Refas
10 anos
Revista Fatec Zona Sul
1ª. Edição Outubro de 2014

ANÁLISE TEMPORAL DOS ACIDENTES DE TRÂNSITO NO ESTADO DE SÃO PAULO: UM ESTUDO DE CASO NA CIDADE DE CARAPICUÍBA NO PERÍODO DE 2015 A 2024

TEMPORAL ANALYSIS OF TRAFFIC ACCIDENTS IN THE STATE OF SÃO PAULO: A CASE STUDY IN THE CITY OF CARAPICUÍBA IN THE PERIOD FROM 2015 TO 2024

ANÁLISIS TEMPORAL DE ACCIDENTES DE TRÁFICO EN EL ESTADO DE SÃO PAULO: UN ESTUDIO DE CASO EN LA CIUDAD DE CARAPICUÍBA DE 2015 A 2024

Drausio de Castro[1]
Enrico Ferreira dos Santos[2]
Murilo Martins Alves[3]
Samuel Henrique Ricomini Souza[4]

Artigo recebido em: agosto de 2024
Artigo aceito em dezembro de 2024

DOI: 10.26853/Refas_ISSN-2359-182X_v11n03_05

RESUMO

O presente artigo realiza uma análise temporal dos acidentes de trânsito em Carapicuíba, na região metropolitana de São Paulo, utilizando uma abordagem exploratória e preditiva. O objetivo é identificar padrões, tendências e fatores associados a esses incidentes. O estudo também compara os resultados com as cidades vizinhas de Osasco e Barueri. Foram utilizados dados públicos do Sistema de Informações Gerenciais de Acidentes de Trânsito (Infosiga SP) para analisar acidentes entre 2015 e 2024, incluindo sinistros não fatais, perfil das vítimas, tipos de acidentes e características das vias. Entre 2019 e 2024, Carapicuíba registrou 5.794 sinistros, com um pico de 1.187 acidentes em 2022. Osasco apresentou o maior número, totalizando 10.922 acidentes, enquanto Barueri teve 7.815 no mesmo período. Em 2024, os números foram menores, com 360 acidentes em Carapicuíba, 608 em Osasco e 506 em Barueri, devido à disponibilidade de dados apenas até abril de 2024. Foi utilizada a técnica de mineração de dados Random Forest para identificar padrões de acidentes, como horários críticos e locais problemáticos, incluindo a Avenida Inocêncio Seráfico em Carapicuíba e a Avenida dos Autonomistas

[1] Mestre Profissional em Matemática pela Universidade Estadual de Campinas - UNICAMP. Docente da Faculdade de Tecnologia de Carapicuíba. E-mail: drausio.castro@fatec.sp.gov.br. Lattes: http://lattes.cnpq.br/3513196793674686. OrcId: https://orcid.org/0009-0000-7597-6364.
[2] Graduando em Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Carapicuíba. E-mail: enrico.santos@fatec.sp.gov.br. Lattes: http://lattes.cnpq.br/4837768993386085. OrcId: https://orcid.org/0009-0008-5006-8544.
[3] Graduando em Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Carapicuíba. E-mail: murilo.alves6@fatec.sp.gov.br. Lattes: http://lattes.cnpq.br/7564106905030035. OrcId: https://orcid.org/0009-0003-3460-1656.
[4] Graduando em Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Carapicuíba. E-mail: samuel.souza41@fatec.sp.gov.br. Lattes: http://lattes.cnpq.br/7766849240971200. OrcId: https://orcid.org/0009-0005-3252-7039.

em Osasco. O estudo oferece insights valiosos para a formulação de políticas públicas voltadas à segurança viária. A análise comparativa sugere que Carapicuíba, Osasco e Barueri compartilham padrões semelhantes, com maior incidência de acidentes à tarde e à noite.

Palavras-chave: Acidentes de Trânsito; Carapicuíba; Estado de São Paulo; Análise Temporal; Mineração de Dados; Random Forest.

ABSTRACT

This paper provides a temporal analysis of traffic accidents in Carapicuíba, a municipality in the metropolitan region of São Paulo State, through an exploratory and predictive data approach. The objective is to identify patterns, trends, and factors associated with these incidents. The study also compares the results with neighboring cities such as Osasco and Barueri. Public datasets from the Traffic Accident Management Information System (Infosiga SP) were used to analyze accidents between 2015 and 2024, covering non-fatal accidents, victim profiles, accident types, and road characteristics. Between 2019 and 2024, Carapicuíba recorded 5,794 accidents, with a peak of 1,187 accidents in 2022. Osasco had the highest number, totaling 10,922 accidents, while Barueri recorded 7,815 in the same period. In 2024, the numbers were lower across all cities, with 360 accidents in Carapicuíba, 608 in Osasco, and 506 in Barueri, a decline attributed to the availability of data only up to April 2024. The data mining technique used was the Random Forest algorithm, which helped identify accident patterns related to critical times of day and locations, such as Avenida Inocêncio Seráfico in Carapicuíba and Avenida dos Autonomistas in Osasco. The study provides valuable insights into formulating public policies aimed at improving road safety. Comparative analysis suggests that Carapicuíba, Osasco, and Barueri share similar patterns, with most accidents occurring in the afternoon and evening periods.

Keywords: Traffic Accidents; Carapicuíba; São Paulo State; Temporal Analysis; Data Mining; Random Forest.

RESUMEN

Este artículo analiza los accidentes de tráfico en Carapicuíba, un municipio de la región metropolitana de São Paulo, utilizando un enfoque exploratorio y predictivo de los datos. El objetivo es identificar patrones, tendencias y factores asociados a estos incidentes. El estudio también compara los resultados con las localidades vecinas de Osasco y Barueri. Se utilizaron datos públicos del Sistema de Información de Gestión de Accidentes de Tránsito (Infosiga SP) para analizar los accidentes entre 2015 y 2024, abarcando siniestros no fatales, perfiles de las víctimas, tipos de accidentes y características de las vías. Entre 2019 y 2024, Carapicuíba registró 5.794 accidentes, con un pico de 1.187 accidentes en 2022. Osasco tuvo el mayor número, con un total de 10.922 accidentes, mientras que Barueri registró 7.815 en el mismo período. En 2024, los números fueron menores en todas las ciudades, con 360 accidentes en Carapicuíba, 608 en Osasco y 506 en Barueri, una disminución que se puede atribuir a la disponibilidad de datos hasta abril de 2024. Se utilizó el algoritmo Random Forest para identificar patrones de accidentes en función de la hora del día y lugares críticos, como la Avenida Inocêncio Seráfico en Carapicuíba y la Avenida dos Autonomistas en Osasco. El estudio proporciona valiosos conocimientos para la formulación de políticas públicas dirigidas a mejorar la seguridad vial. El análisis comparativo sugiere que Carapicuíba, Osasco y Barueri comparten patrones similares, con más accidentes en la tarde y noche.

Palabras clave: Accidentes de Tránsito; Carapicuíba; Estado de São Paulo; Análisis Temporal; Minería de Datos; Random Forest.

1 INTRODUÇÃO

Segundo a Organização Mundial da Saúde (OMS) e o Relatório Global sobre o Estado da Segurança Viária em 2021, os acidentes e sinistros de trânsito representam uma das maiores preocupações globais em termos de saúde pública e segurança viária. Anualmente, milhões de pessoas são impactadas, tanto pelas fatalidades quanto pelos ferimentos causados por essas ocorrências. Cerca de 1,3 milhão de pessoas perdem a vida em acidentes de trânsito a cada ano, enquanto aproximadamente 50 milhões sofrem ferimentos, muitos dos quais resultam em incapacitações permanentes. Essa realidade faz dos acidentes de trânsito a principal causa de morte entre crianças e jovens de 5 a 29 anos em todo o mundo. Além das tragédias humanas, esses sinistros têm um impacto profundo no desenvolvimento socioeconômico dos países, principalmente nas nações de baixa e média renda, onde ocorrem cerca de 93% das mortes no trânsito, apesar de possuírem apenas 60% da frota mundial de veículos.

A previsão é que, se medidas eficazes de segurança viária não forem amplamente implementadas, os acidentes de trânsito poderão causar mais de 13 milhões de mortes e deixar cerca de 500 milhões de feridos ao longo da próxima década. Esse quadro agrava os desafios enfrentados pelos Objetivos de Desenvolvimento Sustentável (ODS), em especial o ODS 3, que visa garantir uma vida saudável e promover o bem-estar para todos. Nos últimos 20 anos, apesar de várias campanhas e esforços das Nações Unidas e de outros órgãos internacionais voltados para a segurança no trânsito, os números de acidentes permanecem praticamente inalterados, revelando a complexidade do problema e a necessidade de ações mais eficazes.

Diante desse cenário, a pesquisa e análise de dados sobre acidentes de trânsito se tornam fundamentais para embasar políticas públicas, melhorar a infraestrutura viária e aumentar a conscientização sobre comportamentos seguros no trânsito. Com base em dados históricos e o uso de técnicas de mineração de dados, como a *Random Forest*, é possível identificar padrões e fatores de risco que podem ser abordados por intervenções específicas, auxiliando na redução de acidentes e salvando vidas.

Define-se o problema de pesquisa: Como os padrões, tendências e fatores associados aos acidentes de trânsito em Carapicuíba podem ser identificados e analisados de maneira a contribuir para a redução de sinistros e a promoção da segurança viária municipal? Como objetivo geral realizar uma análise temporal abrangente dos dados de acidentes de trânsito ocorridos em Carapicuíba, utilizando técnicas de mineração de dados para identificar padrões, tendências e fatores associados, no período de 2015 a 2024. E, como objetivos específicos:

a) Identificar os principais fatores associados aos acidentes de trânsito em Carapicuíba;
b) Analisar os padrões e tendências dos acidentes ao longo do tempo;
c) Comparar os dados de Carapicuíba com os de localidades vizinhas, como Osasco e Barueri;
d) Propor recomendações para a formulação de políticas públicas voltadas à segurança viária.

