MÉTODO DE ANÁLISE DA CARTEIRA DE EMPREENDIMENTOS EM MOBILIDADE URBANA E SUA CORRELAÇÃO NO TERRITÓRIO NACIONAL COM INDICADORES DE SINISTROS: UMA ABORDAGEM BASEADA EM CIÊNCIAS DE DADOS

Kaio Mesquita, Aleandro Matteoni, Bárbara Ferreira, Frederico Rodrigues, Henrique Carvalho, Kleberson da Rocha, Ludmila Lima, Yuri Marim

Imtraff Engenharia e Mobilidade

Abstract. *The study presents an in-depth analysis of the portfolio of projects undertaken by the Ministry of Cities and the National Secretariat for Mobility and Regional and Urban Development (MCID/SEMOB), correlating it with traffic accident indicators in Brazil using advanced data science techniques. The main objective is to investigate the relationship between investments in mobility infrastructure and road safety, with an emphasis on reducing fatal and injury accidents. The method adopted includes four stages: extraction, transformation, and loading of databases; characterization of variables; supervised and unsupervised modeling using machine learning and deep learning algorithms; and clustering. The analysis revealed that greater investments in infrastructure, particularly in paving and road safety, correlate with a significant decrease in the number of serious accidents. State capitals showed specific spatial patterns, with critical areas concentrated in border regions, highlighting the need for targeted interventions. The predictive models used, such as random forests and neural networks, demonstrated high accuracy in predicting accidents, highlighting the effectiveness of these techniques for public management. It is concluded that strategic investments in urban mobility can act as an important accident mitigation mechanism, with direct implications for the formulation of public policies aimed at road safety.*

Resumo. *O estudo apresenta uma análise aprofundada da carteira de empreendimentos do Ministério das Cidades e Secretaria Nacional de Mobilidade e Desenvolvimento Regional e Urbano (MCID/SEMOB), correlacionando-a com indicadores de sinistros de trânsito no Brasil por meio de técnicas avançadas de ciência de dados. O objetivo principal é investigar a relação entre os investimentos em infraestrutura de mobilidade e a segurança viária, com ênfase na redução de acidentes fatais e com feridos. O método adotado inclui quatro etapas: extração, transformação e carregamento das*

Cadernos do IME - Série Informática
e-ISSN: 2317-2193 (online)
DOI: 10.12957/cadinf.2025.89397

bases de dados; caracterização das variáveis; modelagem supervisionada e não supervisionada utilizando algoritmos de machine learning e deep learning; e clusterização. A análise revelou que maiores investimentos em infraestrutura, particularmente em pavimentação e segurança viária, correlacionam-se com uma diminuição significativa no número de acidentes graves. As capitais estaduais mostraram padrões espaciais específicos, com áreas críticas concentradas nas regiões limítrofes, destacando a necessidade de intervenções direcionadas. Os modelos preditivos empregados, como florestas aleatórias e redes neurais, demonstraram alta acurácia na previsão de sinistros, evidenciando à eficácia dessas técnicas para a gestão pública. Conclui-se que os investimentos estratégicos em mobilidade urbana podem atuar como um importante mecanismo de mitigação de sinistros, com implicações diretas na formulação de políticas públicas voltadas à segurança viária.

1. INTRODUÇÃO E CONTEXTUALIZAÇÃO DA PROBLEMÁTICA

A mobilidade urbana é um dos desafios centrais para as grandes metrópoles, especialmente em um contexto de rápido crescimento urbano e aumento da população. A Organização Mundial da Saúde (OMS) estima que cerca de 1,3 milhão de pessoas morrem anualmente em acidentes de trânsito em todo o mundo, e entre 20 e 50 milhões sofrem ferimentos não fatais, com muitos tornando-se incapazes como resultado de seus ferimentos (WHO, 2021). Este cenário impõe a necessidade de políticas públicas eficazes que integrem planejamento urbano e transporte, visando à redução desses índices alarmantes.

No Brasil, os empreendimentos de mobilidade urbana são vistos como instrumentos essenciais para melhorar a qualidade de vida nas cidades, reduzir congestionamentos e, principalmente, mitigar os impactos dos acidentes de trânsito. Diversos estudos apontam que investimentos em infraestrutura de transporte, como a implementação de ciclovias, faixas exclusivas para ônibus e melhorias nas condições das vias, estão associados à diminuição do número de sinistros e à gravidade dos acidentes, porém com avaliações locais apenas (Vasconcellos, 2001; Pucher *et al*., 2010).

Com o avanço das tecnologias da informação, a aplicação da ciência de dados no planejamento urbano tem se tornado cada vez mais relevante. Técnicas de *machine learning* e análise espacial permitem a identificação de padrões complexos e a predição de sinistros com uma precisão que métodos tradicionais não conseguem alcançar (Batty, 2013). Estas ferramentas oferecem aos gestores públicos a capacidade de tomar decisões mais embasadas, alocando recursos de maneira mais eficiente e priorizando áreas críticas para intervenções.

Durante o processo de planejamento urbano deve-se levar em consideração aspectos dos subsistemas de transportes, uso do solo e atividades (Sousa *et al*. 2019). Tradicionalmente os padrões de viagem e comportamentos são identificados em Pesquisas de campo, ferramentas mais comuns para se obter informações sobre a mobilidade de uma cidade. Essas pesquisas podem ser demoradas, requerer grande

logística com pesquisadores e apresentarem custos elevados para execução. Porém, na era da quarta revolução industrial, métodos mais robustos para análises descritivas, diagnósticas, preditivas e prescritivas estão disponíveis para auxiliar na identificação de padrões e correlações nos dados de mobilidade (Cats; Ferrati, 2022; Mesquita, 2023).

O presente trabalho tem por objetivo apresentar um método de análise dos empreendimentos em mobilidade urbana da carteira de empreendimentos do Ministério das Cidades e Secretaria Nacional de Mobilidade e Desenvolvimento Regional e Urbano (MCID/SEMOB). Além da sua correlação no território nacional com indicadores de mortos e feridos utilizando a base de dados do Atlas de violência e da Polícia Rodoviária Federal - PRF. Este trabalho está dividido em 5 capítulos. O primeiro, conforme apresentado, descreve a problemática e o objetivo do trabalho. O segundo elucida outros trabalhos que focaram na utilização de técnicas de *machine learning* para análise de sinistros e aspectos da mobilidade urbana. O terceiro capítulo detalha o método de análise. No capítulo 4 são discutidos os principais resultados de cada modelo empregando a possível relação entre os tipos de empreendimentos e os sinistros de trânsito, e por fim, no capítulo 5 são apresentadas as considerações finais, hipóteses alcançadas e proposta de trabalhos futuros.

2. REVISÃO DA LITERATURA

A ciência de dados tem se tornado uma ferramenta essencial para a análise de sinistros de trânsito, especialmente em sua capacidade de processar grandes volumes de dados e identificar padrões que poderiam passar despercebidos em análises tradicionais. Abdel-Aty e Haleem (2011) destacam que a análise de dados de acidentes com enfoque em variáveis espaciais e temporais pode melhorar significativamente a compreensão dos fatores que contribuem para acidentes graves.

