior número de vítimas.

É possível observar que a parcela da população que mais se vitima nos acidentes é a masculina, com ênfase na faixa de 20 a 24 anos, com o número diminuindo ao passar da idade. A mesma observação é possível ser feita à população feminina, entretanto segue uma proporção menor da masculina.

3.3.3 Engenharia de *Features*

Esta subseção trata sobre a Engenharia de *Features*, que consiste na criação e reformulação de colunas existentes em informações que podem ser mais valiosas e importantes para a

Figura 15 – Faixa etária por acidente



Fonte: Elaborada pelo autor

etapa da implementação do Aprendizado de Máquina.

3.3.3.1 DBSCAN

Para auxiliar no processo de criação de novas *features* foi utilizado o algoritmo de DBSCAN a fim de identificar possíveis zonas de perigo (*hotspots*) através das incidências de acidente de trânsito.

Apesar de existir alguns métodos para definir o valor de ε, como o gráfico de distância-K, para esse projeto os parâmetros foram definidos através de tentativa e erro com visualização por um mapa utilizando a biblioteca Folium.

Uma coluna chamada *zona_acidente* é criada e atribuída um valor. Se o valor for -1 indica que o dado é ruído, logo nâo pertence a nenhum *cluster*. Se o valor for positivo indica que o algoritmo atribuiu o dado a um *cluster* com essa numeração.

Após algumas rodadas de teste e alteração nos parâmetros estes foram definidos como: $\varepsilon = 0{,}024$, que consegue focalizar uma região de perigo num tamanho aproximado de algumas quadras; e *min_samples* $= 18$, que indica uma média de incidência de 3 acidentes no local por ano, o que pode ser considerado como um local de atenção.

A Figura 16 indica a numeração da zona e quantidade de vítimas encontradas pelo algoritmo DBSCAN. Aproximadamente metade da localidade dos acidentes das vítimas são consideradas de caso isolado, pois um dos dois parâmetros do algoritmo não foram atendidos,

o que indica baixa reincidência de vítima na determinada região.

Figura 16 – Numeração de zona e quantidade de vítimas por zona

zona_acidente	
-1	5840
1	497
3	274
10	260
40	242
4	176
22	166
103	152
14	129
64	126
60	125
25	107
53	103
93	101
7	97
27	97
0	90
37	88
20	81
59	80

Fonte: Elaborada pelo autor

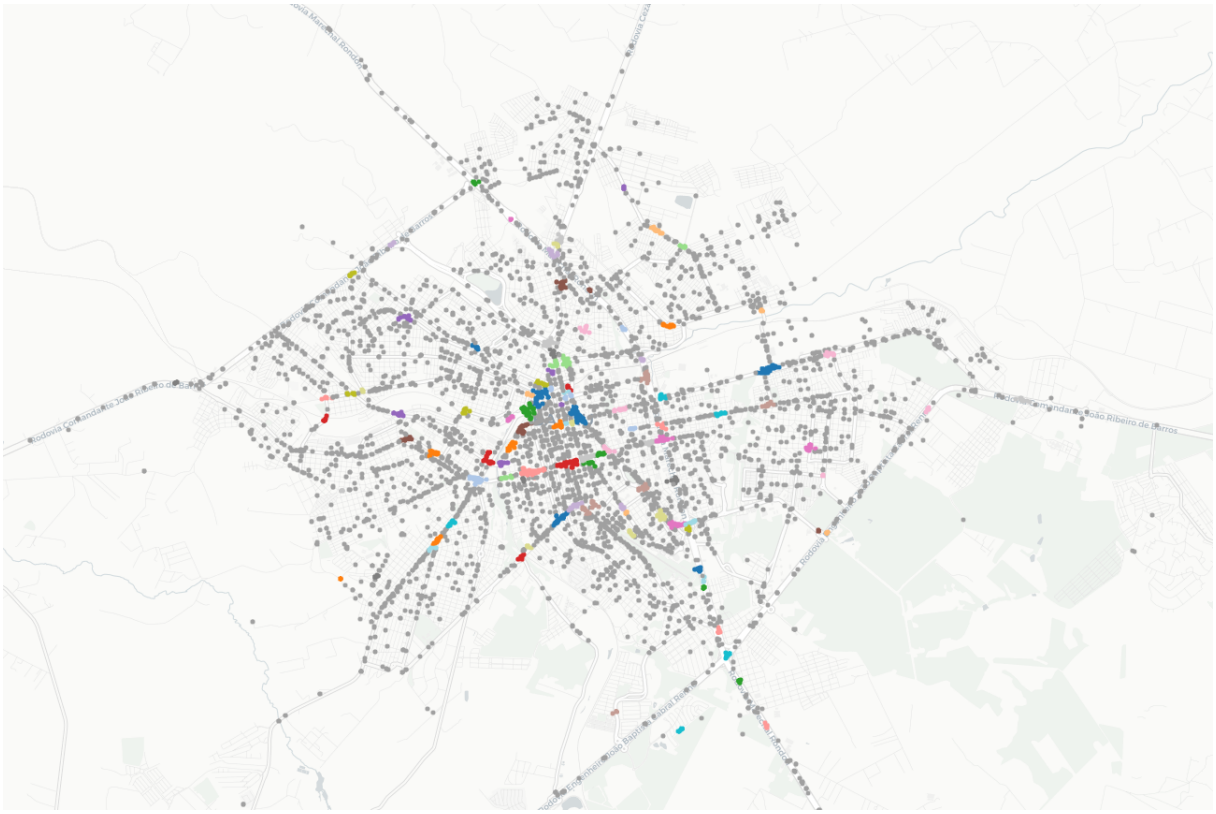
A Figura 17 mostra o mapa de Bauru com alguns dos *clusters* criados, representados por cores distintas.

3.3.3.2 Tipos de Via

Apesar do *DataFrame* de sinistros do Infosiga ter a coluna *tipo_via* ela não possui granularidade para auxiliar no treinamento dos modelos, com suas classificações sendo apenas como "VIAS URBANAS", "ESTRADAS E RODOVIAS"e "NAO DISPONIVEL". Para solucionar esse problema e buscar aumentar a granularidade dos dados foi utilizada a biblioteca OSMnx.

O mapa de Bauru é importado de OSMnx e cada ponto do acidente é mapeado a partir dos valores das coordenadas e atribuído à aresta mais próxima. Após isso, esse mapa adquirido é convertido num *GeoDataFrame*, que contém informações como grafo de origem, destino e tipo de via em inglês. Essa informação é convertida por meio de um dicionário montado para cada tipo de via. A Figura 18 indica o dicionário utilizado, que converte o tipo de via original para o adaptado em português.

Figura 17 – Mapeamento dos acidentes classificadas pelas zonas



Fonte: Elaborada pelo autor

E por último, é feito o *merge()* ao *DataFrame* original, passando as colunas com os tipos de via específicos classificados. Com isso, se torna possível classificar e observar em que tipo de vias urbanas os acidentes tendem a se concentrar e sua fatalidade

3.3.3.3 Categorias de Profissões

A coluna *profissao* do *DataFrame* de pessoas possui um ótimo potencial de encontrar correlação entre a ocupação e o nível de risco associado aos acidentes de trânsito. Entretanto, realizando o filtro e contagem de valores existem mais de 300 profissões diferentes listadas, e algumas com ligeira diferença na escrita mesmo sendo funções semelhantes, como por exemplo motoboy e mototaxista, ou policial e policial militar.