Entende-se como delimitações que o estudo se direciona aos acidentes de trânsito registrados em Carapicuíba entre 2015 e 2024. Serão analisados dados disponibilizados no site Governo Aberto pelo Sistema de Informações Gerenciais de Acidentes de Trânsito do Estado de São Paulo (Infosiga SP), abrangendo óbitos, perfil das vítimas, tipos de acidentes e características da região. E ainda, como hipóteses: a análise de dados de acidentes de trânsito em Carapicuíba, utilizando a técnica de *Random Forest*, permitirá identificar padrões

específicos e fatores determinantes que contribuem para a alta incidência de sinistros na cidade.

Com isso, será possível desenvolver predições mais precisas sobre a ocorrência desses acidentes e fornecer subsídios para a criação de medidas de intervenção mais eficazes, com base em comparações com cidades vizinhas e em uma compreensão aprofundada das variáveis envolvidas.

2 REFERENCIAL TEÓRICO

Nesta seção, serão apresentados os principais conceitos referentes à mineração de dados, ao processo KDD *(Knowledge Discovery in Databases),* à técnica *Random Forest*, à análise exploratória e preditiva, e à análise de acidentes de trânsito.

2.1 Mineração de Dados

A Mineração de Dados (*Data Mining*) é um processo que envolve não apenas a identificação de padrões, mas também a descoberta de correlações e anomalias em um determinado conjunto de dados (*Dataset*). Segundo Han, Kamber e Pei (2011), no livro *"Data Mining: Concepts and Techniques"*, são abordados conceitos e etapas fundamentais do processo de *Data Mining*:

Mineração é um termo vívido que caracteriza o processo que encontra um pequeno conjunto de pepitas preciosas a partir de uma grande quantidade de matéria-prima. Assim, esse nome impróprio que transporta "dados" e "mineração" tornou-se uma escolha popular. Além disso, muitos outros termos têm um significado semelhante ao da mineração de dados – por exemplo, mineração de conhecimento a partir de dados, extração de conhecimento, análise de dados/padrões, arqueologia de dados e dragagem de dados.

Muitas pessoas tratam a mineração de dados como sinônimo de outro termo popularmente usado, descoberta de conhecimento a partir de dados, ou processo KDD, enquanto outras veem a mineração de dados apenas como uma etapa essencial no processo de descoberta de conhecimento. O processo de descoberta de conhecimento é mostrado como uma sequência iterativa das seguintes etapas (Han; Kamber; Pei, 2011):

a) Limpeza de dados: Eliminação de ruídos e inconsistências, assegurando que os dados sejam confiáveis e adequados para a análise;
b) Integração de dados: Combinação de dados provenientes de diversas fontes, criando uma visão unificada e consistente das informações;
c) Seleção de dados: Extração dos dados mais relevantes para a análise, filtrando apenas as informações necessárias a partir dos bancos de dados disponíveis;
d) Transformação de dados: Ajuste e consolidação dos dados em formatos apropriados para mineração, utilizando operações como agregação ou resumo;
e) Mineração de dados: Processo fundamental no qual técnicas avançadas são aplicadas para descobrir padrões ocultos e relevantes nos dados;
f) Avaliação de padrões: Análise dos padrões extraídos para identificar aqueles que realmente representam conhecimento valioso, com base em critérios de interesse;
g) Apresentação do conhecimento: Uso de técnicas de visualização e representação para comunicar os resultados de forma clara e acessível aos usuários.

2.2 Processo KDD (*Knowledge Discovery in Databases)*

Segundo os acadêmicos Usama Fayyad, David Haussler e Paul Stolorz, as técnicas de mineração de dados e descoberta de conhecimento em bancos de dados (Da Língua Inglesa: *Knowledge Discovery in Databases*) para análise automatizada de dados desempenham um papel crucial como interface entre cientistas e grandes conjuntos de dados. Eles argumentam que, enquanto as máquinas ainda estão longe de alcançar as habilidades humanas nas áreas de síntese de novos conhecimentos, formação de hipóteses e modelagem criativa, os processos de obtenção de insights e realização de análises investigativas continuam sendo tarefas mais adequadas para humanos. No entanto, a automação do procedimento de redução de dados é um nicho significativo adequado para computadores. A redução de dados envolve catalogação, classificação, segmentação, partição dos dados, entre outras tarefas (Fayyad et al., 1996).

Mitchell (1997) e Piatetsky-Shapiro (1991) também contribuíram para o entendimento do KDD, enfatizando que este processo oferece uma abordagem sistemática para transformar dados em conhecimento, particularmente relevante em áreas como aprendizado de máquina, inteligência artificial e ciência de dados.

A literatura sobre KDD é vasta e continua a evoluir à medida que novas técnicas e tecnologias emergem. Por exemplo, Gaber, Zaslavsky e Krishnaswamy (2005) abordam a aplicação de técnicas de mineração de dados em fluxos de dados, enfatizando a importância da extração de conhecimento na análise de dados em tempo real, particularmente em ambientes com grandes volumes de dados dinâmicos.

Além disso, a importância desse processo se torna ainda mais evidente na era do *Big Data*, onde o volume e a complexidade dos dados excedem a capacidade das análises tradicionais. Chaudhuri, Dayal e Narasayya (2011) destacam que, com a evolução do processamento de dados em grande escala, permite que as organizações extraíam valor de dados que, de outra forma, permaneceriam subutilizados.

Portanto, o processo KDD não apenas automatiza tarefas complexas, mas também proporciona uma maneira estruturada de transformar grandes volumes de dados em insights valiosos, aplicáveis a diversos campos como: Finanças, Saúde, Segurança Viária e Marketing.

2.3 Técnica Floresta Aleatória (*Random Forest)*

A técnica de Floresta Aleatória (*Random Forest*) é um algoritmo de aprendizado de máquina amplamente utilizado, registrado por Leo Breiman e Adele Cutler. Este método combina a saída de várias árvores de decisão para alcançar um único resultado. Sua facilidade de uso e flexibilidade alimentaram sua adoção, pois lida eficazmente com problemas de classificação e regressão (Breiman, 2001; IBM, 2021). O modelo de floresta aleatória é composto por múltiplas árvores de decisão, que começam com uma pergunta básica e, a partir daí, fazem uma série de perguntas subsequentes que ajudam a dividir os dados e chegar a uma decisão final, indicada pelo nó folha. Essas árvores são treinadas usando algoritmos como a árvore de classificação e regressão (CART), avaliando a qualidade das divisões por meio de métricas como impureza de Gini, ganho de informação ou erro quadrático médio (MSE) (Breiman et al., 1986).

Embora as árvores de decisão sejam úteis para aprendizado supervisionado, elas podem sofrer com problemas de viés e superajuste. No entanto, quando combinadas no algoritmo de floresta aleatória, essas árvores de decisão produzem previsões mais precisas, especialmente

quando as árvores individuais não estão correlacionadas entre si. Os métodos de combinação de aprendizado, como *bagging* e boosting, utilizam conjuntos de classificadores cujas previsões são agregadas para determinar o resultado mais popular. No método de *bagging*, uma amostra aleatória de dados de treinamento é selecionada com substituição, permitindo que os pontos de dados sejam escolhidos mais de uma vez. Os modelos treinados com essas amostras independentes geram previsões que são agregadas para fornecer uma estimativa mais precisa, reduzindo a variância em conjuntos de dados ruidosos (Hastie et al., 2009).

O algoritmo de floresta aleatória (*Random Forest*) expande o método de *bagging* ao introduzir aleatoriedade de recursos, criando uma floresta de árvores de decisão não correlacionadas. Esse processo, também conhecido como *bagging* de recursos, gera subconjuntos aleatórios de recursos que garantem baixa correlação entre as árvores. O algoritmo é composto por uma coleção de árvores de decisão, cada uma treinada com uma amostra bootstrap do conjunto de dados de treinamento. Para problemas de regressão, a média das previsões das árvores individuais é calculada; para problemas de classificação, a classe mais frequente é determinada por votação majoritária. A amostra *out-of-bag* (OOB), uma porção dos dados de treinamento não utilizada no treinamento de uma árvore específica, é empregada para validação cruzada, melhorando a precisão das previsões finais (Breiman, 2001).

2.4 Análise Exploratória e Preditiva

A análise exploratória e preditiva de dados são componentes fundamentais em qualquer estudo de ciência de dados. A análise exploratória de dados (EDA) é a etapa inicial de análise, onde os dados são examinados para descobrir padrões, detectar anomalias, testar hipóteses e verificar pressupostos utilizando métodos estatísticos e gráficos. Esta abordagem foi popularizada pelo matemático John Tukey na década de 1970 (Tukey, 1977) e continua a ser uma prática essencial no processo de descoberta de dados atualmente (IBM, 2021).

Os métodos de EDA ajudam a determinar a melhor forma de manipular as fontes de dados para obter as respostas desejadas, proporcionando uma melhor compreensão das variáveis do conjunto de dados e das relações entre elas. Técnicas comuns incluem visualizações univariadas, bivariadas e multivariadas, como histogramas, diagramas de dispersão, gráficos de barras e mapas de calor (Hastie et al., 2009; IBM, 2021).

A análise preditiva, por outro lado, utiliza modelos estatísticos e algoritmos de aprendizado de máquina para prever resultados futuros com base em dados históricos. Essa forma de análise é crucial para tomar decisões informadas em várias áreas, como negócios, saúde e ciências sociais. Ferramentas de análise preditiva podem incluir regressão linear, redes neurais, árvores de decisão, entre outras (Hastie et al., 2009; IBM, 2021).