Além disso, a integração de dados de diferentes fontes, como registros de acidentes, dados de tráfego e informações meteorológicas, permite uma análise mais abrangente. De acordo com Bíl *et al*. (2013), o uso de Sistemas de Informação Geográfica (SIG) em conjunto com modelos estatísticos e de aprendizado de máquina possibilita a criação de mapas de risco e a identificação de *hotspots* de acidentes (áreas geográficas onde a concentração de um fenômeno específico é significativamente maior do que em áreas vizinhas), facilitando a implementação de medidas preventivas.

A modelagem com aprendizado de máquina tem ganhado destaque em estudos relacionados à mobilidade urbana e à segurança no trânsito. A análise preditiva, em particular, tem sido uma área de grande interesse. O estudo de Huang *et al*. (2010) mostrou que técnicas de *machine learning*, como *Random Forest*, são eficazes na previsão de acidentes em rodovias, permitindo que as autoridades priorizem intervenções em trechos com maior risco.

Outro aspecto importante da modelagem com aprendizado de máquina é a inclusão de variáveis geoespaciais. Geurts *et al.* (2006) argumentam que a combinação de técnicas de machine learning com dados espaciais pode melhorar significativamente a precisão das previsões de sinistros. Essa abordagem permite que modelos preditivos capturem a influência do ambiente construído, como a proximidade de cruzamentos perigosos ou a densidade de tráfego, sobre a frequência e gravidade dos acidentes.

Ademais, o uso de técnicas de *clustering*, como o *K-means*, tem se mostrado útil para agrupar áreas com características semelhantes em termos de risco de acidentes.

Mansourihanis *et al*. (2016) demonstram que o *clustering* pode ajudar a identificar regiões que requerem atenção especial, permitindo uma alocação mais eficiente de recursos para prevenção de acidentes.

Nos últimos anos, a aplicação de técnicas de aprendizado de máquina (ML) e aprendizado profundo (*Deep Learning*) à análise de acidentes de trânsito tem se intensificado. Uma revisão abrangente de 2024 examinou 191 estudos sobre previsão de risco, frequência, gravidade e duração de acidentes, destacando a eficácia da integração de múltiplas fontes de dados e modelos avançados para melhorar a acurácia preditiva (Behboudi; Moosavi; Ramnath, 2024).

Apesar dos avanços, a aplicação de *machine learning* em estudos de mobilidade urbana enfrenta diversos desafios. De acordo com Biau e Scornet (2016), um dos principais obstáculos é o risco de *overfitting*, especialmente quando se trabalha com conjuntos de dados muito pequenos ou desequilibrados. Isso pode resultar em modelos que apresentam bom desempenho em dados de treinamento, mas falham ao serem generalizados para novos dados.

Finalmente, a qualidade dos dados é uma preocupação constante. De acordo com Elvik *et al*. (2013), a confiabilidade das análises depende fortemente da precisão e integridade dos dados coletados. Isso inclui a necessidade de lidar com dados faltantes, *outliers* e variáveis latentes. Estas variáveis são fatores ou características que não podem ser medidos ou observados diretamente, mas que exercem uma influência significativa sobre os dados e os resultados da análise. Elas representam conceitos abstratos ou "ocultos" que, apesar de não serem visíveis ou registrados nos dados brutos, afetam as variáveis observáveis como o comportamento do motorista e a qualidade da infraestrutura.

3. MATERIAIS E MÉTODOS

O método deste trabalho (Figura 1) é composto por 4 macroetapas com a finalidade de transformar dados brutos em informações que possibilitem a adequada avaliação da relação entre os empreendimentos e sinistros geolocalizados. As etapas são: (i) Extração, transformação e carregamento das bases de dados; (ii) Caracterização; (iii) Modelagem supervisionada com *machine learning* e *deep learning*; e (iv) Clusterização.

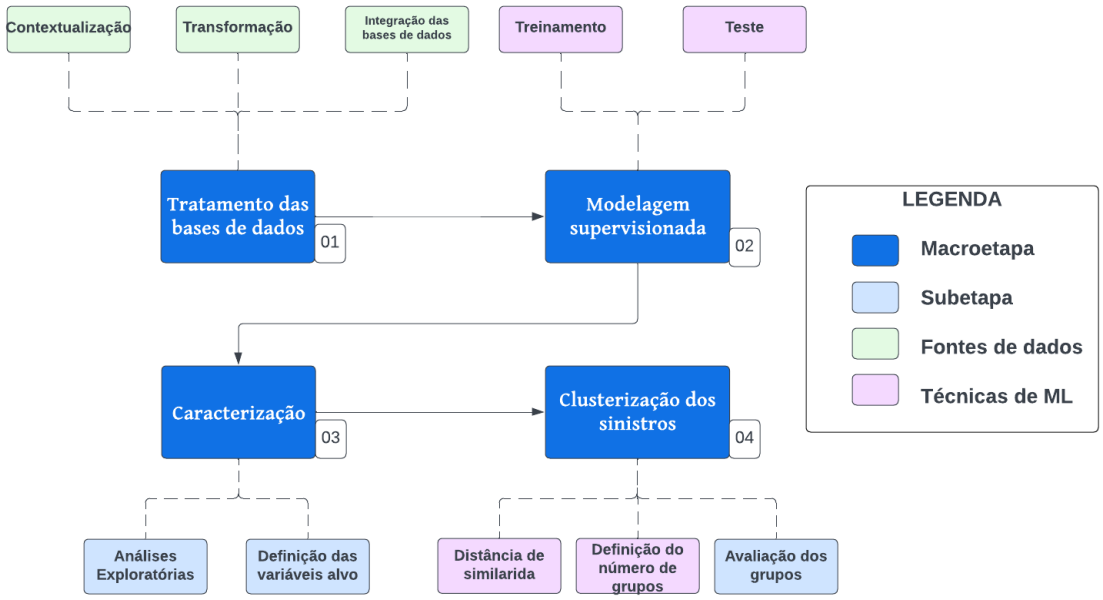

Figura 1 – Proposta Metodológica.

Fonte: Autores (2025).

A primeira etapa consiste na visualização prévia e descrição de todas as bases de dados disponíveis e reconhecimentos das variáveis para possível transformação em indicadores. Seguido do tratamento de valores ausentes ou inconsistências, criação de novas variáveis baseadas nos dados brutos para melhor capturar padrões e ajustar os valores das variáveis pela média para melhorar a performance de técnicas avançadas. Ainda nesta etapa, foi criado uma proposta de estrutura relacional para armazenar os dados. As bases de dados utilizadas correspondem a Carteira de Empreendimentos MCID/SEMOB, base de acidentes da PRF e base de Mortos e Feridos (Atlas Violência) entre os anos de 2014 e 2019.

Durante o processo de tratamento das bases, foi identificada a presença de valores ausentes principalmente nas seguintes variáveis: status da obra (4,2% dos registros), descrição do contrato (3,1%), valores de financiamento total (1,8%) e tipos de empreendimentos (9,1%). Em termos absolutos, as lacunas correspondiam a menos de 10% do total de observações em cada caso, o que permitiu adotar estratégias de imputação sem comprometer a representatividade da amostra.

Optou-se pela imputação pela média apenas para variáveis numéricas contínuas que apresentavam distribuição simétrica, como os valores de financiamento. Essa escolha se justifica por ser um método simples e eficiente para manter a escala original da variável sem introduzir viés significativo. Já para as variáveis categóricas ou com alta dispersão, os registros foram mantidos como ausentes e tratados via codificação binária (*dummy*) com inclusão de uma categoria específica para valores faltantes. Essa abordagem balanceia simplicidade e robustez, evitando distorções na modelagem supervisionada, especialmente nos modelos baseados em árvores, que são menos sensíveis a variáveis com pequenas proporções de dados ausentes.