Para tornar esse tipo de dado, que está extremamente granulado, a melhor opção é agrupar em um conjunto, assim diminuindo sua granularidade e encontrar possíveis padrões em determinados setores específicos.

Foi montado um dicionário com categorias de diferentes tipos de profissões para a chave, como saúde, gestão, comércio, transporte, entre outros, e para o valor foi feito um *array* de diferentes empregos da área. Esse método foi capaz de diminuir de 300 variáveis diferentes para apenas 19, permitindo que os modelos de Aprendizado de Máquina tenham mais facilidade

Figura 18 – Dicionário das vias da biblioteca OSMnx ao equivalente às vias brasileiras

```
mapa_tipo_via = {
    'motorway': 'Autoestrada',
    'trunk': 'Rodovia Principal',
    'primary': 'Avenida Principal',
    'secondary': 'Avenida Secundária',
    'tertiary': 'Via Coletora',
    'unclassified': 'Via Local',
    'residential': 'Rua Residencial',
    'service': 'Via de Serviço',
    'motorway_link': 'Acesso de Autoestrada',
    'trunk_link': 'Acesso de Rodovia',
    'primary_link': 'Acesso de Avenida',
    'secondary_link': 'Acesso de Avenida',
    'tertiary_link': 'Acesso de Via',
    'living_street': 'Área Residencial'
}
```

Fonte: Elaborada pelo autor

para encontrar padrões.

3.3.3.4 Faixas Etárias

Outro caso em que a alta granularidade pode não influenciar positivamente no treinamento do modelo é a coluna *idade*, já que em teoria podemos ter mais de 100 variações de valor. Para contornar isso, foram feitas três novas colunas: *faixa_etaria_geral*, *faixa_etaria_especifica* e *faixa_etaria_demografica*.

A *faixa_etaria_geral* possui variáveis categóricas nominais separadas em três grupos: *JOVEM* (0–17), *ADULTO* (18–59) e *IDOSO* (60+). Permite a classificação de idade em grandes grupos principais.

A *faixa_etaria_especifica* também possui variáveis categóricas nominais porém divididas em sete grupos: *CRIANÇA* (0–11), *ADOLESCENTE* (12–17), *JOVEM ADULTO* (18–29), *ADULTO* (30–59), *IDOSO* (60–74), *ANCIÃO* (75–89), *VELHICE EXTREMA* (90+). Permite a classificação de idade em faixas de idade menores, subdividindo a primeira divisão realizada.

Por último, há a *faixa_etaria_demografica*, que faz a separação das idades em grupos de cinco em cinco anos. Por existir uma ordem e proporção na distância de um valor a outro dentro da coluna a variável pode ser representada em números, apesar de ser uma variável categórica ordinal. Como 0 (0–4), 1 (5–9), 2 (10–14) e assim sucessivamente.

Na Figura 19 são mostradas algumas linhas do *DataFrame* principal e tratado com a coluna *idade* já existente com as novas colunas adicionadas.

Figura 19 – Trecho do *DataFrame* principal com a coluna *idade* e as novas adicionadas

	idade	faixa_etaria_geral	faixa_etaria_especifica	faixa_etaria_demografica
0	45.0	ADULTO	ADULTO	9.0
1	28.0	ADULTO	JOVEM ADULTO	5.0
2	22.0	ADULTO	JOVEM ADULTO	4.0
3	16.0	JOVEM	ADOLESCENTE	3.0
4	27.0	ADULTO	JOVEM ADULTO	5.0
5	19.0	ADULTO	JOVEM ADULTO	3.0
6	40.0	ADULTO	ADULTO	8.0
7	31.0	ADULTO	ADULTO	6.0
8	19.0	ADULTO	JOVEM ADULTO	3.0
9	34.0	ADULTO	ADULTO	6.0

Fonte: Elaborada pelo autor

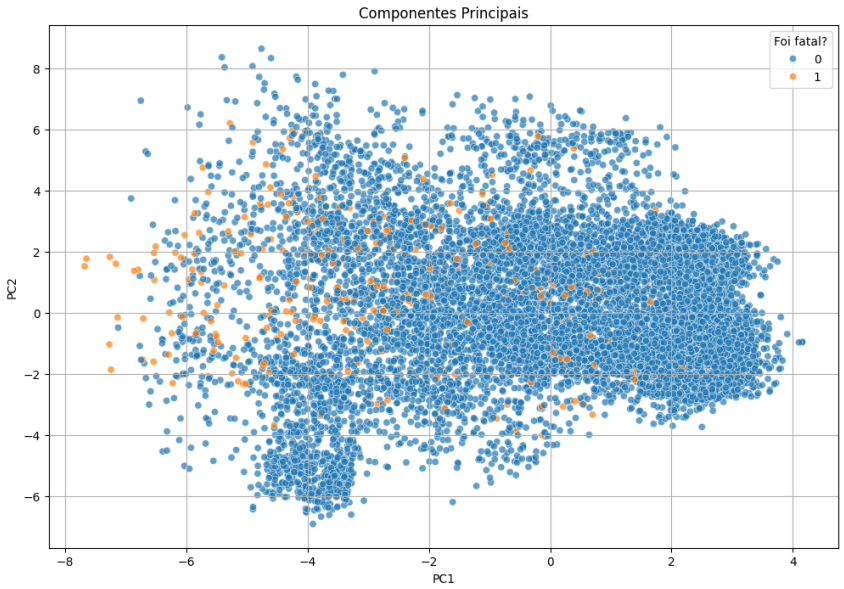
3.3.3.5 *Principal Component Analysis*

Nessa etapa é realizada *PCA*, a fim de reduzir a dimensionalidade das variáveis e encontrar possíveis componentes que correlacionem entre si e indiquem o perfil de fatalidade.

Na primeira análise utilizam-se as variáveis numéricas e categóricas que foram tratadas, inclusive a variável-alvo *foi_fatal*, pois o objetivo é encontrar uma possível correlação dentro do gráfico de dispersão que o modelo fornece.

