

IDENTIFICAÇÃO DE PERFIS EM SINISTROS DE TRÂNSITO EM FORTALEZA-CE USANDO CLUSTERIZAÇÃO COM APRENDIZADO DE MÁQUINA

Mayra Miranda

Discente - Centro Universitário Fametro - Unifametro

Mayra.miranda@aluno.unifametro.edu.br

Renato Mateus

Discente - Centro Universitário Fametro - Unifametro

Renato.souza01@aluno.unifametro.edu.br

Kaio Mesquita

Docente - Centro Universitário Fametro - Unifametro

kaio.mesquita@professor.unifametro.edu.br

Área Temática: Inovação e inteligência artificial.

Área de Conhecimento: Ciências Tecnológicas.

Encontro Científico: XII Encontro de Monitoria.

RESUMO

Este estudo analisa sinistros de trânsito em Fortaleza-CE, utilizando clusterização para identificar padrões e perfis de risco. Entre 2015 e 2020, mais de 10 mil sinistros foram registrados, destacando a importância de entender os fatores que influenciam a gravidade dos acidentes. **Objetivo:** Categorizar perfis de sinistros através de métodos de aprendizado de máquina não supervisionado e identificar variáveis críticas, como iluminação, controle de tráfego e condições climáticas. **Métodos:** Foram coletados dados da plataforma VIDA, aplicadas técnicas de clusterização e Análise de Componentes Principais (PCA). **Resultados:** A análise identificou dois grupos principais: um com sinistros de menor gravidade, onde predominam feridos e ilesos, e outro com sinistros mais graves, com maior proporção de mortos. **Considerações finais:** A análise de dados é fundamental para desenvolver políticas de prevenção mais eficazes. A clusterização facilita a identificação de grupos de risco, orientando intervenções que possam melhorar a segurança viária.

Palavras-chave: Sinistros de trânsito; Clusterização; Análise de Componentes Principais; Segurança viária.

INTRODUÇÃO

Os sinistros de trânsito representam um problema grave, tanto em termos de danos materiais quanto em perda de vidas. A Norma ABNT NBR 10697:2020 define sinistro como qualquer evento que cause danos ao veículo, à carga ou resulte em lesões a pessoas e/ou animais. De acordo com a Secretaria Municipal de Saúde (2023), em 2014, Fortaleza-CE

ocupava a sexta posição entre as cidades com maiores índices de morte no trânsito, contudo, verificou-se uma queda significativa, atingindo a vigésima oitava posição em 2022. No período entre 2015 e 2020, foram registrados mais de 10 mil sinistros, dos quais 150 foram fatais e 10.114 resultaram em vítimas feridas, totalizando 158 vítimas fatais e 11.846 feridos. Isso revela uma média alarmante de uma morte a cada 65 sinistros.

A literatura destaca que a análise de dados de sinistros é fundamental para desenvolver políticas públicas mais eficazes (HAUER; PERSAUD; VINER, 2002). Técnicas como a clusterização permitem identificar padrões e agrupamentos de dados que facilitam a compreensão de fenômenos complexos (CASSIANO, 2014). Assim, a aplicação dessas técnicas na análise de sinistros pode apoiar a formulação de estratégias direcionadas para reduzir acidentes e salvar vidas.

Desta forma, este trabalho visa analisar os dados de sinistros de trânsito em Fortaleza-CE, utilizando técnicas de clusterização para identificar perfis de risco e variáveis críticas que influenciam a gravidade dos acidentes.

METODOLOGIA

Este estudo foi dividido em três etapas principais: coleta e tratamento dos dados, definição de hipóteses e análise dos agrupamentos. Os dados foram extraídos da plataforma VIDA, que contém informações sobre sinistros de trânsito em Fortaleza entre 2015 e 2020. A linguagem Python, com bibliotecas como Pandas e NumPy, foi utilizada para tratar e organizar esses dados, assegurando a integridade e a uniformidade das informações. As variáveis categóricas foram convertidas em dummies, facilitando a modelagem em aprendizado de máquina.

As variáveis analisadas incluíram aspectos como interseções, tipo de acidente, iluminação, condições climáticas, estado da pista, tipo de cruzamento, controle de tráfego, uso do solo e período do dia. Após a etapa de tratamento, foram aplicadas técnicas de imputação para corrigir dados ausentes e eliminar outliers. A clusterização foi escolhida como método para identificar padrões nos sinistros.

