Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Sistemas de Informação

Mapeamento do Quadro Histórico de Acidentes e Predições nas Rodovias Federais Brasileiras através de Técnicas de Mineração de Dados e Modelos Estatísticos

Pedro Lourenço de Carvalho

JUIZ DE FORA
JULHO, 2026

Mapeamento do Quadro Histórico de Acidentes e Predições nas Rodovias Federais Brasileiras através de Técnicas de Mineração de Dados e Modelos Estatísticos

Pedro Lourenço de Carvalho

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Sistemas de Informação

Orientador: Fabrício Martins Mendonça

JUIZ DE FORA

JULHO, 2026

Mapeamento do Quadro Histórico de Acidentes e Predições nas Rodovias Federais Brasileiras através de Técnicas de Mineração de Dados e Modelos Estatísticos

Pedro Lourenço de Carvalho

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM SISTEMAS DE INFORMAÇÃO.

Aprovada por:

Fabrício Martins Mendonça
Doutor em Ciência da Informação

Ronney Moreira de Castro
Doutor em Informática

Wagner Antonio Arbex
Doutor em Engenharia de Sistemas e Computação

JUIZ DE FORA

15 DE JULHO, 2026

Resumo

Os acidentes em rodovias federais brasileiras exigem análises baseadas em dados para subsidiar ações de saúde pública e segurança viária. Contudo, os registros da Polícia Rodoviária Federal (PRF) apresentam dispersão estrutural e problemas crônicos de qualidade, como duplicidades, valores omissos e erros geográficos. Diante desse problema, este estudo investiga como o uso de técnicas de mineração de dados e modelos estatísticos pode propor soluções alternativas e, em paralelo, revelar padrões espaço-temporais e predizer sinistros. O objetivo geral é mapear o quadro histórico dos acidentes e prever novas ocorrências, por meio de técnicas de mineração de dados e de modelos estatísticos aplicados às bases da PRF, evidenciando os fatores de risco, os perfis de severidade e os padrões espaço-temporais dos sinistros. Sobre os registros da PRF, aplicaram-se as seguintes técnicas de mineração de dados: análise descritiva; detecção de trechos críticos (*hotspots*); mineração de regras de associação (FP-Growth); e previsão de novas ocorrências por séries temporais (SARIMA). Os resultados evidenciam uma clara dissociação entre frequência e letalidade: desfechos mais graves concentram-se em colisões frontais, atropelamentos, madrugadas e pistas simples de rodovias longitudinais. O algoritmo FP-Growth isolou combinações de fatores até seis vezes mais letais que a média, e o modelo SARIMA forneceu estimativas úteis ao planejamento operacional. Todos os artefatos foram integrados em um painel interativo. Como limitações, pontuam-se a imputação de coordenadas pela mediana do trecho, que introduz aproximações espaciais, e a restrição das regras de associação a variáveis categóricas e ao desfecho fatal.

Palavras-chave: Acidentes de trânsito, Mineração de dados, Segurança viária, Rodovias federais.

Abstract

Accidents on Brazilian federal highways require data-driven analyses to support public health and road safety actions. However, the records of the Federal Highway Police (PRF) present structural dispersion and chronic quality problems, such as duplicates, missing values, and geographic errors. Faced with this problem, this study investigates how the use of data mining techniques and statistical models can propose alternative solutions to it and, at the same time, reveal spatio-temporal patterns and predict accidents. The general objective is to map the historical picture of accidents and to predict new occurrences, by means of data mining techniques and statistical models applied to the PRF databases, highlighting the risk factors, severity profiles, and spatio-temporal patterns of the crashes. On the PRF records, the following data mining techniques were applied: descriptive analysis; detection of critical segments (*hotspots*); association rule mining (FP-Growth); and the forecasting of new occurrences through time series (SARIMA). The results reveal a clear dissociation between frequency and lethality: the most severe outcomes are concentrated in head-on collisions, pedestrian run-overs, late-night hours, and single-lane stretches of longitudinal highways. The FP-Growth algorithm isolated factor combinations up to six times more lethal than the average, and the SARIMA model provided estimates useful for operational planning. All artifacts were integrated into an interactive dashboard. As limitations, we note the imputation of coordinates by the segment median, which introduces spatial approximations, and the restriction of the association rules to categorical variables and to the fatal outcome.

Keywords: Traffic accidents, Data mining, Road safety, Federal highways.

Agradecimentos

Agradeço, primeiramente, aos meus pais, Fabrício e Vanessa, alicerce de tudo o que sou, pelo amor incondicional e pelo apoio constante em cada etapa da minha vida.

Aos meus irmãos, Arthur e João, pela parceria, pelo incentivo e pela cumplicidade que carregamos desde sempre.

À Raquel, meu amor, pelo carinho, pela paciência, por tornar essa caminhada mais leve e por estar ao meu lado nos momentos mais desafiadores desta jornada.

Ao professor Fabrício Martins Mendonça pela orientação, amizade e, principalmente, pela paciência, sem a qual este trabalho não se realizaria.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos contribuíram de algum modo para o meu enriquecimento pessoal e profissional.

Por fim, aos meus parentes e amigos que me acompanharam ao longo dessa trajetória, pelo apoio e pela torcida em todos os momentos.

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

Arthur Conan Doyle

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ADF	Teste *Augmented Dickey-Fuller*
AIC	*Akaike Information Criterion*
ARCH	*Autoregressive Conditional Heteroscedasticity*
BIC	*Bayesian Information Criterion*
BIT	Banco de Informações de Transportes
CNT	Confederação Nacional do Transporte
DATATRAN	Base de dados de acidentes de trânsito da PRF
DCC	Departamento de Ciência da Computação
ETL	Extração, Transformação e Carga (*Extract, Transform and Load*)
FDR	*False Discovery Rate* (taxa de falsas descobertas)
FP-Growth	*Frequent Pattern Growth*
GARCH	*Generalized Autoregressive Conditional Heteroscedasticity*
IPEA	Instituto de Pesquisa Econômica Aplicada
KDD	*Knowledge Discovery in Databases*
KPSS	Teste *Kwiatkowski-Phillips-Schmidt-Shin*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MLP	*Multilayer Perceptron*
PICO	População, Intervenção, Comparação e Desfecho (*Outcome*)
PRF	Polícia Rodoviária Federal
PRISMA	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
RMSE	*Root Mean Squared Error*
SARIMA	*Seasonal AutoRegressive Integrated Moving Average*
SIGMA	Sistema Integrado de Gestão de Acidentes
SVM	*Support Vector Machine*
UFJF	Universidade Federal de Juiz de Fora

1 Introdução

Este capítulo apresenta a contextualização do problema abordado nesta pesquisa, delimitando o tema, as motivações e os objetivos do trabalho. A estrutura está dividida em subseções que detalham a importância da análise de dados de segurança viária e a abordagem proposta para este projeto.

1.1 Apresentação do Tema

Os acidentes de trânsito configuram-se como uma das principais mazelas sociais e de saúde pública no cenário brasileiro, gerando impactos socioeconômicos profundos e ceifando dezenas de milhares de vidas anualmente. Para compreender a real magnitude desse problema, torna-se imperativo analisar a configuração logística nacional, visto que o Brasil se caracteriza estruturalmente como um país com uma matriz de transporte predominantemente rodoviária. Segundo dados oficiais da Confederação Nacional do Transporte (CNT), o modal rodoviário atua como o grande protagonista da movimentação interna, sendo o responsável por aproximadamente 65% do transporte de cargas e por cerca de 95% do deslocamento total de passageiros em território nacional (Confederação Nacional do Transporte, 2024). Essa dependência extrema em relação às rodovias eleva exponencialmente a exposição de veículos, condutores e pedestres a riscos inerentes ao tráfego viário, tornando a segurança das estradas um tema de relevância estratégica nacional.

Essa primazia do modal rodoviário não é fortuita, mas resultado de uma opção histórica de política de transportes (PAULA, 2010). A prioridade conferida às estradas remonta ao governo de Washington Luís (1926–1930), cuja máxima "governar é abrir estradas" inaugurou a ênfase na construção rodoviária em detrimento da expansão ferroviária. Esse direcionamento foi consolidado nas décadas seguintes e atingiu seu ápice no governo de Juscelino Kubitschek (1956–1961), cujo Plano de Metas implantou a indústria automobilística nacional e expandiu vigorosamente a malha rodoviária, em sintonia com a construção de Brasília e o ideário desenvolvimentista. Como consequência, o trans-

porte ferroviário foi progressivamente relegado, e o país consolidou uma matriz logística estruturalmente dependente das rodovias, condição que, até hoje, amplifica a exposição da população aos riscos do tráfego viário e confere caráter estratégico à segurança nas estradas.

As estimativas históricas e os indicadores epidemiológicos reforçam a severidade da violência no trânsito ao longo dos anos. De acordo com estudos do Instituto de Pesquisa Econômica Aplicada (IPEA) baseados no Sistema de Informações sobre Mortalidade (SIM/DATASUS), o Brasil sustentou uma média alarmante de aproximadamente 40 mil óbitos anuais decorrentes de sinistros de transporte terrestre ao longo da última década (Instituto de Pesquisa Econômica Aplicada, 2023). Embora legislações punitivas e campanhas educativas busquem arrefecer esses índices, estatísticas recentes de saúde apontam que o país registrou uma nova tendência de alta na mortalidade entre os anos de 2019 e 2024, culminando em 38.253 mortes consolidadas sob essa causa no ano de 2024, o que equivale a uma taxa expressiva de 18 óbitos para cada 100 mil habitantes (SANTOS; SANTOS, 2026).

Esse panorama evidencia que a letalidade nas rodovias brasileiras demanda abordagens metodológicas inovadoras e cientificamente robustas. Diante do volume massivo e da natureza complexa dos registros de ocorrências gerados continuamente por órgãos fiscalizadores, as ferramentas tradicionais de análise descritiva tornam-se limitadas para desvelar os fatores latentes que propiciam as colisões. Neste cenário, a aplicação de técnicas avançadas de Ciência de Dados e, especificamente, de Mineração de Dados (*Data Mining*) apresenta-se como um caminho viável. O uso de algoritmos inteligentes permite processar grandes volumes de dados brutos, mapear padrões de comportamento, identificar pontos críticos de acidentabilidade e extrair conhecimento implícito que possa subsidiar de maneira pragmática a formulação de estratégias de prevenção e engenharia de tráfego.

Para mitigar esse cenário, a análise de grandes volumes de dados surge como uma alternativa indispensável para entender a dinâmica dos sinistros. Nesse contexto, a Ciência de Dados e, especificamente, as técnicas de Mineração de Dados (*Data Mining*) oferecem ferramentas robustas para extrair conhecimento implícito, descobrir padrões

ocultos e identificar perfis de risco que não são facilmente observáveis em análises estatísticas tradicionais.

1.2 Problema de Pesquisa e Justificativa

Apesar da ampla disponibilidade de dados abertos fornecidos periodicamente pela Polícia Rodoviária Federal, a mera existência de grandes volumes de registros não garante a extração imediata de inteligência preditiva ou de diagnósticos precisos para a segurança viária. O primeiro grande desafio reside na fragmentação governamental dos ecossistemas informacionais. Atualmente, os dados estruturados sobre sinistros coexistem em fontes com diferentes níveis de agregação, como o Sistema Integrado de Gestão de Acidentes (SIGMA), focado no preenchimento detalhado de boletins de ocorrência, e o Portal de Dados Abertos da PRF[1], estruturado para a divulgação de matrizes tabulares consolidadas. A falta de um identificador unificado e de um protocolo de sincronização automática entre essas frentes gera assimetrias informacionais, nas quais variáveis ambientais importantes de uma base não se conectam nativamente com as descrições de severidade de outra.

Além da dispersão estrutural, as bases de dados públicas sobre acidentes enfrentam problemas crônicos relacionados à qualidade dos dados brutos. São frequentes as ocorrências de registros duplicados, valores omissos em campos críticos, tais como o tipo de pista, o traçado da via e as condições meteorológicas, além de erros severos de digitação nas coordenadas geográficas de latitude e longitude informadas pelos agentes em campo. Sob a ótica da Ciência de Dados, o fenômeno conhecido como *"garbage in, garbage out"* explicita que a aplicação direta de algoritmos de mineração de dados sobre bases corrompidas ou inconsistentes resulta na descoberta de padrões espúrios e em regras de associação desprovidas de valor estatístico real (HAN; KAMBER; PEI, 2011). Portanto, a ausência de um tratamento metodológico rigoroso atua como uma barreira que impede os órgãos de fiscalização e a comunidade acadêmica de extraírem o real valor desses ativos informacionais.

[1]Disponível em: ⟨https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf⟩

A justificativa para o desenvolvimento desta pesquisa fundamenta-se na necessidade científica e social de mitigar essas lacunas por meio de engenharia de dados aplicada. Nos arquivos anuais da base DATATRAN, distribuída pelo Portal de Dados Abertos da PRF, encontram-se registros duplicados pelo identificador da ocorrência, valores omissos em campos críticos como o tipo de pista, o traçado da via e as condições meteorológicas, descrições categóricas não padronizadas (por exemplo, múltiplas grafias para a ingestão de álcool e de substâncias psicoativas) e, sobretudo, erros nas coordenadas de latitude e longitude, que vão de vírgulas decimais e sinais invertidos a pontos situados fora do território nacional, somando-se a isso os arquivos mais antigos, que sequer dispõem de campos de georreferenciamento. Ao projetar e executar um *pipeline* estruturado de Extração, Transformação e Carga (ETL), este trabalho viabiliza, de forma autônoma e reprodutível, a remoção de duplicatas, a uniformização das categorias e a higienização das coordenadas, com imputação dos valores ausentes, condição sem a qual qualquer análise subsequente estaria comprometida pela baixa qualidade dos dados brutos.

Sobre essa base tratada, justifica-se a aplicação de um conjunto complementar de técnicas de mineração de dados, cada qual respondendo a uma necessidade analítica distinta (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A análise descritiva e exploratória traduz o panorama da acidentalidade em gráficos de causas, tipos, perfis de severidade e padrões temporais; a detecção de anomalias por escore-z evidencia os trechos críticos (*hotspots*) da malha; a mineração de regras de associação pelo algoritmo FP-Growth, com seleção por significância estatística, revela as combinações de fatores mais letais; e a previsão por série temporal, com o modelo SARIMA, estima a quantidade mensal futura de acidentes. A conversão desses resultados em painéis interativos cumpre, por fim, um papel social estratégico, ao traduzir métricas densas em visualizações acessíveis que podem nortear investimentos em infraestrutura e a alocação de policiamento ostensivo nas rodovias federais.

Diante desse cenário de fragmentação e de necessidade de tratamento analítico, esta pesquisa estrutura-se em torno de um conjunto de questões que explicitam como a Computação pode contribuir com soluções informacionais para o problema dos dados públicos de acidentes da PRF. O levantamento de trabalhos relacionados subsidia a dis-

cussão dessas questões:

1. Que tipos de inconsistências podem ser identificadas na base de dados DATATRAN que impedem uma visão histórica e geral dos acidentes nas rodovias federais?

2. Que fatores de risco, perfis de severidade e padrões espaço-temporais (como os trechos críticos, a sazonalidade e os turnos de maior risco) a mineração de dados permite revelar, de modo a explicar onde, quando e em que condições ocorrem os acidentes mais graves nas rodovias federais?

3. Em que medida a mineração de dados da série histórica de acidentes permite prever novas ocorrências nas rodovias federais?

1.3 Objetivos

Esta seção delimita o escopo prático e metodológico do trabalho, estabelecendo a meta principal a ser alcançada e desdobrando-a nos procedimentos técnicos necessários para a sua plena execução.

1.3.1 Objetivo Geral

O objetivo geral desta pesquisa consiste em mapear o quadro histórico dos acidentes nas rodovias federais brasileiras e prever novas ocorrências, por meio de técnicas de mineração de dados e de modelos estatísticos aplicados aos registros consolidados e tratados da Polícia Rodoviária Federal, de modo a evidenciar os fatores de risco, os perfis de severidade e os padrões espaço-temporais dos sinistros.