A Figura 20 apresenta o primeiro gráfico de dispersão de PCA:

Figura 20 – Gráfico do PCA nas variáveis de entrada



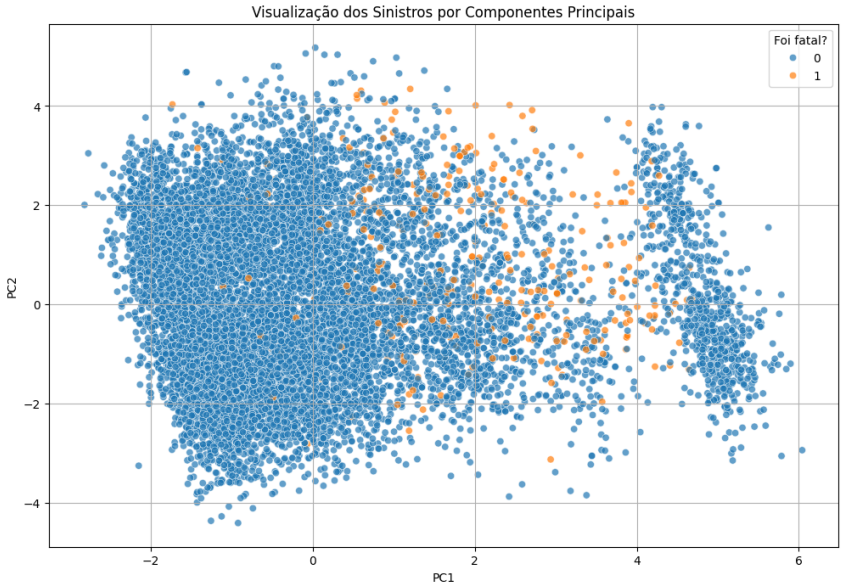
Fonte: Elaborada pelo autor

É possível observar que a variável-alvo não se encontra aglomerada e focada em algum ponto específico, o que pode indicar que algumas colunas criadas a fim de melhorar o

treinamento da árvore pela etapa de Engenharia de *features* podem estar redundantes, como no caso do grupo das colunas de faixa_etaria, que representam a mesma ideia porém em granularidades distintas.

Refinando ao máximo as *features* para que o PCA trabalhe apenas com dados mais essenciais a Figura 21 apresenta o gráfico de dispersão refinado:

Figura 21 – Gráfico do PCA com variáveis reduzidas



Fonte: Elaborada pelo autor

A variável-alvo ainda não se encontra focalizada numa região, entretanto a análise serve como um estudo prévio para o caso do estudo e o funcionamento de suas variáveis.

3.3.4 Implementação de Modelos de Aprendizado de Máquina

A subseção abordará sobre a implementação e a avaliação dos dois modelos de Aprendizado de Máquina deste projeto: Árvore de Decisão e XGBoost.

3.3.4.1 Árvore de Decisão

Com o conjunto de dados tratados, filtrados e adaptados para o modelo, realiza-se as seguintes etapas:

- preparo dos dados: aplica-se a técnica OneHotEncoder para tratar colunas categóricas;
- separação dos dados: 80% dos dados é utilizado para realizar o treinamento do modelo, enquanto o restante é para teste;
- treinamento do modelo: a árvore é instanciada e treinada com o conjunto de dados de treino e depois avaliada com o conjunto de dados de teste; e

- avaliação do modelo: utilizando métricas de avaliação e matriz de confusão avalia-se o modelo gerado.

3.3.4.2 XGBoost

Para o modelo de XGBoost são realizadas as mesmas etapas da Árvore de Decisão, entretanto utiliza-se um conjunto de técnicas e métodos mais sofisticados a fim de assegurar uma performance mais eficiente, precisa e acurada. São utilizadas:

- SMOTEENN, técnica híbrida de *undersample* e *oversample* a fim de lidar com o desbalanceamento dos dados;
- StratifiedKFold, técnica de validação cruzada aplicada no conjunto de dados de treino a fim de melhorar ainda mais a generalização e diminuir o *overfitting*;
- utilização de pipeline a fim de que a instanciação do modelo e aplicação de SMOTEENN sejam feitas de forma ordenada e segura; e
- otimização de hiperparâmetros por busca, como *Grid Search* e *Random Search*.

3.3.4.3 Métricas de Avaliação

Para avaliar o desempenho da Árvore de Decisão e do XGBoost, utilizam-se funções próprias das bibliotecas do Scikit-learn e XGBoost, respectivamente, que incluem:

- *classification_report*
- *precision_score, recall_score, accuracy_score f1_score*

4 Resultados

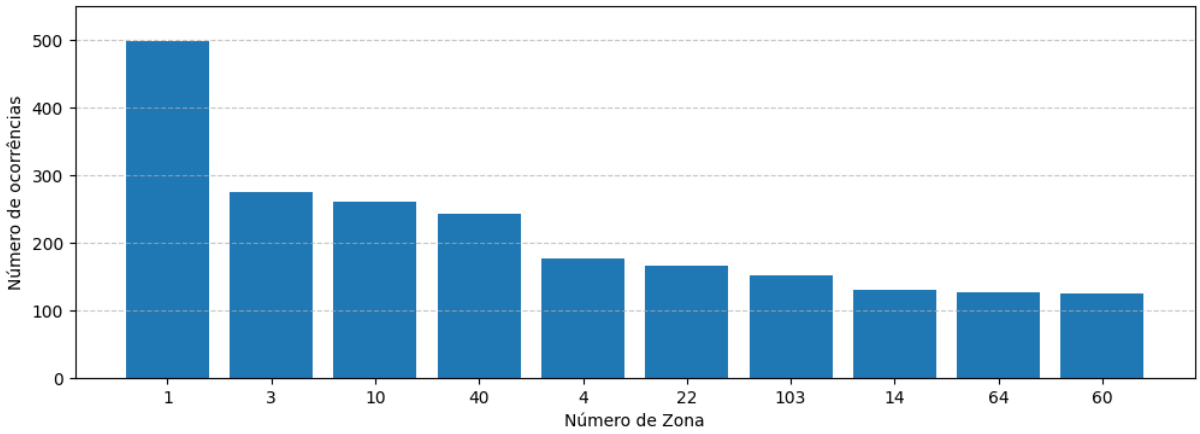
4.1 Zonas de acidente

A fim de encontrar possíveis locais de atenção por conta de reincidência de acidentes foi aplicado o algoritmo de DBSCAN para que zonas fossem gerados em forma de *cluster*, como mencionado e demonstrado na subseção 3.3.3.2.

No total foram criados 110 *clusters* diferentes pela cidade de Bauru, classificando 5.905 de 11.745 vítimas de acidentes, enquanto os 5.840 foram considerados como ruído, ou seja, vítimas de acidentes isolados.

No gráfico da Figura 22 apresentam-se as 10 zonas com os maiores números de vítimas. A média nas 110 zonas é 54 acidentes, o que demonstra como nessas regiões geram muitas vítimas.

Figura 22 – Gráfico de número de vítimas nas zonas de maior perigo



Source: Elaborada pelo autor

É possível pontuar e perceber que a Zona 1 possui um alto número de vítimas nessa região. A seguir a Tabela 2 apresenta os números exatos de vítimas das 10 regiões com mais vítimas.

A seguir serão apresentados os dados e as localizações das quatro Zonas com maiores número de vítimas totais. O Quadro 2 apresenta os dados e localizações das quatro Zonas com o maior número de vítimas, e as Figuras 23, 24, 25 e 26 indicam o mapa e seu *cluster*.

Apesar da alta incidência de acidentes, a taxa de fatalidade dos locais é baixa. O Quadro 3 apresenta os dados e as localizações das quatro Zonas com o maior índice de fatalidade e as Figuras 27, 28, 29 e 30 indicam seu mapa e *cluster*.

