Análise de Dados de Acidentes Rodoviários Federais (2007-2024): Uma Abordagem Híbrida de Engenharia de Dados e Clusterização Não-Supervisionada

Vinícius Santos Monteiro

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
São Carlos – SP – Brazil

vini.mon@usp.br

Abstract. *Road safety is a critical public health issue in Brazil. This work presents an end-to-end data science project analyzing 18 years of Federal Highway Police (PRF) accident data. We developed an automated ETL pipeline using n8n and Python to extract, clean, and harmonize legacy datasets with inconsistent schemas. The methodology employs a dual clustering strategy: (1) Behavioral Clustering using K-Prototypes to identify accident profiles (e.g., 'Mass Casualty Tragedies' vs. 'Urban Conflicts') based on mixed categorical and numerical data; and (2) Geospatial Clustering using DBSCAN to detect high-lethality hotspots along highways. Results reveal that while most accidents are minor rear-end collisions (Cluster 3), a small fraction (Cluster 4, 5.4%) accounts for a disproportionate number of deaths. Furthermore, geospatial analysis identified specific critical segments, such as Km 207 of BR-116/SP, guiding targeted public policies.*

Resumo. *A segurança viária é um problema crítico de saúde pública no Brasil. Este trabalho apresenta um projeto de ciência de dados ponta a ponta analisando 18 anos de dados de acidentes da Polícia Rodoviária Federal (PRF). Desenvolvemos um pipeline de ETL automatizado utilizando n8n e Python para extrair, limpar e harmonizar datasets legados com esquemas inconsistentes. A metodologia emprega uma estratégia de clusterização dupla: (1) Clusterização Comportamental usando K-Prototypes para identificar perfis de acidentes (ex: 'Tragédias de Múltiplos Veículos' vs. 'Conflitos Urbanos') baseados em dados mistos categóricos e numéricos; e (2) Clusterização Geoespacial usando DBSCAN para detectar hotspots de alta letalidade ao longo das rodovias. Os resultados revelam que, embora a maioria dos acidentes sejam colisões traseiras leves (Cluster 3), uma pequena fração (Cluster 4, 5.4%) é responsável por um número desproporcional de mortes. Além disso, a análise geoespacial identificou segmentos críticos específicos, como o Km 207 da BR-116/SP, orientando políticas públicas direcionadas.*

1. Introdução

O Brasil enfrenta desafios significativos na gestão de sua malha rodoviária federal, onde milhares de acidentes ocorrem anualmente. A Polícia Rodoviária Federal (PRF) disponibiliza dados abertos sobre ocorrências desde 2007, constituindo um rico repositório para análise. No entanto, o volume e a heterogeneidade desses dados - que sofreram mudanças

de metodologia de coleta ao longo de quase duas décadas - impõem barreiras técnicas significativas para a extração de conhecimento acionável.

Este trabalho propõe uma abordagem completa de Ciência de Dados para investigar os padrões ocultos nos acidentes no território brasileiro. Diferente de abordagens puramente estatísticas, utilizamos Aprendizado de Máquina Não-Supervisionado para descobrir agrupamentos (*clusters*) naturais nos dados, tanto sob a ótica comportamental (causas e tipos) quanto geoespacial (localização).

Os objetivos específicos são:

- Implementar um pipeline de Engenharia de Dados automatizado para ingestão e normalização de dados históricos.
- Aplicar algoritmos de clusterização para dados mistos (*K-Prototypes*) a fim de taxonomia de acidentes.
- Identificar zonas de alta letalidade (*hotspots*) utilizando algoritmos baseados em densidade (*DBSCAN*).
- Fornecer *insights* para políticas públicas baseados em evidências.

O código-fonte, os notebooks de análise e a documentação completa estão disponíveis no repositório do projeto: [1]

2. Metodologia e Engenharia de Dados

A metodologia seguiu o ciclo de vida clássico de projetos de dados: Coleta, Limpeza, Modelagem e Avaliação. Uma ênfase especial foi dada à etapa de Engenharia de Dados, crítica para lidar com a ”sujeira”vinda dos dados legados do governo.