2.5 Análise de Acidentes de Trânsito

A análise de acidentes de trânsito é uma área vital que se beneficia grandemente da mineração de dados e da análise preditiva. Estudos nesta área buscam identificar padrões e fatores de risco que contribuem para a ocorrência de acidentes, permitindo a implementação de medidas preventivas eficazes.

Um artigo relevante que aborda este tema é "*Traffic Accident Analysis Using Data Mining Techniques*" publicado na revista *Transportation Research*. Este estudo utiliza técnicas de mineração de dados para analisar grandes volumes de dados de acidentes de trânsito, identificando fatores como condições meteorológicas, hora do dia, tipo de veículo e

comportamento do motorista que estão associados a um maior risco de acidentes (Wang et al., 2018).

Na análise de segurança viária, é importante distinguir entre "sinistros de trânsito" e "acidentes de trânsito". Embora os termos sejam frequentemente usados de forma intercambiável, eles têm significados distintos:

a) Acidentes de Trânsito: Refere-se a eventos inesperados e não intencionais que resultam em danos a veículos, pessoas ou propriedades. Estes eventos são geralmente considerados aleatórios e inevitáveis. De acordo com estudiosos como Haddon (1970), reforça a noção de que esses eventos estão fora do controle humano e são parte do acaso natural no trânsito. No entanto, essa visão tem sido cada vez mais criticada por estudiosos da segurança viária, pois implica em menor responsabilização dos envolvidos.

b) Sinistros de Trânsito: Um termo mais abrangente que inclui qualquer ocorrência no trânsito que resulte em danos ou perdas, mas que pode também englobar eventos previsíveis e preveníveis. Sinistros podem resultar de negligência, erro humano, falhas mecânicas, ou condições ambientais adversas. Este termo é amplamente utilizado por seguradoras e autoridades de trânsito para incluir uma gama mais ampla de eventos além dos meramente acidentais. Segundo Peden et al. (2004), essa terminologia é amplamente utilizada por seguradoras e autoridades de trânsito, pois engloba não apenas incidentes acidentais, mas também aqueles que podem ser resultado de negligência, erro humano, falhas mecânicas ou condições ambientais adversas. A adoção desse termo sugere uma abordagem mais ativa para a mitigação de riscos, incentivando uma análise mais ampla das causas e prevenções de tais eventos.

3 MÉTODO

Esta análise utilizou conjuntos de dados coletados ao longo de nove anos, provenientes de fontes confiáveis, como o Governo Aberto SP, o Sistema de Informações Gerenciais de Acidentes de Trânsito do Estado de São Paulo (Infosiga SP), e páginas on-line elaboradas pela Fundação Sistema Estadual de Análise de Dados (SEADE). Essas iniciativas, mantidas pelo Governo do Estado de São Paulo, garantem o acesso público a documentos e informações governamentais, permitindo que a sociedade reutilize esses dados para gerar novas informações e aplicações digitais.

As bases de dados abrangem informações detalhadas sobre acidentes de trânsito no Estado de São Paulo, filtradas especificamente para Carapicuíba, Barueri e Osasco. Esses dados incluem registros de delegacias, números de boletins de ocorrência, datas e locais dos acidentes, tipos de veículos envolvidos, condições meteorológicas, sexo e faixa etária das vítimas, meio de locomoção, tipo de acidente, e informações populacionais e administrativas dos municípios. A atualização dos dados ocorre com uma defasagem temporal de aproximadamente um mês (Infosiga SP; SEADE).

3.1 Definição do Estudo de Caso

Um estudo de caso é uma metodologia de pesquisa que se concentra em uma análise detalhada e contextualizada de um fenômeno específico. No presente artigo, as cidades de Carapicuíba, Barueri e Osasco foram selecionadas como unidades de análise para explorar a dinâmica dos acidentes de trânsito, fornecendo uma perspectiva detalhada sobre fatores locais.

3.2 Dimensão das Cidades

A cidade de Carapicuíba possui aproximadamente 400 mil habitantes, Barueri cerca de 270 mil, e Osasco, aproximadamente 700 mil habitantes, somando mais de 1,37 milhão de pessoas. Esses três municípios representam uma parcela significativa da Região Metropolitana de São Paulo (RMSP), que abriga mais de 22 milhões de habitantes, o que torna a análise dessas cidades relevante para entender as dinâmicas de trânsito na região como um todo (SEADE).

A principal técnica utilizada nesta análise foi o método KDD *(Knowledge-Discovery in Databases)*, que permite a descoberta de conhecimento útil a partir de grandes volumes de dados. Por meio deste método, foi possível identificar padrões recorrentes, tendências temporais e fatores de risco associados aos acidentes de trânsito em Carapicuíba.

3.3 Processo KDD *(Knowledge-Discovery in Databases)*

Para o processo de KDD e desenvolvimento do projeto será utilizado a linguagem de programação Python e as tecnologias: Google Colab; Excel; Pandas; Matplotlib; Scikit-learn; Numpy; Seaborn; e Prophet.

3.3.1 Seleção

Para possibilitar a visualização dos dados do *Dataset* referente aos acidentes fatais e a importação de todas as bibliotecas que serão úteis para o estudo. As bibliotecas necessárias para executar o código são importadas, incluindo Pandas para manipulação de dados, *Scikit-learn* para modelagem e avaliação de *Machine Learning*, e *Matplotlib* e *Seaborn* para visualização de dados., utilizou-se o código mostrado na Figura 1.

Figura 1- Código

```
import pandas as pd
import matplotlib.pyplot as plt

url = './sinistros_fatais.xlsx'
df = pd.read_excel(url)
df
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from sklearn.preprocessing import LabelEncoder
```

```
import matplotlib.pyplot as plt
import seaborn as sns
```

Fonte: autores

Após a importação das bibliotecas, que serão utilizadas posteriormente no código, é criada uma variável para receber os dados do arquivo. Usando o método "*read_excel*", o algoritmo transforma os dados contidos no arquivo que antes estavam no formato *Excel Open XML Spreadsheet* (XLSX) para *DataFrame*, o qual trata-se de uma estrutura que irá organizar os valores em: linhas e colunas, semelhante a uma planilha Excel. Logo em seguida, o *DataFrame* é impresso no terminal do Google Colab para visualização de todos os dados contidos dentro do Dataset, ver Figura 2.

Figura 2- Visualização do DataFrame (Acidentes fatais)

	ID	Id Delegacia (RDO)	Número BO (RDO)	Ano BO (RDO)	Data do Acidente	Dia do Acidente	Mês do Acidente	Ano do Acidente	Ano/Mês do Acidente	Dia da semana	...	Iluminação da via (SIOPM)	Superfície da Via (SIOPM)	Tipo de pista (SIOPM)	Outro Veículo Envolvido	Tipo de via	Condições Climáticas (SIOPM)	Sentido da Via (SIOPM)	Limite da velocidade da via (SIOPM)	Quantidade de vítimas	Tempo entre o Acidente e as Mortes
0	3503124	10211	1009696694	2024	2024-03-31	31	MARÇO	2024	2024.03	DOMINGO	...	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	NÃO HÁ	Vias Municipais	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	1	Até 30 dias
1	3503145	10226	1009700010	2024	2024-03-31	31	MARÇO	2024	2024.03	DOMINGO	...	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	NÃO HÁ	Rodovias	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	1	Até 30 dias
2	3503127	10247	1009697544	2024	2024-03-31	31	MARÇO	2024	2024.03	DOMINGO	...	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	NÃO HÁ	Vias Municipais	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	1	Até 30 dias
3	3503121	10362	1009697665	2024	2024-03-31	31	MARÇO	2024	2024.03	DOMINGO	...	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	NÃO HÁ	Vias Municipais	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	1	Até 30 dias
4	3503153	30111	1009697374	2024	2024-03-31	31	MARÇO	2024	2024.03	DOMINGO	...	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	ONIBUS	Rodovias	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	1	Até 30 dias
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26292	3455478	120607	36	2019	2019-01-01	1	JANEIRO	2019	2019.01	TERÇA	...	LUZ SOLAR	SECA	DUPLA	NAO DISPONIVEL	Vias Municipais	BOM	SUL	Entre 50 e 59km/h	1	Até 30 dias
26293	3460813	120300	3	2019	2019-01-01	1	JANEIRO	2019	2019.01	TERÇA	...	LUZ SOLAR	SECA	DUPLA	NÃO HÁ	Rodovias	BOM	OESTE	Entre 90 e 100km/h	1	Até 30 dias
26294	3468538	130323	1	2019	2019-01-01	1	JANEIRO	2019	2019.01	TERÇA	...	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	NÃO HÁ	Vias Municipais	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	1	Até 30 dias
26295	3470095	130315	72	2019	2019-01-01	1	JANEIRO	2019	2019.01	TERÇA	...	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	Rodovias	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	1	Até 30 dias
26296	3468678	120609	2	2019	2019-01-01	1	JANEIRO	2019	2019.01	TERÇA	...	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	NÃO HÁ	Vias Municipais	NAO DISPONIVEL	NAO DISPONIVEL	NAO DISPONIVEL	1	Até 30 dias

Fonte: Autores

As colunas do *DataFrame* estão separadas em: ID da Delegacia, Número do BO, Ano do BO, Nome da Delegacia, Data do Sinistro, Dia do Sinistro, Mês do Sinistro, Ano do Sinistro, Ano/Mês do Sinistro, Dia da Semana, ..., Choque, Colisão frontal, Colisão lateral, Colisão transversal, Colisão traseira, outras colisões, Engavetamento, Tombamento, outros sinistros, Quantidade de vítimas fatais.

3.3.2 Pré-Processamento

A fim de preservar a qualidade dos dados coletados, realizou-se uma análise para o preenchimento de valores na tabela, constatando que não há indícios de dados nulos, conforme demonstrado na Figura 3.