Tabela 2 – Tabela com o número de acidentes nas zonas de maior perigo

Zona	Vítimas
1	497
3	274
10	260
40	242
4	176
22	166
103	152
14	129
64	126
60	125

Source: Elaborada pelo autor

Quadro 2 – Informações sobre os *hotspots* de maior número de acidentes totais

Zona	Vítimas	Vítimas fatais	Fatalidade	Local
1	497	6	1,20%	Av. Nações Unidas (trecho entre Av. Rodrigues Alves e R. Ezequiel Ramos)
3	274	0	0%	Av. Rodrigues Alves e R. São Patrício (rotatória com a interseção Av. Eng. Hélio Police)
10	260	1	0,38%	Av. Castelo Branco e R. Felicíssimo Antônio Pereira (interseção com a Av. Ambleto Bertolucci)
40	242	5	2,07%	Av. Duque de Caxias e R. Dr. Lisboa Jr. (trecho entre R. Rubens Arruda e R. Azarias Leite)

Fonte: Elaborado pelo autor

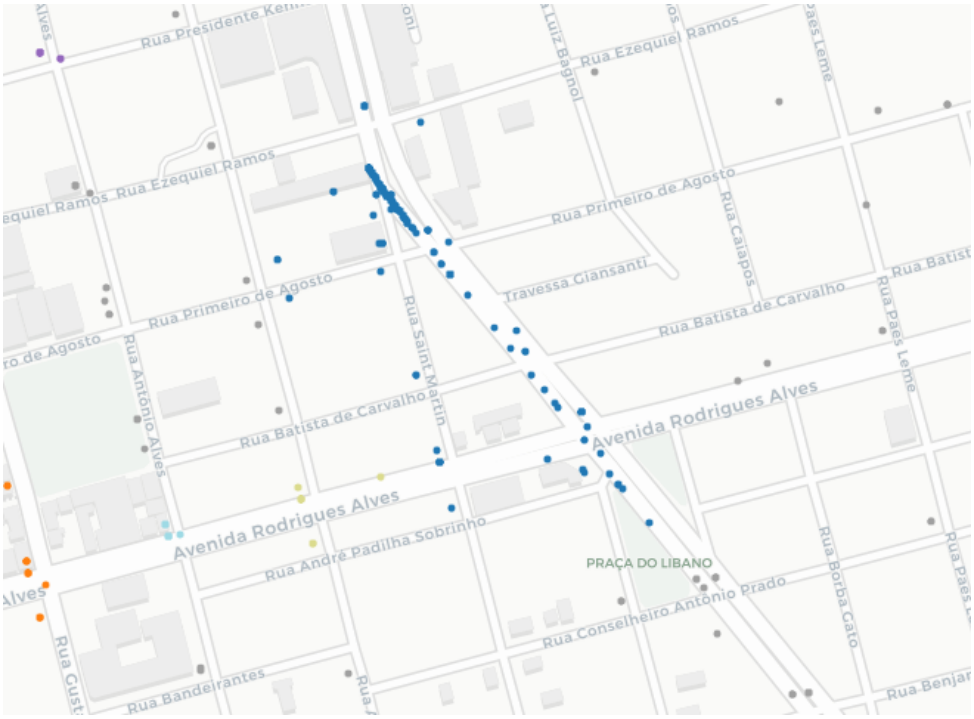
4.2 Desempenho da Árvore de Decisão e XGBoost

A seção abordará sobre as técnicas aplicadas para a escolha dos hiperparâmetros do XGBoost e em seguida sobre os resultados dos modelos da Árvore de Decisão e XGBoost

4.2.1 Árvore de Decisão

A Árvore de Decisão treinada sem uma restrição de profundidade, portanto o modelo ficou com *overfitting* e, consequentemente, queda de desempenho no conjunto de dados de teste. A Figura 31 representa o modelo visual da Árvore de Decisão do trabalho. Por conta de

Figura 23 – Localização de Zona 1



Fonte: Elaborada pelo autor

Quadro 3 – Informações sobre os *hotspots* de maior fatalidade

Zona	Vítimas	Vítimas fatais	Fatalidade	Local
86	33	6	18,18%	Rodovia SP 225 - km 227 Rod. Cmte. João Ribeiro de Barros (Antes da Saída 226)
54	25	4	16,0%	Rodovia SP 294 - km 349 Rod. Cmte. João Ribeiro de Barros (Próximo da Saída 348)
24	30	4	13,33%	Rodovia SP 300 - km 336 Rod. Marechal Rondon (Antes da Saída 336B)
95	17	2	11,76%	Rodovia SP 321 - km 344 Rod. Cesário José de Castilho (Antes da bifurcação)

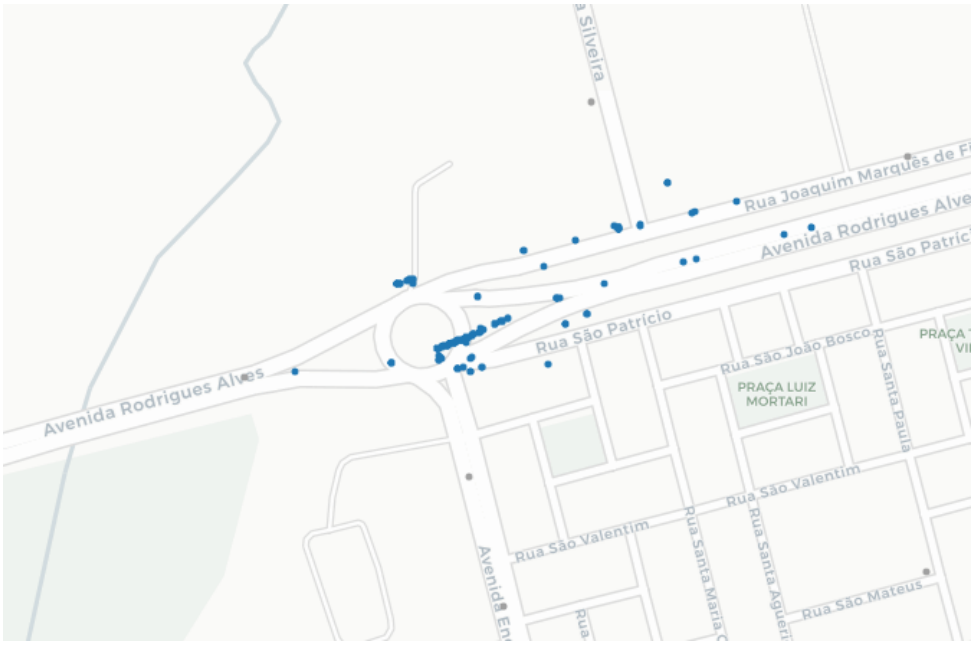
Fonte: Elaborado pelo autor

sua complexidade é impraticável a interpretação dos parâmetros para a classificação do modelo.