Baseado na revisão da literatura e nas análises exploratórias iniciais, foram formuladas quatro perguntas de pesquisa que orientam este estudo (Géron, 2019; Du et. Al, 2020, Cats e Ferrati, 2022): (i) Existe correlação entre os investimentos em

infraestrutura de mobilidade urbana e a ocorrência de acidentes fatais no trânsito?; (ii) As capitais estaduais apresentam maiores índices de sinistros fatais devido à sua infraestrutura mais densa e população elevada?; (iii) Há padrões de autocorrelação espacial entre os sinistros viários que evidenciem áreas críticas influenciadas por regiões vizinhas?; (iv) Os indicadores de sinistros apresentaram tendência de redução ao longo do tempo em função dos tipos de empreendimentos implementados nos municípios?

A segunda etapa realizada foi a caracterização, de modo a compreender de forma descritiva os tipos de variáveis e seus aspectos médios e de dispersão. Dentre as principais análises estão modelos de correlação de Pearson, distribuição simples e espacial com índice de Moran e análises gráficas de séries temporais. Também foi calculado e espacializado o indicador de Unidade Padrão de Severidade (UPS), considerando os coeficientes 1, 5 e 13 para os acidentes sem feridos, com feridos e com vítimas fatais, respectivamente, para cada estado brasileiro. Por fim, nesta etapa foram selecionadas as variáveis alvo e atributos para as modelagens posteriores.

As categorias dos empreendimentos foram tratadas através de técnica de Processamento de Linguagem Natural (PNL). Embora essa técnica adotada tenha se mostrado eficaz na categorização automática dos empreendimentos, é importante reconhecer possíveis limitações associadas a esse processo. A categorização depende diretamente da qualidade e padronização das descrições textuais presentes na base de dados original, que por vezes apresentam inconsistências, siglas ambíguas ou termos genéricos. Além disso, o modelo pode sofrer viés em razão da frequência com que certos termos aparecem em diferentes tipos de empreendimentos, o que pode levar a associações indevidas ou pouco representativas. Para mitigar esses riscos, foram aplicadas técnicas de normalização textual e validação amostral manual, mas reconhece-se que abordagens mais robustas (como o uso de *embeddings* semânticos ou modelos pré-treinados em domínios específicos) podem oferecer ganhos adicionais de precisão e generalização em estudos futuros.

Durante a terceira etapa foram criados, treinados e avaliados os modelos supervisionados com a finalidade de identificar a relação direta das categorias de empreendimentos com variáveis de sinistros, além de avaliar a capacidade preditiva de sinistros com vítimas feridas e fatais. Os dados da base de empreendimentos foram correlacionados com a base de sinistros com uma agregação espacial pelas coordenadas. A base final foi dividida na proporção 7:2:1, em conjunto de dados de treinamento, validação e teste. Para a modelagem supervisionada, foram considerados modelos de regressão múltipla e florestas aleatórias, e para avaliação da predição foram considerados os indicadores de Erro Quadrático Média (MSE) e o coeficiente de determinação (R^2).

Para a análise de correlação entre as variáveis de sinistros e os atributos dos empreendimentos, foi utilizado o coeficiente de correlação de Pearson. Esta medida estatística avalia a intensidade e a direção da relação linear entre duas variáveis contínuas, sendo apropriada para dados com distribuição aproximadamente normal. Os valores do coeficiente variam de -1 a 1, onde valores próximos de 1 indicam forte correlação positiva, próximos de -1 indicam forte correlação negativa, e valores próximos de 0 indicam ausência de correlação linear significativa. Os coeficientes

apresentados na Tabela 2 referem-se às correlações de Pearson entre as variáveis preditoras e os diferentes tipos de mortalidade analisados.

Em seguida, foi aplicado a distância euclidiana e a padronização dos dados (z) pela seguinte fórmula: (x – média)/desvio padrão, em que x seria o valor original do atributo normalizado. Por fim, foi definido o número de grupos pelo método do cotovelo. O método do cotovelo é uma técnica utilizada em Análise de Agrupamento (*Clustering*), especialmente no algoritmo *K-means*, para determinar o número ideal de clusters (grupos) a serem formados nos dados. Vale destacar que variáveis categóricas podem ter influência na separação dos grupos, portanto optou-se por transformar elas em variáveis *dummy* (binárias por tipo de categoria), mantendo-as na clusterização dos grupos na base tratada e correlacionada entre categorias de empreendimentos e indicadores de mortes e feridos. Cada categoria distinta foi transformada em uma nova variável binária (0 ou 1). Essa abordagem foi aplicada, por exemplo, à variável que descreve o tipo de empreendimento. É importante ressaltar que, entre as variáveis categóricas utilizadas, nenhuma possuía natureza ordinal explícita. Todas representavam classes nominais, sem hierarquia natural entre os seus níveis (por exemplo, "Transporte Público", "Segurança e Acessibilidade", "Urbanização"), o que justifica o uso da codificação binária ao invés de técnicas como *label encoding* ou *ordinal encoding*.

Contudo, reconhece-se que, em modelos baseados em distância, como o algoritmo K-means, o uso de variáveis *dummy* pode influenciar a forma como os grupos são formados, sobretudo quando há desbalanceamento na frequência das categorias. Para mitigar esse efeito, as variáveis foram previamente padronizadas (z-score), reduzindo a distorção causada pela presença de múltiplas variáveis binárias com diferentes escalas ou relevâncias.

Por fim, foram aplicados o método *k-means* para clusterização e analisados os resultados. Ainda nesta etapa os dados gerados foram convertidos em análise gráfica interativa com a biblioteca *python folium*, de modo a possibilitar interação direta entre os resultados de forma espacial.

3.1. Fundamentos do Aprendizado de Máquina no Contexto do Estudo

O aprendizado de máquina (Machine Learning – ML) é uma área da ciência da computação que desenvolve algoritmos capazes de identificar padrões em grandes volumes de dados e realizar previsões ou classificações com base nesses padrões. Ao contrário dos métodos estatísticos tradicionais, que exigem modelagem explícita das relações entre variáveis, os algoritmos de ML aprendem essas relações diretamente a partir dos dados.

No presente estudo, o ML foi empregado com o objetivo de modelar e prever a ocorrência de sinistros viários a partir de um conjunto de atributos extraídos da carteira de empreendimentos em mobilidade urbana. A aplicação dessas técnicas permite avaliar como diferentes características dos empreendimentos (como tipo, status de execução, localização, entre outros) estão associadas a indicadores de mortes e feridos no trânsito. Entre os algoritmos utilizados, destacam-se (Géron, 2019):

- Florestas Aleatórias (*Random Forests*): Um método do tipo *ensemble* baseado em múltiplas árvores de decisão. Cada árvore realiza uma predição e o resultado final é obtido pela média ou maioria das árvores. Essa técnica é eficaz para

dados com muitas variáveis e com interações complexas entre elas, como é o caso das bases de mobilidade e sinistros.