2.1. Pipeline de ETL com n8n e Docker

Para a extração dos dados, utilizamos a ferramenta de automação de fluxo de trabalho *n8n*, executada em um contêiner *Docker*. O objetivo foi simular um ambiente de produção robusto, capaz de orquestrar o download de múltiplos arquivos anuais (2007-2024) do portal de dados abertos da PRF.

O fluxo, ilustrado na Figura 1, automatiza as seguintes tarefas:

1. Requisição HTTP à página de dados abertos.
2. *Scraping* e extração dos links de download dos arquivos CSV/ZIP.
3. Download e descompressão controlada dos arquivos.
4. Padronização inicial de nomenclatura e armazenamento em estrutura de diretórios (*Data Lake* local).

[1] `https://github.com/vini-mon/PRF-Accident-Clustering`

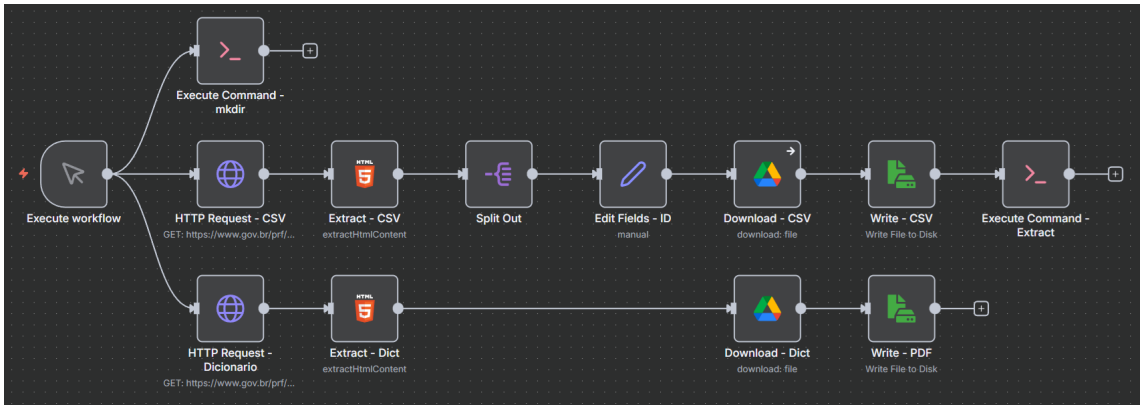

Figura 1. Fluxo de automação ETL no *n8n*. O pipeline realiza a varredura do site, download dos arquivos zipados e organização no sistema de arquivos.

Esta abordagem permitiu reprodutibilidade e escalabilidade, eliminando a necessidade de download manual propenso a erros.

2.2. Limpeza e Harmonização de Dados Legados

A etapa mais desafiadora foi a harmonização dos *schemas* (estrutura das tabelas). Os dados da PRF sofreram uma mudança drástica de metodologia em 2017.

- **Dados Antigos (2007-2016):** Não possuíam coordenadas geográficas precisas (Latitude/Longitude), utilizavam separadores decimais inconsistentes e apresentavam alta cardinalidade em campos de texto livre (ex: variações de grafia para ”Falta de Atenção”).
- **Dados Novos (2017-2024):** Introduziram geolocalização e novas categorias de tipos de acidente.

Utilizando a biblioteca *Pandas* em Python, desenvolvemos funções de tratamento para unificar esses períodos em um *DataFrame* mestre (”Megazord”). Isso envolveu a normalização de texto (remoção de acentos, padronização de caixa alta/baixa), tratamento de valores nulos ocultos (ex: strings ”(null)”que na verdade se referiam à um dado nulo) e a criação de *features* de engenharia, como *flags* binárias para características da via (ex: `feat_curve`, `feat_relief`) que era um campo multivalorado em muitas ocorrências extraídas de descrições textuais complexas.

2.3. Estratégia de Modelagem: Clusterização Dupla

Optamos por uma estratégia de ”Clusterização Dupla”para capturar diferentes dimensões do problema.