4.2.2 Otimização de Hiperparâmetros do XGBoost

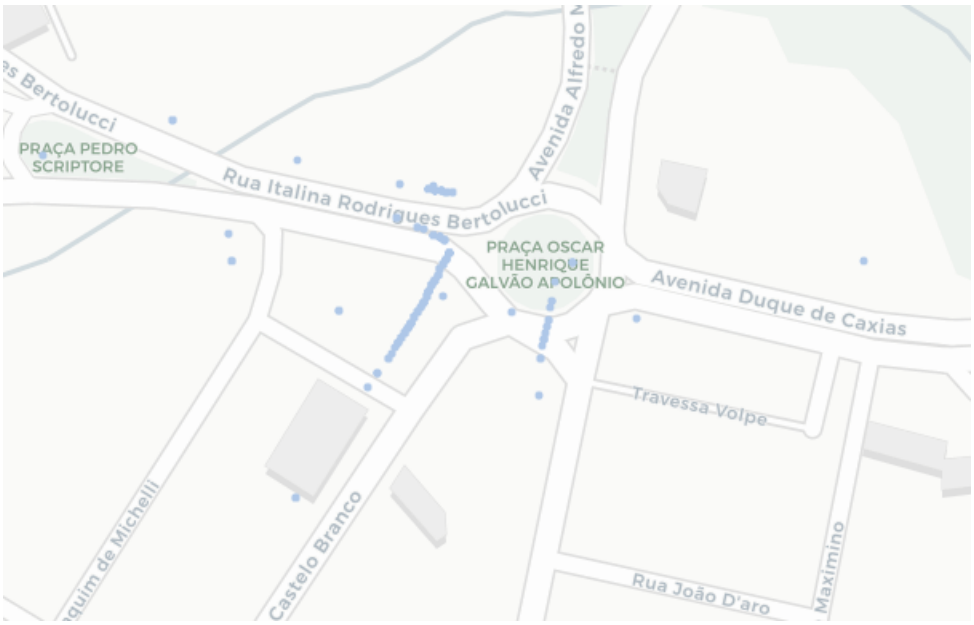
Primeiro, a fim de aumentar o desempenho do XGBoost, é necessário realizar a busca da combinação dos hiperparâmetros do modelo de XGBoost que mais se adapta e adequa

Figura 24 – Localização de Zona 3



Fonte: Elaborada pelo autor

Figura 25 – Localização de Zona 10



Fonte: Elaborada pelo autor

no conjunto de dados e no padrão que se quer encontrar, pois cada modelo irá responder e trabalhar de uma forma extremamente específica de acordo com cada caso

Depois, é preciso definir qual pontuação se deseja maximizar (Precisão, Acurácia, Recall, F1-Score). Para o escopo do projeto é desejável diminuir o número de acidentes fatais classificados como não fatais pelo modelo, o que significa um recall alto. Entretanto se o modelo for muito rigoroso e classificar muitas vítimas como fatais quando estas não são a precisão irá

Figura 26 – Localização de Zona 40

Fonte: Elaborada pelo autor

Figura 27 – Localização de Zona 86

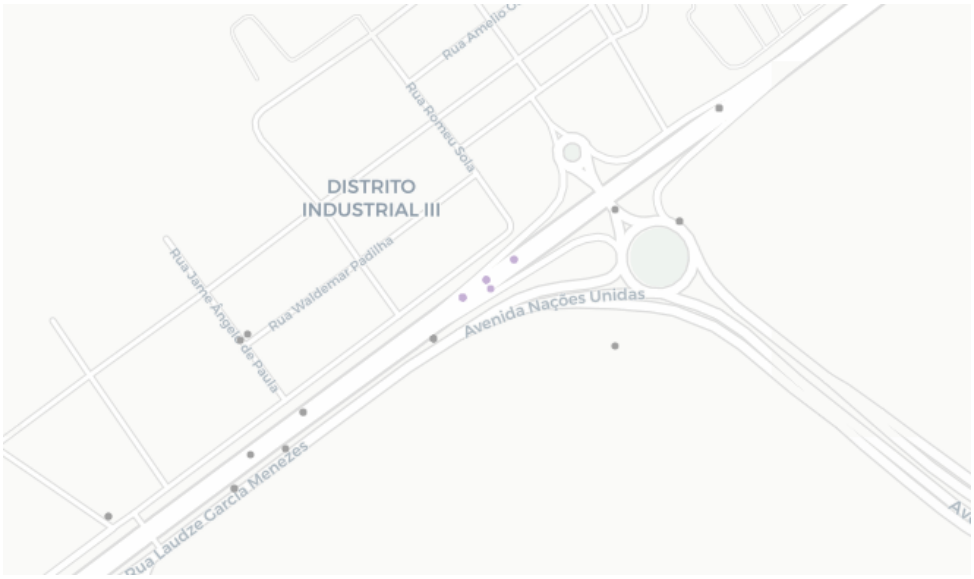
Fonte: Elaborada pelo autor

diminuir, o que também não é desejável. Para equilibrar isso o foco será na maximização da pontuação F1-Score.

Foram utilizados dois tipos de busca: *Random Search* e *Grid Search*. Considere que *scoring* = f1 e *n_estimators* = 50. O Quadro 4 indica o hiperparâmetro e o intervalo do conjunto de diferentes valores para cada respectivo parâmetro. Em negrito corresponde sãos valores encontrados pelas ambas buscas que maximizem a pontuação F1-Score.

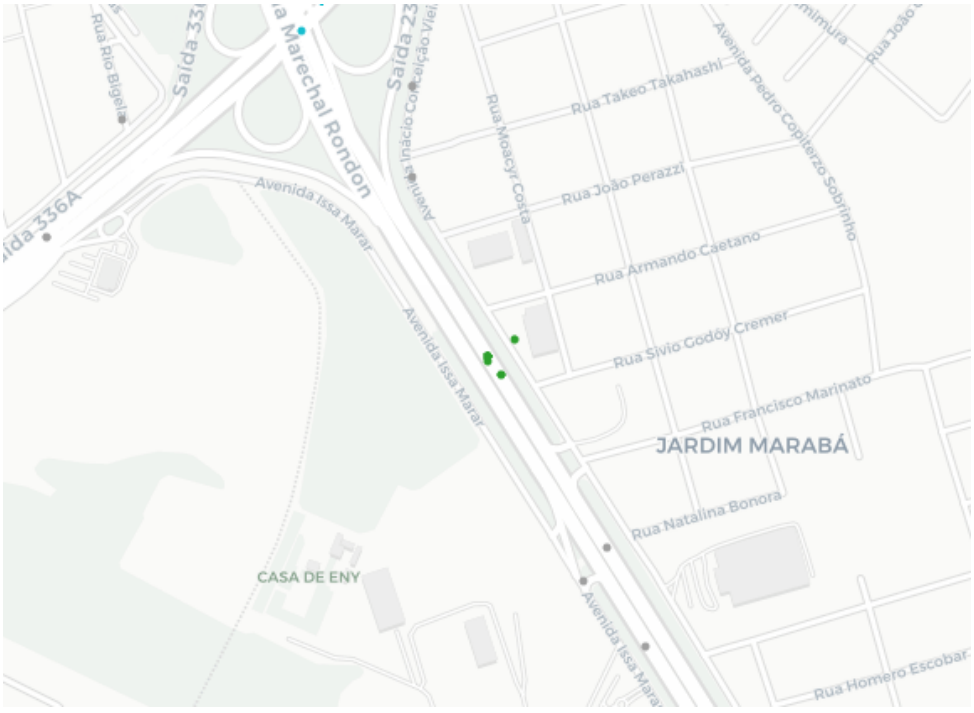
Com as combinações encontradas os modelos de XGBoost também são automaticamente treinados e testados. Para avaliar os modelos de Árvore de Decisão e XGBoost serão utilizadas