Figura 3 – Dados coletados

```
# Preenchendo valores ausentes
for column in df.columns:
    if df[column].dtype == 'object':
        df[column] = df[column].fillna(df[column].mode()[0])
    else:
        df[column] = df[column].fillna(df[column].mean())
```

Os valores ausentes nas colunas do tipo objeto são preenchidos com o valor mais frequente (moda) e os valores ausentes em colunas numéricas são preenchidos com a média dos valores existentes.

3.3.3 Transformação

Nesta seção, são detalhados os processos de transformação aplicados aos dados, essenciais para preparar o *dataset* para análise e modelagem. A transformação dos dados inclui a conversão de colunas de data/hora em *timestamps* e a codificação de variáveis categóricas, além da divisão dos dados em conjuntos de treinamento e teste. Estas etapas são cruciais para assegurar que os algoritmos de aprendizado de máquina possam trabalhar de forma eficaz com os dados fornecidos (números representando a quantidade de segundos desde uma data de referência).

As colunas do tipo *datetime64*[ns] são convertidas para *timestamps*. Esta conversão facilita a manipulação dos dados temporais, transformando as datas em números representando a quantidade de segundos desde uma data de referência. Este processo é realizado iterando sobre as colunas do *DataFrame* e aplicando a conversão apenas às colunas com o tipo de dado *datetime64*[ns], ver Figura 4.

Figura 4 - Iterações

```
# Convertendo colunas de data/hora para timestamp
for column in df.columns:
    if df[column].dtype == 'datetime64[ns]':
        df[column] = df[column].apply(lambda x: x.t
```

Fonte: autores

As variáveis categóricas são codificadas utilizando o *LabelEncoder* do *Scikit-learn*. Este método converte as categorias em números inteiros, o que é necessário para que os algoritmos de aprendizado de máquina possam processar estas variáveis. A codificação é realizada iterando sobre as colunas do *DataFrame* e aplicando o *LabelEncoder* às colunas do tipo *object* (Figura 5).

Figura 5 – Iterações colunas

```
# Codificando variáveis categóricas
le = LabelEncoder()
for column in df.columns:
    if df[column].dtype == 'object':
        df[column] = le.fit_transform(df[column])
```

Fonte: autores

Em seguida, os dados são divididos em conjuntos de treinamento e teste utilizando a função *train_test_split* do *scikit-learn*. O conjunto de teste é configurado para conter 20% dos dados, e uma semente aleatória (*random_state*=42) é utilizada para garantir a reprodutibilidade dos resultados, ver Figura 6.

Figura 6 – Divisão dos dados

```
# Dividindo os dados em conjuntos de treinamento e teste
```

```
X = df.drop('Atropelamento', axis=1)  # Substituir 'Atropelamento' por 'alvo'
y = df['Atropelamento'] # Substituir 'Atropelamento' por 'alvo'
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Fonte: autores

3.3.4 Mineração de Dados

Na etapa de Mineração de Dados, a técnica *Random Forest* foi usada para extrair os padrões significativos dos dados e determinar as características mais relevantes.

3.3.4.1 Floresta Aleatória (*Random Forest*)

Este Algoritmo foi utilizado para realizar previsões e classificações baseadas nos dados transformados. Esse método é conhecido por sua robustez e precisão em tarefas de classificação e regressão, sendo adequado para lidar com grandes volumes de dados complexos. Neste estudo, o modelo de *Random Forest* foi aplicado para prever o (adicionar) no munícipio de Carapicuíba.

Na Figura 7 vê-se o classificador *Random Forest*, o qual é inicializado com 100 árvores de decisão e treinado com os dados de treinamento para o *dataset* de sinistros não fatais ocorridos no munícipio em questão, com o objetivo de fazer previsões sobre os dados de teste

Figura 7 – Classificador Random

```
# Treinando o modelo Random Forest
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)

# Fazendo previsões
y_pred = clf.predict(X_test)

# Avaliando o modelo
print(classification_report(y_test, y_pred))
# Identificando padrões
# Visualizando a importância das características
feature_importances = pd.Series(clf.feature_importances_, index=X.columns)
feature_importances.nlargest(10).plot(kind='barh')
plt.show()
```

Fonte: autores

Para a avaliação do modelo, é gerado um relatório de classificação que inclui métricas como precisão (*precision*), *recall*, *F1-score* e suporte para cada classe.

3.3.5 Métricas de Avaliação

As métricas de avaliação são fundamentais para medir o desempenho dos modelos de *Machine Learning*. Neste contexto, quatro métricas populares são frequentemente utilizados (Precisão, *Recall*. *F1-Score* e *Support*):

1. Precisão (*precision*): A proporção de observações positivas classificadas corretamente dentre todas as observações classificadas como positivas;
2. Recall: A proporção de observações positivas classificadas corretamente dentre todas as observações que realmente são positivas;
3. F1-score: A média harmônica entre precisão e *recall*, é uma métrica que combina as duas medidas;
4. Support: O número de ocorrências de cada classe no conjunto de teste;
5. Identificação de padrões: É gerado um gráfico de barras horizontais que mostra a importância das características (*features*) no modelo treinado. Isso permite identificar quais características têm maior impacto na classificação.

Segundo Mateus o autor Padua, as métricas de avaliação, incluindo Acurácia, Precisão, Recall e F1-score, são cruciais para avaliar o desempenho de modelos de *Machine Learning* (Padua, 2020). O Autor Mario Filho ressalva a importância do uso da precisão, *recall* e *F1-score* em *Machine Learning*, destacando a sua aplicação e relevância na avaliação de modelos (Filho, M, 2023).

Em síntese, a acurácia corresponde a proporção de fatos e não fatos que foram corretamente classificados. A precisão corresponde caso a proporção dos dados classificados como fatos eram realmente fatos. Já o *recall* corresponde entre todas as amostras que realmente eram de fatos, a proporção classificada como fatos. E Por fim, o *F1-Score* trata-se de uma maneira de observar em um único número a precisão e o *recall*.

3.3.6 Análise e Interpretação do Modelo

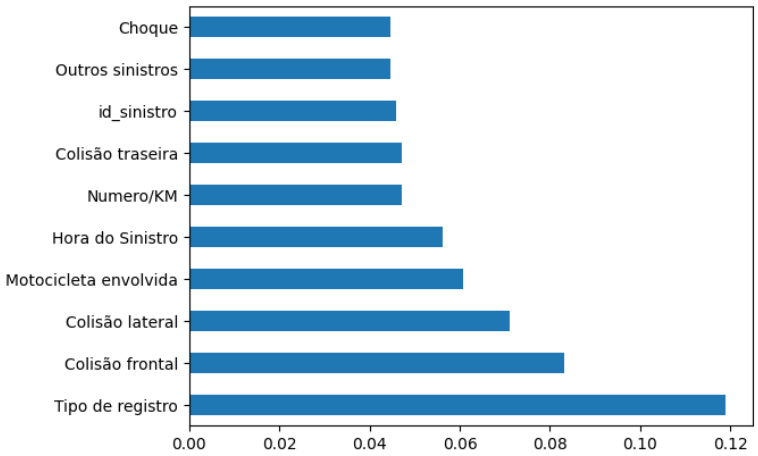
Os resultados do relatório de classificação indicam que o modelo tem uma alta precisão (0.99) e *recall* (0.90) para a classe 1 (Atropelamento), com uma acurácia geral de 0.99 (Figura 8 e Gráfico 1). Isso sugere que o modelo é capaz de identificar corretamente a maioria das instâncias de atropelamento no conjunto de teste. A macro média e a média ponderada das métricas também são elevadas, indicando um bom desempenho geral do modelo.

Figura 8 - Visualização dos níveis estatísticos do modelo

	precision	recall	f1-score	support
0	0.99	1.00	1.00	1075
1	0.97	0.90	0.94	84
accuracy			0.99	1159
macro avg	0.98	0.95	0.97	1159
weighted avg	0.99	0.99	0.99	1159

Fonte: autores

Gráfico 1 - Visualização gráfica da Importância das features



Fonte: autores

4 RESULTADOS E DISCUSSÃO

Dividem-se os resultados da pesquisa em duas subseções: análise exploratória dos dados e, em seguida, mostram-se as análises preditivas do munícipio de Carapicuíba e as cidades situadas ao redor deste.

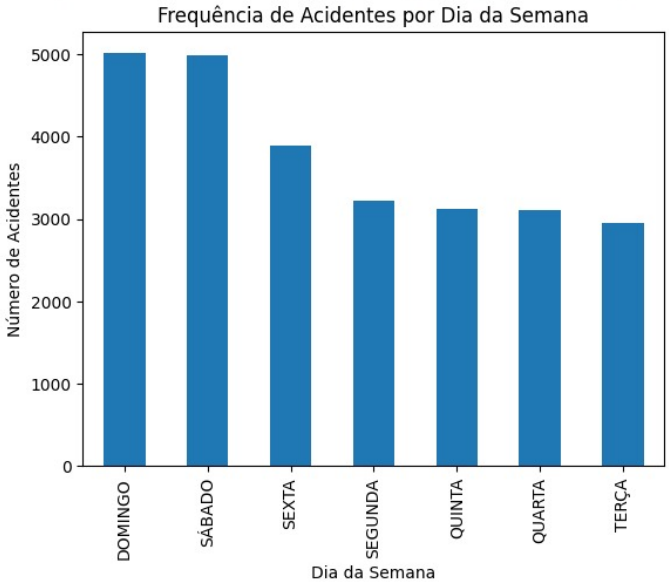
4.1 Análise Exploratória

Observando apenas a visualização do *Dataset* em forma de linhas e colunas, não é possível determinar quaisquer informações relevantes sobre a base de dados. Portanto, para uma análise mais aprofundada, criou-se um gráfico de barras que ilustra quais dias da semana ocorrem maiores quantidades de acidente foi utilizado os seguintes comandos (Gráfico 2):