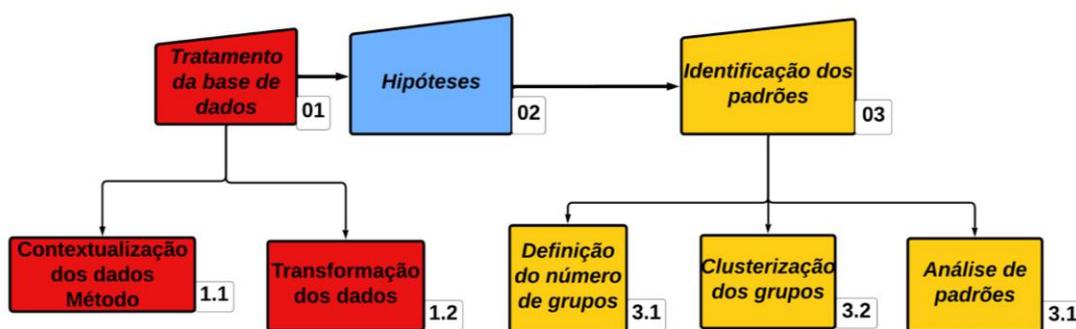


Figura 1 – Fluxograma da metodologia

Fonte: Autores

No tratamento de dados, estes foram concatenados em um único arquivo, e aplicou-se a técnica de transformação para uniformizar formatos e preencher lacunas. O método dummy foi usado para as variáveis categóricas, permitindo uma melhor interpretação dos modelos de aprendizado de máquina. As variáveis essenciais analisadas incluíram “interseção”, “natureza do sinistro”, “iluminação” e “tempo”.

Tabela 1 – Variáveis com legendas

| Variável | Número de categorias | Descrição |
|------------------------|----------------------|--|
| Interseção | 2 | Existência ou não de interseções |
| Natureza dos sinistros | 22 | Tipo de acidente, como colisão, atropelamento, capotamento, entre outros. |
| Iluminação | 7 | Descreve as condições de iluminação no momento do acidente, como luz do dia, via iluminada, ou mal iluminada, entre outros |
| Tempo | 5 | Refere-se às condições climáticas no momento do acidente, como bom, chuva, nublado, etc. |
| Superfície da pista | 7 | Indica as condições da superfície da pista durante o acidente, como seca, molhada, ou com material derrapante. |
| Tipo de cruzamento | 10 | Define o tipo de cruzamento onde ocorreu o acidente, como cruz, rotatória, T, Y, entre outros. |
| Controle de tráfego | 13 | Especifica o tipo de controle de tráfego presente no local do acidente, como semáforo, pare, cancela, ou outros. |
| Uso do solo | 9 | Descreve o tipo de uso do solo onde ocorreu o acidente, como comercial, residencial, industrial, rural, entre outros. |
| Hora | Inteiro | Faixa Horária da ocorrência |
| Ilesos | Inteiro | Número de ilesos envolvidos no registro |
| Feridos | Inteiro | Número de feridos envolvidos no registro |
| Mortes | Inteiro | Número de mortes envolvidas no registro |

Fonte: Autores

O tratamento dos dados incluiu a eliminação de inconsistências e a consolidação de variáveis para garantir um modelo robusto. Após essa preparação, as variáveis foram submetidas ao processo de clusterização.

Na definição de hipóteses duas hipóteses foram formuladas: (i) a existência de padrões distintos nos sinistros relacionados às características das vias e condições climáticas; (ii) a maior influência de fatores operacionais (como controle de tráfego) em comparação a fatores externos (como clima) na gravidade dos acidentes.

A clusterização foi realizada com o algoritmo K-Means, e o número ideal de clusters foi determinado pelo método da silhueta. Com um score de 0.8, concluiu-se que a divisão em dois clusters era a mais adequada, indicando uma boa separação entre os grupos.

Após a identificação dos clusters, aplicou-se a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados, facilitando a visualização dos grupos. As duas principais componentes explicaram a maior parte da variação nos dados. As variáveis "natureza do sinistro" e "controle de tráfego" foram identificadas como as mais influentes na separação dos clusters.

A técnica de clusterização permitiu identificar perfis distintos de sinistros, e a PCA destacou variáveis críticas para a gravidade dos acidentes, facilitando o planejamento de intervenções de segurança viária.