1.3.2 Objetivos Específicos

Para viabilizar o cumprimento do objetivo geral, estabelecem-se os seguintes objetivos específicos, cada qual materializado em uma das abas do painel interativo entregue ao final da pesquisa:

- Caracterizar o panorama georreferenciado da acidentalidade nas rodovias federais brasileiras, quantificando a distribuição espacial das ocorrências e de sua severidade;

- Identificar os padrões temporais do risco, como a sazonalidade, os turnos e os dias de maior letalidade, evidenciando quando ocorrem os acidentes mais graves;
- Descobrir os fatores de risco e as combinações de condições desproporcionalmente letais associadas aos desfechos fatais;
- Detectar os trechos rodoviários com concentração anômala de acidentes graves (*hotspots*);
- Prever novas ocorrências de acidentes em base mensal, com modelo validado frente a alternativas.

1.4 Estrutura do Trabalho

Para alcançar os objetivos propostos, este documento está organizado em seis capítulos. O Capítulo 2 apresenta a Fundamentação Teórica, revisando os conceitos de segurança viária e as principais tarefas de mineração de dados empregadas no trabalho. O Capítulo 3 discute os Trabalhos Relacionados, analisando individualmente os estudos existentes e comparando-os com esta pesquisa. O Capítulo 4 detalha a Metodologia adotada, incluindo os métodos de pesquisa empregados no levantamento bibliográfico, as fontes de dados, a arquitetura do *pipeline* de tratamento e os modelos analíticos aplicados. O Capítulo 5 expõe e discute os Resultados obtidos, incluindo a caracterização do panorama de acidentalidade, os padrões espaço-temporais identificados e a avaliação do modelo preditivo. Por fim, o Capítulo 6 traz as Considerações Finais, sintetizando as contribuições, as limitações e as perspectivas de trabalhos futuros.

2 Fundamentação Teórica

Este capítulo reúne os fundamentos conceituais que sustentam a pesquisa, apresentados na mesma ordem em que são mobilizados pela metodologia. Inicialmente, discute-se a segurança viária e a natureza dos dados de acidentes de trânsito. Em seguida, apresentam-se os fundamentos de engenharia de dados que orientam o tratamento dessas bases, os conceitos de descoberta de conhecimento em bases de dados e as tarefas de mineração de dados aplicadas, os modelos preditivos empregados e, por fim, a visualização da informação que orienta a construção dos painéis analíticos.

2.1 Segurança Viária e Dados de Acidentes de Trânsito

A segurança viária constitui um campo interdisciplinar que articula engenharia de tráfego, saúde pública, comportamento humano e políticas públicas, com o objetivo de reduzir a frequência e a severidade dos sinistros de trânsito. Segundo Elvik et al. (2009), as ocorrências de trânsito resultam da interação entre três elementos fundamentais: o fator humano, o veículo e a via, incluindo as condições ambientais que a circundam. Essa tríade é amplamente adotada na literatura para classificar os fatores contribuintes de um acidente e fundamenta a seleção das variáveis explicativas analisadas neste trabalho, tais como o traçado da via, o tipo de pista, a condição meteorológica e a causa atribuída ao condutor.

No contexto brasileiro, a Polícia Rodoviária Federal (PRF) é o órgão responsável pelo registro dos sinistros ocorridos na malha rodoviária federal. Esses registros são consolidados na base DATATRAN e disponibilizados publicamente por meio do Portal de Dados Abertos da instituição (Polícia Rodoviária Federal, 2025), em conformidade com os princípios de transparência e governo aberto. Cada registro corresponde a uma ocorrência e contém atributos temporais (data e horário), espaciais (unidade federativa, rodovia, quilômetro e coordenadas geográficas), descritivos (causa, tipo, traçado e condições ambientais) e de severidade (número de mortos, feridos e ilesos).

A relevância de bases de dados confiáveis decorre da própria magnitude do problema de saúde pública que os sinistros de trânsito representam. Essas ocorrências estão entre as principais causas de morte por fatores externos no Brasil e impõem elevado custo assistencial e socioeconômico, sobretudo pela sobrecarga ao Sistema Único de Saúde e pela perda de vidas em idade produtiva (Instituto de Pesquisa Econômica Aplicada, 2023; SANTOS; SANTOS, 2026). Dimensionar o fenômeno, identificar seus fatores determinantes e avaliar políticas de prevenção dependem, antes de tudo, da existência de dados íntegros e comparáveis ao longo do tempo, o que faz da qualidade das bases uma questão central, e não acessória, para a segurança viária.

Apesar de sua riqueza, os dados de acidentes apresentam desafios estruturais amplamente reconhecidos. O primeiro é a fragmentação informacional: os registros distribuem-se por fontes distintas e pouco integradas, como os boletins da PRF, restritos às rodovias federais, e os dados de mortalidade do SIM/Datasus, de origem hospitalar, sem um identificador comum que viabilize o cruzamento direto entre elas (Instituto de Pesquisa Econômica Aplicada, 2023). O segundo é a subnotificação e o preenchimento incompleto, frequentes nos campos manuais. O terceiro, e mais crítico para a mineração, é a baixa qualidade intrínseca dos registros: ao analisar os boletins das rodovias federais, Costa, Bernardini e Filho (2014) concluiu que essa baixa qualidade é o principal obstáculo à extração de conhecimento confiável, manifestando-se em duplicidades, categorias não padronizadas, valores omissos em campos críticos e erros nas coordenadas geográficas. Soma-se a isso a ausência de padronização entre os formatos publicados ao longo dos anos e as mudanças nos próprios procedimentos de registro: a partir de 2018, a PRF passou a registrar parte das ocorrências por meio de boletim eletrônico, o que introduziu uma redução de nível na série, ainda que moderada, aspecto tratado em detalhe no Capítulo 4 e que orienta a definição da janela de modelagem temporal.

É nesse cenário que a Computação oferece o instrumental capaz de converter dados brutos e dispersos em conhecimento acionável. A engenharia de dados, por meio de *pipelines* de Extração, Transformação e Carga, automatiza a unificação das fontes, a remoção de duplicatas, a padronização das categorias e a correção e imputação das coordenadas, entregando uma base íntegra e reprodutível. Sobre ela, o processo de Descoberta

de Conhecimento em Bases de Dados e suas técnicas de mineração permitem caracterizar padrões, isolar fatores de risco, detectar concentrações espaciais anômalas e prever tendências, tarefas inviáveis pela inspeção manual de milhões de registros (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; HAN; KAMBER; PEI, 2011). Por fim, a visualização da informação traduz esses resultados em representações acessíveis, como painéis interativos, aproximando o conhecimento técnico dos gestores e da sociedade. É essa cadeia, da engenharia de dados à visualização, que estrutura a presente pesquisa e organiza os fundamentos apresentados nas seções seguintes.

2.2 Engenharia de Dados e Arquitetura em Camadas

A engenharia de dados reúne os métodos e as ferramentas para coletar, integrar, transformar e disponibilizar grandes volumes de dados de forma confiável, escalável e reprodutível, constituindo a fundação sobre a qual qualquer análise consistente se apoia. Quando os dados provêm de fontes heterogêneas, mudam de formato ao longo do tempo ou apresentam defeitos de qualidade, torna-se essencial organizar o seu processamento em estágios bem definidos, em vez de aplicar transformações *ad hoc* diretamente sobre os arquivos originais. Uma prática hoje consolidada para essa organização é a arquitetura em camadas conhecida como *medallion*, que dispõe o fluxo de dados em níveis de refinamento progressivo, do dado bruto ao dado pronto para consumo analítico (INMON, 2005).

A arquitetura *medallion* estrutura o processamento em três camadas sucessivas, cada uma com uma responsabilidade própria. A camada *bronze*, também chamada de *raw*, preserva os dados brutos exatamente como recebidos da fonte, sem qualquer modificação de conteúdo; seu papel é assegurar a fidelidade da ingestão e servir de ponto de partida imutável para reprocessamentos, acomodando a heterogeneidade de codificações de caracteres e a evolução dos esquemas (*schema drift*) ao longo do tempo. A camada *silver* concentra a limpeza e a padronização ao nível de registro: é nela que ocorrem a remoção de duplicatas, a conversão de tipos, a normalização de categorias, o tratamento de valores omissos e a integração entre diferentes origens, produzindo uma base consistente e confiável. Por fim, a camada *gold* consolida os dados já tratados em tabelas agregadas e orientadas ao consumo, organizadas segundo as dimensões e os indicadores de interesse,

de modo a alimentar diretamente os painéis, os relatórios e os modelos analíticos com baixo custo de leitura.

A principal vantagem dessa separação em camadas é a delimitação clara de responsabilidades entre os estágios, o que favorece a rastreabilidade da origem de cada dado (*data lineage*), o reprocessamento seletivo de uma etapa sem a necessidade de refazer as demais e, sobretudo, a reprodutibilidade do experimento, já que cada camada pode ser regenerada de forma determinística a partir da anterior. A persistência intermediária costuma adotar formatos colunares, como o Parquet, que comprimem os dados e aceleram as leituras analíticas ao permitirem o acesso seletivo apenas às colunas relevantes. Esses princípios fundamentam a arquitetura do *pipeline* adotado nesta pesquisa, cuja implementação concreta sobre os dados da PRF, com as operações específicas de cada camada, é detalhada no Capítulo 4.

2.3 Descoberta de Conhecimento em Bases de Dados

A Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* — KDD) é definida por Fayyad, Piatetsky-Shapiro e Smyth (1996) como o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis a partir de dados. O processo de KDD é tipicamente decomposto em etapas iterativas: seleção, pré-processamento, transformação, mineração de dados e, por fim, interpretação e avaliação do conhecimento extraído. Nesse arcabouço, a mineração de dados é a etapa central, responsável pela aplicação de algoritmos que efetivamente extraem os padrões, ao passo que as etapas que a antecedem são responsáveis por garantir a qualidade dos dados de entrada.

Cabe distinguir, nesse ponto, a mineração de dados do próprio KDD, termos por vezes empregados como sinônimos, mas que designam níveis distintos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O KDD é o processo completo e iterativo que parte dos dados brutos e culmina no conhecimento interpretado, abrangendo desde a seleção e a preparação dos dados até a avaliação dos padrões obtidos. A mineração de dados, por sua vez, é a etapa específica desse processo na qual algoritmos são efetivamente aplicados para extrair padrões, modelos ou regras a partir dos dados já preparados. Em

outras palavras, toda mineração de dados ocorre dentro de um processo de KDD, mas o KDD não se reduz à mineração: o resultado da etapa de mineração depende criticamente da qualidade das etapas que a antecedem, e o valor dos padrões extraídos só se concretiza na etapa de interpretação que a sucede. É por isso que a mineração, embora seja o núcleo algorítmico do processo, não pode ser avaliada isoladamente das demais etapas (HAN; KAMBER; PEI, 2011).

A relevância das etapas de preparação é frequentemente sintetizada pelo princípio *garbage in, garbage out*: a aplicação de algoritmos sofisticados sobre dados de baixa qualidade conduz à descoberta de padrões espúrios e a conclusões inválidas (HAN; KAMBER; PEI, 2011). Estima-se que as atividades de seleção, limpeza e transformação consumam a maior parte do esforço de um projeto de mineração de dados, o que justifica a centralidade do *pipeline* de tratamento descrito nesta monografia.

No âmbito desta pesquisa, a correspondência entre as etapas do KDD e as suas componentes é direta. As etapas de seleção, pré-processamento e transformação materializam-se no *pipeline* de tratamento em camadas (Seção 2.2), que unifica, limpa e padroniza os registros da PRF. A etapa de mineração propriamente dita corresponde à aplicação da análise descritiva, das regras de associação, da detecção de anomalias e da previsão de séries temporais, descritas no Capítulo 4. Por fim, a etapa de interpretação e avaliação concretiza-se na consolidação dos padrões em painéis interativos, que tornam o conhecimento extraído compreensível e acionável. Essa leitura integrada reforça que a contribuição do trabalho não reside em uma técnica isolada, mas no encadeamento completo do processo de descoberta de conhecimento.

2.4 Tarefas de Mineração de Dados

A mineração de dados engloba um conjunto de tarefas que, segundo Tan, Steinbach e Kumar (2009), podem ser agrupadas em dois grandes paradigmas. As tarefas *preditivas* estimam o valor de um atributo-alvo a partir dos demais e abrangem a classificação, quando o alvo é categórico, e a regressão, quando o alvo é numérico, caso em que se inclui a previsão de séries temporais. As tarefas *descritivas*, por sua vez, buscam padrões interpretáveis que sumarizam as relações latentes nos dados e compreendem a análise de

associação, que identifica regras de coocorrência entre atributos; a análise de agrupamentos (*clustering*), que reúne instâncias semelhantes sem rótulos predefinidos; e a detecção de anomalias, que isola observações atípicas. A essas tarefas costuma anteceder a análise descritiva e exploratória que, embora não constitua um algoritmo de mineração em sentido estrito, caracteriza a distribuição dos dados e orienta a formulação de hipóteses.

Este trabalho não emprega todas essas tarefas, mas um subconjunto complementar, escolhido em função dos seus objetivos. A análise de agrupamentos não é utilizada; em contrapartida, mobilizam-se a análise descritiva e exploratória, a previsão de séries temporais, a análise de associação e a detecção de anomalias, além de a classificação ser tomada como referência conceitual para a análise de severidade. As subseções a seguir descrevem cada uma dessas tarefas, com ênfase no papel que desempenham nesta pesquisa.

2.4.1 Análise Descritiva e Exploratória

A análise exploratória de dados consiste no uso de medidas-resumo e de representações visuais para caracterizar a distribuição das variáveis, detectar anomalias e formular hipóteses (WITTEN; FRANK; HALL, 2011). No domínio dos acidentes de trânsito, essa análise permite quantificar a frequência relativa de causas e tipos de acidente, comparar a severidade entre diferentes condições e identificar concentrações espaciais e temporais de ocorrências, subsidiando as etapas analíticas subsequentes.

2.4.2 Classificação e Análise de Severidade

A classificação é uma tarefa preditiva supervisionada cujo objetivo é atribuir uma instância a uma entre um conjunto de classes predefinidas, a partir de um modelo aprendido sobre exemplos rotulados (HAN; KAMBER; PEI, 2011). Na segurança viária, a classificação é tipicamente empregada para estimar a severidade de um acidente, por exemplo, distinguir ocorrências com vítimas fatais das demais, em função de atributos da via, do ambiente e do condutor. Ainda que este trabalho privilegie a caracterização descritiva da severidade, os fundamentos da classificação orientam a investigação das combinações de fatores associadas aos acidentes fatais.

2.4.3 Previsão de Séries Temporais

A previsão de séries temporais visa estimar valores futuros de uma grandeza a partir de suas observações passadas, exigindo a modelagem de componentes como tendência e sazonalidade (BOX et al., 2015). Duas grandes famílias de métodos são empregadas para essa tarefa: os modelos estatísticos clássicos, como o SARIMA, detalhado na Seção 2.5, e os modelos de aprendizado de máquina, como as redes neurais, que tratam a previsão como um problema de regressão supervisionada sobre uma janela de atrasos (*lags*). A escolha entre as duas famílias depende da natureza da série, em especial da força de sua sazonalidade e da linearidade da relação entre as observações passadas e futuras (WITTEN; FRANK; HALL, 2011).

2.4.4 Análise de Associação

A análise de associação é uma tarefa descritiva que busca regras do tipo *se-então* que expressam a coocorrência frequente de itens em um conjunto de transações (HAN; KAMBER; PEI, 2011). Cada regra, da forma $A \rightarrow B$, é avaliada por métricas como o suporte, que mede a frequência relativa da combinação; a confiança, que estima a probabilidade do consequente dado o antecedente; e o *lift*, que indica quantas vezes a ocorrência conjunta supera o esperado sob a hipótese de independência. Algoritmos como o *Apriori* e o FP-Growth percorrem o espaço de combinações de forma eficiente, sendo o FP-Growth particularmente vantajoso por dispensar a geração explícita de candidatos. No domínio dos acidentes de trânsito, essa tarefa permite revelar quais combinações de fatores, como o tipo de acidente, a fase do dia e a condição da via, ocorrem associadas a desfechos de maior gravidade.