- Redes Neurais Artificiais (*Deep Learning*): Inspiradas no funcionamento do cérebro humano, as redes neurais são compostas por camadas de unidades chamadas "neurônios artificiais", que transformam e transmitem sinais. Neste estudo, foi utilizada uma rede do tipo *Multilayer Perceptron* (MLP), com múltiplas camadas ocultas, adequada para capturar relações não lineares entre variáveis. O modelo foi treinado para prever o número de mortes com base nos atributos dos empreendimentos e nos sinistros observados.

A escolha dessas técnicas se deve à sua robustez e à capacidade de lidar com variáveis categóricas, numéricas e até mesmo geoespaciais, tornando-as particularmente úteis para problemas complexos como a análise da segurança viária em diferentes contextos urbanos. Com isso, busca-se oferecer suporte à tomada de decisão baseada em evidências para gestores públicos, promovendo uma alocação mais eficiente de recursos em infraestrutura de mobilidade urbana.

3.2. Aprendizado supervisionado e não-supervisionado

No aprendizado supervisionado, os dados de treinamento fornecidos ao algoritmo incluem as soluções desejadas, chamadas de rótulos. A classificação é uma tarefa típica de aprendizado supervisionado. Sendo utilizado também para prever um alvo de valor numérico, através de um modelo de regressão. Alguns algoritmos de regressão também podem ser usados para classificação e vice-versa. Os principais algoritmos supervisionados são: (i) *K-Nearest Neighbours*; (ii) Máquinas de Vetores de Suporte (SVM); (iii) Árvore de Decisão e Florestas Aleatórias; e (iv) Redes Neurais. No aprendizado não-supervisionado, utilizado para identificar padrões, os dados de treinamento não são rotulados. O Sistema tenta aprender sem um “professor”. A seguir estão alguns dos principais algoritmos de aprendizado não-supervisionado: (i) Clustering (*K-Means*, *Clustering* Hierárquico, Maximização de Expectativa); (ii) Visualização e redução de dimensionalidade (Análise de Componentes Principais, *Locally Linear Embedding*, *t-distributed Stochastic Neighbor Embedding*); e (iii) Aprendizado da regra de associação.

3.3. Modelagem para identificação de padrões

A distância de similaridade utilizada foi a distância Euclidiana, se aplicando melhor a dados padronizados, e devido a isso o resultado é invisível a outliers (exceções, ou dados com uma diferença muito grande em relação à média). Uma desvantagem sobre essa medida de distância pode acontecer se houver diferença de escala entre as dimensões; por isso a importância de se normalizar os dados. Essa distância é calculada como a soma da raiz quadrada da diferença entre coordenadas de dois pontos, onde P é número de registros:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

(1)

onde x e y são dois pontos que se está comparando. Uma vez que a distância de similaridade é definida, o próximo passo é padronizar os dados para garantir que todas as dimensões (ou atributos) tenham a mesma importância na determinação da distância (Japkowicz; Shah, 2011; Géron, 2019). Neste caso, optou-se pela padronização por z-score, também conhecida como padronização, em que se transforma cada valor em um valor z, representando o número de desvios padrões acima ou abaixo da média dos dados. Isso é realizado subtraindo a média dos dados e dividindo pelo desvio padrão, conforme segue:

$$z = \frac{(x - média)}{Desvio\ Padrão}$$

(2)

Este tipo de padronização, além de garantir a mesma importância dos atributos, também evita que atributos com valores extremos dominem a distância de similaridade. Posteriormente, foi necessário definir o número de grupos, o qual é o parâmetro do método de clusterização. Dessa forma é possível executar o *k-means* variando o número de clusters, k, de 1 a 10, por exemplo (Du *et al*., 2020). Em cada rodada, calculou a variação dentro dos grupos pela soma dos desvios quadráticos em relação ao centroide de cada grupo (MacQueen, 1967; Jain *et. al*, 1999). A partir do gráfico relacionando as variações dentro dos grupos com os valores de k, identifica-se o "ponto de cotovelo" como sendo o ponto em que a variação começa a se estabilizar, ou seja, a curva começa a ficar mais suave.

4. **RESULTADOS E DISCUSSÃO**

Para as análises descritivas, foi espacializado o indicador de UPS por município para cada Unidade Federativa (UF) do Brasil, considerando o intervalo temporal de uma década para evidenciar os efeitos dos sinistros em larga escala. Na Figura 2, está representado o estado de São Paulo e Rio de Janeiro, variando o UPS até 10000 e 8000 unidades de severidade, respectivamente. Os pontos críticos estão em torno das capitais dos estados (para todas as UF). Vale destacar o fenômeno de maior ponto crítico nos municípios limítrofes às capitais, indicando que as entradas das cidades são pontos mais críticos que a própria capital, devendo ter uma maior atenção em relação aos empreendimentos construídos nessas regiões.

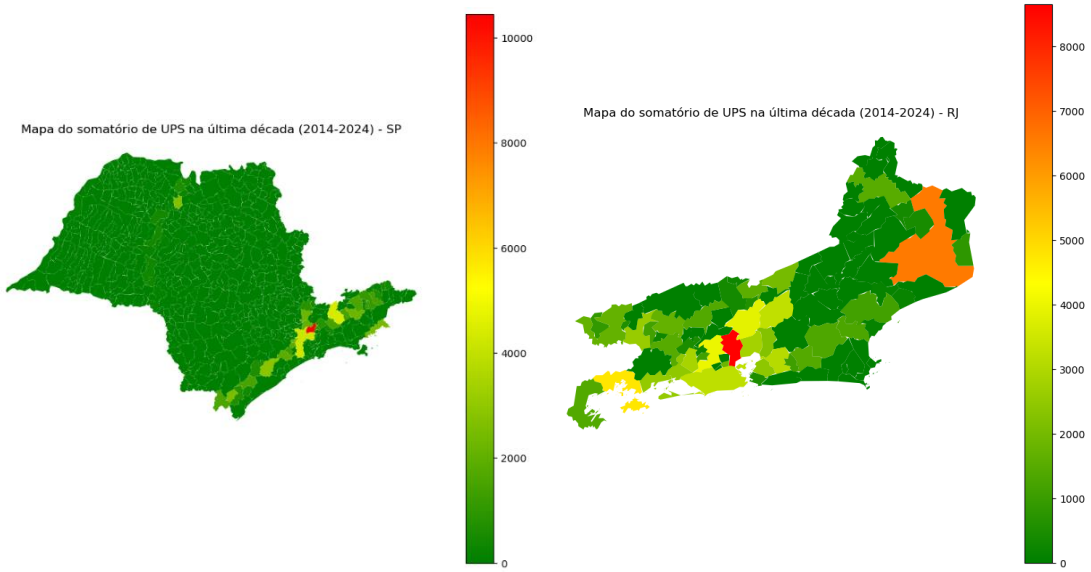

Figura 2 – Unidade Padrão de Severidade por município e UF

Fonte: Autores (2025)

Além das análises geoespaciais, foi realizada uma avaliação comparativa da média anual de mortes por veículo em grandes municípios brasileiros entre 2010 e 2019. A Tabela 1 apresenta os 20 municípios com maiores médias de fatalidades associadas a sinistros veiculares nesse período. Observa-se que Brasília (DF), São Paulo (SP) e Fortaleza (CE) lideram o ranking, com médias de 144,6, 117,9 e 90,3 mortes por veículo, respectivamente. Este resultado evidencia a concentração dos sinistros letais em centros urbanos com alta densidade populacional e tráfego intenso. Ainda que a média de mortes tenha diminuído em muitos desses municípios ao longo da década, os valores absolutos permanecem elevados. Por outro lado, cidades como Campo Grande (MS), Manaus (AM) e Contagem (MG) apresentaram médias inferiores a 35 mortes anuais, o que pode estar relacionado a políticas locais de mobilidade, investimentos em infraestrutura viária ou características territoriais e demográficas. Essa análise complementar reforça a necessidade de intervenções direcionadas em regiões com alta exposição ao risco, especialmente nas capitais e suas zonas de transição urbana.