RESULTADOS E DISCUSSÃO

Para determinar o número ideal de clusters, utilizamos o método da silhueta, que avalia a coesão interna e a separação entre os grupos. Os resultados indicaram que dois clusters são a estrutura mais adequada, com um score de 0.8. Este valor sugere uma boa segmentação dos dados, confirmando que dois clusters são suficientes para distinguir padrões relevantes.



Figura 2: Método da silhueta

Fonte: Autores

Após a definição dos clusters, aplicamos a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados e facilitar a visualização. A PCA revelou duas componentes principais que explicam a maior parte da variância dos dados: 18.17% e 14.33%, respectivamente. Embora haja alguma sobreposição, a separação entre os clusters indica diferenças claras nas características dos sinistros.

Tabela 2 - Variância explicada para os componentes 1 e 2

| Componente | Variância explicada (%) |
|------------------------|-------------------------|
| Componente Principal 1 | 18,17 |
| Componente Principal 2 | 14,33 |

Fonte: Autores

O Cluster 0, com menor dispersão, está associado a sinistros menos graves, enquanto o Cluster 1, mais disperso, está relacionado a sinistros mais graves, com maior incidência de mortes e feridos graves. A PCA destacou que as variáveis "natureza do sinistro" e "controle de tráfego" são as mais influentes na formação dos clusters.

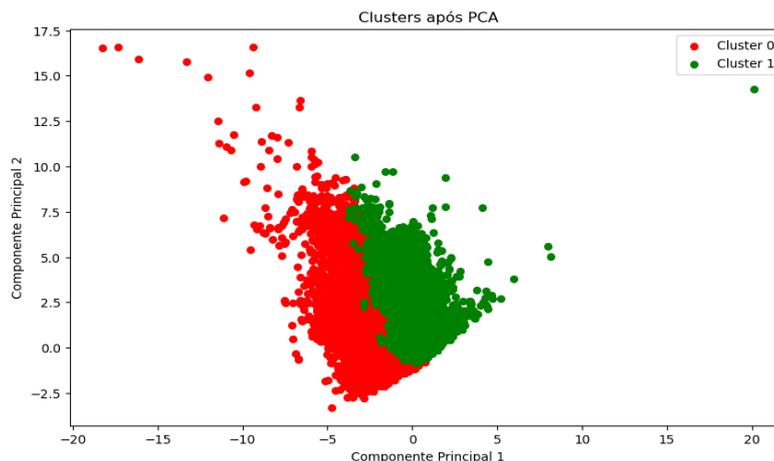


Figura 3 - Aplicação do PCA

Fonte: Autores

A análise dos atributos revelou uma clara distinção entre os clusters. O Cluster 0 apresentou menor incidência de mortes e feridos graves, sendo predominante em sinistros de menor gravidade. As variáveis "superfície da pista" e "iluminação" foram relevantes para os acidentes deste grupo, indicando que sinistros menos graves tendem a ocorrer em condições de pista e iluminação mais favoráveis.

Em contraste, o Cluster 1 mostrou uma maior concentração de sinistros graves, com 65.7% dos acidentes resultando em mortes. As variáveis "iluminação" e "tempo" demonstraram forte correlação neste cluster, sugerindo que condições adversas de visibilidade e clima estão associadas a uma maior gravidade dos sinistros.