2.4.5 Detecção de Anomalias

A detecção de anomalias busca identificar instâncias que se desviam significativamente do comportamento esperado de um conjunto de dados (TAN; STEINBACH; KUMAR, 2009). Uma técnica simples e robusta para essa finalidade é a padronização estatística pelo escore-z, que mede, em unidades de desvio-padrão, a distância de uma observação

em relação à média da distribuição. Trechos rodoviários cujo indicador de severidade apresenta escore-z elevado podem ser interpretados como pontos críticos (*hotspots*) que se destacam estatisticamente do padrão geral, orientando a priorização de intervenções.

2.5 Modelos Estatísticos de Séries Temporais: ARIMA e SARIMA

Os modelos da família ARIMA (*AutoRegressive Integrated Moving Average*), formalizados por Box e Jenkins, constituem a abordagem estatística clássica e de referência para a previsão de séries temporais (BOX et al., 2015). Eles assumem que o valor presente de uma série pode ser explicado pela combinação de três componentes: a parte autorregressiva (AR), que expressa o valor atual como função linear de seus p valores passados; a parte de média móvel (MA), que o relaciona aos q erros de previsão anteriores; e a integração (I), que aplica d diferenciações sucessivas para remover tendências e tornar a série estacionária, condição necessária à modelagem. A combinação desses elementos é denotada por ARIMA(p, d, q).

Quando a série exibe padrões que se repetem em um período fixo, como a sazonalidade anual dos acidentes mensais, emprega-se a extensão sazonal SARIMA (*Seasonal ARIMA*), denotada por $(p, d, q)(P, D, Q)_s$, que acrescenta componentes autorregressivos, de média móvel e de diferenciação aplicados no atraso sazonal s (igual a 12, para dados mensais). A construção do modelo segue a metodologia de Box-Jenkins, organizada em três etapas iterativas. Na *identificação*, a estacionariedade da série é verificada por testes formais, como o ADF (*Augmented Dickey-Fuller*) (DICKEY; FULLER, 1979) e o KPSS (KWIATKOWSKI et al., 1992), e a ordem dos componentes é orientada pelas funções de autocorrelação e de autocorrelação parcial e por critérios de informação, como o AIC (*Akaike Information Criterion*) e o BIC (*Bayesian Information Criterion*), que equilibram a qualidade do ajuste e a parcimônia. A *estimação* dos coeficientes é feita por máxima verossimilhança. Por fim, a *verificação diagnóstica* examina os resíduos, que devem comportar-se como ruído branco, sem autocorrelação (avaliada pelo teste de Ljung-Box), e idealmente apresentar normalidade e variância constante (homocedastici-

dade). Entre as vantagens do SARIMA para este trabalho destacam-se a parcimônia, a adequação a séries de forte componente sazonal e a obtenção nativa de intervalos de confiança para as previsões, propriedade ausente nos modelos de aprendizado de máquina e relevante para a comunicação da incerteza ao tomador de decisão.

Embora o SARIMA seja um modelo de natureza estatística, ele se insere plenamente na mineração de dados como o instrumento de uma de suas tarefas preditivas: a previsão de séries temporais (Seção 2.4). A mineração de dados não se restringe a algoritmos de aprendizado de máquina, mas incorpora também métodos estatísticos consolidados sempre que estes melhor respondem ao problema. Uma mesma tarefa preditiva, como estimar a quantidade mensal de acidentes, pode ser abordada tanto por modelos estatísticos, a exemplo do SARIMA, quanto por modelos de aprendizado de máquina, como o *Perceptron* Multicamadas, razão pela qual ambos são apresentados nesta fundamentação e posteriormente confrontados. A escolha do SARIMA como modelo principal decorre das características da série de acidentes, marcada por forte sazonalidade anual e por dinâmica predominantemente linear, condições para as quais ele oferece parcimônia, interpretabilidade e intervalos de confiança nativos. No arcabouço do KDD, portanto, o SARIMA ocupa a etapa de mineração, convertendo o histórico tratado de acidentes em previsões que, ao final, alimentam os gráficos temporais e a projeção exibidos no painel.

2.6 Redes Neurais Artificiais e o Perceptron Multicamadas

As redes neurais artificiais são modelos computacionais inspirados na organização do sistema nervoso biológico, capazes de aproximar relações não lineares complexas entre variáveis (HAYKIN, 2009). O *Perceptron* Multicamadas (*Multilayer Perceptron* — MLP) é uma arquitetura do tipo *feedforward* composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída, na qual cada neurônio aplica uma função de ativação não linear à combinação ponderada de suas entradas. O treinamento ajusta os pesos da rede por meio do algoritmo de retropropagação do erro, minimizando uma função de perda sobre os exemplos disponíveis. Pelo teorema da aproximação universal,

uma rede com ao menos uma camada oculta e número suficiente de neurônios é capaz de aproximar qualquer função contínua, o que torna o MLP adequado à modelagem de séries temporais com dinâmica não linear (HAYKIN, 2009).

À semelhança do SARIMA, o MLP figura nesta fundamentação por ser um dos instrumentos da tarefa preditiva de previsão de séries temporais (Seção 2.4), agora pela via do aprendizado de máquina. Para aplicá-lo a uma série, a previsão é reformulada como um problema de regressão supervisionada, no qual o valor futuro é estimado a partir de uma janela de observações passadas (*lags*) tomadas como atributos de entrada. Sua principal virtude é a capacidade de capturar relações não lineares que escapam aos modelos lineares clássicos. Neste trabalho, contudo, o MLP não é o modelo principal, mas um modelo de comparação: sua função é servir de referência para avaliar o desempenho do SARIMA, de modo que a escolha do modelo final seja empiricamente justificada, e não apenas presumida. Confrontam-se, assim, as duas grandes famílias de métodos preditivos, a estatística e a de aprendizado de máquina, sobre a mesma tarefa de mineração, o que confere robustez à etapa preditiva do processo de KDD.

2.7 Visualização da Informação

A visualização da informação utiliza representações gráficas para ampliar a capacidade humana de compreender grandes conjuntos de dados, explorando o sistema perceptual visual para revelar padrões, tendências e exceções que permaneceriam ocultos em tabelas numéricas (WARE, 2012). Painéis interativos (*dashboards*) consolidam múltiplas visualizações coordenadas e permitem que o usuário filtre e detalhe os dados de forma exploratória. No domínio da segurança viária, mapas coropléticos e gráficos temporais traduzem métricas densas em representações acessíveis, cumprindo o papel social de tornar o conhecimento extraído utilizável por gestores e pela sociedade.

3 Trabalhos Relacionados

Este capítulo analisa os trabalhos relacionados identificados pelo mapeamento sistemático descrito na Seção 4.2. Diferentemente de uma simples revisão, cada trabalho é examinado individualmente e confrontado com a presente pesquisa, destacando-se as principais diferenças de escopo, de método e de contribuição, bem como os aspectos em que esta pesquisa se sobressai. A Tabela 3.1 sintetiza os cinco trabalhos primários selecionados, discutidos individualmente nas seções seguintes; ao final do capítulo, a Tabela 3.2 consolida a comparação entre esses trabalhos e a presente pesquisa.

Tabela 3.1: Lista de trabalhos selecionados.

Título	**Autoria**
A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras	(COSTA; BERNARDINI; FILHO, 2014)
Modelagem da informação para cidades inteligentes: aplicação em acidentes de trânsito de Belo Horizonte	(CUNHA, 2019)
Aplicação de técnicas de aprendizado de máquina na análise de ocorrências de trânsito de Belo Horizonte-MG	(MARTINS; ANDRADE, 2021)
Técnicas de Aprendizado de Máquina para Predição de Gravidade de Acidentes em Rodovias do Estado do Rio de Janeiro	(MAFORT; KAPPEL, 2025)
Predição de acidentes de trânsito em Santa Catarina: avaliando o impacto do pré-processamento dos dados	(FüHR; INACIO, 2026)

Fonte: Elaborado pelo autor (2026).

3.1 Trabalhos Acadêmicos

Esta seção discute individualmente os cinco trabalhos primários selecionados pelo mapeamento sistemático, sintetizados na Tabela 3.1.

3.1.1 A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras

O artigo de Costa, Bernardini e Filho (2014) aplica regras de associação e classificação simbólica, por meio do *software* Weka, aos boletins de ocorrência da Polícia Rodoviária Federal referentes ao ano de 2012, extraindo 38 regras com confiança superior a 0,8. Sua principal contribuição é o pioneirismo na aplicação de técnicas de mineração de dados aos registros da PRF em âmbito nacional, ainda pouco explorados à época, demonstrando a viabilidade da descoberta de regras de associação nesse domínio. Soma-se a isso o alerta, fundamentado empiricamente, de que a baixa qualidade dos dados é o principal obstáculo à extração de conhecimento confiável, acompanhado da recomendação de aprimoramentos desde a etapa de coleta, o que estabeleceu a qualidade dos dados como questão central para as pesquisas subsequentes.

As limitações do trabalho residem sobretudo na abrangência e na profundidade analítica. A análise restringe-se a um único ano (2012) e a uma única tarefa de mineração, as regras de associação, sem explorar as dimensões espacial e temporal dos acidentes. Além disso, a extração das regras baseia-se em limiares fixos de suporte e confiança definidos a priori, sem qualquer controle de significância estatística, o que limita a garantia de que os padrões reportados não decorram do acaso. A classificação empregada é igualmente simples e de natureza simbólica, sem confronto com modelos preditivos mais robustos.

Em relação a esta pesquisa, o artigo de Costa, Bernardini e Filho (2014) funciona como ponto de partida cuja preocupação central, a qualidade dos dados, é aqui retomada e ampliada. Enquanto o artigo se limita a um ano, este trabalho estende o tratamento dos registros a um *pipeline* de Extração, Transformação e Carga que cobre o período de 2007 a 2025 e integra, além das regras de associação, a análise espacial de trechos críticos, a análise de severidade e a previsão de séries temporais. A própria mineração de associação é aqui conduzida sem limiares fixos: os parâmetros são determinados pelos dados, com seleção por significância estatística e controle da taxa de falsas descobertas, o que confere maior rigor à identificação dos padrões letais e supera a principal fragilidade metodológica

desse artigo.

3.1.2 Modelagem da informação para cidades inteligentes: aplicação em acidentes de trânsito de Belo Horizonte

A dissertação de Mestrado de Cunha (2019) é a que mais se aproxima desta pesquisa em filosofia de implementação. O autor propõe um *pipeline* de ETL associado a painéis de visualização sobre os dados abertos de acidentes de trânsito do município de Belo Horizonte, no contexto da modelagem da informação para cidades inteligentes. Sua principal contribuição é evidenciar, de forma sistematizada, padrões descritivos relevantes, como a concentração de ocorrências em dias úteis e no período da tarde e a forte predominância de condutores do sexo masculino nos casos fatais, demonstrando o valor dos dados abertos e das ferramentas de visualização como instrumentos de apoio à gestão pública da segurança viária.

As limitações dessa dissertação concentram-se no escopo e na natureza das tarefas. A análise opera em escala estritamente municipal e possui caráter essencialmente descritivo, sem incorporar modelagem preditiva nem mineração de regras de associação. Dessa forma, embora o trabalho caracterize bem o que ocorreu, ele não avança na antecipação de tendências futuras nem na descoberta automática de combinações de fatores de risco, permanecendo restrito à observação retrospectiva dos padrões.

A relação com esta pesquisa é de continuidade e de ampliação. Compartilha-se com essa dissertação a filosofia de tratar os dados por um *pipeline* estruturado e de entregar os resultados em painéis interativos, mas aqui a abordagem opera em escala nacional, sobre toda a malha rodoviária federal, e acrescenta uma camada analítica e preditiva ausente naquela dissertação: a previsão mensal por modelos de série temporal, a detecção de trechos críticos por escore de anomalia e a mineração estatística de combinações letais. Com isso, o presente trabalho transforma a abordagem descritiva de painéis em um sistema de apoio à decisão que, além de caracterizar o passado, antecipa tendências.

3.1.3 Aplicação de técnicas de aprendizado de máquina na análise de ocorrências de trânsito de Belo Horizonte-MG

O artigo de Martins e Andrade (2021) combina técnicas estatísticas e de aprendizado de máquina para prever a ocorrência de óbitos a partir das ocorrências de trânsito de Belo Horizonte registradas entre 2011 e 2019. Entre suas contribuições, destacam-se a integração de abordagens estatísticas e de aprendizado de máquina em uma mesma análise e a quantificação da evolução do fenômeno no período, com ênfase na redução de 51,6% nos acidentes com vítimas fatais, resultado que evidencia o potencial dos dados de ocorrência para subsidiar a avaliação de políticas de segurança viária.

As limitações do trabalho dizem respeito ao escopo geográfico e ao recorte da tarefa. A análise restringe-se ao município de Belo Horizonte e concentra-se na classificação da ocorrência de óbitos, sem abordar a previsão da quantidade de acidentes ao longo do tempo nem a dimensão espacial dos trechos mais perigosos. O alcance dos achados fica, assim, circunscrito a uma realidade municipal específica, com limitada generalização para a malha rodoviária federal.

Em relação a esta pesquisa, esse artigo é complementar quanto à tarefa preditiva. Enquanto aquele artigo classifica a ocorrência de óbitos em nível municipal, este trabalho prevê novas ocorrências de acidentes, em base mensal e em escala nacional, uma tarefa de série temporal de natureza distinta e mais sensível à continuidade dos registros. Além disso, a previsão aqui desenvolvida integra-se à análise espacial de trechos críticos e à mineração de regras de associação, compondo um conjunto analítico mais amplo do que o foco em classificação de óbitos adotado nesse artigo.

3.1.4 Técnicas de Aprendizado de Máquina para Predição de Gravidade de Acidentes em Rodovias do Estado do Rio de Janeiro

O artigo de Mafort e Kappel (2025) é, entre os analisados, o mais alinhado a esta pesquisa quanto à fonte de dados, pois também utiliza os registros da Polícia Rodoviária Federal. Os autores comparam sete algoritmos supervisionados para classificar a gravidade dos

acidentes nas rodovias federais do Rio de Janeiro entre 2020 e 2023. Sua principal contribuição está no caráter comparativo e na identificação dos preditores mais influentes da severidade, a saber, o tipo de acidente, a fase do dia, as condições meteorológicas, a localização e a densidade de acidentes por quilômetro, tendo a Regressão Logística alcançado a melhor acurácia, de 73,85%, o que oferece um referencial útil sobre quais variáveis pesam na gravidade das ocorrências.

As limitações desse artigo residem no recorte geográfico e na amplitude das tarefas. O artigo restringe-se a um único estado e a uma janela temporal curta, de 2020 a 2023, e concentra-se exclusivamente na classificação de severidade, sem contemplar a previsão temporal de novas ocorrências, a detecção de trechos críticos ou a entrega dos resultados em uma ferramenta interativa de apoio à decisão. O escopo estadual e o período reduzido limitam a representatividade dos achados para o conjunto das rodovias federais.

A relação com esta pesquisa é, ao mesmo tempo, de proximidade e de ampliação. Compartilha-se a fonte de dados da PRF e o interesse pela severidade dos acidentes, mas aqui a cobertura abrange todas as rodovias federais ao longo de 2007 a 2025, e a análise de severidade é complementada pela previsão temporal, pela detecção de *hotspots* por escore de anomalia e por um painel interativo de apoio à decisão. O presente trabalho aproveita, assim, as lições desse artigo sobre os fatores associados à gravidade e as insere em um arcabouço analítico de escopo nacional e multitarefa.

3.1.5 Predição de acidentes de trânsito em Santa Catarina: avaliando o impacto do pré-processamento dos dados

O artigo de Führ e Inacio (2026) treina os algoritmos *Random Forest*, *Support Vector Machine* e *Perceptron* Multicamadas sobre um trecho da BR-101 em Santa Catarina, no período de 2017 a 2024, enriquecendo os dados da PRF com informações meteorológicas e de tráfego organizadas em cinco conjuntos distintos. Sua principal contribuição é demonstrar empiricamente que a correção de inconsistências na etapa de pré-processamento foi o fator que mais elevou o desempenho dos modelos, evidência que reforça o papel decisivo da qualidade e do tratamento dos dados nas tarefas de mineração aplicadas a acidentes de trânsito.