Tabela 1 - Média de mortes em sinistros de trânsito com veículos por capital

ano_y	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	Média Total de Mortes por veículo
Municipio_UF											
Brasília (DISTRITO FEDERAL)	206.0	247.0	207.0	147.0	166.0	97.0	117.0	86.0	86.0	87.0	144.6
São Paulo (SAO PAULO)	173.0	168.0	185.0	185.0	186.0	157.0	16.0	10.0	10.0	89.0	117.9
Fortaleza (CEARA)	30.0	140.0	151.0	119.0	117.0	139.0	150.0	21.0	21.0	15.0	90.3
Belo Horizonte (MINAS GERAIS)	140.0	134.0	79.0	106.0	87.0	93.0	52.0	50.0	50.0	29.0	82.0
Montes Claros (MINAS GERAIS)	75.0	69.0	55.0	75.0	92.0	97.0	75.0	64.0	64.0	52.0	71.8
Curitiba (PARANA)	99.0	95.0	80.0	75.0	77.0	49.0	57.0	61.0	61.0	32.0	68.6
Goiânia (GOIAS)	56.0	58.0	58.0	57.0	75.0	64.0	63.0	71.0	71.0	54.0	62.7
Rio de Janeiro (RIO DE JANEIRO)	93.0	99.0	86.0	49.0	42.0	32.0	35.0	57.0	57.0	16.0	56.6
Salvador (BAHIA)	52.0	75.0	105.0	73.0	63.0	55.0	31.0	31.0	31.0	23.0	53.9
Campos dos Goytacazes (RIO DE JANEIRO)	53.0	40.0	67.0	47.0	42.0	44.0	42.0	28.0	28.0	23.0	41.4
Recife (PERNAMBUCO)	45.0	40.0	40.0	29.0	51.0	62.0	30.0	34.0	34.0	38.0	40.3
Cuiabá (MATO GROSSO)	50.0	54.0	25.0	34.0	51.0	55.0	40.0	35.0	35.0	21.0	40.0
Uberlândia (MINAS GERAIS)	36.0	22.0	14.0	7.0	17.0	59.0	73.0	43.0	43.0	67.0	38.1
Feira de Santana (BAHIA)	59.0	65.0	66.0	47.0	32.0	20.0	10.0	22.0	22.0	17.0	36.0
Cascavel (PARANA)	38.0	37.0	43.0	44.0	37.0	35.0	26.0	23.0	23.0	36.0	34.2
Serra (ESPIRITO SANTO)	37.0	44.0	50.0	46.0	46.0	29.0	21.0	16.0	16.0	27.0	33.2
Itabuna (BAHIA)	21.0	27.0	32.0	44.0	43.0	31.0	36.0	40.0	40.0	14.0	32.8
Contagem (MINAS GERAIS)	45.0	42.0	35.0	31.0	46.0	41.0	16.0	25.0	25.0	20.0	32.6
Manaus (AMAZONAS)	35.0	22.0	32.0	39.0	35.0	47.0	31.0	25.0	25.0	31.0	32.2
Campo Grande (MATO GROSSO DO SUL)	42.0	33.0	23.0	39.0	30.0	33.0	18.0	33.0	33.0	33.0	31.7

Fonte: Autores (2025)

Foram avaliados também a autocorrelação espacial do número de mortes e feridos graves para as unidades federativas. A Figura 3, exemplifica os resultados para a variável espacial do número de mortes. Aparentemente as capitais apresentam valores *Low-low*, ou seja, são valores baixos cercados por valores baixos, indicando algum grau de eficiência devido a fatores que podem estar incluídos as obras de empreendimentos, fato que é deixado em evidência quando avaliado regiões com baixos índices de empreendimentos, na qual apresentam valores *High-high* (valores altos cercados por valores altos de número de mortes) e *Low-High* (Áreas com baixos valores cercadas por áreas com altos valores) indicando *outliers* espaciais, ou seja, não deveriam ocorrer esse fenômeno nessas regiões. As regiões em cinza, tiveram os valores do indicador igual a 0, indicando uma distribuição aleatória, e portanto, não seguindo algum padrão reconhecível. As áreas em vermelho devem ser colocadas como prioritárias em políticas públicas de cada estado para viabilizar obras de infraestrutura de segurança viária, mobilidade e acessibilidade urbana. Também foram verificadas nesta etapa se existia alguma relação entre as categorias de mortes e a população da região. O resultado do R^2 para o número de mortes de ciclistas em relação a população foi de 0.36, indicando baixa correlação. Enquanto a variável de mortes de pedestres apresentou o valor de R^2 de 0.73, dando indícios de que possa haver relação direta.

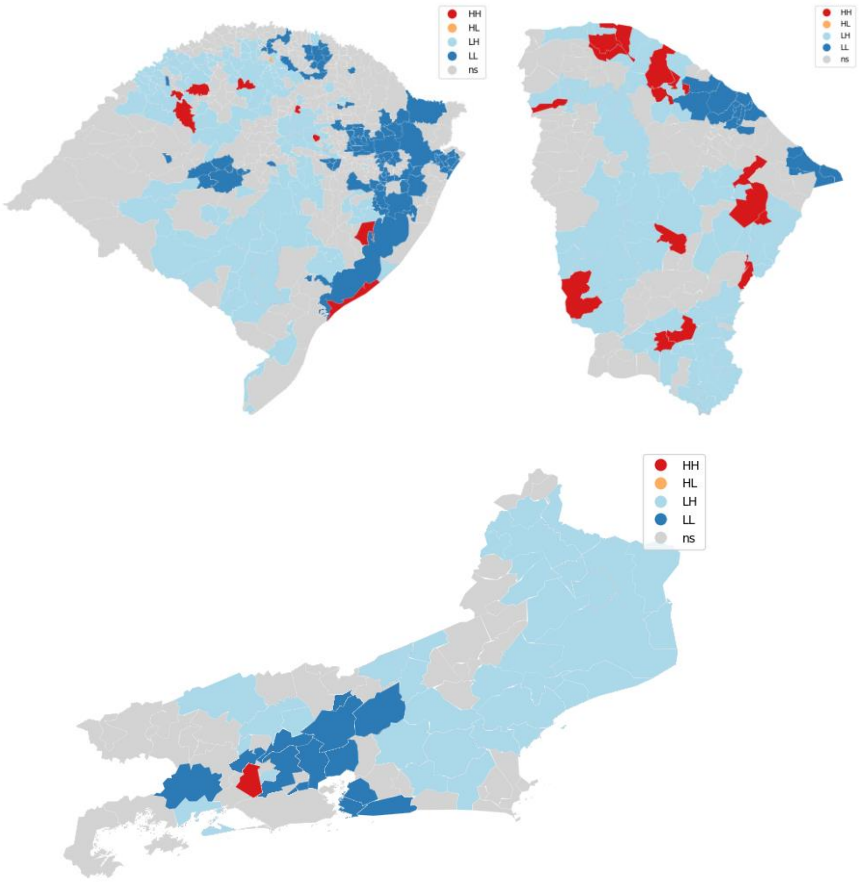

Figura 3 – Lisamap para autocorrelação espacial de n° de mortes – Uf CE, RS e RJ

Fonte: Autores (2025)