2.3.1. Cluster A: Análise Comportamental (K-Prototypes)

Para entender ”O QUE”acontece, utilizamos o algoritmo *K-Prototypes* [1]. Este algoritmo é uma extensão do clássico *K-Means*, projetado para lidar com dados mistos (numéricos e categóricos), o que é essencial para dados de acidentes que misturam contagens (mortos, feridos) com categorias (clima, tipo de pista). O número ideal de clusters ($K = 6$) foi determinado através do Método do Cotovelo (*Elbow Method*), buscando o ponto de inflexão na curva de custo (Figura 2).

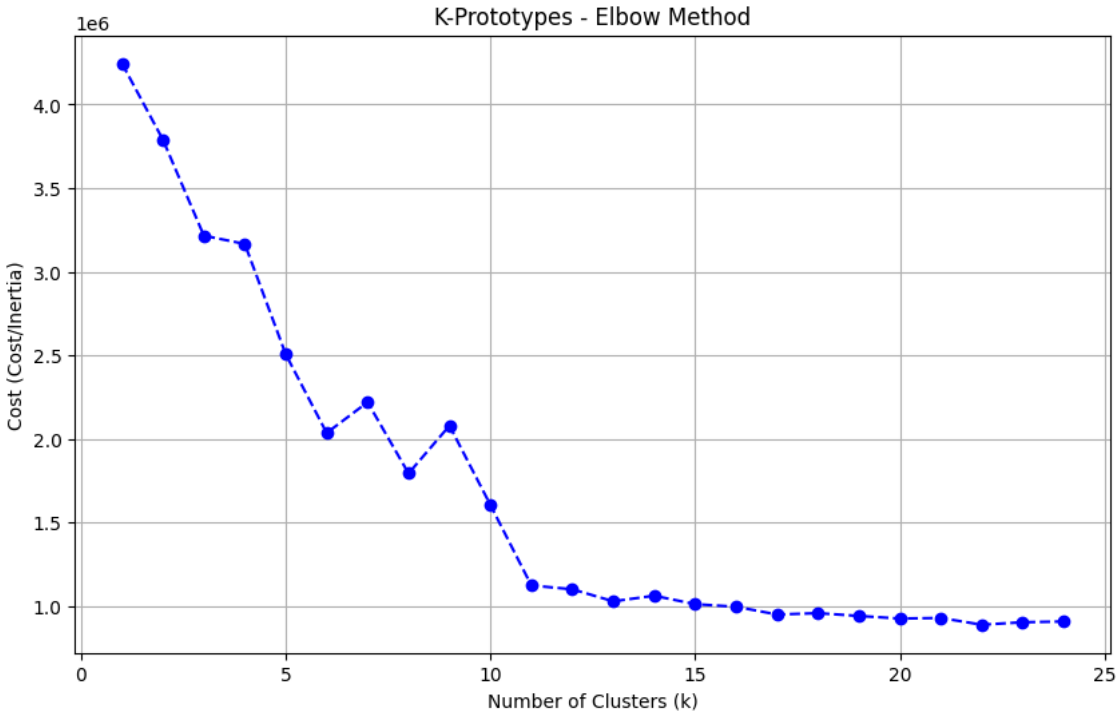

Figura 2. Método do Cotovelo para determinação do K ideal. A curva indica que $K = 6$ oferece o melhor equilíbrio entre compactação dos clusters e interpretabilidade.

2.3.2. Cluster B: Análise Geoespacial (DBSCAN)

Para entender ”ONDE”acontece, utilizamos o *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*) [2]. Diferente do *K-Means*, o *DBSCAN* não requer a definição prévia do número de grupos e é capaz de identificar clusters de formatos arbitrários (como o traçado sinuoso de uma rodovia), além de isolar ruídos (acidentes esporádicos). Utilizamos a distância de Haversine para cálculos geodésicos precisos, definindo múltiplos cenários de raio (*epsilon*) para detectar desde ”curvas da morte”(Micro: 200m) até ”regiões caóticas”(Macro: 5km). Vale ressaltar que, para esse cluster específico, apenas dados posteriores a 2017 foram utilizados pois até essa data, não eram registrados os dados da latitude e longitude do acidente, e utilizar apenas o dado de qual pista e o km em que ocorreu o acidente não seria preciso o suficiente.