Figura 28 – Localização de Zona 54



Fonte: Elaborada pelo autor

Figura 29 – Localização de Zona 24



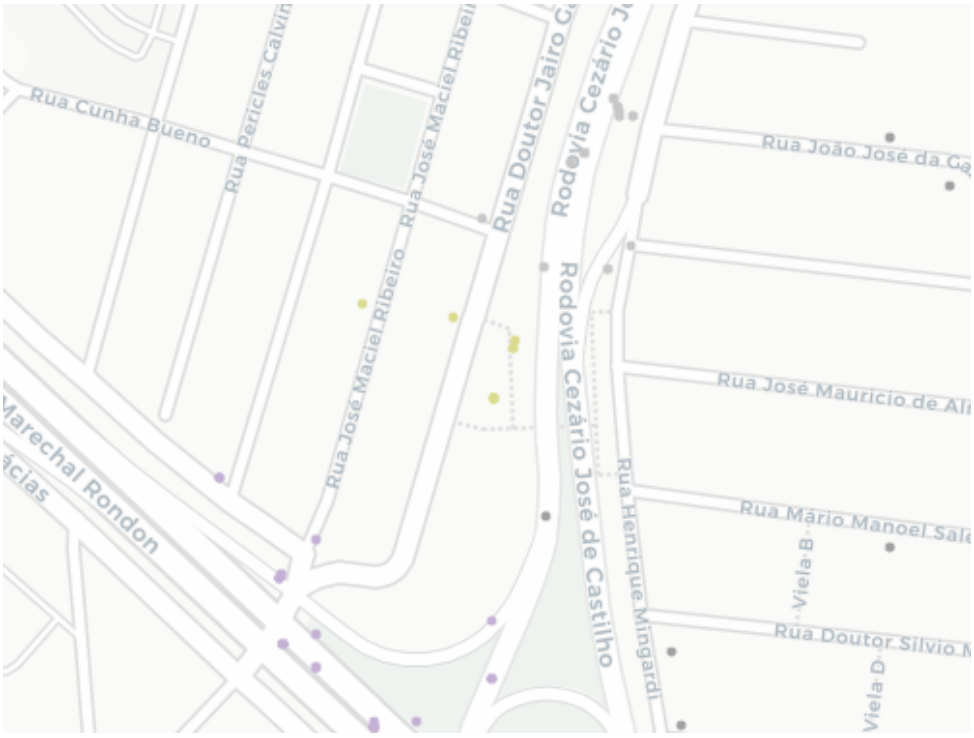
Fonte: Elaborada pelo autor

as métricas de avaliação citadas na subsubseção 3.3.4.3 para verificar se os algoritmos estão sendo capazes de prever a fatalidade de um acidente.

As Tabelas 3 e 4 mostram a pontuação em cada métrica no conjunto de dados de treino e teste, respectivamente.

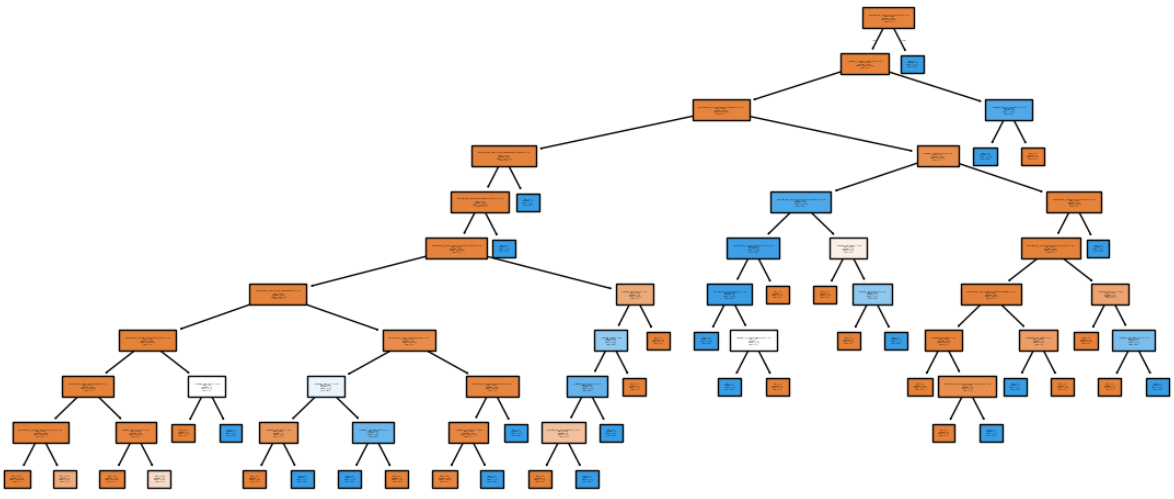
Realizando o comparativo de performance entre os dados de treino e teste, bem como

Figura 30 – Localização de Zona 95



Fonte: Elaborada pelo autor

Figura 31 – Modelo visual da Árvore de Decisão



Fonte: Elaborada pelo autor

com outras técnicas é possível observar que nos modelos de Árvore de Decisão e de XGBoost utilizando *Random Search* tiveram performances exageradamente altas no conjunto de dados de treino e uma, o que significa que os modelos decoraram o padrão dos dados e que há o *overfitting*. Já no modelo de XGBoost utilizando *Grid Search* sua performance foi mais modesta, o que indica que o modelo tenha de fato aprendido e captado algum padrão no conjunto de

Quadro 4 – Combinações de Hiperparâmetros testadas

Hiperparâmetro	***Random Search***	***Grid Search***
xgboost__learning_rate	[0.05, 0.0625, 0.075, 0.0875, **0.1**, 0.1125, 0.125]	[0.04, 0.045, **0.05**, 0.055]
xgboost__max_depth	[2, 4, 6, **8**, 10, 12]	[3, 4, **5**]
xgboost__colsample_bytree	[0.5, 0.6, 0.7, **0.8**, 0.9, 1.0]	[0.55, **0.65**, 0.75, 0.85, 0.95]
xgboost__subsample	[0.5, **0.6**, 0.7 0.8, 0.9, 1.0]	[0.75, **0.85**, 0.95]
xgboost__n_estimators	[50, 150, 250, **350**, 450, 550, 650, 750, 850, 950]	[**45**, 60, 75, 90, 105]

Fonte: Elaborado pelo autor

Tabela 3 – Métricas de Avaliação da Árvore de Decisão e XGBoost com combinações de hiperparâmetros diferentes com dados de treino

Métrica	**Árvore de Decisão**	**XGBoost *Random Search***	**XGBoost *Grid Search***
Precisão	0,98	0,98	0,89
Recall	1,00	1,00	0,77
F1-Score	0,99	0,99	0,82

Fonte: Elaborada pelo autor

Tabela 4 – Métricas de Avaliação da Árvore de Decisão e XGBoost com combinações de hiperparâmetros diferentes com dados de teste

Métrica	**Árvore de Decisão**	**XGBoost *Random Search***	**XGBoost *Grid Search***
Precisão	0,66	0,93	0,86
Acurácia	0,98	0,79	0,80
Recall	0,74	0,74	0,64
F1-Score	0,70	0,83	0,73