Conforme especificado no método, os empreendimentos foram categorizados com uma técnica de Processamento de Linguagem Natural – PNL em 8 categorias: (i) infraestrutura urbana; (ii) Transporte Ativo; (iii) Pavimentação e Drenagem; (iv) Projetos específicos; (v) Segurança e Acessibilidade; (vi) Transporte Público; (vii) Urbanização e qualificação urbana e (viii) Outros tipos de obras. Os municípios com maior quantidade de empreendimentos entre 2014 e 2019, que receberam mais investimentos, foram Campo Grande (301), Salvador (165), Teresina (138), Cuiabá (104) e Recife (95). Avaliando as Figuras 4 e 5, a média de mortes associadas por tipo de empreendimento (associação feita de forma geoespacial pela localização) decaiu na última década em todas as categorias. Outro fator interessante é que a relação entre o número de mortes e as categorias seguiu a mesma tendência em todos os anos, tendo uma maior associação aos empreendimentos de transporte público e uma menor associação aos empreendimentos de segurança e acessibilidade, dando evidências que esse tipo de empreendimento pode reduzir o número de mortes no trânsito. Por fim, aspectos de infraestrutura e qualificação urbana aparentam reduzir este indicador em menor grau.

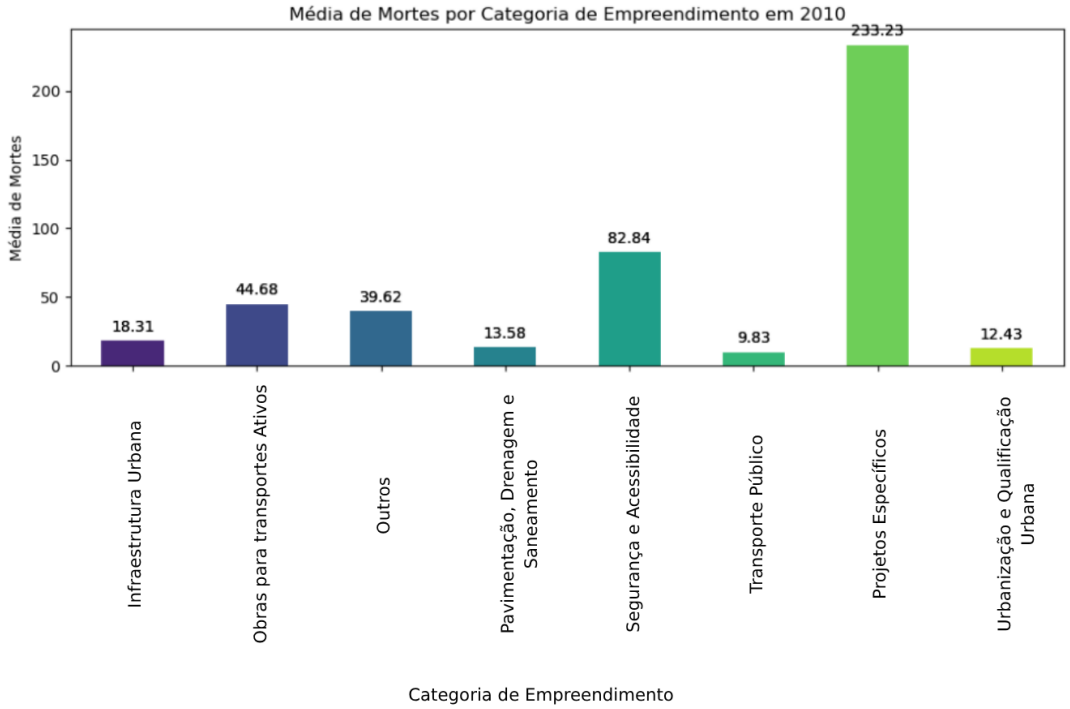

Figura 4 – Relação do número médio de mortes por empreendimento e ano de 2010

Fonte: Autores (2025)

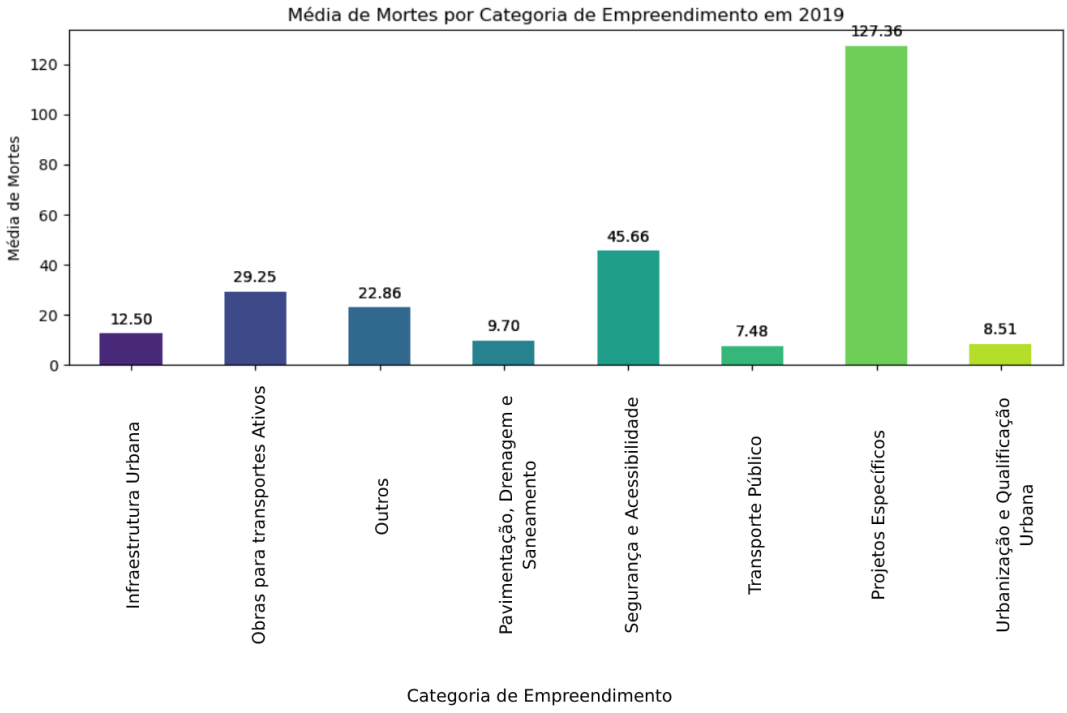

Figura 5 – Relação do número médio de mortes por empreendimento e ano de 2019

Fonte: Autores (2025)

Partindo para a modelagem por intermédio do aprendizado de máquina (*Machine Learning*) e aprendizado profundo (*Deep learning*) foram treinados modelos de regressão múltipla e floresta aleatória para avaliar a capacidade de predição dos principais indicadores de mortes e feridos, além da influência de outros tipos de sinistros e dos empreendimentos sobre a variável alvo. A Tabela 2 resume os resultados de correlação para as variáveis de mortes. Os modelos obtiveram R^2 acima de 0,95 indicando alta relação entre as variáveis preditoras e a variável alvo. O total de feridos influencia diretamente a predição de mortes no trânsito, enquanto os usuários vulneráveis (pedestres e ciclistas) têm influência direta de sinistros envolvendo automóveis. Em todos os modelos existiu uma relação inversamente proporcional em relação ao status da obra estar concluído e as obras de pavimentação e drenagem existirem na região, ou seja, quanto maior esses investimentos menores poderão ser os impactos de morte no trânsito. Também foi treinado um modelo sem variáveis de mortes para garantir que não haveria *overfitting* devido à autocorrelação entre essas variáveis, utilizando uma rede neural *multilayer perceptron* com 3 camadas de 64, 32 e 1 neurônio, respectivamente. O otimizador escolhido foi o Adam, e a função de perda utilizada foi o erro quadrático médio. Com 50 épocas de treinamento obteve-se um Erro Quadrático Médio - MSE de 53,44. O uso de práticas padrões como divisão de dados, normalização e validação durante o treinamento fortalece a robustez do modelo.