As limitações do trabalho relacionam-se à granularidade do objeto e à finalidade da análise. Esse artigo examina um único segmento rodoviário, com foco exclusivo na classificação, o que restringe a generalização dos resultados para outras rodovias e contextos. Ademais, embora destaque a importância do pré-processamento, o artigo não estende essa lição a outras tarefas analíticas, como a previsão temporal ou a análise espacial em larga escala.

Em relação a esta pesquisa, esse artigo fornece respaldo direto a uma das premissas centrais deste trabalho: a de que o tratamento cuidadoso dos dados é determinante para a qualidade dos resultados. Aqui, essa lição é generalizada da escala de um único trecho para a escala nacional, por meio de um *pipeline* completo, e estendida a tarefas além da classificação. O presente trabalho confirma empiricamente, no contexto de um modelo preditivo de série temporal, a importância do pré-processamento, mas vai além ao adotar o SARIMA como modelo principal, validado contra o MLP, o Holt-Winters e uma linha de base sazonal, e ao incorporar a dimensão espacial e a entrega final em uma ferramenta interativa, elementos ausentes nesse artigo.

3.2 Iniciativas de Código Aberto

Além dos trabalhos primários, cabe registrar uma iniciativa de código aberto que, embora não constitua publicação acadêmica revisada por pares e, por isso, não atenda aos critérios de inclusão do mapeamento, guarda forte semelhança de implementação com este trabalho. O projeto de Lima (2023) processa os dados públicos de acidentes em rodovias federais por meio de uma arquitetura de *data lakehouse* no padrão *medallion* e disponibiliza os resultados em um painel desenvolvido com *Streamlit*.

A diferença, novamente, está na profundidade analítica. Aquele projeto concentra-se na engenharia de dados e na visualização descritiva, enquanto este acrescenta o tratamento explícito da mudança de nível de 2018, a análise de severidade e de trechos críticos por escore de anomalia, a mineração estatística de regras de associação e a previsão por modelos de série temporal. Registra-se ainda que a camada geográfica das rodovias utilizada no mapa coroplético deste trabalho foi obtida a partir do arquivo GeoJSON simplificado disponibilizado por Lima (2022), cujos dados originais provêm do

Banco de Informações de Transportes (BIT) do Ministério da Infraestrutura (Ministério da Infraestrutura, 2025).

3.3 Síntese Comparativa

A análise individual dos trabalhos revela lacunas recorrentes que delimitam a contribuição desta pesquisa. A Tabela 3.2 sintetiza a comparação entre os trabalhos relacionados e a presente pesquisa, segundo o escopo geográfico, o período analisado, as tarefas de mineração empregadas e a entrega dos resultados em uma ferramenta interativa.

Tabela 3.2: Comparação entre os trabalhos relacionados e esta pesquisa.

Trabalho	**Escopo**	**Período**	**Tarefas empregadas**
(COSTA; BERNARDINI; FILHO, 2014)	Rodovias federais (nacional)	2012	Regras de associação; classificação simbólica
(CUNHA, 2019)	Município (Belo Horizonte)	2012–2018	ETL; análise descritiva; visualização
(MARTINS; ANDRADE, 2021)	Município (Belo Horizonte)	2011–2019	Estatística; classificação de óbitos
(MAFORT; KAPPEL, 2025)	Rodovias federais (RJ)	2020–2023	Classificação de gravidade (sete algoritmos)
(FüHR; INACIO, 2026)	Trecho da BR-101 (SC)	2017–2024	Pré-processamento; classificação
Esta pesquisa	Rodovias federais (nacional)	2007–2025	ETL em camadas; análise descritiva e espaço-temporal; trechos críticos (escore-z); regras de associação (FP-Growth); previsão de novas ocorrências (SARIMA)

Fonte: Elaborado pelo autor (2026).

Dessa comparação emergem as lacunas que esta pesquisa preenche. Em primeiro lugar, predomina o escopo restrito: os trabalhos limitam-se a um único estado (MAFORT; KAPPEL, 2025), a um único trecho rodoviário (FüHR; INACIO, 2026), a um município (CUNHA, 2019; MARTINS; ANDRADE, 2021) ou a um único ano (COSTA; BERNARDINI; FILHO, 2014), ao passo que este trabalho abrange todas as rodovias federais ao longo de dezenove anos. Em segundo lugar, nenhum dos trabalhos trata explicitamente

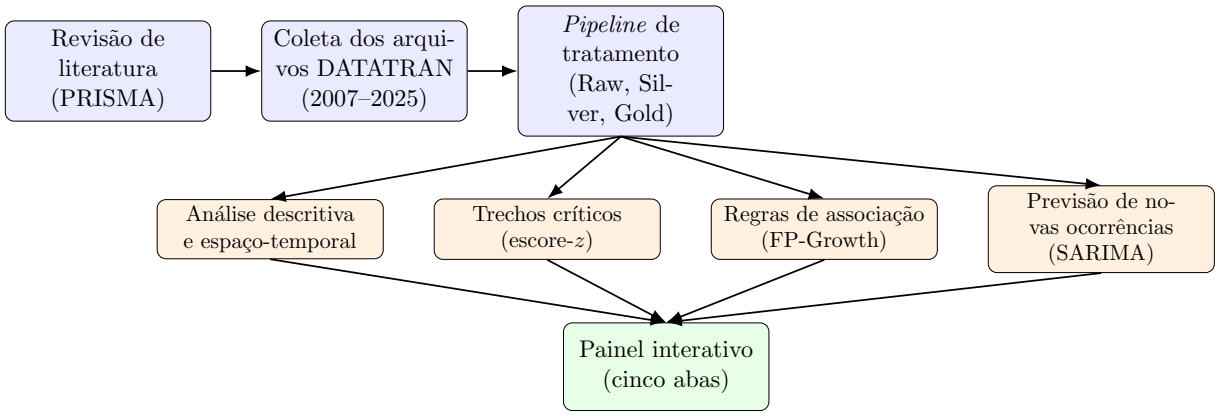
da mudança de nível introduzida pelo boletim eletrônico em 2018, fator que afeta a comparabilidade temporal e a confiabilidade das previsões. Em terceiro lugar, as tarefas de classificação, descrição e visualização são abordadas de forma isolada, raramente integradas a uma previsão temporal de novas ocorrências de acidentes.

Este trabalho posiciona-se exatamente nessas lacunas. É o único, dentre os analisados, a combinar, em escala nacional e sobre todo o período de 2007 a 2025, um *pipeline* de tratamento e georreferenciamento, a análise descritiva e espacial de padrões de severidade, a mineração de regras de associação com limiares estatisticamente fundamentados, a previsão de séries temporais com tratamento explícito da mudança de nível de 2018 e a consolidação de tudo em um painel interativo de apoio à decisão. É dessa amplitude de escopo, do rigor metodológico no tratamento da série e da integração em uma ferramenta utilizável que decorre o diferencial da presente pesquisa.

4 Metodologia

Este capítulo descreve os procedimentos metodológicos adotados para alcançar os objetivos da pesquisa. A abordagem é de natureza aplicada e quantitativa, organizada segundo as etapas do processo de Descoberta de Conhecimento em Bases de Dados. Apresentam-se as fontes de dados, a arquitetura do *pipeline* de tratamento, as técnicas de mineração empregadas e as ferramentas computacionais utilizadas. A Figura 4.1 sintetiza a visão geral da metodologia: a partir da revisão de literatura e da coleta dos arquivos da base DATATRAN, os dados percorrem o *pipeline* de tratamento em camadas, alimentam as quatro frentes de mineração e análise e, por fim, têm seus resultados consolidados nas abas do painel interativo.

Figura 4.1: Visão geral da metodologia: da revisão de literatura e coleta dos dados ao painel interativo.



Fonte: Elaborado pelo autor (2026).

4.1 Caracterização da Pesquisa

Adotando a classificação metodológica proposta por Lakatos e Marconi (2021), esta pesquisa pode ser caracterizada sob quatro aspectos. Quanto à natureza, classifica-se como aplicada, pois visa gerar conhecimento de uso prático para a segurança viária. Quanto à abordagem, é quantitativa, baseada na análise estatística e computacional de um grande

volume de registros. Quanto aos objetivos, é descritiva e exploratória. Quanto ao procedimento, adota como fio condutor o processo de KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), percorrendo as etapas de seleção, pré-processamento, transformação, mineração e interpretação dos dados.

4.2 Métodos de Pesquisa

A identificação e a seleção dos trabalhos relacionados, analisados no Capítulo 3, foram conduzidas por meio da revisão de literatura, seguindo as diretrizes de Kitchenham e Charters (2007) e o protocolo de reporte PRISMA (PAGE et al., 2021). As subseções a seguir descrevem o modelo PICO(T), as palavras-chave, as *strings* de busca, os critérios de inclusão e exclusão e o processo de seleção e avaliação que operacionalizam as questões de pesquisa definidas na Seção 1.2.

4.2.1 PICO(t)

Para estruturar a busca por trabalhos relacionados, em alinhamento com as questões de pesquisa definidas na Seção 1.2, foi adotado o modelo PICO(T), que auxilia na definição dos principais elementos do problema investigado:

P (População / Problema): registros públicos de acidentes de trânsito em rodovias federais brasileiras (bases da PRF/DATATRAN).

I (Intervenção): aplicação de técnicas de mineração de dados e de aprendizado de máquina para análise, classificação e predição de acidentes.

C (Comparação): abordagens estatísticas descritivas tradicionais ou ausência de tratamento e mineração sistemáticos dos dados.

O (Outcome / Desfecho): maior capacidade de identificar padrões, fatores de risco e pontos críticos, e de prever a ocorrência e a gravidade dos acidentes.

t (Tempo): estudos publicados entre 2014 e 2026.

Assim, a pergunta de pesquisa pode ser formulada da seguinte forma: *"Sobre registros públicos de acidentes em rodovias brasileiras (P), o uso de técnicas de mineração de dados e aprendizado de máquina (I), em comparação com a análise estatística descritiva*

tradicional (C), resulta em maior capacidade de identificar padrões e prever a gravidade dos acidentes (O)?"

4.2.2 Palavras-chave

Os principais termos que refletem os temas centrais do trabalho podem ser listados como: acidentes de trânsito, acidentes rodoviários, mineração de dados, aprendizado de máquina, rodovias federais, PRF, DATATRAN, e seus correspondentes em inglês *traffic accidents*, *road crashes*, *data mining*, *machine learning* e *highway*.

4.2.3 String de Busca

A partir da elaboração da pergunta de pesquisa e da escolha das palavras-chave, foram feitas buscas em três bibliotecas digitais (Scopus, IEEE *Xplore* e Google Acadêmico), escolhidas por sua abrangência e por capturarem tanto a produção internacional quanto a produção nacional sobre o tema, frequentemente publicada em eventos e periódicos brasileiros. Cabe destacar que a utilização integral de todos os termos resultou em restrição excessiva do escopo; por esse motivo, foi necessário um processo iterativo de refinamento das *strings*, visando equilibrar abrangência e relevância. Como cada base adota uma sintaxe própria, foram elaboradas variações específicas, mantendo-se as palavras-chave e as relações semânticas entre elas, conforme a Tabela 4.1.

Tabela 4.1: Strings de busca por biblioteca.

Biblioteca	**String de busca**
Scopus	`TITLE-ABS-KEY ( ("traffic accidents"OR "road crashes") AND ("data mining"OR "machine learning") AND ("highway"OR "Brazil") ) AND ( LIMIT-TO ( DOCTYPE , "ar") )`
IEEE *Xplore*	`(("All Metadata":"traffic accident") AND ("All Metadata":"machine learning"OR "All Metadata":"data mining") AND ("All Metadata":"highway"))`
Google Acadêmico	`("acidentes de trânsito"OR "acidentes rodoviários") AND ("mineração de dados"OR "aprendizado de máquina") AND (rodovias OR PRF OR DATATRAN)`

Fonte: Elaborado pelo autor (2026).

4.2.4 Critérios de Inclusão e Exclusão

Os seguintes critérios de inclusão (CI) e exclusão (CE) foram definidos para este mapeamento sistemático:

CI1: o estudo aplica técnicas de mineração de dados ou de aprendizado de máquina a dados de acidentes de trânsito.

CI2: o estudo tem como objeto rodovias e/ou dados públicos brasileiros (preferencialmente PRF/DATATRAN).

CI3: o estudo é primário e foi publicado entre 2014 e 2026.

CE1: o estudo não está disponível para leitura na íntegra.

CE2: o estudo não está escrito em português ou inglês.

CE3: o estudo não é primário (revisões, editoriais, pôsteres).

CE4: o estudo não está relacionado a acidentes de trânsito ou não emprega mineração de dados ou aprendizado de máquina.

4.2.5 Seleção e Avaliação

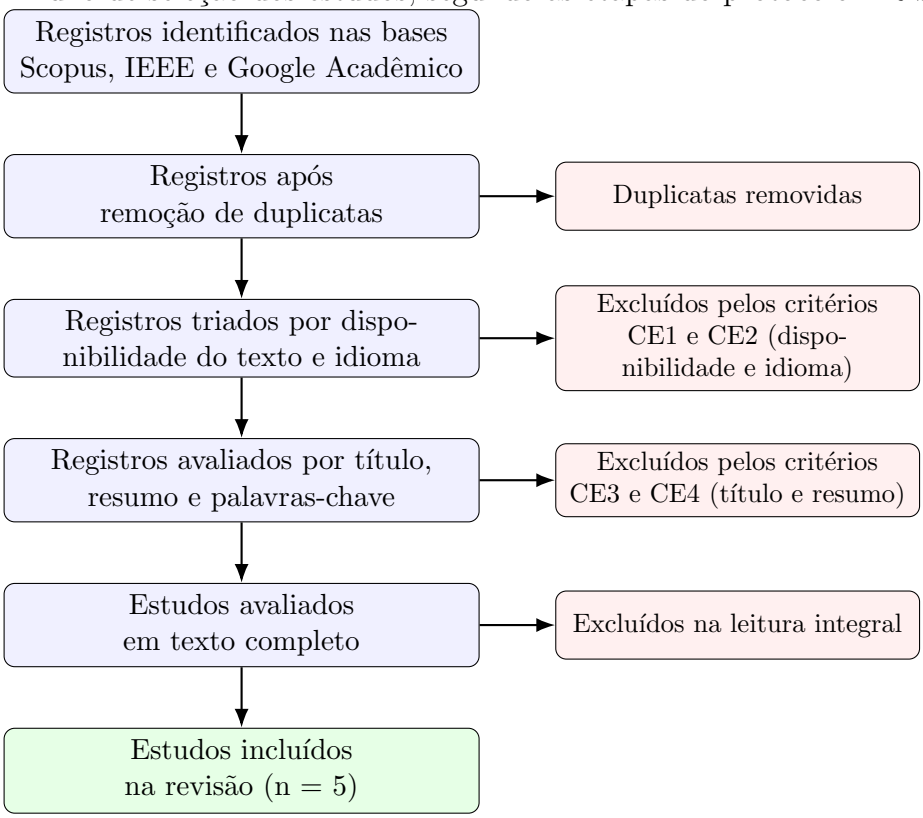
Para o processo de seleção e avaliação, foi adotada uma abordagem em quatro estágios, ilustrada na Figura 4.2. No primeiro estágio, os registros retornados pela aplicação das *strings* de busca nas bases foram reunidos e tiveram suas duplicatas removidas. No segundo estágio, realizou-se a triagem por disponibilidade do texto completo e idioma de publicação, aplicando-se os critérios de exclusão CE1 e CE2. No terceiro estágio, procedeu-se à leitura de título, resumo e palavras-chave, com base nos critérios CE3 e CE4. Por fim, no quarto estágio, os estudos remanescentes foram lidos na íntegra e avaliados quanto ao atendimento aos critérios de inclusão, resultando na seleção final de cinco estudos primários para análise.

4.3 Fonte de Dados

A fonte primária de dados é a base DATATRAN, disponibilizada pela Polícia Rodoviária Federal por meio de seu Portal de Dados Abertos[2] (Polícia Rodoviária Federal, 2025).

[2]Base DATATRAN, disponível em: ⟨https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf⟩. Acesso em: 25 jun. 2026.