```
import pandas as pd
import matplotlib.pyplot as plt

# Carrega os dados do arquivo Excel
url = '/content/acidentes_fatais.xlsx'
df = pd.read_excel(url)

# Gera um gráfico de barras para a frequência de acidentes por dia da semana
df['Dia da semana'].value_counts().plot(kind='bar')
plt.title('Frequência de Acidentes por Dia da Semana')
plt.xlabel('Dia da Semana')
```

```
plt.ylabel('Número de Acidentes')
plt.show()
```

Gráfico 2 - Filtragem de Acidentes por Dias da Semana em Carapicuíba

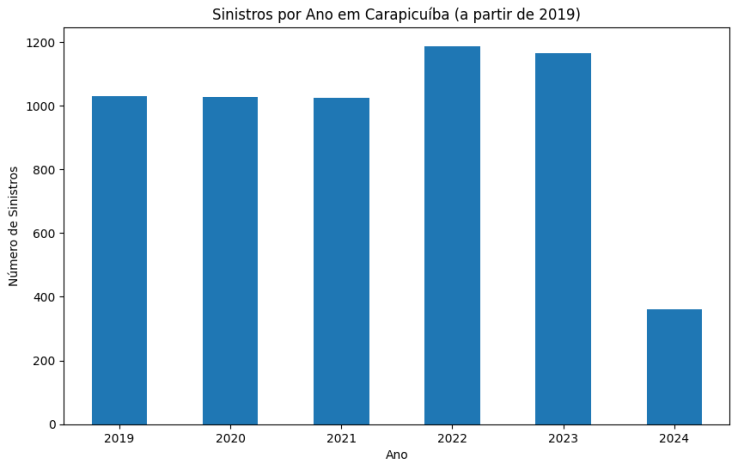


Fonte: autores

Com base no Gráfico 2, é possível identificar que os dias com as maiores taxas de acidentes são: sexta-feira, sábado e domingo. Após análise, verificamos que esses dias registram mais acidentes durante a noite, possivelmente devido ao fato de serem dias de folga, quando as pessoas frequentemente participam de eventos sociais. Infelizmente, é comum que algumas dessas pessoas saiam desses eventos alcoolizadas e tendam a desrespeitar as leis de trânsito, o que contribui significativamente para a ocorrência desses acidentes.

Para obtermos os dados referentes aos anos com mais acidentes, ver Gráfico 3.

Gráfico 3 - Taxa de acidentes anuais em Carapicuíba de 2019 (não fatais)



Fonte: autores

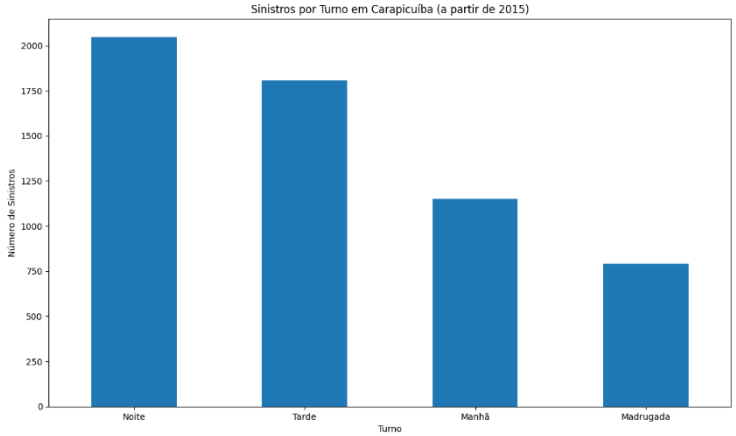
A análise do gráfico revela que nos anos de 2020 a 2021 houve uma redução significativa no número de acidentes. Esse fenômeno pode ser atribuído à pandemia de COVID-19, que resultou na diminuição do número de veículos em circulação, consequentemente reduzindo a incidência de acidentes durante esse período.

4.2 Comparativos entre munícipios vizinhos

A fim de entender melhor os padrões e tendências de sinistros viários, realizaram-se comparações entre os municípios vizinhos de Barueri, Osasco e São Paulo. Estas análises comparativas oferecem uma visão abrangente das características dos acidentes em diferentes contextos urbanos (Gráficos 4, 5 e 6).

Em primeiro lugar, foram gerados gráficos que representam a quantidade de sinistros por turno nas três cidades.

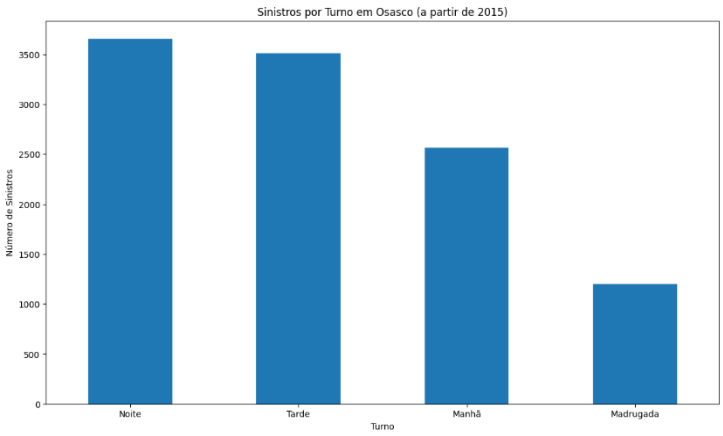
Gráfico 4 - Visualização da quantidade de sinistros em Carapicuíba



Fonte: autores

Com base no Gráfico 4, é perceptível que a quantidade de sinistros de trânsito está associada ao período noturno, em sua grande maioria, na cidade de Carapicuíba.

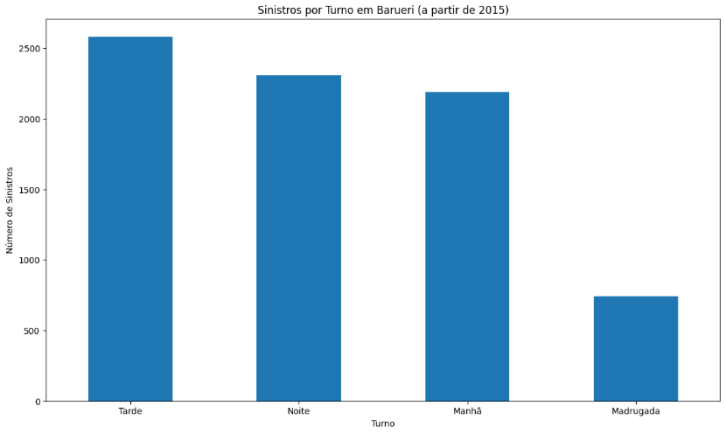
Gráfico 5 - Visualização da quantidade de sinistros em Osasco



Fonte: autores

Em comparação com o munícipio de Carapicuíba, a quantidade de sinistros de trânsito cresce significativamente no período da tarde e manhã, em Osasco.

Gráfico 6 - Visualização da quantidade de sinistros em Barueri

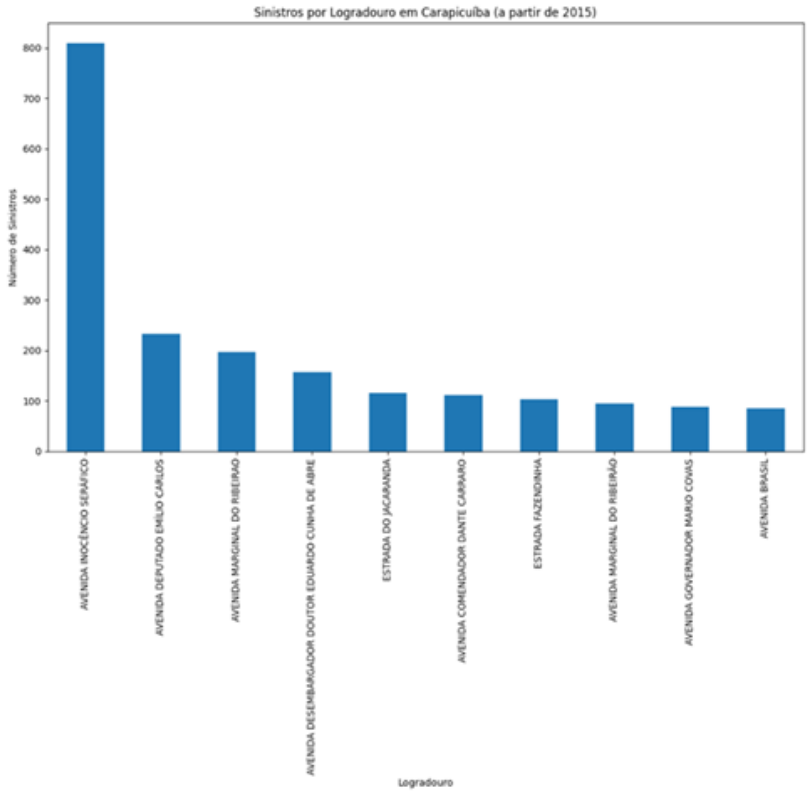


Fonte: autores

Em comparação com os munícipios de Carapicuíba e Osaco, a quantidade de sinistros de trânsito no período da tarde e manhã ficam próximas, em Barueri.

Devido à similaridade observada entre os gráficos referentes à quantidade de sinistros por turno em Carapicuíba, Osasco e Barueri, é possível identificar um padrão claro na ocorrência de acidentes de trânsito. Em todos os gráficos, os períodos com maior taxa de acidentes são a Tarde e a Noite, seguidos pela Manhã e, por último, a Madrugada. A principal diferença entre eles reside no número de registros desses acidentes. Em Carapicuíba, o intervalo dos gráficos varia de 0 a 2000, em Barueri, de 0 a 2500, e em Osasco, que apresenta o maior número de registros, o intervalo vai até 3500.

Gráfico 7 - Visualização de sinistros por logradouro em Carapicuíba

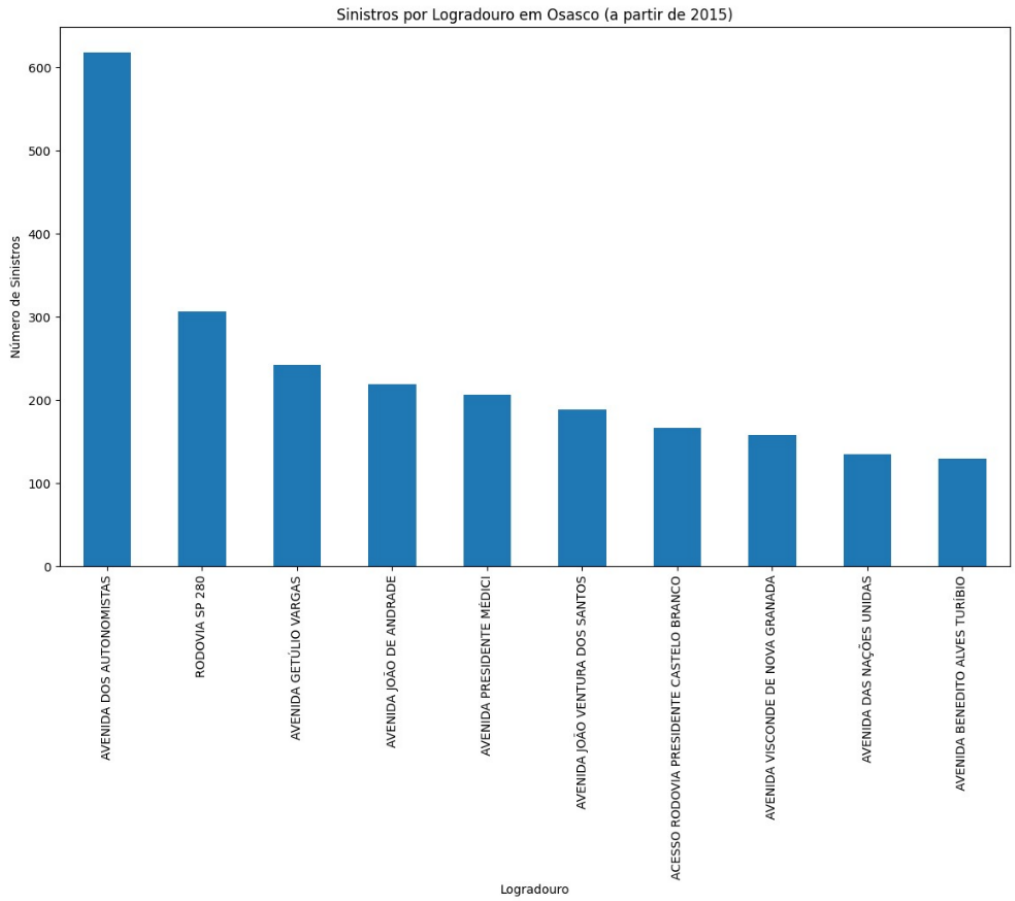


Fonte: autores

O Gráfico 7 registra os sinistros por logradouro em Carapicuíba a partir de 2015. Observamos que o intervalo do gráfico vai de 0 a 800 sinistros. O logradouro com o maior número de acidentes de trânsito é a Avenida Inocêncio Seráfico, uma das vias mais conhecidas e com maior fluxo de veículos, sendo a maior do município. Os demais logradouros apresentam menos da metade dos registros de sinistros em comparação à Avenida Inocêncio Seráfico. O gráfico segue uma ordem decrescente no número de acidentes por logradouro, sendo eles, em ordem: Avenida Emílio Carlos, Avenida Marginal do Ribeiro, Avenida Desembargador Doutor Eduardo Cunha de Abreu, Estrada do Jacarandá, Avenida Comendador Dante Carraro, Estrada da Fazendinha, Avenida Marginal do Ribeirão, Avenida Governador Mario Covas, e Avenida Brasil.

Vale destacar que os gráficos referentes aos sinistros por logradouros e por turnos dos respectivos municípios foram baseados na base de dados de sinistros não fatais disponibilizada pelo sistema Infosiga SP, do Governo Aberto do Estado de São Paulo.

Gráfico 8 - Visualização de sinistros por logradouro em Osasco



Fonte: Autoria autores

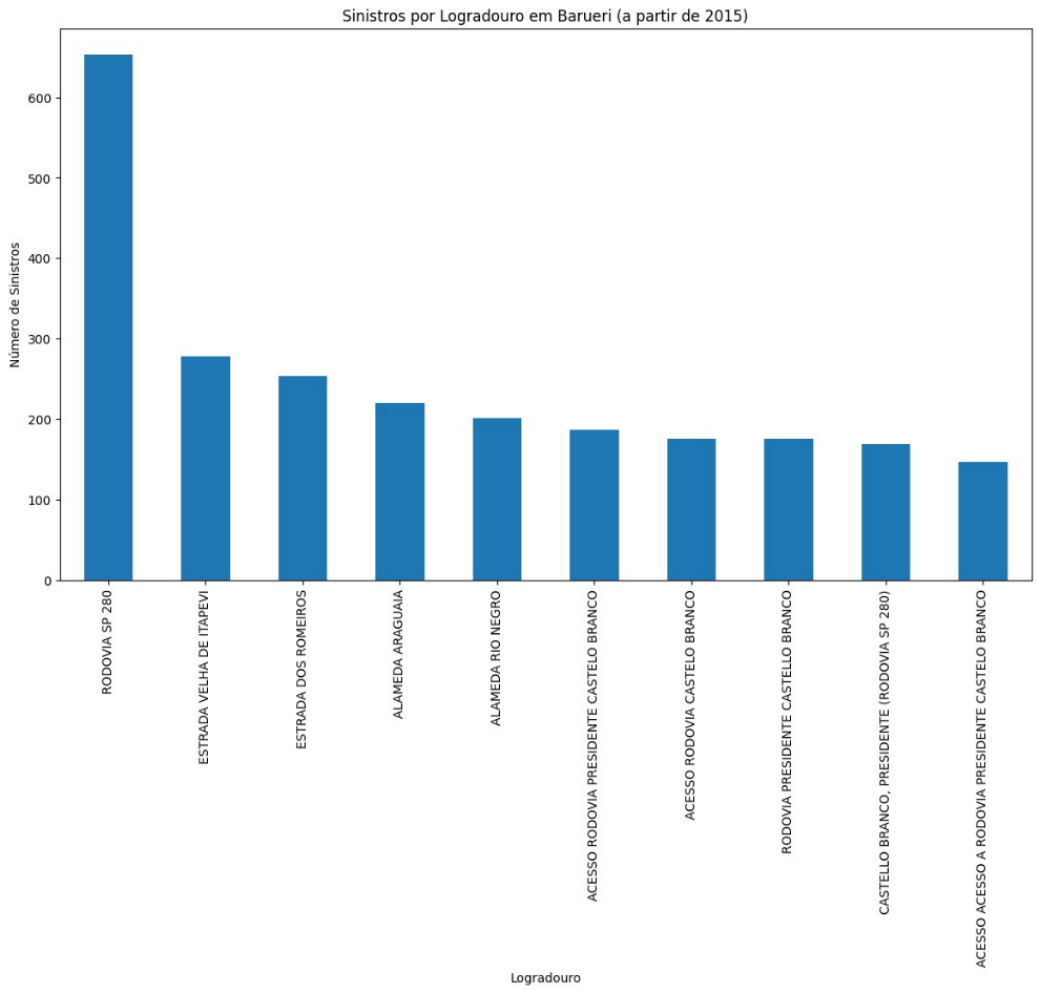