Tabela 2 - Coeficientes de desempenho veicular.

Mortes de ciclistas		**Mortes com veículos**		**Mortes com pedestres**		**Total de mortes**	
Variável	Correlação	Variável	Correlação	Variável	Correlação	Variável	Correlação
Mortes_motociclistas	0,86	Total_mortes	0,87	Total_mortes	0,94	Pedestres_Mortes	0,94
Total_mortes	0,83	Motociclistas_mortes	0,82	Mortes_motociclistas	0,84	Motociclistas_mortes	0,93
Ocup_automovel_mortes	0,77	Pedestres_Mortes	0,81	Pedestres_feridos	0,82	Ocup_automovel_mortes	0,87
Pedestres_Mortes	0,73	Ciclistas_mortes	0,77	Total_feridos	0,81	Total_feridos	0,86
Total_feridos	0,72	Total_feridos	0,76	Ocup_automovel_mortes	0,81	Ciclistas_mortes	0,83
Desc_unidade_SEMOB	-0,22	Desc_unidade_SEMOB	-0,25	Desc_unidade_SEMOB	-0,29	Desc_unidade_SEMOB	-0,31
Desc_fonte_OGU	-0,19	Desc_fonte_OGU	-0,21	Desc_fonte_OGU	-0,21	Desc_fonte_OGU	-0,22
Situacao_obra_concluida	-0,09	Situacao_obra_concluida	-0,10	Situacao_obra_concluida	-0,11	Situacao_obra_concluida	-0,11
Categoria_emp_pavimentação	-0,08	Categoria_emp_pavimentação	-0,08	Categoria_emp_pavimentação	-0,09	Categoria_emp_pavimentação	-0,10
Desc_contrato_concluido	-0,04	Desc_contrato_concluido	-0,05	Desc_contrato_concluido	-0,05	Desc_contrato_concluido	-0,06
R^2	0,98	**R^2**	0,96	**R^2**	0,99	**R^2**	0,99
MSE	0,10	**MSE**	4,16	**MSE**	1,31	**MSE**	3,62

Fonte: Autores (2025)

Analisando o gráfico, foi indicado a existência de possíveis 4 grupos (Figuras 6 e 7) em relação às categorias de empreendimentos e dados de acidentes. O grupo 1 apresentou maior distribuição média de investimento em obras de saneamento e acessibilidade, enquanto o grupo 0 apresentou maior investimento médio em obras de pavimentação, drenagem e saneamento. O grupo 3 foi o único que apresentou relação direta significativa com apenas uma categoria de empreendimentos que foi a de pavimentação e drenagem. Por fim, o grupo 2 com menor quantidade de representantes, correspondem às capitais com maior número de acidentes médios por ano, sendo (em ordem decrescente): São Paulo, Rio de Janeiro, Fortaleza, Goiânia e Brasília. Esse grupo também apresentou baixos índices médios de investimento em transporte ativo em relação aos outros.

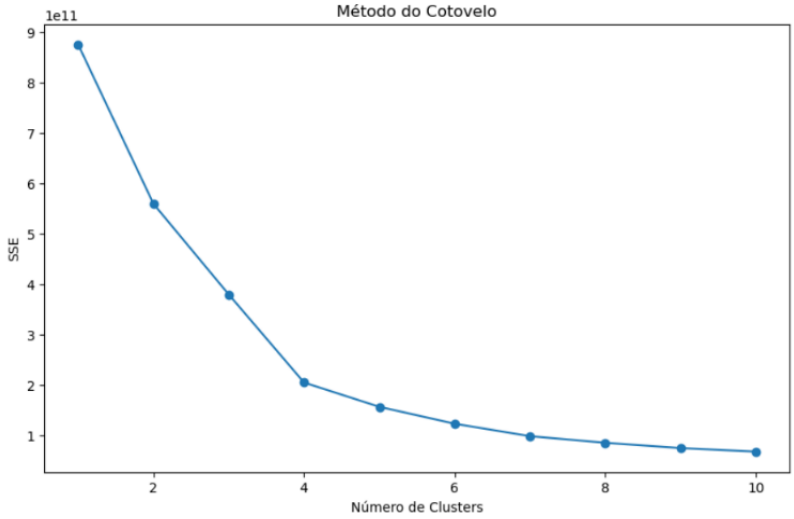

Figura 6 – Método do cotovelo para indicação do número de clusters

Fonte: Autores (2025)