3. Resultados e Discussão

A análise resultou na identificação de perfis de acidentes distintos e zonas de risco críticas.

3.1. Cluster A: Taxonomia dos Acidentes

A aplicação do *K-Prototypes* revelou 6 perfis comportamentais distintos (Tabela 1).

Tabela 1. Perfis de Acidentes Identificados (Cluster A)

ID	Nome do Perfil	% Total	Letalidade	Característica Chave
3	O Acidente Padrão	63.9%	Baixa (0.05)	Retas, Colisão Traseira
2	A Tangente da Curva	19.3%	Baixa (0.05)	100% em Curvas
5	Conflito Urbano	7.1%	Baixa (0.02)	Cruzamentos/Retornos
4	Tragédia de Massa	5.4%	**Alta (0.21)**	Múltiplos Veículos/Vítimas
0	Risco do Relevo	3.4%	Média (0.10)	Serras/Declives
1	Gargalo de Obra	0.9%	Média (0.07)	Obras/Pontes

Destacamos que o **Cluster 3 (Padrão)** representa a maioria esmagadora dos casos: acidentes de baixo impacto em retas, sugerindo que a maior parte das ocorrências é fruto de fluxo intenso ou desatenção, e não de falhas na via. Em contraste, o **Cluster 4 (Tragédia de Massa)**, embora raro (5.4%), apresenta uma taxa de letalidade 4 vezes superior à média, envolvendo múltiplos veículos e pessoas, típico de colisões frontais ou envolvendo transporte coletivo.

3.2. Cluster B: Mapeamento de Hotspots

A análise geoespacial com *DBSCAN* permitiu identificar pontos críticos que os números agregados escondem. A Figura 3 ilustra a distribuição dos clusters ao longo da malha viária.