No Gráfico 8, que registra ocorrências de sinistros por Logradouro no Município de Osasco, observamos uma semelhança notável com o padrão identificado em Carapicuíba. Isso é evidenciado pela predominância da Avenida dos Autonomistas, a principal via, que apresenta o maior número de registros de acidentes de trânsito. No entanto, é importante notar que, embora haja essa semelhança, os valores no gráfico de Osasco variam de 0 a 600, com o pico de registros sendo inferior aos 800 registrados em Carapicuíba.

Destaca-se que a Rodovia SP-280 em Osasco é o segundo logradouro com maior número de ocorrências, embora represente aproximadamente metade dos registros da principal

via do município. Além disso, o gráfico revela uma tendência decrescente nos números de acidentes ao longo de outros logradouros, como a Avenida Getúlio Vargas, Avenida João de Andrade, Avenida Presidente Médici, Avenida João Ventura dos Santos, o Acesso à Rodovia Castelo Branco, Avenida Visconde de Nova Granada, Avenida das Nações Unidas e Avenida Benedito Alves Turíbio.

Esses logradouros apresentam um menor número de registros de sinistros, atribuível ao seu menor fluxo de veículos. Algumas dessas vias são menos conhecidas em comparação com a movimentada Avenida dos Autonomistas, o que pode contribuir para a redução dos incidentes.

Gráfico 9 - Visualização de sinistros por logradouro em Barueri



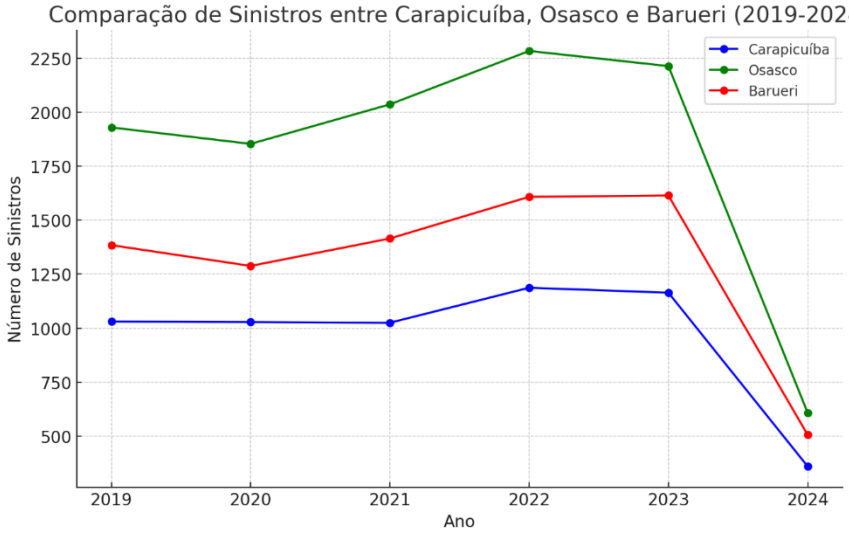
Fonte: autores

Ao analisarmos o Gráfico 9, de sinistros por logradouro em Barueri, notamos que sua similaridade com o gráfico do município de Osasco é um pouco mais próxima, uma vez que os valores em ambos os gráficos variam de 0 até 600.

Destaca-se que, em Barueri, os logradouros com os maiores registros de acidentes de trânsito são o trecho da Rodovia SP-280, com mais de 600 ocorrências de sinistros, e a Estrada Velha de Itapevi, que também está entre os logradouros com mais registros, contabilizando pouco menos de 300. O gráfico segue uma sequência decrescente de valores pelos seguintes logradouros: Estrada dos Romeiros, Alameda Araguaia e Alameda Rio Negro. Após esses logradouros, há diferentes registros referentes à mesma rodovia, a Rodovia Presidente Castelo Branco, porém, representam locais e registros distintos do mesmo logradouro. Isso significa

que, conforme o que pode ser interpretado no gráfico, o logradouro com o maior número de sinistros é a Rodovia Presidente Castelo Branco.

Gráfico 10 - Visualização da evolução dos sinistros em Carapicuíba, Osasco e Barueri



Fonte: autores

O Gráfico 10 apresenta a evolução dos sinistros de trânsito nas cidades de Carapicuíba, Osasco e Barueri durante o período de 2019 a 2024. Ao analisar os dados de cada município, observam-se as seguintes tendências:

a. Carapicuíba: Em 2019, foram registrados 1.030 sinistros; O número permaneceu estável em 2020 com 1.028 sinistros e em 2021 com 1.025 sinistros; em 2022, houve um aumento significativo, totalizando 1.187 sinistros, seguido por uma leve redução em 2023 com 1.164 sinistros; no ano de 2024, foram contabilizados 360 sinistros até o momento. Vale destacar que essa queda acentuada no número de sinistros se deve ao fato de que a base de dados utilizada contém registros apenas até o mês de abril de 2024.

b. Osasco: Em 2019, o número de sinistros foi de 1.929; em 2020, houve uma leve diminuição para 1.853 sinistros, seguida por um aumento em 2021, com 2.036 sinistros; o pico de sinistros ocorreu em 2022, com 2.283 ocorrências; em 2023, Osasco registrou 2.213 sinistros, mantendo-se em um patamar elevado; assim como em Carapicuíba, a queda para 608 sinistros em 2024 deve-se à disponibilidade parcial dos dados até abril.

c. Barueri: A cidade teve 1.384 sinistros em 2019, seguidos por uma queda para 1.288 sinistros em 2020; em 2021, os registros aumentaram para 1.415 sinistros, e essa tendência de crescimento continuou em 2022, com 1.608 sinistros; no ano de 2023, Barueri registrou seu maior número de sinistros no período, totalizando 1.614 sinistros; no entanto, em 2024, foram registrados 506 sinistros até abril, o que explica a aparente queda.

Esses números revelam que, embora cada município tenha suas particularidades, há uma tendência geral de aumento nos sinistros até 2023, seguida de uma aparente redução em 2024, que é explicada pela limitação dos dados disponíveis até o mês de abril. Esses padrões reforçam a importância de uma análise contínua e de medidas preventivas para reduzir os acidentes nas vias urbanas.

4.3 Análise Preditiva

A análise preditiva é uma ferramenta importante para antecipar eventos futuros com base em padrões identificados em dados históricos. Neste contexto, foi aplicada a análise preditiva para prever o número de sinistros fatais em anos subsequentes. Este processo, fundamentado na utilização do modelo *Prophet* e dados históricos de sinistros, no período de 2015 a 2024.

Na Figura 9 é possível verificar o código desenvolvido para o treinamento do modo *Prophet*:

Figura 9 – Código - Treinamento

