Análise e Pré-processamento de Dados de Acidentes de Trânsito para Cidades Inteligentes: Um Estudo de Caso em São Paulo

Andrei Inoue Hirata
UNESP
andreihirata@unesp.br

RESUMO

No contexto das Cidades Inteligentes e Gêmeos Digitais, a análise de dados de trânsito é crucial para o aprimoramento da segurança viária e o planejamento urbano. Este artigo apresenta um pipeline completo de pré-processamento e análise exploratória de dados com o objetivo de preparar um conjunto de dados para a predição da gravidade de acidentes. Utilizando a base de dados públicos da Polícia Rodoviária Federal (PRF) referente aos acidentes de 2023 no estado de São Paulo, aplicamos uma série de técnicas de mineração de dados. Estas incluem limpeza, visualização, discretização de variáveis contínuas (horário do acidente em períodos do dia), encoding de atributos categóricos (One-Hot Encoding) e seleção de atributos utilizando o método de Informação Mútua. A análise exploratória identificou os fins de semana como os períodos de maior ocorrência de acidentes fatais. A seleção de atributos revelou que, além de variáveis óbvias como o número de feridos graves, fatores contextuais como o período do dia (madrugada/noite) e a causa do acidente são preditores de alta relevância para a fatalidade, superando até mesmo algumas condições meteorológicas. Este trabalho demonstra um roteiro prático para a preparação de dados de trânsito complexos e heterogêneos, resultando em uma base de dados robusta e pronta para a modelagem preditiva, servindo como alicerce para sistemas de apoio à decisão em segurança pública.

Palavra Chaves

Mineração de Dados, Cidades Inteligentes, Acidentes de Trânsito, Pré-processamento de Dados, Seleção de Atributos, Modelagem Preditiva

ABSTRACT

In the context of Smart Cities and Digital Twins, the analysis of traffic data is crucial for improving road safety and urban planning. This paper presents a complete pipeline for data pre-processing and exploratory analysis aimed at preparing a dataset for accident severity prediction. Using the 2023 public dataset from the Brazilian Federal Highway Police (PRF) for the state of São Paulo, a series of data mining techniques were applied. These include data cleaning, visualization, discretization of continuous variables (transforming accident times into periods of the day), encoding of categorical attributes (One-Hot Encoding), and feature selection using the Mutual Information method. Exploratory analysis identified weekends as the period with the highest occurrence of fatal accidents. Feature selection revealed that, in addition to obvious variables such as the number of serious injuries, contextual factors like the period of the day (dawn/night) and the cause of the accident are highly relevant predictors of fatality, even outranking some weather conditions. This work demonstrates a practical roadmap for preparing complex, heterogeneous traffic data, resulting in a robust dataset ready for predictive modeling and serving as a foundation for decision support systems in public safety.

Keywords

Data Mining, Smart Cities, Traffic Accidents, Data Preprocessing, Feature Selection, Predictive Modeling.

1. INTRODUÇÃO

O crescente volume de dados gerados em ambientes urbanos impulsiona o desenvolvimento de Cidades Inteligentes, onde a informação é utilizada para otimizar a gestão de recursos e melhorar a qualidade de vida. Nesse cenário, a segurança no trânsito emerge como uma área de grande impacto, e a análise de acidentes pode fornecer subsídios para a criação de políticas públicas mais eficazes. A aplicação de técnicas de Mineração de Dados permite a extração de conhecimento implícito e potencialmente útil a partir de grandes volumes de dados históricos.

Este trabalho apresenta um estudo de caso focado na análise e pré-processamento de dados de acidentes de trânsito ocorridos nas rodovias federais do estado de São Paulo em 2023, disponibilizados pela Polícia Rodoviária Federal (PRF) [1]. O objetivo principal é aplicar um pipeline de técnicas de pré-processamento, conforme os conceitos de Data Quality e Data Preprocessing, para transformar dados brutos em um formato estruturado e otimizado para tarefas de modelagem preditiva, especificamente a classificação da gravidade do acidente (com ou sem fatalidades). Este processo é um passo fundamental para a construção de um componente de um Gêmeo Digital do sistema viário, capaz de simular e prever cenários de risco.