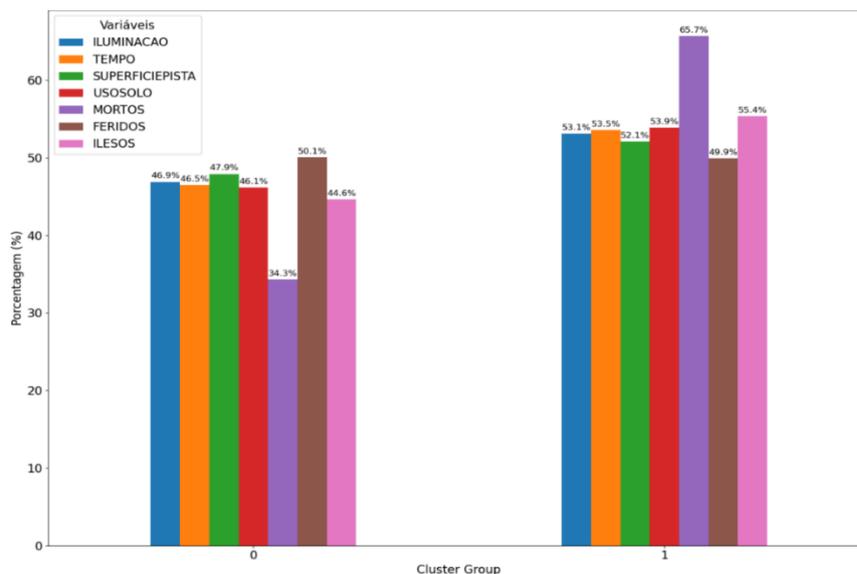


Figura 4: Porcentagem das médias das variáveis por grupo de cluster

Fonte: Autores

Os resultados destacam o papel crucial das variáveis "natureza do sinistro" e "controle de tráfego" na gravidade dos acidentes, enquanto fatores climáticos como "tempo" e "iluminação" contribuem significativamente para a diferenciação entre os clusters. No Cluster 0, as condições de pista e iluminação mais favoráveis estão associadas a acidentes menos graves. No Cluster 1, condições adversas de visibilidade e controle de tráfego inadequado parecem ser fatores determinantes para a gravidade dos sinistros.

Tabela 3 - Score de importância para os componentes principais 1 e 2

| CP1 | | CP2 | |
|---------------------|----------------------|---------------------|----------------------|
| Variáveis | Score de importância | Variáveis | Score de importância |
| Natureza | 0.566806 | Ilesos | 0.587339 |
| Controle do tráfego | 0.457552 | Feridos | 0.557737 |
| Tempo | 0.193644 | Interseção | 0.323399 |
| Superfície da pista | 0.189381 | Uso do solo | 0.266891 |
| Interseção | 0.129737 | Tempo | 0.213531 |
| Uso do solo | 0.116198 | Natureza | 0.211521 |
| Ilesos | 0.102400 | Superfície da pista | 0.187265 |
| Feridos | 0.060801 | Tipo de cruzamento | 0.114607 |
| Iluminação | 0.053140 | Controle de tráfego | 0.089734 |
| Hora | 0.037457 | Hora | 0.011479 |
| Mortos | 0.023638 | Mortos | 0.006380 |
| Tipo de cruzamento | 0.020305 | Iluminação | 0.004087 |

Fonte: Autores

A técnica de clusterização revelou-se eficaz para identificar perfis de risco distintos entre os sinistros de trânsito. Esses achados enfatizam a necessidade de intervenções específicas em áreas com iluminação deficiente e controle de tráfego inadequado, além de uma maior atenção às condições climáticas no planejamento de políticas de segurança viária.

CONSIDERAÇÕES FINAIS

Este estudo demonstrou a utilidade da clusterização na identificação de padrões em sinistros de trânsito em Fortaleza-CE. Foram identificados dois perfis distintos: um com sinistros de menor gravidade, caracterizado por condições mais favoráveis de pista e iluminação; e outro com sinistros mais graves, onde a iluminação inadequada e condições climáticas adversas foram fatores significativos.

A análise confirmou a influência de fatores como "natureza do sinistro", "controle de tráfego" e "tempo" na gravidade dos acidentes. A identificação desses padrões pode orientar a formulação de políticas públicas mais eficazes, focando em intervenções que melhorem a visibilidade e o controle de tráfego, especialmente em áreas críticas.

Sugere-se, como trabalhos futuros, a incorporação de dados geoespaciais para mapear pontos críticos de sinistros e a análise de variações temporais para avaliar o impacto de políticas de segurança implementadas ao longo do tempo. A inclusão dessas informações pode auxiliar no desenvolvimento de estratégias preventivas mais eficientes para a redução de sinistros graves e melhoria da segurança viária em Fortaleza.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 10697:2020 - Pesquisa de sinistros de trânsito - Terminologia. Rio de Janeiro: ABNT, 2020.

CASSIANO, Keila Mara. Análise de séries temporais usando análise espectral singular (SSA) e clusterização de suas componentes baseada em densidade. 2014. Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2014.

HAUER, E.; PERSAUD, B.; VINER, J. The role of data in traffic safety management. *Traffic Engineering & Control*, v. 43, n. 8, p. 298-302, 2002.

SECRETARIA MUNICIPAL DE SAÚDE. Relatório Anual de Segurança Viária. Fortaleza: Secretaria Municipal de Saúde, 2023.