**Caracterização de Clusters de Acidentes de Trânsito no Paraná: Identificando Padrões e Fatores de Risco**

Regis Oscar Land;Auberth Henrik Venson

**Caracterização de Clusters de Acidentes de Trânsito no Paraná: Identificando Padrões e Fatores de Risco**

**Resumo**

O trabalho analisa as principais causas que influenciam a gravidade dos acidentes de trânsito no Paraná entre 2019 e 2023. Utilizam-se dados da Polícia Rodoviária Federal e aplicam-se técnicas de clusterização geográfica para dividir o estado em quatro regiões. Em seguida, emprega-se a regressão logística multinomial para cada região, identificando que comportamentos de pedestres são os principais fatores associados à gravidade dos acidentes. O estudo aponta a necessidade de intervenções regionais específicas, como campanhas educativas, melhorias na infraestrutura viária e maior fiscalização, visando à redução da gravidade dos acidentes e à promoção da segurança viária.

**Palavras-chave:** Acidentes de trânsito; Gravidade dos acidentes; Clusterização geográfica; Regressão logística multinomial; Segurança viária

1. **Introdução**

Estudos demonstram que os acidentes de trânsito geram custos sociais e econômicos consideráveis, impactando não apenas as vítimas e seus familiares, mas também toda a sociedade. O estudo "Estimativa dos custos associados aos acidentes de trânsito" do Observatório Nacional de Segurança Viária (2020) detalha que estes custos podem aumentar em mais de 10% até 2027.

Diante desses números, é evidente a necessidade de políticas públicas eficazes para lidar com os desafios relacionados ao trânsito, visando garantir a segurança viária e a qualidade de vida da população.

Dados do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2022 demonstram um aumento significativo na frota de veículos em relação ao crescente número de habitantes, o que consequentemente eleva as chances de ocorrência de acidentes de trânsito. No Brasil, conforme levantamento do IBGE em 2022, havia aproximadamente 2 habitantes para cada veículo registrado. Além disso, dados do IBGE apontam que a frota nacional passou de 45.029.257 veículos em 2006 para 115.116.532 em 2022, representando um aumento de 155,65% ao longo de 16 anos (IBGE, 2022).

Os setores governamentais que administram as rodovias enfrentam barreiras ao identificar onde e como implementar medidas que possam ter um impacto na segurança viária (Ansari & Al-Shabi, 2012). Atualmente, existe uma falta de material que realize uma caracterização dos acidentes de trânsito em território nacional e que apresente resultados de forma simples de ser analisada. O processo de identificação de intervenções corretivas e preventivas inicia-se com a identificação dos locais com maior incidência de acidentes e a análise dos elementos que colaboram para essa frequência (Li & Chang, 2019; Su *et al*., 2019).

Espera-se que este estudo contribua para a redução do número de acidentes de trânsito nas rodovias federais brasileiras, promovendo uma melhor segurança viária, preservação de vidas e redução de gastos públicos nessa área.

A pergunta de pesquisa que norteia este estudo é: Quais são os padrões distintivos e os fatores de risco associados à clusters de acidentes de trânsito identificados por dados abertos da Polícia Rodoviária Federal no estado do Paraná?

1. **Material e Métodos**

A Polícia Rodoviária Federal disponibiliza dados abertos de acidentes de trânsito desde 2007 em formato CSV. Nestes relatórios, são encontradas informações sobre data, local, causa do acidente, tipos de acidente, gravidade, sentido da via, condição meteorológica, entre outras características de acidentes que foram usadas como referência para este estudo (Dados Abertos da PRF: Acidentes de Trânsito, 2024).