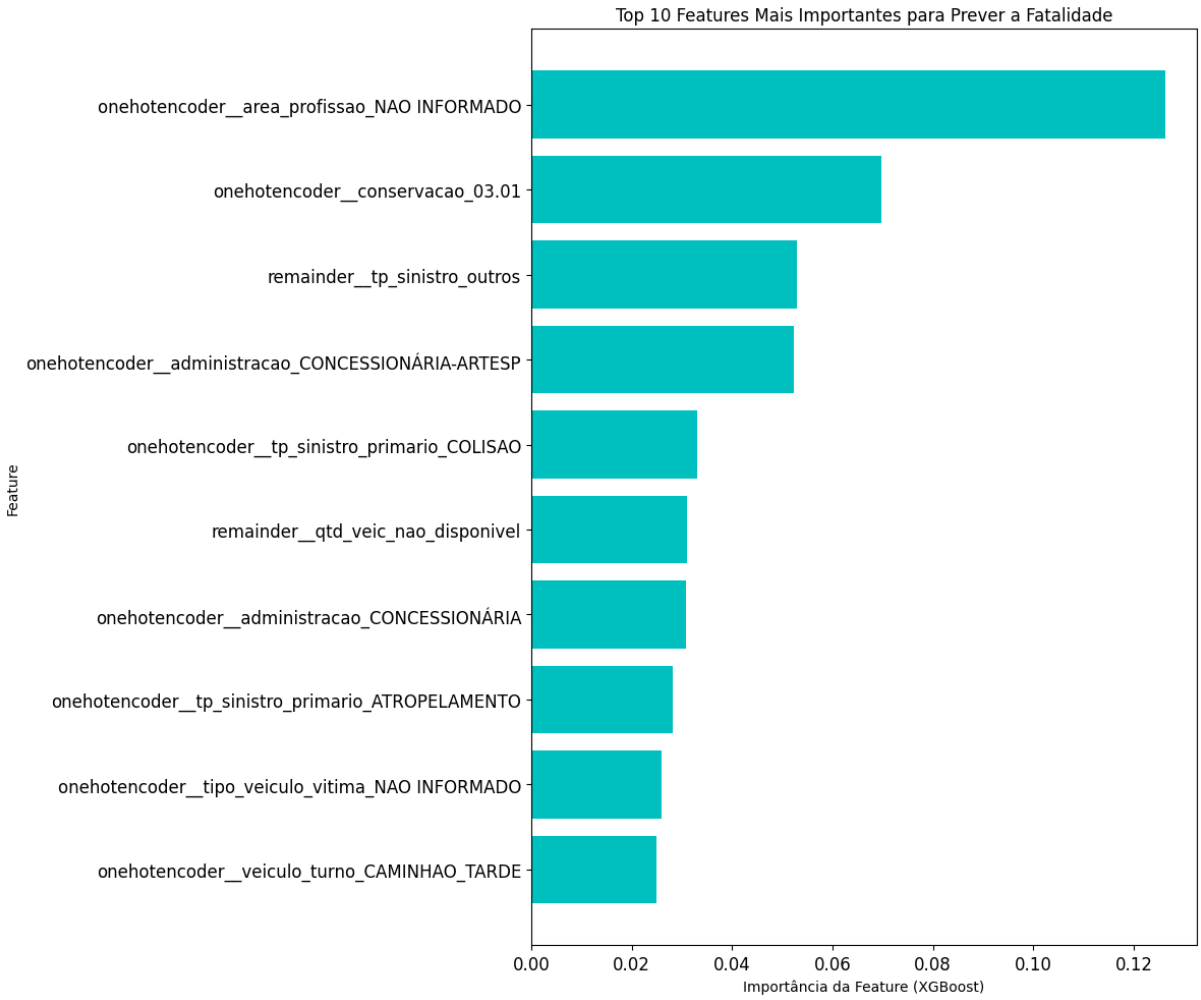
Fonte: Elaborada pelo autor

dados de treino, o que o torna a escolha mais ideal.

As Figuras 32 e 33 mostram as 10 *features* com maior peso para a construção dos modelos de XGBoost que utilizaram os hiperparâmetros encontrados por *Random Search* e *Grid Search*, respectivamente.

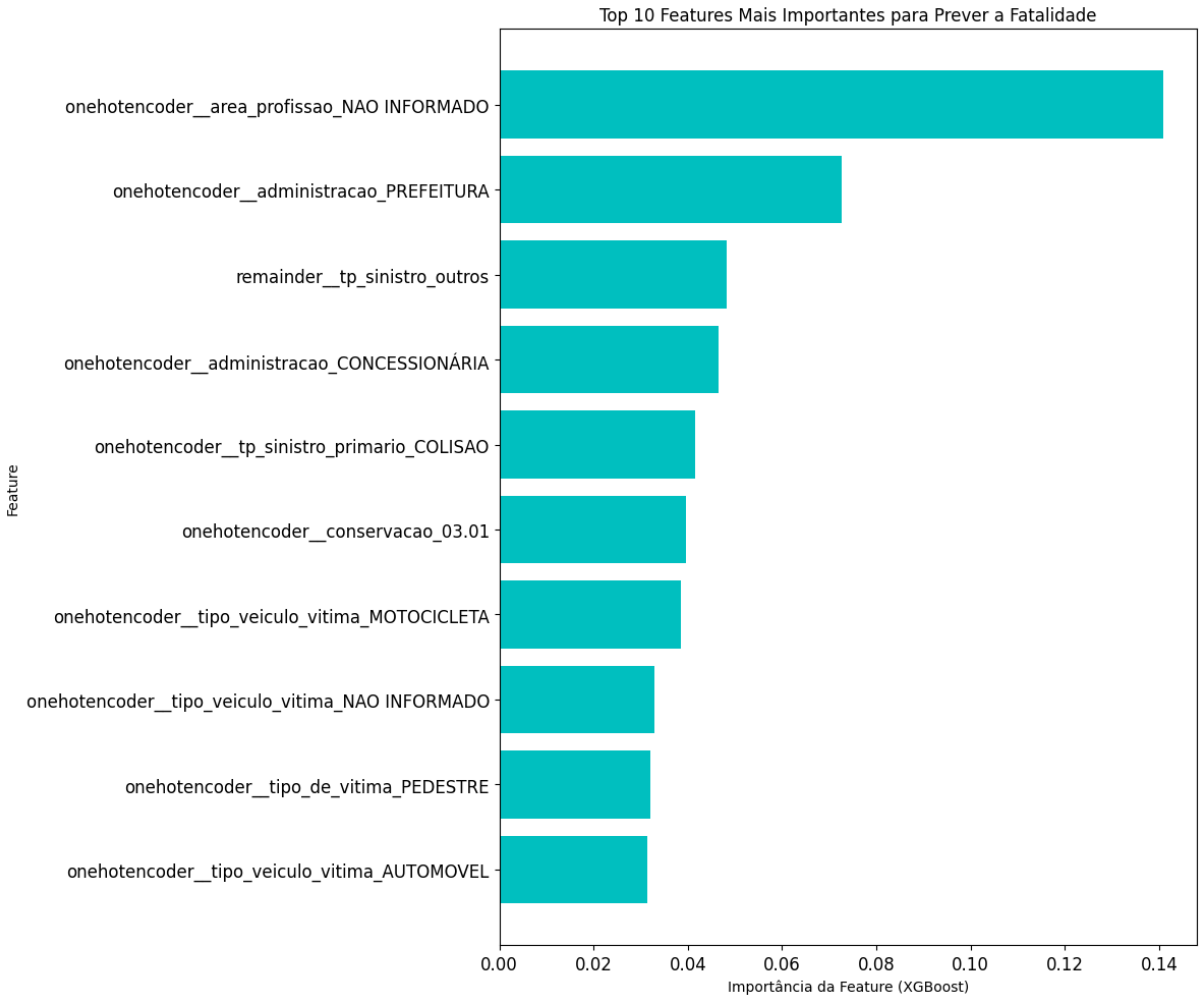
Apesar de trabalharem com o mesmo conjunto de dados e buscarem maximixar o F1-Score cada modelo trabalhou com *features* e pesos diferentes, identificando diferentes características para que um acidente seja determinado como fatal.