O dataset selecionado é caracterizado por sua heterogenei-

dade, contendo atributos mistos (numéricos e categóricos), o que o torna um excelente objeto de estudo para a aplicação de técnicas como visualização, discretização, encoding e seleção de atributos.[1]

2. METODOLOGIA E DESENVOLVIMENTO

O processo de análise seguiu etapas fundamentais da Mineração de Dados, alinhadas a metodologias como o CRISP-DM, que preconizam a importância da compreensão e preparação dos dados antes da modelagem.

2.1 Análise Exploratória de Dados (AED)

A etapa inicial consistiu no carregamento dos dados e em uma exploração para compreender suas características. O dataset, referente ao ano de 2023, foi filtrado para conter apenas ocorrências no estado de São Paulo, resultando em 4.754 registros. A variável alvo, 'houve_morte', foi criada a partir da coluna 'mortos' e se mostrou altamente desbalanceada, com apenas 4,5% dos acidentes resultando em fatalidades. Esta característica é crucial e deve ser considerada em futuras etapas de modelagem.

As estatísticas descritivas revelaram a escala variada dos atributos numéricos, como 'pessoas' e 'feridos_graves', indicando a necessidade de escalonamento para algoritmos sensíveis à magnitude dos dados.

2.2 Visualização dos Dados:

Foram gerados gráficos para explorar a relação entre os atributos e a variável alvo.

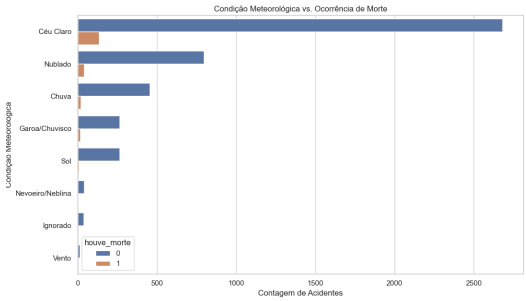

Figure 1: Relação entre Condição Meteorológica e a Ocorrência de Morte.

A Figura 2 evidencia que os acidentes fatais têm um pico de ocorrência nos fins de semana, especialmente aos domingos.

[1]O código-fonte, os dados e os resultados gerados neste trabalho estão disponíveis em: `https://github.com/gitandreihirata/Unesp_Doutorado/tree/main/Mineracao%20de%20Dados/Exercicios/Ex1`

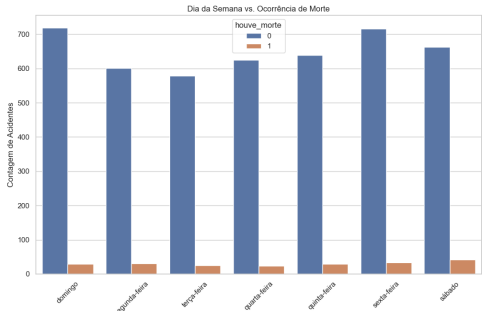

Figure 2: Distribuição de Acidentes por Dia da Semana.

A matriz de correlação (Figura 3) confirma que o número de 'feridos_graves' e de 'pessoas' envolvidas possui a maior correlação positiva com a ocorrência de mortes. A correlação é uma técnica de filtro fundamental na seleção de atributos.

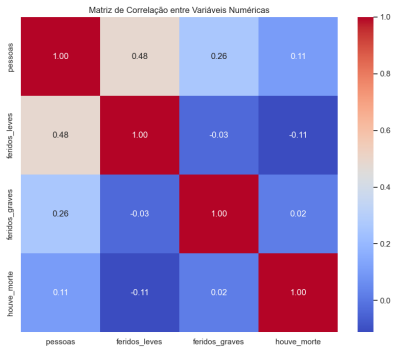

Figure 3: Matriz de Correlação entre Variáveis Numéricas.

Com base na análise dos gráficos, é possível extrair as seguintes observações:

- **Condição Meteorológica:** A análise revela que a grande maioria dos acidentes, inclusive aqueles com vítimas fatais, ocorre em condições de tempo bom, com céu claro ou nublado. Este dado sugere que, embora condições adversas como a chuva sejam um fator de risco conhecido, o excesso de confiança dos condutores em dias de boa visibilidade pode ser um fator contribuinte para a ocorrência de sinistros;
- **Dia da Semana:** Observa-se uma maior frequência de acidentes na sexta-feira e no sábado. Em relação aos acidentes fatais, o pico de ocorrências se concentra no final de semana, com destaque para o domingo;
- **Matriz de Correlação:** Conforme esperado, a matriz de correlação demonstra que o número total de pessoas envolvidas em um acidente e o número de feridos graves apresentam a correlação mais forte com a ocorrência de mortes. Isso indica que acidentes com

maior número de envolvidos e com ferimentos de maior gravidade têm maior probabilidade de resultar em fatalidades.