```
import pandas as pd
import matplotlib.pyplot as plt
from prophet import Prophet

# Carregar os dados do arquivo XLSX
file_path = '/content/sinistros_fatais.xlsx'
df = pd.read_excel(file_path)

# Converter a coluna 'Data do Sinistro' para datetime
df['Data do Sinistro'] = pd.to_datetime(df['Data do Sinistro'])

# Filtrar os dados a partir de 2015
df = df[df['Data do Sinistro'] >= '2015-01-01']

# Preparar os dados para Prophet
df_prophet = df[['Data do Sinistro']].copy()
df_prophet['y'] = 1  # Adicionar uma coluna de contagem
df_prophet = df_prophet.groupby('Data do Sinistro').count().reset_index()
df_prophet.columns = ['ds', 'y']

# Treinar o modelo Prophet
model = Prophet()
model.fit(df_prophet)
# Fazer previsões até 2029 (365 dias * (2029 - 2024 + 1) = 2190 dias)
future = model.make_future_dataframe(periods=2190)
forecast = model.predict(future)
```

```
# Plotar as previsões
fig = model.plot(forecast)
plt.title('Previsão de Sinistros de 2015 a 2029')
plt.xlabel('Data')
plt.ylabel('Número de Sinistros')
plt.show()
```

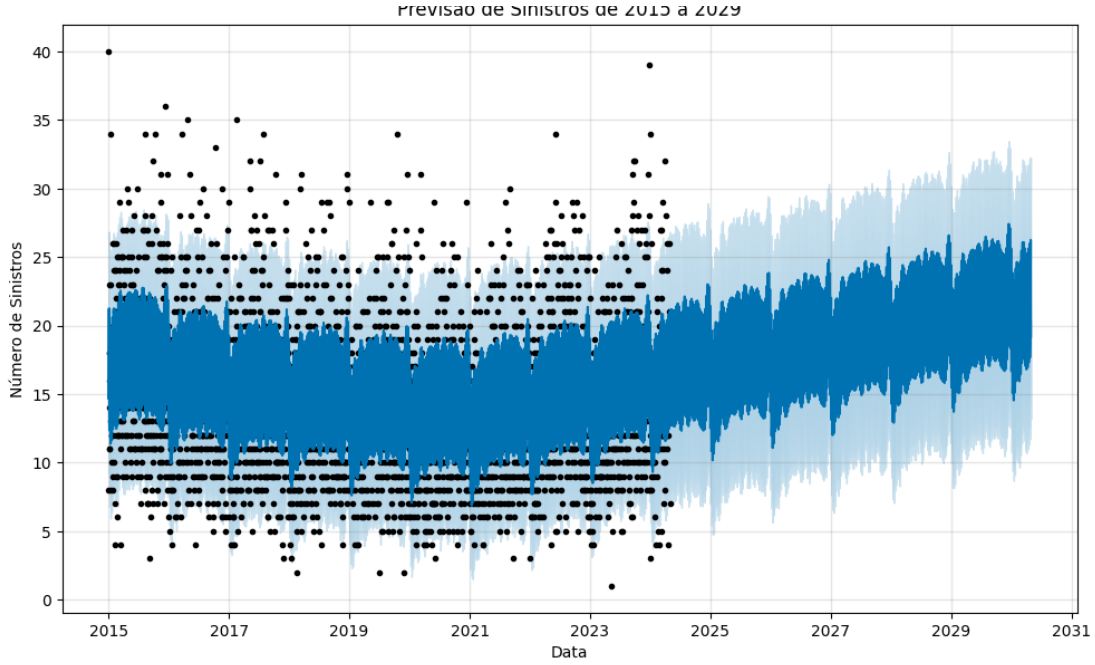
Fonte: autores

Este código foi desenvolvido para aplicar técnicas de análise preditiva ao problema específico de prever o número de sinistros fatais em anos futuros. Começando com a instalação da biblioteca *Prophet* usando o gerenciador de pacotes pip, o código continua carregando os dados históricos dos sinistros de um arquivo *Excel*. Em seguida, os dados são processados e preparados para serem usados pelo modelo *Prophet*.

O modelo *Prophet* é então treinado com os dados históricos preparados. Este modelo é uma ferramenta poderosa para lidar com séries temporais, como é o caso dos sinistros fatais, e é capaz de identificar padrões e tendências nos dados (ver Figura 10 e Gráfico 12).

Após o treinamento do modelo, são geradas previsões para o número de sinistros até o ano de 2029. Essas previsões são então visualizadas em um Gráfico (11), permitindo uma análise visual das tendências projetadas ao longo do período previsto.

Gráfico 11 - Previsão do Número de Sinistros até 2029



Fonte: Autores

Figura 10 – Análise preditiva

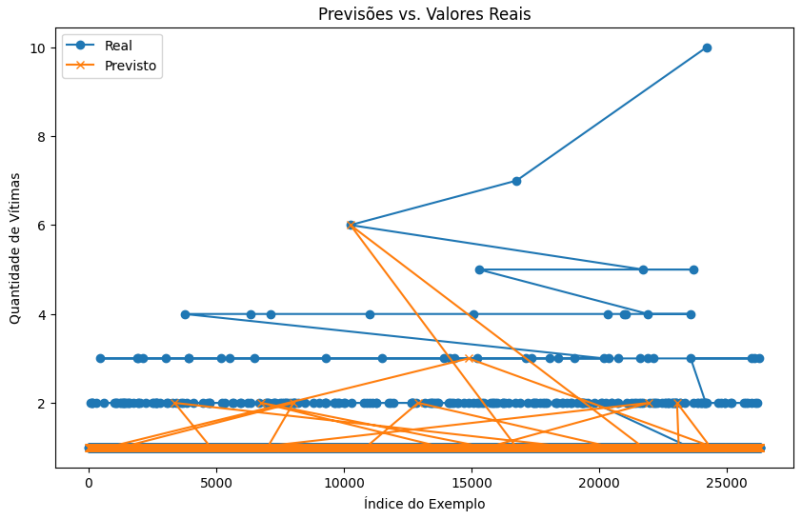
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt

file_path = '/content/acidentes_fatais.xlsx'
df = pd.read_excel(file_path)

# Selecionar características (features) e variável alvo (target)
features = df[['Dia do Acidente', 'Mês do Acidente', 'Ano do Acidente', 'Hora do Acidente',
               'Município', 'Tipo de via', 'Condições Climáticas (SIOPM)']]
target = df['Quantidade de vítimas']

features_encoded = pd.get_dummies(features)

X_train, X_test, y_train, y_test = train_test_split(features_encoded, target, test_size=0.2,
random_state=42)

model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

# Avaliar o modelo
accuracy = accuracy_score(y_test, y_pred)
print("Acurácia do modelo:", accuracy)

resultados = pd.DataFrame({'Real': y_test, 'Previsto': y_pred})

resultados = resultados.sort_values(by='Real')

# Gerar o gráfico
plt.figure(figsize=(10,6))
plt.plot(resultados.index, resultados['Real'], label='Real', marker='o')
plt.plot(resultados.index, resultados['Previsto'], label='Previsto', marker='x')
plt.xlabel('Índice do Exemplo')
plt.ylabel('Quantidade de Vítimas')
```

```
plt.title('Previsões vs. Valores Reais')
plt.legend()
plt.show()
```