Entre as informações disponíveis nessa base de dados tem-se:

* **id**: Identificador único de cada acidente.
* **data\_inversa**: Data do acidente no formato inverso (provavelmente AAAA-MM-DD).
* **dia\_semana**: Dia da semana em que o acidente ocorreu.
* **horario**: Horário específico do acidente.
* **uf**: Unidade Federativa (estado) onde o acidente aconteceu.
* **br**: Número da rodovia federal (BR) envolvida.
* **km**: Marcador de quilômetro onde ocorreu o acidente.
* **municipio**: Município onde o acidente ocorreu.
* **causa\_acidente**: Motivo ou causa principal do acidente.
* **tipo\_acidente**: Tipo de acidente (ex.: colisão frontal, capotamento).
* **classificacao\_acidente**: Caracterização do acidente (ex.: com vítimas, sem vítimas).
* **fase\_dia**: Período do dia (manhã, tarde, noite) em que ocorreu o acidente.
* **sentido\_via**: Sentido da via no momento do acidente.
* **condicao\_metereologica**: Condições climáticas durante o acidente (ex.: chuva, tempo claro).
* **tipo\_pista**: Tipo de pista (simples, dupla).
* **tracado\_via**: Características da via (reta, curva).
* **uso\_solo**: Tipo de área onde o acidente ocorreu (urbano, rural).
* **pessoas**: Número total de pessoas envolvidas.
* **mortos**: Número de óbitos resultantes do acidente.
* **feridos\_leves**: Número de pessoas com ferimentos leves.
* **feridos\_graves**: Número de pessoas com ferimentos graves.
* **ilesos**: Número de pessoas que saíram ilesas.
* **ignorados**: Dados desconhecidos ou não informados.
* **feridos**: Número total de feridos.
* **veiculos**: Número de veículos envolvidos no acidente.
* **latitude**: Coordenada geográfica de latitude do local.
* **longitude**: Coordenada geográfica de longitude do local.
* **regional**: Região administrativa responsável.
* **delegacia**: Delegacia que registrou o acidente.
* **uop**: Unidade Operacional da Polícia Rodoviária Federal responsável.

A Análise de Agrupamentos é um conjunto de técnicas multivariadas que possuem o objetivo principal de agrupar objetos com base em suas características compartilhadas. Os algoritmos dedicados a essa tarefa exploram padrões ou formam agrupamentos. A clusterização envolve a análise desses conjuntos (clusters), avaliando o grau de semelhança ou diferença entre eles e seus elementos, geralmente através de uma função de distância, na qual menor distância indica maior similaridade entre os dados analisados (Fávero & Belfiore, 2017).

Para o delineamento da pesquisa, este trabalho seguiu a estrutura proposta por Hair *et al*. (2009) ao realizar a análise de agrupamento de dados:

1. Definir o objetivo da análise;
2. Realizar estudo de padronização e de detecção de observações atípicas;
3. Medir a similaridade entre as variáveis;
4. Padronizar as variáveis;
5. Selecionar o algoritmo de agrupamento;
6. Identificar se será necessária uma reespecificação da análise de agrupamentos;
7. Validar e caracterizar os agrupamentos.

A partir das etapas detalhadas, e de Gil (2008), o trabalho pode ser classificado da seguinte maneira:

**Quadro 1 ­- Classificação da pesquisa**

|  |  |  |
| --- | --- | --- |
| **Aspecto** | **Classificação** | **Justificativa** |
| Abordagem | Quali­-Quanti | Neste trabalho foram  estudados aspectos quantitativos para realizar o agrupamento dos dados e qualitativos para analisar os resultados |
| Natureza | Aplicada | O trabalho busca que autoridades possam identificar características de acidentes e tomem as devidas providências para evitá-los |
| Objetivos | Descritiva | O trabalho expõe métricas e temas de situações que já ocorreram |
| Procedimentos | Análise de dados descundários | O presente estudo foi baseado em dados abertos dispobilizados pela polícia rodoviária federal |

Fonte: Autoria Própria (2024)

* 1. Método de clusterização

K-means foi o algoritmo utilizado para o agrupamento de dados deste trbalaho. Ele é caracterizado como um método “guloso”, onde em cada iteração o ponto que parece mais "apetitoso" é selecionado. Os passos do algoritmo, ao lidar com um conjunto de dados, são os seguintes:

1. Selcionar k objetos para serem os centroides iniciais que serão o ponto de partida para a criação dos k clusters. A escolha desses centroides pode ser feita de forma aleatória, baseada em alguma técnica específica ou utilizando heurísticas (DE AMO, 2004).
2. Para cada objeto no conjunto disponível, o algoritmo calcula a distância em relação aos k centroides existentes, utilizando distância euclidiana. Em seguida, o objeto é atribuído ao cluster cujo centroide possui a menor distância euclidiania em relação a ele (De KAMBER, 2001).
3. Para cada um dos k clusters, um novo centroide é calculado. Isso é feito obtendo a média dos valores de cada atributo dos elementos que pertencem ao cluster. Essa média é então atribuída como o novo centro de gravidade do cluster, ajudando a balanceá-lo. O cálculo do novo centroide é definido por:

sendo o valor médio dos objetos do cluster k, o número de objetos no agrupamento k e o valor do objeto n para os atributos q, dentro do somatório, pertencendo a k (MAIMON e ROKACH, 2010).

1. Após realocar os objetos para os clusters, verifica-se se algum objeto foi realocado para um cluster diferente. Se isso ocorrer, o algoritmo retorna ao Passo 2 e repete o processo. Esse ciclo de recalculo do centroide é repetido até o ponto em que não haja objeto realocados, o que indica que os clusters estão esáveis (DE AMO, 2004).
   1. Regressão Logística

A regressão logística é uma técnica amplamente utilizada em estatística para modelar dados em que a variável dependente é categórica. Essa técnica é especialmente útil quando o objetivo é prever a probabilidade de ocorrência de um determinado evento, com base em uma ou mais variáveis independentes (MONTGOMERY; PECK; VINING, 2012). Diferentemente da regressão linear, onde a variável dependente é contínua, na regressão logística, essa variável assume valores binários ou multiclasses, sendo comum sua aplicação em áreas como medicina, biologia, marketing e ciências sociais (HOSMER; LEMESHOW, 2000).

A equação da regressão logística pode ser descrita como uma transformação da regressão linear que permite que o resultado seja interpretado como uma probabilidade. A função logística, também conhecida como função sigmoide, limita a saída da equação para valores entre 0 e 1, o que é interpretado como a probabilidade de pertencimento a uma classe. Para ajustar um modelo de regressão logística, utiliza-se o método de máxima verossimilhança, que estima os coeficientes que maximizam a probabilidade de observar os dados amostrais. Além disso, a interpretação dos coeficientes na regressão logística é diferente da regressão linear. Cada coeficiente β indica o impacto de uma unidade de mudança na variável X na razão de chances (odds ratio) do evento de interesse (MONTGOMERY; PECK; VINING, 2012).

Segundo Fávero e Belfiore (2017), a regressão logística multinomial também pode ser aplicada quando a variável dependente possui mais de duas categorias, sendo uma extensão natural do modelo binário. A regressão logística múltipla refere-se ao contexto da regressão logística em que a variável dependente é binária ou dicotômica, ou seja, possui duas categorias, e há mais de uma variável independente