Figura 4.2: Fluxo de seleção dos estudos, segundo as etapas do protocolo PRISMA 2020.



Fonte: Elaborado pelo autor (2026).

Foram coletados os arquivos anuais referentes ao período de 2007 a 2025, em formato CSV. Cada arquivo contém os registros de ocorrências atendidas pela PRF nas rodovias federais, com granularidade de um acidente por linha. Ao longo do período, o esquema dos arquivos sofreu alterações, notadamente a introdução das colunas de latitude e longitude nos anos mais recentes, exigindo um procedimento de leitura tolerante a variações de formato e de codificação de caracteres.

4.4 Arquitetura do *Pipeline* de Dados

O tratamento dos dados foi implementado em um *pipeline* estruturado segundo a arquitetura em camadas *medallion* (INMON, 2005), organizado em três estágios sucessivos: *Raw*, *Silver* e *Gold*, conforme ilustrado na Figura 4.3. A persistência intermediária em formato Parquet em cada camada garante reprodutibilidade e permite o reprocessamento seletivo de etapas.

Figura 4.3: Arquitetura em camadas do *pipeline* de dados (Raw, Silver e Gold).

Fonte: Elaborado pelo autor (2026).

4.4.1 Camada *Raw*

Na camada *Raw*, todos os arquivos anuais são lidos e unificados em uma única estrutura tabular, preservando-se os dados em seu estado original. Para acomodar a heterogeneidade dos arquivos, a leitura tenta múltiplas codificações de caracteres (*latin-1* e *utf-8*) e harmoniza esquemas distintos: para os arquivos antigos, o ano da ocorrência é inferido a partir do nome do arquivo e as colunas de georreferenciamento ausentes são criadas com valores nulos, assegurando a compatibilidade entre todos os anos.

4.4.2 Camada *Silver*: Limpeza e Normalização

A camada *Silver* concentra as rotinas de limpeza e padronização ao nível de registro. As principais operações realizadas são:

- **Remoção de duplicatas:** eliminação de registros repetidos a partir do identificador único da ocorrência;
- **Normalização de datas:** conversão dos múltiplos formatos de data presentes nos arquivos para um padrão único, com extração de atributos de ano e mês;
- **Padronização de variáveis categóricas:** uniformização de maiúsculas e minúsculas, remoção de espaços supérfluos e consolidação de categorias semanticamente equivalentes por meio de dicionários de mapeamento (por exemplo, agrupando diferentes descrições de ingestão de álcool e psicoativos em uma única categoria);
- **Tratamento de coordenadas geográficas:** conversão de vírgula decimal para ponto, correção de sinais invertidos de latitude e longitude, descarte de coordenadas situadas fora do retângulo envolvente do território brasileiro e imputação dos valores

ausentes pela mediana das coordenadas válidas agrupadas por unidade federativa, rodovia e quilômetro;

- **Normalização de contagens:** conversão das colunas de número de pessoas, mortos, feridos e veículos para tipos inteiros, com tratamento de valores nulos.

A imputação espacial baseada na mediana por trecho merece destaque por sua importância na análise cartográfica: como descrito no Capítulo 5, essa estratégia elevou substancialmente a cobertura de georreferenciamento da base.

4.4.3 Engenharia de Atributos Temporais

Ainda na camada *Silver*, novos atributos foram derivados para enriquecer as análises temporais. A partir da data, calculou-se o dia da semana e um indicador binário de fim de semana. A partir do horário, definiu-se o turno da ocorrência em quatro categorias: madrugada, manhã, tarde e noite, possibilitando a análise dos padrões de risco ao longo do ciclo diário.

4.4.4 Camada *Gold*: Agregação Analítica

A camada *Gold* consolida os dados limpos em uma tabela agregada, orientada ao consumo analítico. Os registros são agrupados por combinações de dimensões temporais, geográficas e descritivas (data, unidade federativa, rodovia, causa, tipo, classificação, turno e condição meteorológica), e para cada grupo são computadas métricas como a quantidade de acidentes, o número de acidentes com mortos e feridos e os totais de vítimas. São ainda derivados indicadores de severidade, como o percentual de acidentes com vítimas fatais e a média de mortos por acidente, que fundamentam a identificação dos trechos críticos.

4.5 Técnicas de Mineração e Análise Aplicadas

4.5.1 Análise Descritiva de Padrões

A caracterização do panorama de acidentalidade foi conduzida por meio de análise exploratória sobre a base tratada, quantificando a distribuição das causas e tipos de acidente,

a distribuição da severidade e os padrões temporais por turno, dia da semana e sazonalidade. A análise da severidade adotou como métrica principal a média de mortos por acidente que, por ser uma razão, é robusta à mudança de nível da série e permite comparar o risco entre diferentes categorias.

4.5.2 Detecção de Trechos Críticos

A identificação de trechos com concentração anômala de acidentes graves baseou-se na padronização pelo escore-z (TAN; STEINBACH; KUMAR, 2009). Para cada trecho (definido pela combinação de unidade federativa e rodovia), foram computados o total de óbitos e a taxa de mortalidade, e seus respectivos escores-z foram combinados em um indicador de anomalia. Optou-se por métricas baseadas em óbitos, e não no volume bruto de acidentes, justamente por serem menos sensíveis à mudança no procedimento de registro ocorrida ao longo do período.

4.5.3 Regras de Associação para Fatores de Risco

Para descobrir combinações de condições associadas a um desfecho fatal, aplicou-se a mineração de regras de associação (HAN; KAMBER; PEI, 2011). Cada acidente foi modelado como uma transação (uma cesta de itens) composta por seus atributos categóricos (causa, tipo, fase do dia, condição meteorológica e tipo de pista, entre outros), e buscaram-se regras da forma $\{\text{combinação de fatores}\} \rightarrow \text{Fatal}$ por meio do algoritmo FP-Growth. Cada regra foi avaliada por três métricas: a confiança, que estima a probabilidade de óbito dada a combinação; o *lift*, que indica quantas vezes a combinação é mais letal que a taxa média do recorte; e o suporte, que mede a frequência relativa da combinação.

Um cuidado central foi evitar limiares arbitrários. O suporte mínimo foi determinado pelo tamanho da amostra, exigindo que cada combinação ocorresse em um número absoluto mínimo de acidentes para ser estatisticamente estável. A seleção das regras não se baseou em um *lift* fixo, mas em significância estatística: cada regra foi submetida a um teste z unilateral de proporção, verificando se a probabilidade de óbito sob a combinação supera a taxa-base, com correção para múltiplas comparações pelo procedimento de Benjamini-Hochberg para controle da taxa de falsas descobertas (FDR) (BENJAMINI;

HOCHBERG, 1995). Esse procedimento equivale a um *lift* mínimo determinado pelos próprios dados. Como a taxa-base de fatalidade varia com o nível da série, a mineração é conduzida dentro de um recorte homogêneo (a partir de 2018), por meio do filtro de período do painel.

4.5.4 Modelo Preditivo de Série Temporal

Para a previsão da quantidade mensal de acidentes, adotou-se como modelo principal o SARIMA (*Seasonal AutoRegressive Integrated Moving Average*) (BOX et al., 2015), escolhido por seu desempenho superior em uma série de forte componente sazonal. Adotou-se a especificação $(1,1,1)(0,1,1)_{12}$, que corresponde ao clássico modelo do tipo *airline*: uma parte não sazonal com um termo autorregressivo, uma diferenciação e um termo de média móvel, combinada a uma parte sazonal com uma diferenciação e um termo de média móvel no período de doze meses. Essa estrutura é amplamente empregada para séries mensais de sazonalidade anual; seus coeficientes são estimados por máxima verossimilhança, e o modelo fornece nativamente o intervalo de confiança de 95% da previsão. O horizonte de previsão adotado foi de doze meses.

Como referência de desempenho, o SARIMA foi confrontado com três alternativas: um *Perceptron* Multicamadas (MLP) (HAYKIN, 2009), a suavização exponencial de Holt-Winters e uma linha de base sazonal ingênua, na qual cada mês é estimado pelo valor do mesmo mês do ano anterior. No MLP, a série foi transformada em um problema de aprendizado supervisionado por meio de uma janela deslizante de atrasos (*lags* de 1, 2, 3, 6 e 12 meses) e da posição do mês no ciclo anual, codificada por funções seno e cosseno, com a rede em duas camadas ocultas (64 e 32 neurônios), ativação ReLU e otimizador Adam; o Holt-Winters empregou tendência e sazonalidade aditivas com período anual.

Um cuidado metodológico central foi a definição da janela de treino. A série apresenta um nível de longo prazo decrescente (do pico de 192 mil ocorrências em 2011 a cerca de 69 mil em 2018) e, a partir de 2018, com a adoção do boletim eletrônico, estabiliza-se em um patamar homogêneo (entre 63 mil e 73 mil ocorrências anuais). Por essa razão, o treinamento foi restrito ao período a partir de janeiro de 2018, que reúne 96 observações mensais homogêneas e evita a tendência de nível dos anos anteriores.

A avaliação seguiu uma validação retrospectiva (*backtest*), reservando-se os últimos doze meses observados para teste, com o desempenho medido pelas métricas de Erro Médio Absoluto (MAE), Raiz do Erro Quadrático Médio (RMSE) e Erro Percentual Absoluto Médio (MAPE). A incerteza da projeção futura é expressa pelo intervalo de confiança de 95% nativo do SARIMA, decorrente diretamente da estimação do modelo, sem necessidade de aproximações heurísticas.

A construção do modelo seguiu a metodologia de Box-Jenkins. A ordem de diferenciação foi fundamentada em testes formais de estacionariedade, o ADF e o KPSS, aplicados à série em nível e às suas diferenças, e a especificação *airline* adotada foi confrontada com ordens alternativas por meio dos critérios de informação AIC e BIC. A adequação do modelo final foi submetida a uma bateria de diagnósticos sobre os resíduos: o teste de Ljung-Box, para a autocorrelação remanescente; o teste de Jarque-Bera (JARQUE; BERA, 1980), para a normalidade; e o teste ARCH-LM (ENGLE, 1982), para a presença de heterocedasticidade (variância condicional não constante).

Para confirmar de forma robusta a escolha do modelo principal, três mecanismos de validação complementares foram incorporados. Primeiro, o SARIMA foi comparado, sobre a mesma janela retida, à linha de base sazonal ingênua (cada mês igual ao mesmo mês do ano anterior) e aos demais modelos de série temporal avaliados, o *Perceptron* Multicamadas e a suavização exponencial de Holt-Winters (BOX et al., 2015). Segundo, a adequação do modelo final foi avaliada pela bateria de diagnósticos de resíduos descrita anteriormente (Ljung-Box, Jarque-Bera e ARCH-LM), tendo como referência a hipótese de que os resíduos constituem ruído branco; a não rejeição dessa hipótese indica que o modelo capturou a estrutura da série. Terceiro, para além do corte único do *backtest*, aplicou-se uma validação por origem móvel (*walk-forward*) com janela expansível: em cada origem, o modelo é retreinado apenas com os dados até aquele ponto e prevê o horizonte seguinte, gerando uma distribuição de erro (média e desvio-padrão) mais conservadora e informativa que um único corte.

4.5.5 Parâmetros dos Algoritmos e Sua Obtenção

Os algoritmos empregados neste trabalho requerem parâmetros de configuração, sintetizados na Tabela 4.2. Os procedimentos puramente descritivos, como a análise de distribuições e a detecção de trechos críticos por escore-z, não envolvem hiperparâmetros ajustáveis e, por isso, não constam da tabela. A estratégia de obtenção dos valores variou conforme a natureza de cada algoritmo.

Tabela 4.2: Parâmetros utilizados em cada algoritmo.

Algoritmo	**Parâmetros**
MLP (*Perceptron* Multicamadas, comparação)	camadas ocultas = (64, 32); ativação = ReLU; otimizador = Adam; iterações máximas = 2000; atrasos (*lags*) = {1, 2, 3, 6, 12} meses; padronização das variáveis (*z-score*); semente = 42
SARIMA (principal)	ordem $(p, d, q) = (1, 1, 1)$; ordem sazonal $(P, D, Q)_s = (0, 1, 1)_{12}$
Holt-Winters (comparação)	tendência aditiva; sazonalidade aditiva; período sazonal = 12
FP-Growth (regras de associação)	suporte mínimo = máx(25 casos; 0,1% da amostra); itens por regra ≤ 4; pré-filtro de *lift* ≥ 1; seleção por teste z de proporção com FDR ($\alpha = 0{,}05$); 15 categorias mais frequentes por atributo

Fonte: Elaborado pelo autor (2026).

A escolha do SARIMA como modelo principal foi confirmada empiricamente: na validação retrospectiva, a configuração $(1, 1, 1)(0, 1, 1)_{12}$ obteve o menor erro entre todos os modelos avaliados (RMSE de 109,2 e MAPE de 1,6%), superando o *Perceptron* Multicamadas, a suavização de Holt-Winters e a linha de base sazonal. A estrutura adotada corresponde ao clássico modelo do tipo *airline*, adequada a séries com forte sazonalidade anual, e seus coeficientes são estimados por máxima verossimilhança. Os parâmetros dos modelos de comparação seguem escolhas convencionais: no MLP, a arquitetura de duas camadas (64 e 32 neurônios), a ativação ReLU e o otimizador Adam constituem uma configuração compacta e robusta, e os atrasos são selecionados de forma adaptativa em função do comprimento da série; no Holt-Winters, adotam-se tendência e sazonalidade aditivas com período anual. A semente aleatória fixa garante a reprodutibilidade.

Para a mineração de regras de associação, optou-se deliberadamente por não ajustar limiares de forma manual. O suporte mínimo é derivado do tamanho da amostra

e a seleção das regras baseia-se em significância estatística (teste z de proporção com correção FDR de Benjamini-Hochberg), o que equivale a um *lift* mínimo determinado pelos próprios dados. Essa abordagem evita cortes arbitrários e controla a taxa de falsas descobertas, tornando os parâmetros uma consequência da base analisada, e não de uma escolha subjetiva.

Por fim, a adequação do conjunto de parâmetros escolhido é corroborada *a posteriori* pela bateria de validação apresentada no Capítulo 5: o SARIMA superou os modelos de comparação e a linha de base sazonal, e seus resíduos comportam-se como ruído branco no teste de Ljung-Box, evidência de que a configuração captura a estrutura da série.

4.6 Tecnologias Usadas

4.6.1 Visualização e Painéis Interativos

Os resultados foram materializados em um painel interativo desenvolvido com a biblioteca *Streamlit*, organizado em abas temáticas que contemplam o mapa de intensidade por rodovia, a série temporal com a previsão, os padrões temporais, a análise de causas e fatores e o ranking de trechos críticos. Filtros globais de período, unidade federativa e rodovia permitem ao usuário restringir interativamente o escopo da análise, tornando as visualizações exploráveis e adequadas ao apoio à decisão.

4.6.2 Ferramentas Computacionais

O trabalho foi desenvolvido na linguagem Python. A manipulação e agregação dos dados utilizaram a biblioteca *pandas* (MCKINNEY, 2010); o modelo de rede neural e as rotinas de pré-processamento basearam-se na biblioteca *scikit-learn* (PEDREGOSA et al., 2011); a mineração de regras de associação (FP-Growth) empregou a biblioteca *mlxtend* (RASCHKA, 2018); os modelos estatísticos de série temporal e os testes de diagnóstico apoiaram-se na biblioteca *statsmodels* (SEABOLD; PERKTOLD, 2010); as visualizações foram construídas com *Plotly*; e a interface do painel, com *Streamlit*. A persistência das camadas de dados adotou o formato colunar Parquet, eficiente para leitura analítica.

5 Resultados e Discussão

Este capítulo apresenta e discute os resultados obtidos, distinguindo três tipos de contribuição que se complementam. A primeira é a **análise descritiva**, em que gráficos e medidas-resumo convencionais caracterizam o panorama da acidentalidade, como o volume de registros, as causas, os tipos, a severidade e os padrões temporais, sem o emprego de algoritmos de mineração. A segunda reúne os **resultados de mineração de dados**, com os modelos aplicados e as visualizações deles derivadas: as regras de associação, a detecção de trechos críticos e o modelo preditivo de série temporal. A terceira é o **painel interativo**, que integra os resultados anteriores em uma única ferramenta exploratória. Antes dessas três frentes, reporta-se o resultado do tratamento dos dados, que produz a base sobre a qual todas as análises se apoiam; ao final, a síntese articula os achados das três contribuições.