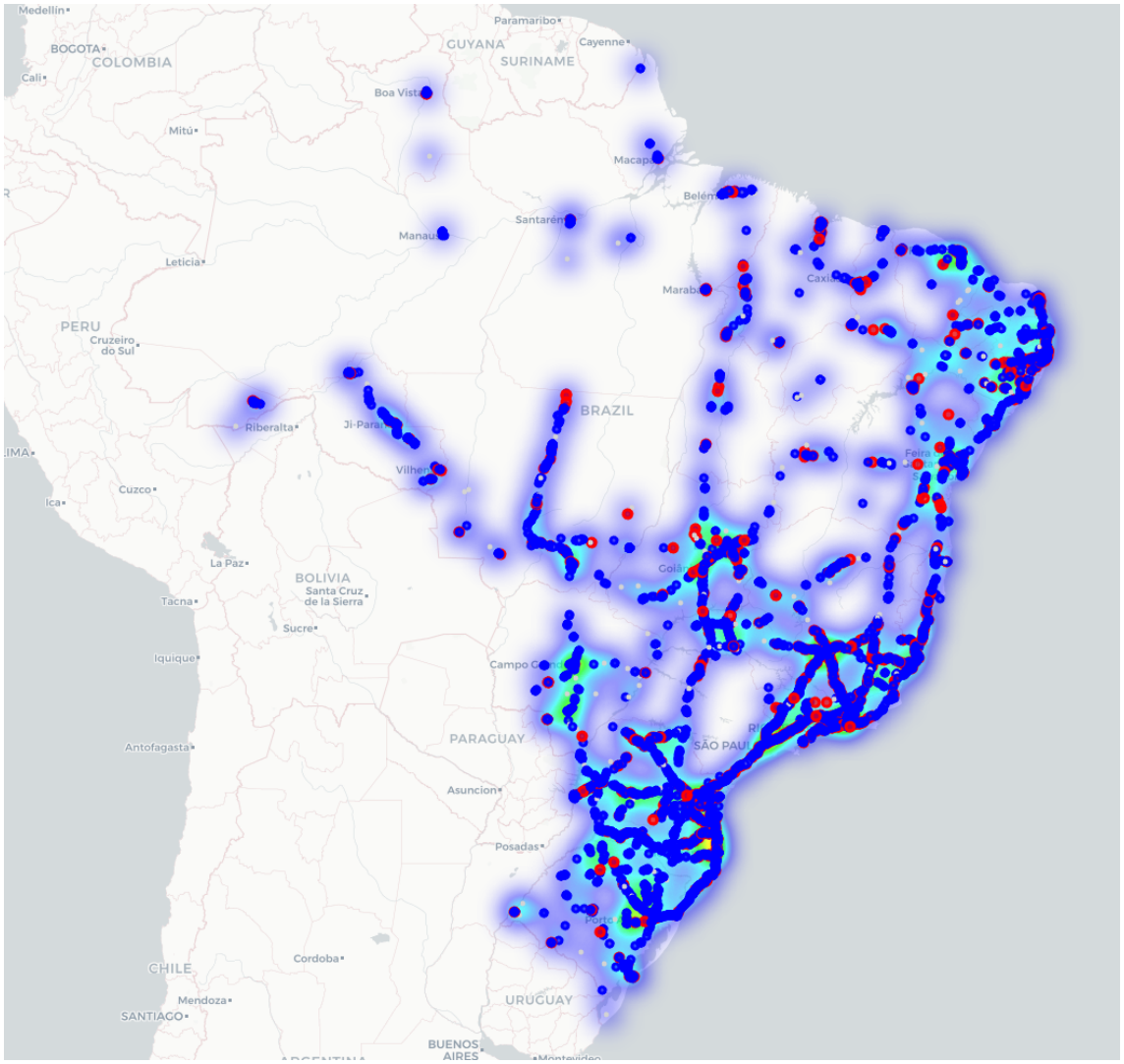

Figura 3. Visualização dos Hotspots identificados pelo DBSCAN. Os pontos vermelhos indicam clusters com alta letalidade, enquanto os azuis indicam alta frequência com menor gravidade.

Ao cruzar os dados geoespaciais com os perfis comportamentais (Análise de DNA do Hotspot), obtivemos *insights* profundos:

- **O Ponto Mais Crítico (Micro):** Localizado no Km 207 da BR-116 em São Paulo (Cluster 162). Este único ponto acumula 278 mortes. A análise de DNA revelou que 75% dos acidentes ali são do tipo ”Padrão”(Cluster 3), indicando que a alta letalidade decorre do volume massivo de tráfego em alta velocidade, não de geometria complexa.
- **A Região do Caos (Macro):** O trecho da BR-101 em Santa Catarina (Km 169) formou um cluster gigantesco com mais de 72.000 acidentes. Apesar do volume, a letalidade é menor que a do cluster de SP, caracterizando um gargalo logístico de saturação.

3.3. Dificuldades e Limitações

O principal obstáculo encontrado foi a inconsistência dos dados históricos. O dicionário de dados nem sempre refletia a realidade das colunas nos arquivos CSV antigos (ex: formatação de datas variando entre anos, colunas de coordenadas preenchidas com zeros ou nulos). Isso exigiu um esforço extensivo de tratamento de exceções no código Python. Além disso, a falta de dados de GPS antes de 2017 limitou a análise geoespacial a um recorte temporal mais recente.

4. Conclusões e Trabalhos Futuros

Este trabalho demonstrou que a aplicação de técnicas de clusterização em dados públicos de tráfego pode revelar padrões vitais para a segurança viária. Concluímos que não existe uma solução única para reduzir acidentes: enquanto ”Acidentes de Curva”(Cluster 2) exigem intervenções de engenharia civil, as ”Tragédias de Massa”(Cluster 4) demandam fiscalização intensiva e gestão de comportamento do condutor.

Como trabalhos futuros, sugerimos:

- **Enriquecimento de Dados:** Integrar dados meteorológicos históricos para validar a correlação entre chuva e o aumento de acidentes do tipo ”Saída de Pista”.
- **Modelagem Preditiva:** Treinar modelos supervisionados (*XGBoost* ou Redes Neurais) para prever a gravidade de um acidente com base nas características da via e do momento.
- **Dashboard em Tempo Real:** Conectar o pipeline do *n8n* a um painel *Power BI* para monitoramento contínuo dos novos dados disponibilizados pela PRF.

Referências

[1] Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3), 283-304.

[2] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD (Vol. 96, No. 34, pp. 226-231).