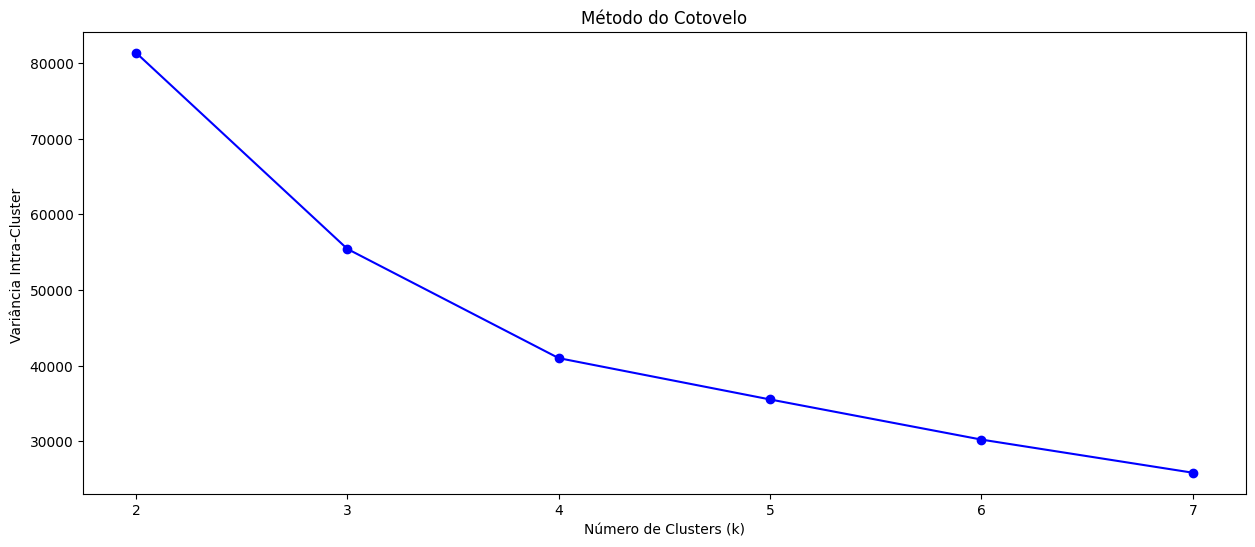
1. **Resultados e Discussão**

Esta seção apresenta os resultados dos algoritmos e suas análises. A ordem dos resultados é apresentada na sequência: tomada de decisão do número ideal de clusters, clusterização dos acidentes, regressão logística para caracterizar esses clusters.

* 1. Tomada de decisão de número de ideal de clusters

Para definir o número de grupos para a análise de agrupamento, uma técnica amplamente utilizada é o método do cotovelo. Este método envolve a execução do algoritmo de agrupamento k-means, variando o número de clusters (k) e calculando a variância intra-cluster para cada valor de k. A variância intra-cluster é medida pela soma das distâncias quadráticas médias dos pontos de dados em relação aos centróides de seus respectivos clusters. Em seguida, é gerado um gráfico com o número de clusters no eixo X e a variância intra-cluster no eixo Y. O 'cotovelo' do gráfico, que representa o ponto de inflexão onde a redução da variância intra-cluster começa a diminuir significativamente, indica o número ideal de clusters.