Figura 32 – *Features* mais relevantes do modelo XGBoost com *Random Search*



Fonte: Elaborada pelo autor

Figura 33 – *Features* mais relevantes do modelo XGBoost com *Grid Search*



Fonte: Elaborada pelo autor

5 Conclusão

Esse trabalho teve como fim realizar a análise de acidentes de trânsito em Bauru aplicando as técnicas de Ciência de Dados e, com isso, gerar informações relevantes e pertinentes sobre a segurança no trânsito, além de atingir os objetivos específicos como: predição da fatalidade por meio de modelos de Aprendizado de Máquina; identificação de locais com maior número de vítimas; análise das características dos acidentes; e caracterização de um perfil mais vulnerável.

A implementação dos modelos Supervisionados (Árvore de Decisão e XGBoost) e do Não Supervisionado (DBSCAN) foram partes fundamentais no desenvolvimento do trabalho e para atingir os principais objetivos do trabalho.

O algoritmo XGBoost utilizando *Grid Search* encontrou a configuração de hiperparâmetros mais balanceada e modesta, e seu resultado na predição da fatalidade dos acidentes teve um desempenho promissor. A utilização de diferentes *features* ao desenvolver ambos os modelos é um indicativo que para a previsão de fatalidade de um acidente são consideradas diversas variáveis com diversos pesos diferentes.

Para o objetivo de identificar os locais de maior incidência a geração dos *clusters* pelo algoritmo de DBSCAN foi fundamental: foi feita a análise focada e detalhada dessas regiões, e foi possível concluir que os acidentes com maior incidência estão localizadas nas vias urbanas, mais especificamente nas principais avenidas da cidade e locais com alta circulação de veículos. Apesar do alto número de vítimas nesses locais, sua fatalidade é extremamente baixa, que contrasta com os locais de alto número de vítimas fatais, que em suma estão todas localizadas nas rodovias, o que sugere que a fatalidade também está associada à velocidade dos veículos da via.

Portanto é possível concluir que existem condições espaciais, sociais e temporais que podem aumentar a chance de uma fatalidade ocorrer, como encontradas por cada algoritmo XGBoost, e que não é apenas um único perfil suscetível, mas vários.

Para concluir, como aprimoramento é possível realizar diferentes melhorias no trabalho, como: teste e busca de predição de diferentes variáveis-alvo, como *foi_fatal* podendo variar de 0 a 1, indicando probabilidade; aplicação de algumas técnicas e abordagens a fim de melhorar significamente a predição, como refinamento e utilização de PCA, que não foi aprofundado nesse trabalho, mas que há potencial de uma melhora significativa; engenharia de *features* mais eficiente, como melhora da qualidade dos dados eliminando dados incoerentes através de imputação lógica e/ou busca de fonte de dados mais consistentes; e adaptar o modelo e tratamento a fim de alterar o estudo de caso para uma cidade maior com maior quantidade de dados.

Referências

AYODELE, T. O. Types of machine learning algorithms. *New advances in machine learning*, v. 3, n. 19-48, p. 5–1, 2010.

Brasil. *Lei nº 13.614, de 11 de janeiro de 2018.* 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13614.htm>. Publicada no Diário Oficial da União de 12.1.2018. Acesso em: 28 de outubro de 2025.

Brasil. *Lei nº 14.129, de 29 de março de 2021.* 2021. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14129.htm>. Publicada no Diário Oficial da União de 30.3.2021. Acesso em: 27 de outubro de 2025.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794.

Departamento Estadual de Trânsito de São Paulo. *INFOSIGA: Óbitos no Estado de São Paulo*. 2025. <https://infosiga.detran.sp.gov.br/#obitos>. Dados referentes a Setembro de 2025. Acesso em: 27 de outubro de 2025.

G1. *SUS gastou R$ 449 milhões com vítimas de trânsito em 2024; fim do DPVAT agrava 'rombo' na Saúde*. 2025. <https://g1.globo.com/carros/noticia/2025/07/27/sus-gastou-r-449-milhoes-com-vitimas-de-transito-em-2024-fim-do-dpvat-agrava-rombo-na-saude.ghtml>. Acesso em: 27 de outubro de 2025.

Instituto de Pesquisa Econômica Aplicada. *Atlas da Violência: Óbitos em acidentes de transporte (1989-2023)*. 2023. <https://www.ipea.gov.br/atlasviolencia/dados-series/85>. Fonte: MS/SVS/CGIAE - Sistema de Informações sobre Mortalidade - SIM. Elaboração: Diest/Ipea. Acessado em: 27 de outubro de 2025.

KHAN, K.; REHMAN, S. U.; AZIZ, K.; FONG, S. Dbscan: Past, present and future. In: IEEE. *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. [S.l.], 2014. p. 232–238.

KINGSFORD, C.; SALZBERG, S. L. What are decision trees? *Nature biotechnology*, Nature Publishing Group US New York, v. 26, n. 9, p. 1011–1013, 2008.

KRIčKOVIć, E.; LUKIć, T.; SREJIć, T.; STOJšIć-MILOSAVLJEVIć, A.; STOJANOVIć, V.; KRIčKOVIć, Z. Spatial-temporal and trend analysis of traffic accidents in ap vojvodina (north serbia). *Open Geosciences*, v. 16, n. 1, p. 20220630, 2024. Acesso em: 27 de outubro de 2025. Disponível em: <https://doi.org/10.1515/geo-2022-0630>.

LOPES, G. R.; ALMEIDA, A. W. S.; DELBEM, A. C.; TOLEDO, C. F. M. Introdução à análise exploratória de dados com python. *Minicursos ERCAS ENUCMPI*, v. 2019, p. 160–176, 2019.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, v. 35, n. 101, p. 85–94, 2021. ISSN 1806-9592.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

Organização Mundial de Saúde. *Global status report on road safety 2023*. Geneva, 2023.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. *Big Data*, Mary Ann Liebert, Inc., v. 1, n. 1, p. 51–59, March 2013.

QUINLAN, J. R. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 28, n. 1, p. 71–72, 1996.