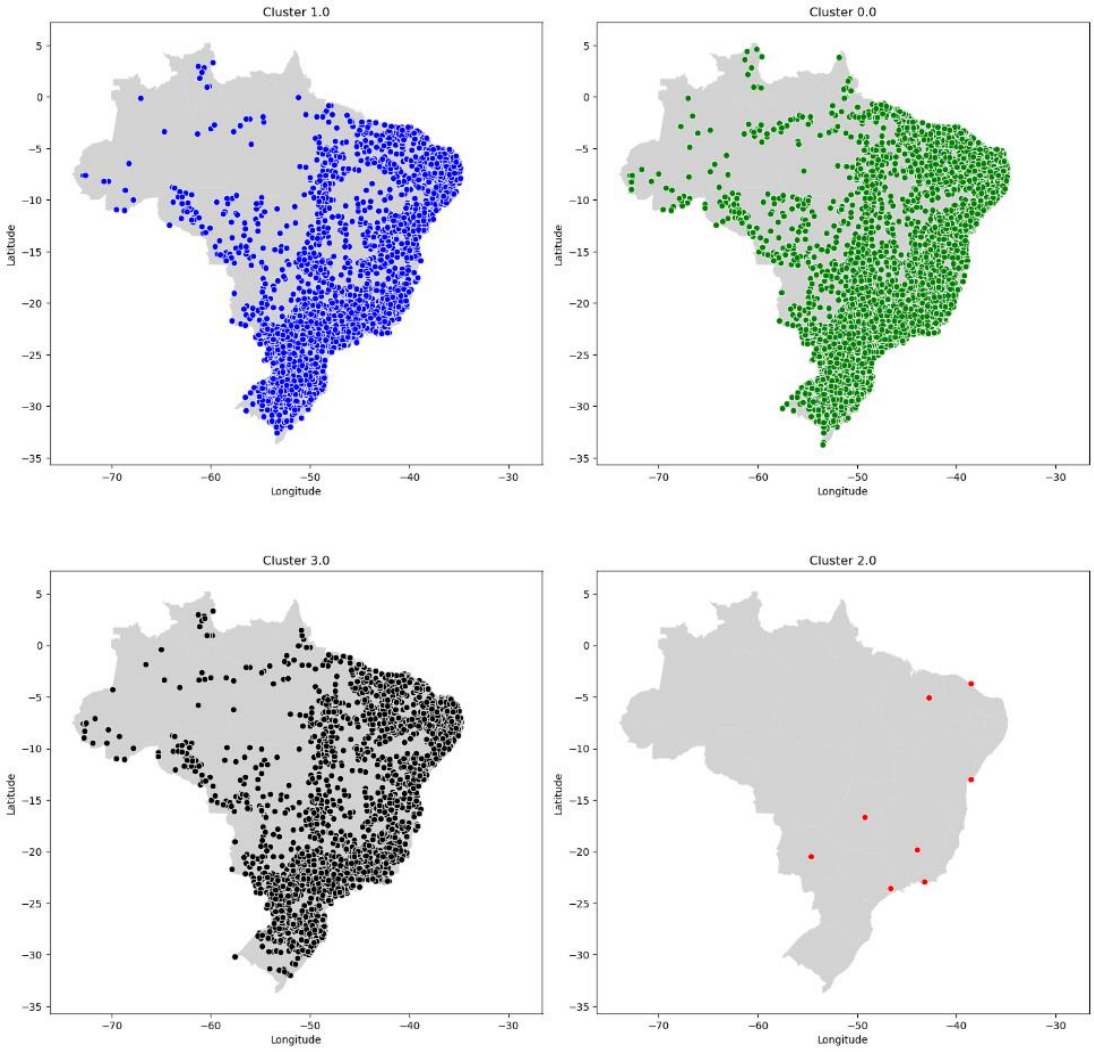

Figura 7 – Clusters de empreendimentos em relação aos sinistros

Fonte: Autores (2025)

5. CONSIDERAÇÕES FINAIS

Com base nos resultados obtidos, este estudo apresenta a possibilidade de existir uma correlação significativa entre os tipos de empreendimentos em mobilidade urbana e a redução no número de acidentes envolvendo vítimas fatais e feridas. Especificamente, verificou-se que maiores investimentos em infraestrutura, como pavimentação, segurança viária e transporte público, estão inversamente relacionados ao número de sinistros, corroborando a hipótese de que essas intervenções são eficazes na mitigação dos acidentes. A análise espacial revelou que as capitais estaduais, devido à sua infraestrutura mais complexa e densidade populacional elevada, apresentam uma distribuição desigual de sinistros, com pontos críticos concentrados principalmente nas regiões limítrofes, o que sugere a necessidade de uma atenção especial a essas áreas nas políticas públicas. Os mapas de Lisa e o índice de Moran identificaram correlações espaciais significativas em regiões específicas, indicando que a proximidade geográfica desempenha um papel crucial na distribuição dos sinistros, e que intervenções localizadas podem ter efeitos positivos em áreas adjacentes.

Os modelos de ensemble learning e aprendizado profundo, aplicados para prever os indicadores de sinistros, demonstraram alta acurácia, reforçando a eficácia do uso de técnicas avançadas de ciência de dados na gestão de empreendimentos em mobilidade urbana. Esses modelos não apenas confirmam as correlações observadas, mas também oferecem uma ferramenta poderosa para predições futuras, podendo ser adaptados e replicados em diferentes contextos regionais e temporais. No entanto, apesar dos avanços, o estudo também destaca a complexidade envolvida na análise de políticas públicas, sugerindo que a relação de causalidade entre os investimentos em infraestrutura e a redução de acidentes precisa ser explorada em maior profundidade. Propõe-se, assim, que pesquisas futuras se concentrem em avaliar o impacto direto de políticas públicas específicas, utilizando métodos mais refinados de análise causal, para entender melhor os mecanismos pelos quais os investimentos em mobilidade urbana podem influenciar os índices de segurança viária.

Referências

ABDEL-ATY, M.; HALEEM, K. Analyzing angle crashes at unsignalized intersections using machine learning techniques. Accident Analysis & Prevention, v. 43, n. 1, p. 461-470, 2011.

BATTY, M. The new science of cities. Cambridge: MIT Press, 2013.

BEHBOUDI, N.; MOOSAVI, S.; RAMNATH, R. Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques. *arXiv*, 2024.

BIAU, G.; SCORNET, E. A random forest guided tour. Test, v. 25, p. 197-227, 2016.

BÍL, M.; ANDRÁSIK, R.; JANOSKA, Z. Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. Accident Analysis & Prevention, v. 55, p. 265-273, 2013.

CATS, O.; FERRANTI, F. Unravelling individual mobility temporal patterns using longitudinal smart card data. Research in Transportation Business & Management,

2022. Disponível em: https://doi.org/10.1016/j.rtbm.2022.100816. Acesso em: 22 de abril de 2024.

DU, Z.; YANG, H.; LIU, H. X. A trajectory clustering-based approach for bus passenger identification. IET Intelligent Transport Systems, v. 14, n. 4, p. 350-357, 2020.

ELVIK, R.; HOYE, A.; VAA, T.; SØRENSEN, M. The handbook of road safety measures. Bingley: Emerald Group Publishing Limited, 2009.

GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 1. ed. Sebastopol: O'Reilly Media, 2019. 856 p.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. Machine Learning, v. 63, p. 3-42, 2006.

HUANG, H.; ABDEL-ATY, M. A.; DARWICHE, A. L. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transportation Research Record, v. 2148, n. 1, p. 27-37, 2010.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. ACM Computing Surveys, v. 31, n. 3, p. 264-323, 1999.

JAPKOWICZ, N.; SHAH, M. Evaluating learning algorithms: a classification perspective. Cambridge: Cambridge University Press, 2011.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, v. 1, n. 14, p. 281-297, 1967.

MANSOURIHANIS, O.; MAGHSOODI, T. M. J.; YOUSEFIAN, S.; ZAROUJTAGHI, A. A computational geospatial approach to assessing land-use compatibility in urban planning. Land, v. 12, n. 11, p. 2083, 2023.

MESQUITA, K. G. A. Método de identificação dos padrões de uso e locais de embarque a partir do Big Data de transporte público: uma abordagem baseada em Machine Learning. 2023. 147 f. Dissertação (Mestrado em Engenharia de Transportes) – Centro de Tecnologia, Universidade Federal do Ceará, Fortaleza, 2023.

PUCHER, J.; DILL, J.; HANDY, S. Infrastructure, programs, and policies to increase bicycling: An international review. Preventive Medicine, v. 50, p. S106-S125, 2010.

SOUSA, F. F. L. M.; MESQUITA, K. G. A.; LOUREIRO, C. F. G. Caracterização da evolução do padrão de mobilidade de Fortaleza a partir da calibração do Tranus. In: 33° CONGRESSO NACIONAL DE PESQUISA EM TRANSPORTES DA ANPET, 2019, Balneário Camboriú, SC. Anais [...]. Balneário Camboriú: ANPET, 2019.

VASCONCELLOS, E. A. Transporte urbano, espaço e eqüidade: análise das políticas públicas. São Paulo: Annablume, 2001.

WHO (World Health Organization). Global status report on road safety. Geneva: WHO Press, 2021.