Figura 1 - Variância Intra-cluster para Feridos e Mortos



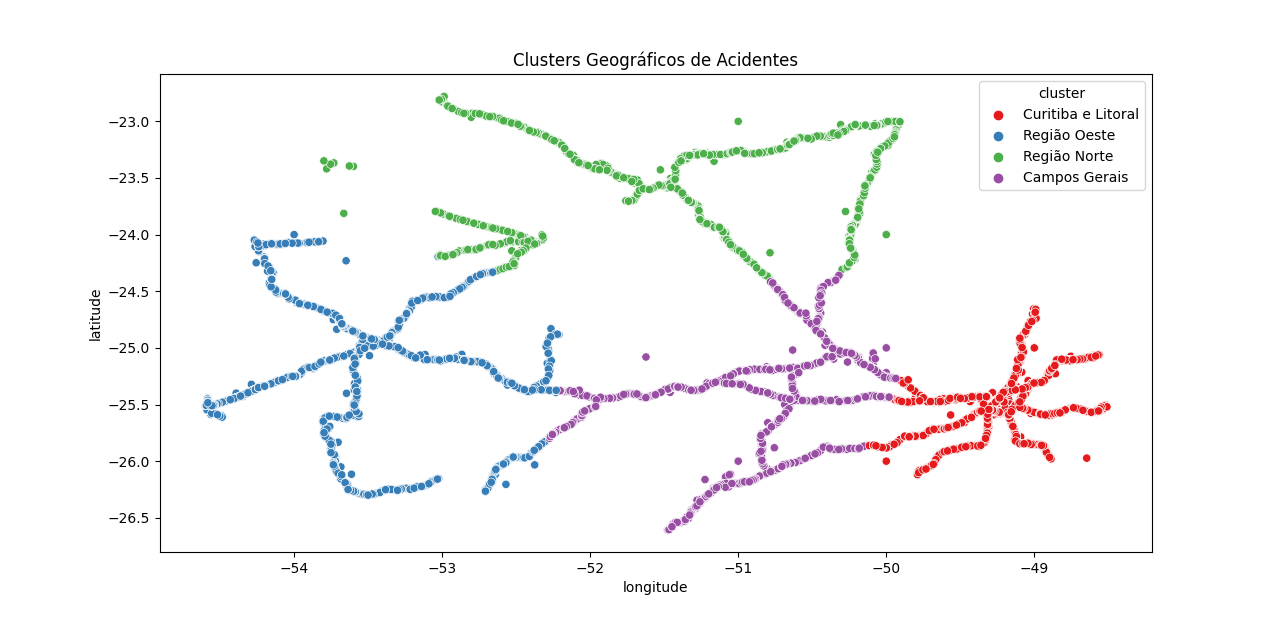
Fonte: Autoria Própria

Ao observar o gráfico, pode ser inferido que 4 grupos pode ser considerado o número ideal de grupos para a aplicação de k-means neste grupo de dados.

* 1. Clusterização dos acidentes

Para visualizar a distribuição espacial dos acidentes, foi elaborado um gráfico com as coordenadas de latitude e longitude dos locais dos acidentes no Paraná. A clusterização geográfica, utilizando as bibliotecas Python scikit-learn, matplotlib e seaborn, dividiu os dados em quatro regiões: Região Norte, Curitiba e Litoral, Região Oeste e Campos Gerais. Essa segmentação agrupou os acidentes por proximidade geográfica, facilitando a identificação de padrões regionais na gravidade dos acidentes. Dessa forma, a análise considerou as características espaciais e associou as causas dos acidentes a áreas específicas do estado.

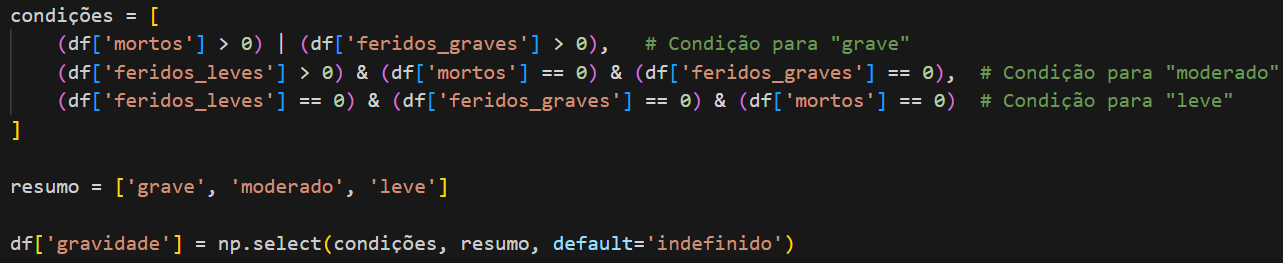
Figura 2 – Distribuição Geográfica de Clusters



Fonte: Autoria Própria

* 1. Regressão Logística

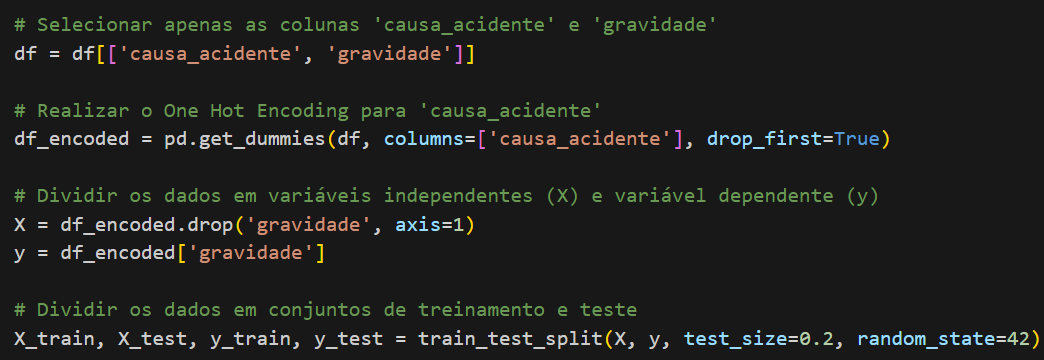
Para compreender melhor a gravidade dos acidentes nas rodovias do Paraná para cada um dos clusters, foi aplicada a técnica de regressão logística multinomial. A análise teve como objetivo categorizar a gravidade dos acidentes em três níveis: grave, moderado e leve, e identificar os principais fatores que influenciam a ocorrência de acidentes graves. Para utilizar este método, foi criada uma nova variável no conjunto de dados chamada “gravidade”. A categoria gravidade classifica os acidentes da seguinte forma:

Figura 3 – Código Para Criar Variálve Gravidade  
Fonte: Autoria Própria

* Grave: acidentes que resultaram em mortes ou feridos graves.
* Moderado: acidentes com apenas feridos leves e sem mortes ou feridos graves.
* Leve: acidentes sem feridos ou mortos.

Esta análise utilzou a variável “causa\_acidente” como uma variável independente significativa, que como o próprio nome infere, é uma string com o motivo que provevlmente ocasionou o acidente. Foi utilizado One Hot Enconding para transformar a variável categórica “causa\_acidente” em variáveis dummy, facilitando a inclusão no modelo de regressão. A divisão dos dados foi realizada em variáveis independentes (X) e a variável dependente (y). Posteriormente, os dados foram divididos em conjuntos de treinamento e teste para a construção e validação do modelo, conforme pode ser visto na figura do código a seguir:

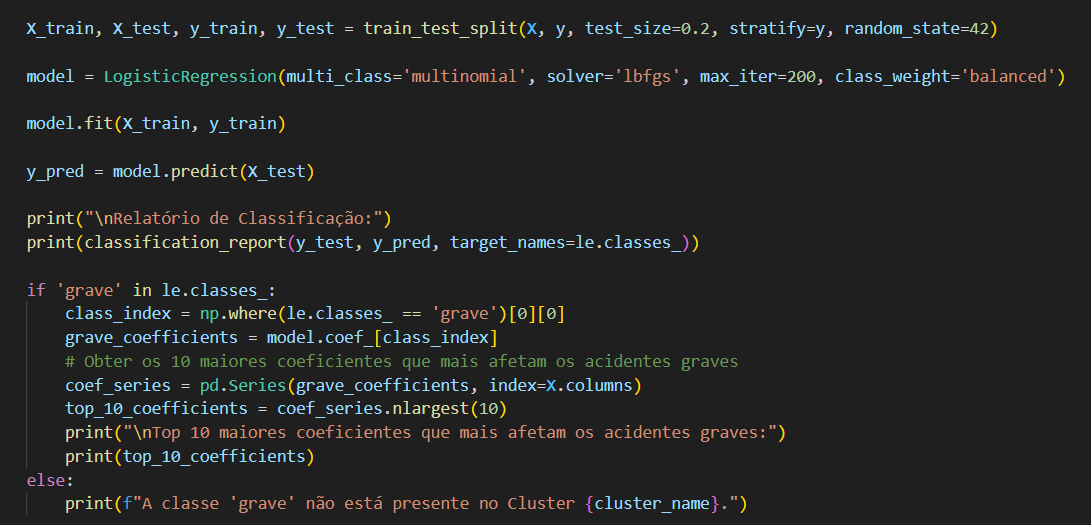
Figura 4 – Código Para Treinar Modelo



Fonte: Autoria Própria

Foi utilizado o modelo de regressão logística multinomial, apropriado para a previsão de variáveis categóricas com mais de duas classes. Os coeficientes do modelo foram analisados para identificar os fatores que mais influenciam nos acidentes graves em cada um dos 4 clusters:

Figura 4 – Código Para Identifcar Principais Causas De Acidentes



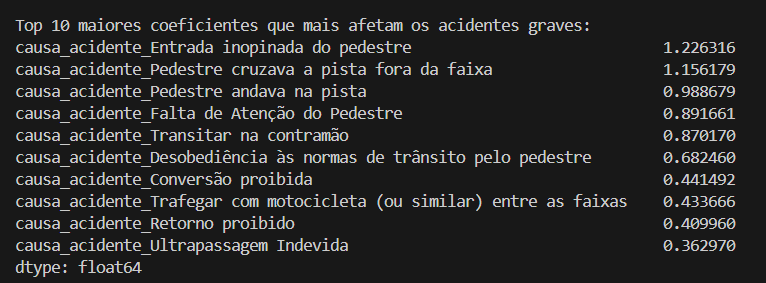
Fonte: Autoria Própria

O modelo de regressão logística multinomial calcula as probabilidades (baseado nos valores dos coeficientes de sua equação) de um acidente pertencer a cada uma das classes de acidente (leve, moderado e grave) com base nas variáveis independentes (as colunas representando as causas do acidente). Porém foram selecionadas apenas as informações dos casos graves, de acordo com o escopo do presente trabalho. Quanto maior o valor do coeficiente, maior a influência desta causa no tipo de acidete.

* + 1. Região Norte

Os coeficientes do modelo do cluster da região Norte são apresentados na imagem a seguir:

Figura 5 – Coeficientes do cluster Região Norte



Fonte: Autoria Própria

Os resultados indicaram que comportamentos inadequados de pedestres são as principais causas associadas a acidentes graves. Fatores como "Entrada inopinada do pedestre", "Pedestre cruzava a pista fora da faixa" e "Pedestre andava na pista" apresentaram os maiores coeficientes. Isso evidencia a necessidade de intervenções focadas na segurança dos pedestres, incluindo campanhas de conscientização e melhorias na infraestrutura viária para travessias seguras.

* + 1. Curitiba e Litoral

Os coeficientes do modelo do cluster Curitiba e Litoeral são apresentados na imagem a seguir:

Figura 6 – Coeficientes do cluster Curitiba e Litoral

A black screen with white text

Description automatically generated

Fonte: Autoria Própria

Em Curitiba e Litoral, observou-se uma tendência similar, com destaque para causas relacionadas ao pedestre. Além das já mencionadas na Região Norte, "Desobediência às normas de trânsito pelo pedestre" e "Falta de Atenção do Pedestre" também tiveram impacto significativo. Esses achados reforçam a importância de ações educativas voltadas tanto para pedestres quanto para condutores, bem como a implementação de medidas de fiscalização mais rigorosas.

* + 1. Região Oeste

Os coeficientes do modelo do cluster da região Oeste são apresentados na imagem a seguir:

Figura 7 – Coeficientes do cluster da Região Oeste

A black screen with white text

Description automatically generated

Fonte: Autoria Própria

Na Região Oeste, além dos fatores envolvendo pedrestres, também são relevantes "Transitar na contramão/acostamento" e "Conversão proibida".

* + 1. Campos Gerais

Os coeficientes do modelo do cluster Campos Gerais são apresentados na imagem a seguir:

Figura 8 – Coeficientes do cluster Campos Gerais

A black screen with white text

Description automatically generated

Fonte: Autoria Própria

No cluster Campos Gerais, "Mal súbito do condutor" e "Restrição de Visibilidade" também influenciaram a gravidade dos acidentes, além do comportamento do pedestre notado em todos os clusters anteriores. Isso sugere a necessidade de abordar não apenas a segurança dos pedestres, mas também questões relacionadas ao comportamento dos condutores e condições da infraestrutura rodoviária.

1. **Conclusão(ões) ou Considerações Finais**

O presente estudo proporcionou uma análise das principais causas que influenciam a gravidade dos acidentes de trânsito no estado do Paraná entre os anos de 2019 e 2023. Ao utilizar dados abertos da PRF e aplicar técnicas de clusterização geográfica, foi possível segmentar o estado em quatro regiões distintas, permitindo uma compreensão mais precisa dos padrões regionais de acidentes.

A aplicação da regressão logística multinomial em cada um dos clusters revelou que comportamentos inadequados de pedestres são os fatores mais associados à gravidade dos acidentes em todas as regiões analisadas. Além disso, em algumas regiões, fatores como "transitar na contramão", "conversão proibida", "mal súbito do condutor" e "restrição de visibilidade" também apresentaram influência relevante. Esses resultados destacam a importância de abordar não apenas o comportamento dos pedestres, mas também questões relacionadas à conduta dos motoristas e às condições da infraestrutura rodoviária.