Fonte: autores

Gráfico 12 - Visualização do comparativo entre valores reais e previsões feitas



Fonte: autores

O gráfico "Previsões vs. Valores Reais" oferece uma visualização clara da performance do modelo *Random Forest* em prever a quantidade de vítimas. A análise detalhada revela:

a. O modelo tende a subestimar a quantidade de vítimas em muitos casos, especialmente para valores mais altos;

b. Há uma maior variação nos valores reais em comparação com as previsões do modelo, sugerindo que o modelo pode não estar capturando toda a complexidade dos dados;

c. A precisão geral do modelo, medida pela acurácia (não especificada no gráfico, mas mencionada no código), não é suficiente para garantir previsões exatas para cada exemplo individual.

Esta visualização é essencial para identificar as limitações do modelo atual e considerar melhorias para capturar melhor a variabilidade dos dados reais.

5 CONSIDERAÇÕES FINAIS

Este estudo realizou uma análise temporal dos acidentes de trânsito em Carapicuíba (2015-2024), aplicando o método Random Forest para identificar padrões e fatores associados. O objetivo foi atingido ao destacar aspectos importantes, como características das vias, condições climáticas, tipos de veículos e perfis das vítimas. A análise revelou variações sazonais e picos em determinados períodos, indicando fatores como condições climáticas e fluxo de tráfego.

A comparação com Osasco e Barueri evidenciou semelhanças e diferenças nos padrões de acidentes, com tendências comuns como o aumento em certos meses e particularidades na

infraestrutura. Esses dados forneceram subsídios valiosos para políticas públicas adaptadas a cada município.

Com base nas diretrizes da OMS (Plano Global para a Década de Ação pela Segurança no Trânsito 2021-2030), recomenda-se melhorar a infraestrutura viária com foco em pedestres e ciclistas, como calçadas, ciclovias e melhor sinalização. A gestão da velocidade, com lombadas, radares e zonas de velocidade reduzida, é crucial para áreas de alta circulação. Campanhas educativas sobre segurança no trânsito, especialmente nas escolas, são fundamentais para promover comportamentos seguros.

O investimento em tecnologias de monitoramento, como câmeras e sistemas de detecção de infrações, permitirá monitorar e ajustar o trânsito com base em dados. No âmbito das políticas públicas, deve-se promover o transporte público e a mobilidade sustentável, integrando o planejamento urbano com corredores de ônibus e maior conectividade entre meios de transporte. Além disso, o fortalecimento dos serviços de emergência garantirá respostas rápidas a acidentes. Essas medidas, alinhadas com a abordagem de sistemas seguros da OMS, podem transformar a segurança viária em Carapicuíba, contribuindo para a redução de mortes e lesões em 50% até 2030, servindo como modelo para outras localidades.

6 REFERÊNCIAS

BREIMAN, L. Random Forests. Machine Learning, v. 45, n. 1, p. 5-32, 2001.

BREIMAN, L.; FREIDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and Regression Trees. Wadsworth and Brooks/Cole Advanced Books & Software, 1986.

CHAUDHURI, S.; DAYAL, U.; NARASAYYA, V. An overview of business intelligence technology. Communications of the ACM, 54(8), 88-98, 2011.

FAYYAD, Usama; HAUSSLER, David; STOLORZ, Paul. Mining and knowledge discovery in databases: Analysis automation interface. Advances in Knowledge Discovery and Data Mining, [s.l.], v. 1, n. 1, p. 30, 1996.

GABER, Mohamed Medhat; ZASLAVSKY, Arkady; KRISHNASWAMY, Shonali. Mining data streams: a review. ACM SIGMOD Record, v. 34, n. 2, p. 18-26, 2005.

GOOGLE. Google Colaboratory. Disponível em: <https://colab.research.google.com/>. Acesso em: 09 maio 2024.

HADDON, W. A Logical Framework for Categorizing Highway Safety Phenomena and Activity. Journal of Trauma, v. 10, n. 3, p. 312-323, 1970.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. San Francisco: Morgan Kaufmann, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, 2009.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. Zenodo, (Versão 3.7.1). Disponível em: <https://matplotlib.org/>. Acesso em: 09 maio 2024.

IBM. Random Forest. Disponível em: <https://www.ibm.com/br-pt/cloud/learn/random-forest>. Acesso em: 6 jun. 2024.

IBM. Data Analysis: Exploratory Data Analysis (EDA). Disponível em: https://www.ibm.com/cloud/learn/exploratory-data-analysis. Acesso em: 6 jun. 2024.

MCKINNEY, Wes. pandas: powerful Python data analysis toolkit. Zenodo, (Versão 2.0.3). Disponível em: <https://pandas.pydata.org/>. Acesso em: 09 maio 2024.

MICROSOFT. Excel. Redmond: Microsoft Corporation, (Versão 2404 Build 16.0.17531.20140) 64 bits.

Ministério da Infraestrutura. OMS lança plano para reduzir acidentes e mortes no trânsito até 2030. Portal Gov.br, 28 out. 2021. Disponível em: <https://www.gov.br/transportes/pt-br/assuntos/noticias/2021/10/oms-lanca-plano-para-reduzir-acidentes-e-mortes-no-transito-ate-2030>. Acesso em: 6 jun. 2024.

MITCHELL, T. Machine Learning. McGraw-Hill, Inc., 1997.

ONU Brasil. Objetivos de Desenvolvimento Sustentável. Nações Unidas Brasil. Disponível em: <https://brasil.un.org/pt-br/sdgs>. Acesso em: 10 maio 2024.

OPAS. Apesar do progresso notável, segurança viária continua sendo um problema mundial. Organização Pan-Americana da Saúde, 13 dez. 2023. Disponível em: <https://www.paho.org/pt/noticias/13-12-2023-apesar-do-progresso-notavel-seguranca-viaria-continua-sendo-um-problema-mundial#:~:text=Este%20relat%C3%B3rio%20abrange%20o%20progresso%20entre%202010%20e,pela%20metade%20as%20mortes%20no%20tr%C3%A2nsito%20at%C3%A9%202030>. Acesso em: 09 maio 2024.

PADUA, Mateus. Machine Learning – Métricas de avaliação: Acurácia, Precisão e Recall, F1-score. Medium, 2020. Disponível em: <https://medium.com/@mateuspdua/machine-learning-m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-e-recall-d44c72307959>. Acesso em: 09 maio 2024.

PEDEN, M.; Scurfield, R.; Sleet, D.; Mohan, D.; Hyder, A.A.; Jarawan, E.; Mathers, C. World report on road traffic injury prevention. World Health Organization, 2004.

PIATETSKY-SHAPIRO, G. Discovery, analysis, and presentation of strong rules. In Knowledge Discovery in Databases, AAAI Press, 1991.

PYTHON SOFTWARE FOUNDATION. Python Language Reference, (Versão 3.10.12). Disponível em: <https://www.python.org/>. Acesso em: 09 maio 2024.

SECRETARIA DE GOVERNO - SEADE. Infosiga SP - Sistema de Informações Gerenciais de Acidentes de Trânsito do Estado de São Paulo. Governo Aberto SP, 2015. Disponível em: <http://catalogo.governoaberto.sp.gov.br/dataset/infosiga-sp-sistema-de-informacoes-gerenciais-de-acidentes-de-transito-do-estado-de-sao-paulo.> Acesso em: 09 maio 2024.

TAYLOR, S. J.; LETHAM, B. Prophet: Forecasting at scale (Versão 1.1). Disponível em: <https://github.com/facebook/prophet>. Acesso em: 09 maio 2024.

TRAFFIC accident analysis using data mining techniques. Transportation Research, [s.l.],

TUKEY, J. W. Exploratory Data Analysis. Addison-Wesley, 1977.

WANG, Y.; SHI, Y.; LI, W.; LI, J. Traffic Accident Analysis Using Data Mining Techniques. Transportation Research Part C: Emerging Technologies, v. 95, p. 99-111, 2018.

WHO. Global Status Report on Road Safety 2023. World Health Organization. Disponível em: <https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023>. Acesso em: 09 maio 2024.