5.1 Resultado do Tratamento dos Dados

A inspeção dos arquivos originais confirmou o conjunto de inconsistências que comprometia uma visão histórica e geral confiável dos acidentes: registros duplicados pelo identificador da ocorrência, valores omissos em campos críticos (como o tipo de pista, o traçado da via e as condições meteorológicas), categorias textuais não padronizadas, erros nas coordenadas de latitude e longitude (vírgula decimal, sinais invertidos e pontos fora do território nacional) e a heterogeneidade do esquema ao longo dos anos, com os arquivos mais antigos sem campos de georreferenciamento. Essas inconsistências foram superadas pelo *pipeline* de Extração, Transformação e Carga organizado em camadas, que removeu duplicatas, normalizou as categorias, corrigiu e imputou as coordenadas e harmonizou os esquemas.

A execução do *pipeline* sobre os arquivos anuais de 2007 a 2025 resultou em uma base *Silver* consolidada com **2.194.832** registros de acidentes, após a remoção de duplicatas e a padronização descrita no Capítulo 4. Os registros distribuem-se por todas

as 27 unidades federativas, além de uma fração residual com unidade não informada, e por 213 rodovias federais distintas. No conjunto do período, contabilizam-se **127.704** óbitos e **1.658.724** feridos.

Destaca-se o efeito da imputação espacial: ao final do tratamento, **99,3%** dos registros dispunham de coordenadas geográficas válidas, ante uma cobertura original muito inferior, sobretudo nos arquivos antigos, que sequer possuíam campos de georreferenciamento. Esse ganho foi decisivo para viabilizar as análises cartográficas. A camada *Gold* resultante, voltada ao consumo analítico, totalizou 2.137.877 linhas agregadas.

5.2 Análise Descritiva

Esta primeira parte do capítulo caracteriza o panorama da acidentalidade por meio de análise descritiva e exploratória, sem o emprego de algoritmos de mineração. A partir de medidas-resumo, proporções e gráficos, descrevem-se a evolução do volume de registros, as principais causas e tipos de acidente, os perfis de severidade e os padrões temporais, estabelecendo o pano de fundo sobre o qual as técnicas de mineração serão posteriormente aplicadas.

5.2.1 Evolução do Volume de Registros

A análise do volume anual de registros, evidenciada na Figura 5.1, revela uma trajetória de declínio real seguida de estabilização. Como mostra a Tabela 5.1, o número de acidentes cresce até o início da década de 2010, atinge o máximo de 192.322 ocorrências em 2011 e, a partir de 2015, recua de forma consistente, de 122.158 ocorrências em 2015 para 89.567 em 2017. Entre 2017 e 2018, com a adoção do boletim eletrônico, observa-se uma redução de nível adicional (para 69.333 registros), de magnitude moderada (cerca de 22%), a partir da qual a série se estabiliza em um patamar homogêneo, oscilando entre 63 mil e 73 mil ocorrências anuais até 2025.

A Figura 5.1 torna esse comportamento visualmente claro: após o declínio real iniciado em 2011, a série apresenta uma redução de nível moderada em 2018, associada à mudança no procedimento de registro, e estabiliza-se a seguir. A diferença de nível entre

Tabela 5.1: Quantidade anual de acidentes registrados (DATATRAN, 2007–2025).

Ano	Acidentes	Ano	Acidentes
2007	127.671	2017	89.567
2008	141.038	2018	69.333
2009	158.646	2019	67.558
2010	183.465	2020	63.585
2011	192.322	2021	64.567
2012	184.561	2022	64.606
2013	186.745	2023	67.766
2014	169.197	2024	73.156
2015	122.158	2025	72.529
2016	96.362		

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

Figura 5.1: Série mensal de acidentes registrados (2007 a 2025). Há um declínio real a partir de 2011, uma redução de nível moderada com a adoção do boletim eletrônico em 2018 e estabilização posterior.

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

os anos exige cautela na comparação de contagens absolutas, mas as análises descritivas de proporções e de severidade (que são razões) permanecem válidas para a totalidade do período, ao passo que a modelagem preditiva utiliza o recorte homogêneo a partir de 2018.

5.2.2 Caracterização das Causas e Tipos de Acidente

A distribuição das causas confirma a predominância do fator humano na gênese dos acidentes. Como apresentado na Tabela 5.2, a *falta de atenção* responde sozinha por 28,0% das ocorrências, seguida por causas igualmente associadas ao comportamento do condutor, como a velocidade incompatível (8,5%) e o não guardar distância de segurança (8,2%). Quanto aos tipos, predominam a colisão traseira (25,2%), a saída de pista (15,2%) e a colisão lateral (13,2%).

Além das frequências marginais, a relação entre causa e tipo de acidente revela associações fortes, sintetizadas no mapa de calor da Figura 5.2, em que cada célula ex-

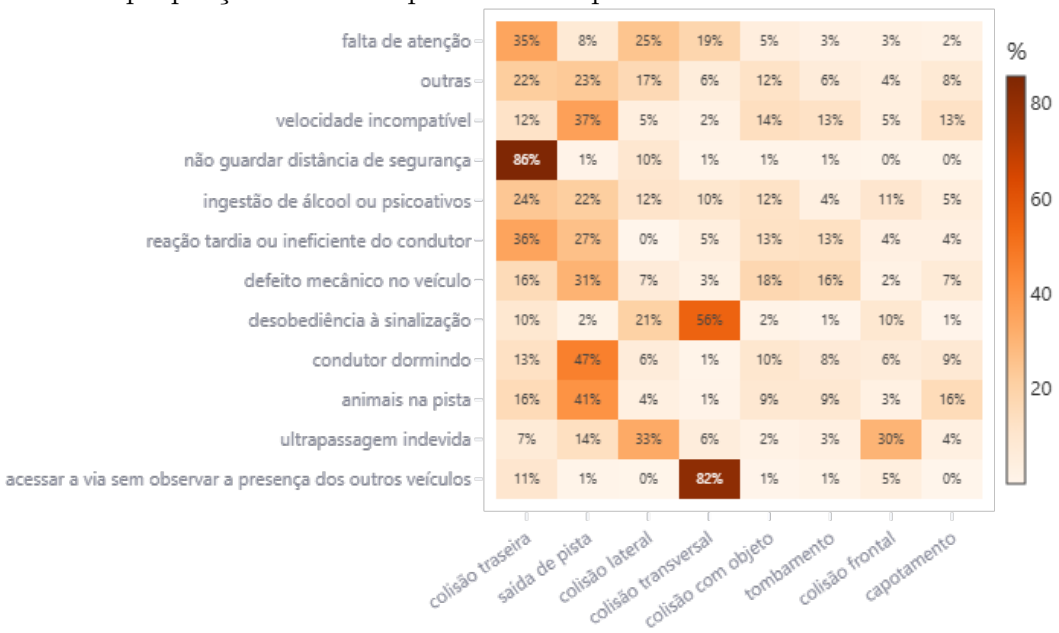
Tabela 5.2: Principais causas e tipos de acidente (participação percentual).

Causa	%	**Tipo**	%
Falta de atenção	28,0	Colisão traseira	25,2
Velocidade incompatível	8,5	Saída de pista	15,2
Não guardar distância	8,2	Colisão lateral	13,2
Ingestão de álcool/psicoat.	4,7	Colisão transversal	10,7
Reação tardia do condutor	4,2	Colisão com objeto	7,2
Defeito mecânico no veículo	3,7	Tombamento	5,3
Desobediência à sinalização	2,9	Colisão frontal	4,6

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

pressa a proporção de um tipo dentro de uma causa. Algumas combinações são quase determinísticas: o ato de não guardar distância de segurança resulta em colisão traseira em 85% dos casos, a desobediência à sinalização leva a colisão transversal em 56%, e a ultrapassagem indevida culmina em colisão frontal em 26%. Já causas como velocidade incompatível (37%), condutor dormindo (46%) e a presença de animais na pista (42%) convergem majoritariamente para a saída de pista. Esse mapeamento conecta diretamente o comportamento do condutor ao desfecho físico do sinistro.

Figura 5.2: Associação entre causa e tipo de acidente. Cada linha (causa) soma 100%, e a cor indica a proporção de cada tipo dentro daquela causa.



Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

5.2.3 Análise de Severidade

Quanto à classificação, 47,3% dos acidentes ocorrem sem vítimas, 47,8% com vítimas feridas e 4,9% com vítimas fatais. Contudo, a frequência de um tipo de acidente não se confunde com sua letalidade. A Tabela 5.3 ordena os tipos pela média de mortos por acidente e revela um contraste fundamental: a **colisão frontal**, embora represente apenas 4,6% das ocorrências, é de longe a mais letal, com 0,391 morto por acidente, muito acima da letalidade dos demais tipos. O atropelamento de pessoa surge como o segundo tipo mais letal (0,288).

Tabela 5.3: Letalidade por tipo de acidente (média de mortos por acidente).

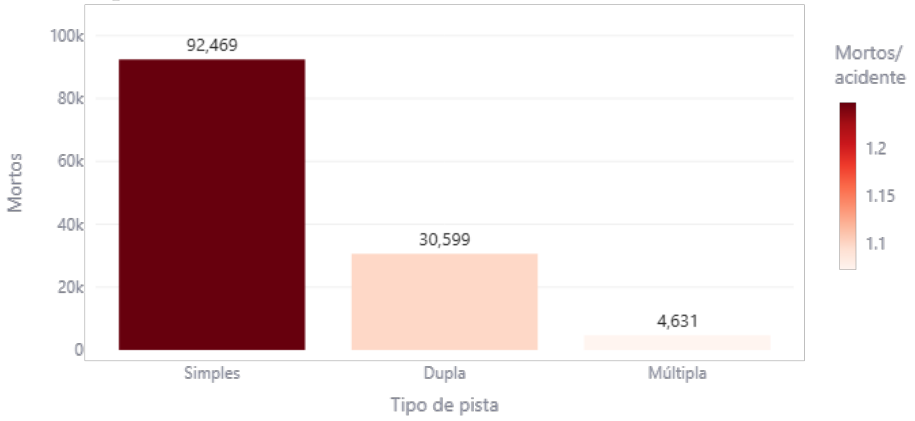
Tipo de acidente	**Mortos / acidente**
Colisão frontal	0,391
Atropelamento de pessoa	0,288
Capotamento	0,049
Colisão transversal	0,047
Saída de pista	0,042
Queda do veículo	0,036

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

Padrão análogo observa-se nas causas: a *ultrapassagem indevida*, ainda que pouco frequente, apresenta a maior letalidade (0,197 morto por acidente), seguida do condutor dormindo (0,083) e da velocidade incompatível (0,078). Esse achado é coerente com a literatura de segurança viária, que associa a maior severidade aos eventos de alta transferência de energia, como as colisões frontais decorrentes de manobras de ultrapassagem.

A severidade também se relaciona com a infraestrutura da via. Como mostra a Figura 5.3, as pistas simples concentram a esmagadora maioria dos óbitos (92.474) e, simultaneamente, a maior taxa de mortos por acidente, comportamento coerente com a letalidade das colisões frontais em vias de fluxo não separado. Já as pistas duplas (30.599 óbitos) e múltiplas (4.631), que separam fisicamente os sentidos de tráfego, registram letalidade substancialmente menor, o que reforça o papel da separação de fluxos como medida de engenharia para a redução de mortes.

Figura 5.3: Óbitos por tipo de pista. A altura da barra indica o total de mortos e a cor, a taxa de mortos por acidente.



Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

5.2.4 Padrões Temporais

O estudo dos padrões temporais, de natureza descritiva e exploratória, foi escolhido por responder, em conjunto com a investigação espacial dos trechos críticos, à terceira questão de pesquisa (Seção 1.2): identificar onde e quando os acidentes mais graves se concentram. Enquanto a detecção de trechos críticos por escore de anomalia localiza o risco no espaço, a caracterização por turno, dia da semana e sazonalidade o localiza no tempo, e ambas complementam a mineração de regras de associação, que revela com quais fatores o risco se combina (QP2). A opção por uma abordagem descritiva, e não preditiva, nesta etapa justifica-se pelo objetivo de caracterizar e tornar visíveis os padrões de risco para gestores e sociedade, e não de classificar ocorrências individuais. A métrica central adotada, a média de mortos por acidente, por ser uma razão, permite comparar a letalidade entre períodos independentemente das diferenças de volume, evidenciando a dissociação entre frequência e gravidade que percorre toda a análise.

Essa dissociação manifesta-se já na distribuição por turno. O maior número de acidentes ocorre à tarde (32,4%) e pela manhã (28,6%), períodos de maior fluxo de tráfego. Entretanto, a **madrugada**, responsável por apenas 11,4% das ocorrências, é o período mais letal, com 0,099 morto por acidente, cerca de 2,3 vezes a letalidade do período da tarde (0,043), conforme a Tabela 5.4. Esse comportamento é consistente com fatores de risco típicos do período noturno, como a menor visibilidade, a maior velocidade praticada

em vias menos congestionadas e a maior incidência de fadiga e de consumo de álcool.

Tabela 5.4: Frequência e letalidade por turno.

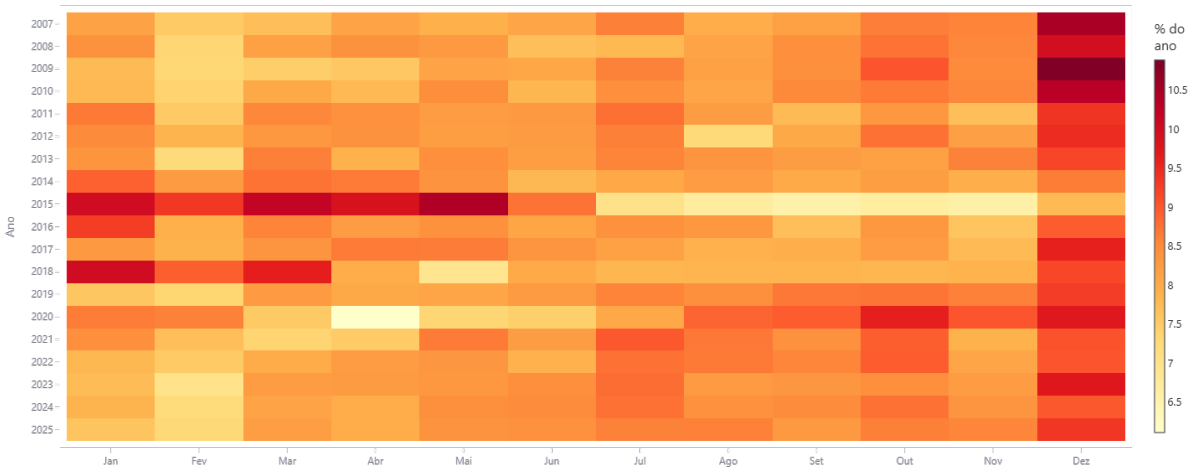
Turno	**% dos acidentes**	**Mortos / acidente**
Madrugada	11,4	0,099
Noite	27,6	0,076
Tarde	32,4	0,043
Manhã	28,6	0,041

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

O mesmo padrão manifesta-se na comparação entre fins de semana e dias úteis: embora os fins de semana concentrem 31,1% das ocorrências, sua letalidade (0,075 morto por acidente) é cerca de 47% superior à dos dias úteis (0,051), reforçando a hipótese de que os deslocamentos de lazer e o consumo de álcool elevam a gravidade dos sinistros nesses períodos.

A dimensão sazonal completa a análise temporal. A Figura 5.4 apresenta o índice sazonal mensal, no qual cada ano é normalizado pela sua própria média (valor 100), recurso que permite comparar a forma da sazonalidade apesar da diferença de nível entre os anos. As curvas dos diversos anos sobrepõem-se em formato, o que confirma que a sazonalidade é uma característica estrutural da série, e não um artefato do volume de registros: o perfil mensal dos acidentes mantém-se essencialmente o mesmo ao longo de todo o período.