2.3 Pré-processamento de Dados

Esta etapa focou em transformar os dados brutos em um formato adequado para algoritmos de machine learning, aplicando técnicas de discretização, encoding e escalonamento.

- **Discretização de Variável Contínua:** A variável 'horario' foi convertida para um valor numérico (hora) e, em seguida, discretizada em quatro períodos categóricos: 'Manhã', 'Tarde', 'Noite' e 'Madrugada'. A discretização transforma atributos numéricos em categóricos, o que pode simplificar a relação com a variável alvo para certos modelos;
- **Encoding de Variáveis Categóricas:** Atributos nominais como 'dia_semana' e 'causa_acidente' foram convertidos em formato numérico utilizando a técnica de One-Hot Encoding. Este método cria novas colunas binárias para cada categoria, evitando a imposição de uma ordem artificial que ocorreria com o Label Encoding. O processo expandiu o número de atributos para 206.
- **Escalonamento de Atributos (Feature Scaling):** As variáveis numéricas foram padronizadas utilizando o 'StandardScaler' (Z-score), que transforma os dados para terem média 0 e desvio padrão 1. Este passo é essencial para algoritmos que se baseiam em medidas de distância, como SVM e k-NN.

3. SELEÇÃO DE ATRIBUTOS (FEATURE SELECTION)

Para reduzir a dimensionalidade e focar nos preditores mais relevantes, foi aplicada a técnica de Informação Mútua. Este é um método de filtro que captura dependências lineares e não-lineares entre os atributos e a variável alvo.

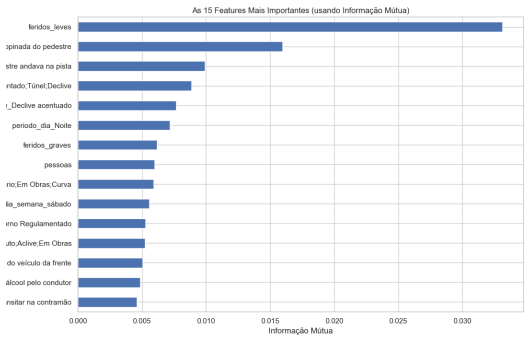

Figure 4: As 15 features mais importantes segundo a Informação Mútua.

A Figura 4 exibe as 15 variáveis mais importantes. Os resultados destacam que 'feridos_leves' e 'feridos_graves' são os principais preditores. Adicionalmente, causas específicas de acidentes ('Entrada inopinada do pedestre', 'Pedestre andava na pista') e o período do dia ('periodo_dia_Noite') emergem como fatores de alta relevância, demonstrando a capacidade da análise em identificar fatores de risco detalhados.

4. DISCUSSÃO E CONCLUSÃO

Este trabalho demonstrou a aplicação de um pipeline de pré-processamento de dados de mineração em um problema prático e relevante de cidades inteligentes. A análise dos dados de acidentes da PRF em São Paulo permitiu não apenas confirmar padrões esperados, mas também extrair insights valiosos sobre os fatores que influenciam a gravidade dos acidentes.

A aplicação rigorosa de técnicas de pré-processamento, como a discretização do horário e o encoding de variáveis categóricas, foi fundamental para transformar dados complexos e heterogêneos em um formato estruturado. A etapa de seleção de atributos validou a importância de variáveis contextuais que vão além das condições climáticas, como o comportamento do condutor/pedestre e o período do dia.

Conclui-se que o pré-processamento de dados é uma etapa indispensável e de grande impacto no ciclo de vida de um projeto de Mineração de Dados. As técnicas aplicadas resultaram em um dataset robusto e informativo, pronto para ser utilizado na próxima fase de modelagem preditiva. Os insights gerados já possuem valor intrínseco, podendo subsidiar políticas de segurança viária e alocação de recursos de fiscalização, contribuindo para um trânsito mais seguro e um planejamento urbano mais eficiente.

5. REFERENCES

[1] Dados Abertos da PRF. `https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf`, 2025. Acessado em 21 de setembro de 2025.