Com base nos achados, torna-se evidente a necessidade de intervenções específicas para cada região. Recomenda-se a implementação de campanhas educativas voltadas tanto para pedestres quanto para condutores, visando conscientizar sobre as normas de trânsito e os riscos associados ao seu descumprimento. Além disso, melhorias na infraestrutura viária, como a adequação de faixas de pedestres e sinalização, bem como o aumento da fiscalização, podem contribuir para a redução da gravidade dos acidentes.

Em suma, este estudo contribui para a compreensão dos padrões e fatores de risco associados aos acidentes de trânsito no Paraná, oferecendo subsídios para a formulação de políticas públicas mais eficazes. Ao direcionar esforços para os fatores identificados, é possível promover uma maior segurança viária, preservando vidas e reduzindo os custos sociais e econômicos decorrentes dos acidentes de trânsito.

1. **Referências**

**ANSARI, G. A.; ALSHABI, M. A.** **Modeling of Traffic Accident Reporting System through GIS**. **International Journal of Advanced Computer Science and Applications**, v. 3, n. 6, p. 26, 2012.

**BANCO MUNDIAL.** **Custo social dos acidentes de trânsito no Brasil**. Brasília: Banco Mundial, 2012.

**BEZERRA, J. A.; ROSENSTOCK, K. I. V.** **Análise dos dados abertos da Polícia Rodoviária Federal do tipo acidente de trânsito no estado da Paraíba**. **Revista Campo do Saber**, v. 2, n. 1, p. 1-18, 2016.

**Dados Abertos da PRF**. Disponível em: < https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf Acesso em: 10 mar. 2024.

DE AMO, Sandra. **Técnicas de Mineração de Dados**. Congresso da Sociedade Brasileira de Computação, 14. 2004, Salvador, Brasil. p.1 - 43.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel, SPSS e Stata**. 1ª ed. Rio de Janeiro: Elsevier, 2017.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. Ediitora Atlas  SA, 2008.

**HAIR, J. F. Jr. et al.** **Análise de Dados Multivariados**. 7. ed. Porto Alegre: Bookman Editora, 2009.

HAN, Jiawei; KAMBER, Micheline. **Data Mining Concepts and Techniques.Second** Edition University Of Illinois At Urbana-champaign: Elsevier, 2001. 772 p.

HOSMER, David W.; LEMESHOW, Stanley. **Applied Logistic Regression**. 2ª ed. New York: John Wiley & Sons, 2000.

**IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA.** **Censo Brasileiro de 2022**. Rio de Janeiro: IBGE, 2022.

**LI, C. C.; CHANG, T. C.** **Applying the GIS to discuss the traffic accidents under environmental analysis:** Case on Taichung city. **Ekoloji**, v. 28, n. 107, p. 3805–3816, 2019.

LUIZ PAULO FÁVERO; BELFIORE, P. **Manual de Análise de Dados**. [s.l.] Elsevier Brasil, 2017.

MAIMON, Oded; ROKACH, Lior. **Data Mining and Knowledge Discovery Handbook**. Second Edition Israel: Springer Science, 2010. 1306 p.

MCCULLAGH, Peter; NELDER, John A. **Generalized Linear Models**. 2ª ed. London: Chapman & Hall, 1989.

**MINISTÉRIO DA SAÚDE.** **DATASUS - Sistema de Informação sobre Mortalidade**. Brasília: Ministério da Saúde, 2022.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to Linear Regression Analysis**. 5ª ed. New York: John Wiley & Sons, 2012.

**OBSERVATÓRIO NACIONAL DE SEGURANÇA VIÁRIA (ONSV).** **Estimativa dos custos associados aos acidentes de trânsito: projeção no período 2018-2027**. Brasília: ONSV, 2023.

**PINTO, C. E. F.; DE ANDRADE, L. O.; DE OLIVEIRA, G. C.; DA SILVA, M. A. C.** **O custo social dos acidentes de trânsito no Brasil**. **Revista Brasileira de Saúde Pública**, v. 49, n. 1, p. 124, 2015.