Figura 5.4: Índice sazonal mensal por ano. Cada ano é normalizado por sua média (100), tornando a forma sazonal comparável entre anos de nível distinto.



Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

5.3 Resultados de Mineração de Dados

Esta segunda parte do capítulo apresenta os resultados obtidos com a aplicação de técnicas de mineração de dados sobre a base tratada, indo além da descrição para extrair padrões não evidentes por simples inspeção. Reúnem-se aqui a mineração de regras de associação, que isola as combinações de fatores mais letais, a detecção de trechos críticos por escore de anomalia e o modelo preditivo de série temporal, cada qual respondendo a uma das questões de pesquisa. Para cada técnica, apresentam-se também as visualizações dela derivadas, que tornam os padrões minerados interpretáveis.

5.3.1 Regras de Associação: Combinações de Fatores Letais

A escolha da mineração de regras de associação para esta etapa responde diretamente à segunda questão de pesquisa (Seção 1.2), que indaga quais fatores de risco e perfis de severidade podem ser revelados a partir da base. Diferentemente da classificação, que atribui um rótulo a cada ocorrência por meio de um modelo frequentemente opaco, a análise de associação produz regras explícitas e interpretáveis da forma {combinação de fatores} $\rightarrow$ Fatal, acompanhadas de métricas que quantificam a força de cada padrão. Essa característica é a mais aderente ao objetivo descritivo do trabalho: não se busca prever a gravidade de um acidente individual, mas evidenciar quais combinações de condições concentram desproporcionalmente os desfechos letais, conhecimento diretamente acionável para campanhas e fiscalização. Acrescente-se que a técnica lida nativamente com os atributos categóricos da base e com a interação entre múltiplos fatores, superando a análise por fator isolado ao expor efeitos conjuntos que passariam despercebidos quando cada variável é examinada separadamente.

A análise de severidade por fator isolado foi aprofundada pela mineração de regras de associação, que identifica combinações de condições associadas a um desfecho fatal. A Tabela 5.5 apresenta as combinações de maior *lift* obtidas no recorte homogêneo a partir de 2018, em que a taxa-base de fatalidade é de 7,04%. Nesse período, 145 combinações mostraram-se significativamente mais letais que a média, segundo o teste de proporção com correção FDR.

Os resultados são coerentes com a análise de severidade e a reforçam quanti-

Tabela 5.5: Combinações de fatores mais letais (regras de associação, recorte de 2018 em diante).

Combinação de fatores	**Confiança**	**Lift**
Transitar na contramão + pista simples + colisão frontal	40,2%	5,81
Ultrapassagem indevida + céu claro + colisão frontal	37,8%	5,46
Ultrapassagem indevida + plena noite + colisão frontal	37,3%	5,38
Transitar na contramão + céu claro + colisão frontal	37,1%	5,36
Transitar na contramão + plena noite + colisão frontal	37,0%	5,35
Plena noite + pista dupla + atropelamento de pessoa	36,9%	5,33

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

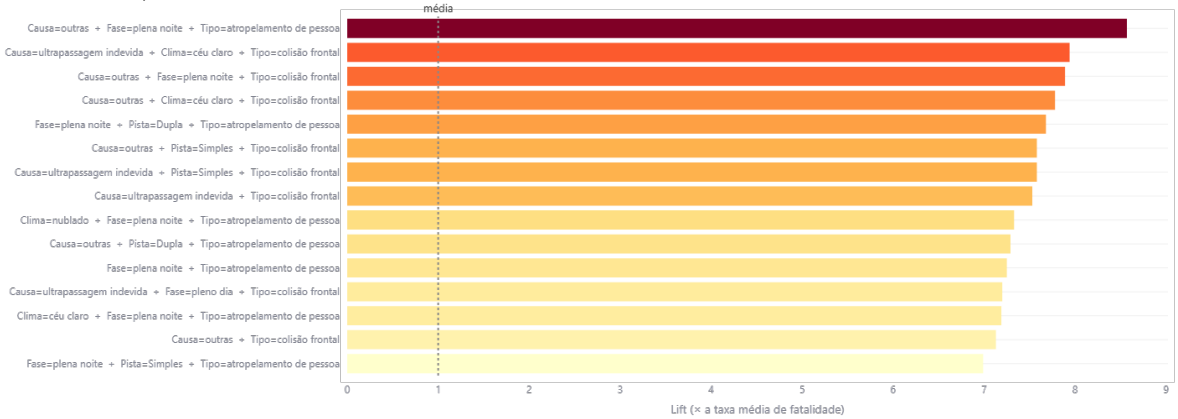
tativamente. Duas famílias de combinações concentram o risco: as **colisões frontais por ultrapassagem indevida ou por transitar na contramão**, cuja probabilidade de óbito chega a cerca de 40% (quase seis vezes a taxa-base), e os **atropelamentos de pessoa em plena noite**. O elevado *lift* dessas regras confirma que não se trata de eventos frequentes, mas de configurações desproporcionalmente letais, exatamente o tipo de conhecimento acionável para campanhas e fiscalização direcionadas. A mesma estrutura de combinações ressurge quando a mineração é repetida sobre o período anterior a 2018, o que indica que esses padrões de letalidade são estáveis ao longo do tempo, ainda que os valores de *lift* variem com a taxa-base de cada recorte.

A Figura 5.5 ilustra a saída da mineração para um recorte da base, com as combinações ordenadas pelo *lift* em relação à taxa média de fatalidade (linha tracejada em 1). Reaparecem as famílias dominantes de risco, as colisões frontais (por ultrapassagem indevida ou por transitar na contramão) e os atropelamentos de pessoa em plena noite, cujos valores de *lift* variam conforme a taxa-base do recorte, mas cuja posição no topo do *ranking* se mantém.

5.3.2 Padrões Espaciais e Trechos Críticos

A distribuição geográfica dos óbitos evidencia forte concentração em poucos estados e rodovias. Minas Gerais lidera em número absoluto de mortos (18.087), seguida pela Bahia

Figura 5.5: Combinações de fatores mais letais ordenadas por *lift* (múltiplo da taxa média de fatalidade). A linha tracejada marca o *lift* unitário, correspondente à média do recorte.



Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

(12.018) e pelo Paraná (11.298). No nível das rodovias, a **BR-101** (17.257 óbitos) e a **BR-116** (17.224 óbitos) destacam-se isoladamente das demais, como mostra a Tabela 5.6. Essas duas rodovias longitudinais, que cortam o litoral e o interior do país conectando importantes polos econômicos, concentram intenso tráfego de longa distância e respondem, juntas, por parcela expressiva da mortalidade.

Tabela 5.6: Estados e rodovias com maior número absoluto de óbitos (2007–2025).

UF	**Óbitos**	**Rodovia**	**Óbitos**
MG	18.087	BR-101	17.257
BA	12.018	BR-116	17.224
PR	11.298	BR-153	5.935
SC	8.862	BR-381	5.020
RJ	7.765	BR-040	4.785
RS	7.147	BR-316	4.707

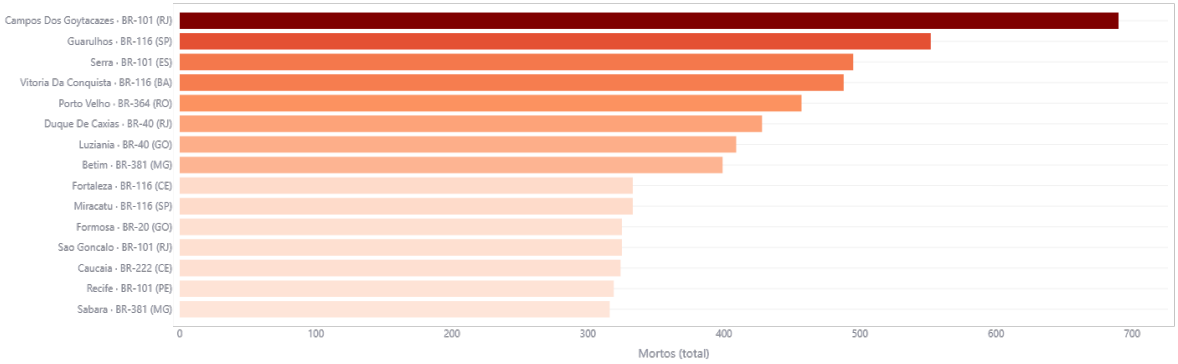
Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

A detecção de trechos críticos pelo escore-z combinado de óbitos e taxa de mortalidade, disponibilizada interativamente no painel (Seção 5.4.5), permite ir além da contagem absoluta e identificar segmentos que se destacam estatisticamente do padrão geral, ainda que não figurem entre os de maior volume, subsídio direto à priorização de fiscalização e de intervenções de engenharia.

Descendo da escala da rodovia para a do trecho municipal, a Figura 5.6 apresenta o *ranking* dos segmentos mais críticos por total de óbitos. Lidera, com folga, o trecho da BR-101 em Campos dos Goytacazes (RJ), seguido pela BR-116 em Guarulhos (SP) e pela

BR-101 em Serra (ES). O predomínio das BR-101 e BR-116 no topo do *ranking* confirma, em escala local, a concentração de mortalidade já observada no nível das rodovias e oferece alvos geográficos concretos para a atuação dos órgãos de fiscalização.

Figura 5.6: Trechos rodoviários mais críticos por total de óbitos, identificados por município, rodovia e unidade federativa.



Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

5.3.3 Avaliação do Modelo Preditivo

O modelo SARIMA foi treinado e avaliado sobre a série mensal nacional do recorte homogêneo, compreendendo 96 meses (janeiro de 2018 a dezembro de 2025). Na validação retrospectiva sobre os últimos doze meses, o SARIMA alcançou um MAPE de **1,6%**, o menor entre todas as alternativas avaliadas, conforme a Tabela 5.7. O modelo superou a linha de base sazonal ingênua e, com folga, os demais modelos de série temporal (o *Perceptron* Multicamadas e o Holt-Winters), reduzindo o RMSE em **18,8%** frente ao segundo colocado (a linha de base sazonal). Esse resultado é coerente com o forte componente sazonal da série, que o SARIMA modela de forma explícita.

Tabela 5.7: Desempenho do SARIMA frente aos modelos de comparação (*backtest* de 12 meses, recorte de 2018 em diante).

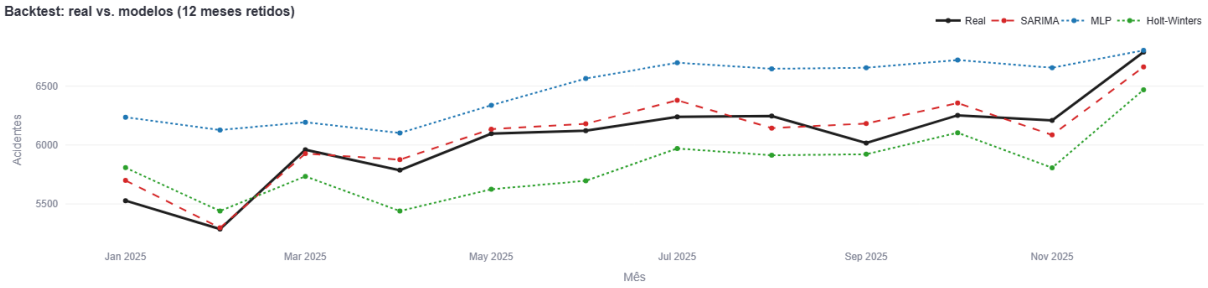
Modelo	**MAE**	**RMSE**	**MAPE**
SARIMA	97,0	109,2	1,6%
Baseline sazonal	112,9	134,5	1,8%
Holt-Winters	289,0	310,0	4,8%
MLP	433,9	484,0	7,4%

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

A Figura 5.7 confronta visualmente, nos doze meses retidos para teste, os valores

reais com as previsões dos modelos avaliados. A curva do SARIMA acompanha de perto a série real, ao passo que o MLP e o Holt-Winters se afastam mais do observado, o que explica seus erros mais elevados na Tabela 5.7.

Figura 5.7: *Backtest* dos doze meses retidos para validação: valores reais frente às previsões do SARIMA, do MLP e do Holt-Winters.



Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

Além de superar os modelos de comparação, o SARIMA passou nos diagnósticos de adequação. O teste de Ljung-Box aplicado aos seus resíduos (até o atraso de doze meses) resultou em um valor-p de 0,110, acima de 0,05: não se rejeita a hipótese de que os resíduos são ruído branco, o que indica que o modelo capturou a estrutura temporal da série, sem autocorrelação relevante remanescente.

A construção do modelo é ainda corroborada pelos demais diagnósticos da metodologia de Box-Jenkins, sintetizados na Tabela 5.8. Os testes de estacionariedade confirmam a necessidade da diferenciação: em nível, a série é não estacionária (ADF com $p = 0{,}85$ e KPSS com $p = 0{,}02$), tornando-se estacionária após a primeira diferença (ADF com $p < 0{,}001$ e KPSS com $p = 0{,}10$), o que justifica empiricamente a ordem de integração adotada. O modelo ajustado apresenta AIC de 991,3 e BIC de 1000,2. Quanto aos resíduos, embora o teste de Ljung-Box confirme a ausência de autocorrelação, o teste de Jarque-Bera rejeita a normalidade ($p < 0{,}001$) e o teste ARCH-LM detecta heterocedasticidade ($p = 0{,}04$). A não normalidade, comum em séries reais, implica que o intervalo de confiança de 95% deve ser interpretado como aproximado; a heterocedasticidade revela que a volatilidade dos resíduos varia ao longo do tempo, o que aponta para a modelagem explícita da variância como extensão natural do trabalho, discutida nas Considerações Finais.

Por fim, a validação por origem móvel (*walk-forward*), com 15 origens e horizonte

Tabela 5.8: Diagnósticos estatísticos do modelo SARIMA, segundo a metodologia de Box-Jenkins.

Diagnóstico	**Resultado**	**Interpretação**
Estacionariedade (nível)	ADF $p = 0{,}85$; KPSS $p = 0{,}02$	não estacionária
Estacionariedade (1ª dif.)	ADF $p < 0{,}001$; KPSS $p = 0{,}10$	estacionária ($d = 1$)
Critérios de informação	AIC = 991,3; BIC = 1000,2	ajuste do modelo
Ljung-Box (resíduos)	$p = 0{,}110$	ruído branco
Jarque-Bera (normalidade)	$p < 0{,}001$	resíduos não normais
ARCH-LM (heterocedast.)	$p = 0{,}04$	variância não constante

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

de doze meses, fornece uma medida de robustez mais conservadora que o *backtest* único. Nesse procedimento, o SARIMA apresentou MAPE mediano de 8,4%, valor que reflete o desempenho típico nas diferentes origens. A média é fortemente inflada por poucas origens iniciais, nas quais o curto histórico de treino torna o ajuste do SARIMA instável; descartadas essas, o modelo mantém erros baixos e estáveis. O resultado confirma que o desempenho melhora à medida que o histórico cresce.

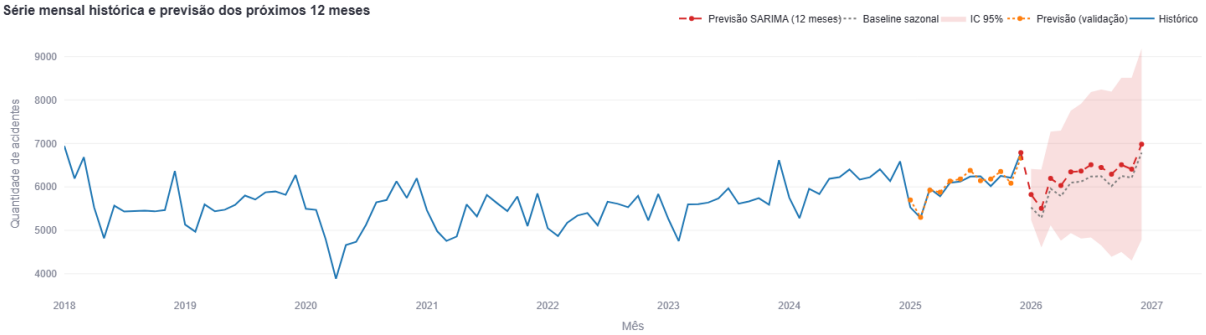
Projetados para o futuro, os modelos divergem, como mostra a Figura 5.8. Por ter apresentado o menor erro na validação, a projeção do SARIMA, modelo principal, é a adotada como referência, enquanto o MLP e o Holt-Winters servem de comparação.

Figura 5.8: Projeção dos próximos doze meses segundo cada modelo avaliado: SARIMA (principal), MLP e Holt-Winters.

Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

A Figura 5.9 apresenta a série histórica, a previsão de validação e a projeção dos doze meses seguintes, acompanhada da banda de incerteza de 95%. Ressalta-se que a restrição do treinamento ao recorte homogêneo a partir de 2018 é o que confere validade a essa projeção, ao evitar a contaminação pela tendência de nível dos anos anteriores.

Figura 5.9: Série mensal histórica, validação e previsão dos próximos 12 meses com banda de incerteza de 95%.



Fonte: Elaborado pelo autor, a partir dos dados abertos da PRF (2026).

5.4 Painel Interativo

Para além da análise estática apresentada nas seções anteriores, os resultados foram consolidados em um painel interativo, desenvolvido com a biblioteca *Streamlit*, que funciona como um sistema de apoio à decisão. O painel organiza as visualizações em cinco abas temáticas e, sobretudo, disponibiliza um conjunto de filtros globais que conferem ao usuário autonomia para conduzir a própria investigação, transformando os indicadores em um instrumento exploratório de planejamento.

5.4.1 Filtros Dinâmicos e Interatividade

A barra lateral do painel reúne três filtros globais, ilustrados na Figura 5.10: o período (intervalo de anos), a unidade federativa (UF) e a rodovia (BR). Esses filtros são aplicados simultaneamente a todas as abas, de modo que qualquer seleção se propaga de forma coordenada para o mapa, as séries temporais, os padrões e os trechos críticos. Essa propagação é o que caracteriza o painel como sistema de apoio à decisão: um gestor pode, por exemplo, selecionar a BR-101 no estado de Santa Catarina e restringir o período ao recorte homogêneo a partir de 2018, obtendo instantaneamente o perfil de severidade, a sazonalidade e o ranking de trechos daquele recorte específico, sem necessidade de reprocessar os dados. A possibilidade de partir do panorama nacional e refinar progressivamente a análise até a escala de uma rodovia em um estado constitui o principal valor prático da ferramenta.

Figura 5.10: Filtros globais da barra lateral, aplicados de forma coordenada a todas as abas.

Fonte: Elaborado pelo autor (2026).

5.4.2 Aba de Mapa e Indicadores

A aba de mapa, exibida na Figura 5.11, combina quatro cartões de indicadores (total de acidentes, mortos, feridos e número de rodovias afetadas) com um mapa coroplético da malha rodoviária federal. Cada rodovia é colorida segundo a intensidade de acidentes do recorte atual, em escala logarítmica, o que evita que poucos trechos de altíssimo volume saturem a paleta e tornem os demais indistinguíveis. A interação por *tooltip* revela, ao passar o cursor, a rodovia, o estado e o número de acidentes do trecho, e o mapa é redesenhado automaticamente a cada mudança de filtro.

Figura 5.11: Aba de mapa: indicadores-resumo e mapa coroplético das rodovias federais por intensidade de acidentes.

Fonte: Elaborado pelo autor (2026).

5.4.3 Aba de Padrões Temporais

A aba de padrões temporais reúne, em um único ambiente interativo, as visualizações de padrões temporais já discutidas (como o índice sazonal da Figura 5.4), todas projetadas para serem robustas à mudança de nível da série. A evolução mensal de longo prazo destaca, por meio de uma faixa sombreada, a transição para o boletim eletrônico em 2018, a partir da qual a série se estabiliza. Complementam a aba um mapa de calor de dia da semana por hora, expresso em percentual do total, o perfil horário que contrasta dias úteis e fins de semana, e um índice sazonal que, ao normalizar cada ano por sua própria média, torna a forma sazonal comparável apesar das diferenças de nível entre os anos. Por fim, a composição de gravidade por turno evidencia visualmente a maior participação de acidentes fatais nos períodos noturnos.

5.4.4 Aba de Causas e Fatores

A aba de causas e fatores aprofunda a investigação sobre os determinantes da severidade, reunindo de forma interativa as análises de fatores de risco já apresentadas (Figuras 5.2 e 5.5). Um gráfico de hierarquia circular (*sunburst*) decompõe os acidentes fatais na sequência causa, tipo e fase do dia, enquanto um mapa de árvore (*treemap*) destaca as combinações de maior taxa de mortalidade. Uma sub-aba dedicada apresenta as regras de associação discutidas na Seção 5.3.1, exibindo de forma interativa as combinações de

fatores mais letais por *lift* e confiança, com os limiares determinados automaticamente a partir dos dados do recorte. Mapas de calor relacionam causa e tipo de acidente, tanto em volume absoluto quanto em proporção dentro de cada causa, e um gráfico divergente compara a frequência das causas entre o dia e a noite. Vale notar que esta aba enfatiza os acidentes com vítimas fatais, que exigem atendimento presencial e, portanto, são registrados de forma consistente ao longo de todo o período, permanecendo comparáveis.

5.4.5 Aba de Trechos Críticos

A aba de trechos críticos materializa a detecção de pontos de concentração de acidentes graves descrita na Seção 4.5.2, cujo *ranking* foi apresentado na Figura 5.6. O usuário escolhe a métrica de ordenação (total de mortos, taxa de mortalidade, escore de anomalia ou feridos graves) e o número de trechos exibidos, obtendo um ranking dinâmico. Um gráfico de dispersão posiciona cada trecho segundo o volume de acidentes e a taxa de mortalidade, dividindo o espaço em quadrantes pelas medianas: os trechos situados no quadrante superior direito reúnem, simultaneamente, alto volume e alta letalidade, representando a concentração máxima de risco. A coloração pelo escore de anomalia, calculado pela combinação dos escores padronizados de óbitos e de taxa de mortalidade, orienta diretamente a priorização de fiscalização e de intervenções de engenharia.

Em conjunto, as cinco abas e os filtros globais conferem ao painel a natureza de um sistema de apoio à decisão: as métricas privilegiadas (proporções, taxas de mortalidade e escores de anomalia) foram escolhidas por serem robustas à mudança de nível da série, e a interatividade permite que autoridades competentes e a sociedade explorem livremente os dados, partindo de uma visão nacional e chegando ao diagnóstico de um trecho específico.

5.5 Síntese dos Resultados

Em conjunto, os resultados confirmam a hipótese norteadora da pesquisa: o tratamento sistemático dos registros da PRF permitiu transformar uma base dispersa e inconsistente em um ativo analítico capaz de revelar padrões acionáveis. Essa inconsistência manifesta-se, sobretudo, quando se compara o início e o fim da série: entre 2007 e 2018, a própria

forma de coletar e de disponibilizar os dados mudou, culminando na adoção do boletim eletrônico em 2018, o que torna os registros desses dois extremos, à primeira vista, pouco comparáveis entre si. É justamente por isso que o processo de limpeza e padronização se mostrou indispensável, pois foi ele que tornou a série consistente ao longo de todo o período, permitindo tratar 2007 a 2025 como um único ativo analítico coerente, ainda que a comparação de contagens absolutas exija cautela e a modelagem preditiva se restrinja ao recorte homogêneo a partir de 2018. Os achados convergem para um perfil de risco bem delimitado: os acidentes de maior gravidade associam-se às colisões frontais e aos atropelamentos, à madrugada e aos fins de semana, e concentram-se nas pistas simples das grandes rodovias longitudinais.

Um aspecto a destacar é a convergência entre análises independentes. A caracterização descritiva da severidade, a mineração de regras de associação e a análise espacial apontam, cada uma por um caminho distinto, para o mesmo conjunto de fatores letais, o que reforça a robustez das conclusões. Particularmente relevante é a dissociação entre frequência e letalidade: os tipos mais comuns de acidente, como as colisões traseiras (25,2%), não são os mais mortais, ao passo que as colisões frontais, apesar de representarem apenas 4,6% das ocorrências, lideram a letalidade (0,391 morto por acidente). Essa distinção tem implicação direta para a gestão pública, pois indica que políticas voltadas a reduzir o número de acidentes e políticas voltadas a reduzir o número de mortes não se confundem: a mitigação da mortalidade exige atuação dirigida a configurações pouco frequentes, porém desproporcionalmente letais, como as combinações de ultrapassagem indevida e colisão frontal isoladas pelas regras de associação, com *lift* de até 5,81.

A dimensão espacial acrescenta um vetor de priorização. A forte concentração da mortalidade em poucas rodovias, com destaque para a BR-101 e a BR-116, e em trechos municipais específicos, identificados pelo escore de anomalia, transforma um diagnóstico nacional em alvos geográficos concretos. Esse achado dialoga com o papel da infraestrutura: a predominância dos óbitos em pistas simples (92.474, ante 30.599 nas duplas) e a menor letalidade das vias de fluxo separado sugerem que a duplicação e a separação física de sentidos figuram entre as intervenções de engenharia de maior potencial sobre a redução de mortes.

No eixo preditivo, o desempenho do SARIMA (MAPE de 1,6% na validação), superior ao do MLP, ao do Holt-Winters e à linha de base sazonal, demonstra que a série tratada comporta previsões de curto prazo confiáveis, úteis ao dimensionamento antecipado de recursos operacionais. Cabe, todavia, uma ressalva metodológica: embora os resíduos não apresentem autocorrelação, os diagnósticos indicaram não normalidade e heterocedasticidade, de modo que as bandas de incerteza devem ser lidas como aproximadas, o que aponta a modelagem explícita da variância como aprimoramento natural. Por fim, é a integração dessas três frentes, a descritiva, a de mineração e a preditiva, em um único painel interativo que confere unidade ao trabalho: o mesmo instrumento que diagnostica o passado e o presente da acidentalidade também antecipa a sua tendência, oferecendo a gestores e à sociedade uma leitura articulada do fenômeno.

6 Conclusões

Esta monografia abordou o problema da organização, do tratamento e da análise de dados públicos de acidentes em rodovias federais brasileiras, com o objetivo de mapear o quadro histórico dos sinistros e prever novas ocorrências por meio de técnicas de mineração de dados e de modelos estatísticos, em apoio à segurança viária.

O objetivo geral foi alcançado por meio do cumprimento de todos os objetivos específicos propostos, cada qual materializado em uma das abas do painel interativo. O panorama georreferenciado da acidentalidade foi caracterizado sobre a base tratada de 2.194.832 registros, com 99,3% de cobertura de coordenadas, e consolidado na aba **Mapa e Indicadores**; os padrões temporais do risco, como a sazonalidade, os turnos e os dias de maior letalidade, foram identificados e sintetizados na aba **Padrões Temporais**; os fatores de risco e as combinações de condições desproporcionalmente letais foram descobertos pelas regras de associação e apresentados na aba **Causas e Fatores**; os trechos rodoviários com concentração anômala de acidentes graves foram detectados por escore de anomalia e ranqueados na aba **Trechos Críticos**; e a previsão de novas ocorrências foi obtida pelo modelo SARIMA, validado frente aos modelos de comparação, integrando a aba **Série Temporal e Previsão**. O detalhamento do valor de cada um desses resultados foi apresentado no Capítulo 5.

Entre as principais contribuições deste trabalho, destacam-se:

- **Base consolidada e *pipeline* reprodutível:** a disponibilização de uma base consolidada, limpa e georreferenciada de 2.194.832 registros de acidentes (2007 a 2025), acompanhada de um *pipeline* de Extração, Transformação e Carga reprodutível e organizado em camadas;
- **Tratamento da mudança de nível de 2018:** o tratamento metodológico da mudança de nível introduzida pelo boletim eletrônico em 2018, por meio do janelamento da série, condição indispensável para a validade da modelagem temporal;
- **Dissociação entre frequência e letalidade:** a evidência empírica da dissociação

entre a frequência e a letalidade dos acidentes, com implicações diretas para a priorização de políticas públicas;

- **Modelo preditivo validado:** o desenvolvimento de um modelo preditivo validado de forma rigorosa, tendo o SARIMA como modelo principal, que superou o *Perceptron* Multicamadas, o Holt-Winters e uma linha de base sazonal, com resíduos que se comportam como ruído branco, demonstrando seu potencial para o planejamento operacional de curto prazo;

- **Regras de associação com rigor estatístico:** a mineração de regras de associação com limiares estatisticamente controlados, que isolou combinações de fatores desproporcionalmente letais;

- **Painel interativo integrador:** a integração de todos esses produtos em um painel interativo de apoio à decisão, dotado de filtros dinâmicos.

O desenvolvimento da pesquisa permite responder, de forma conclusiva, às questões de pesquisa formuladas na Seção 1.2. Os resultados e as análises que fundamentam cada resposta encontram-se detalhados no Capítulo 5.

QP1 (inconsistências da base de dados). A base DATATRAN apresenta inconsistências que, sem tratamento, impedem uma visão histórica e geral confiável dos acidentes: registros duplicados, valores omissos em campos críticos, categorias textuais não padronizadas, erros nas coordenadas geográficas e a heterogeneidade do esquema dos arquivos ao longo dos anos. Todas elas mostraram-se identificáveis e superáveis, sendo eliminadas pelo tratamento aplicado, que consolidou uma base íntegra e georreferenciada (Seção 5.1).

QP2 (fatores de risco, severidade e padrões espaço-temporais). A mineração de dados revelou onde, quando e em que condições ocorrem os acidentes mais graves, evidenciando uma clara dissociação entre frequência e letalidade: os desfechos mais graves não estão nos tipos de acidente mais comuns, mas concentram-se nas colisões frontais e nos atropelamentos, na madrugada e nos fins de semana, e nas pistas simples das grandes rodovias longitudinais, com as combinações de fatores desproporcionalmente letais

isoladas pelas regras de associação e os trechos críticos localizados pela análise espacial (Seções 5.2, 5.3.1 e 5.3.2).

QP3 (previsão de novas ocorrências). A mineração da série histórica mostrou-se capaz de prever novas ocorrências de acidentes com boa acurácia. O modelo adotado como principal superou todas as alternativas avaliadas e forneceu projeções acompanhadas de banda de incerteza, oferecendo estimativas de curto prazo úteis ao planejamento operacional das autoridades competentes (Seção 5.3.3).

Quanto às limitações, a principal decorre de uma mudança no procedimento de registro ocorrida no período analisado: a adoção do boletim eletrônico pela PRF em 2018, que introduziu uma redução de nível na série histórica de acidentes. Essa descontinuidade restringe a janela de dados homogêneos disponíveis para a modelagem preditiva, uma vez que apenas o recorte a partir de 2018 pode ser tomado como comparável, reduzindo o histórico efetivamente utilizável no treinamento do modelo. Além disso, a imputação de coordenadas pela mediana do trecho, embora amplie a cobertura espacial, introduz uma aproximação que pode deslocar pontualmente a localização exata de algumas ocorrências. As regras de associação, por sua vez, embora estatisticamente filtradas, limitam-se ao desfecho fatal e aos atributos categóricos disponíveis na base, não incorporando variáveis contínuas nem a severidade individual graduada.

Como perspectivas de trabalhos futuros, propõem-se:

- **Integração com os boletins eletrônicos:** a integração da base DATATRAN com a base de boletins eletrônicos, de modo a reconstruir uma série contínua e ampliar o histórico de modelagem;
- **Modelo preditivo sobre a série completa:** o desenvolvimento de um modelo preditivo capaz de aproveitar a totalidade da série histórica (2007 a 2025), mediante técnicas de harmonização dos diferentes níveis da série (tais como a estimação de fatores de correção de escala ou o emprego de modelos robustos a mudanças de nível), de forma a aproveitar também os anos anteriores a 2018 e ampliar a base de treinamento sem comprometer a validade das estimativas;
- **Classificação da severidade individual:** a aplicação de modelos de classificação

supervisionada para a predição da severidade individual de cada acidente, complementando as regras de associação já implementadas;

- **Incorporação de variáveis exógenas:** a incorporação de variáveis exógenas, como volume de tráfego, condições climáticas detalhadas e características de infraestrutura, ao modelo preditivo;
- **Modelagem da variância (GARCH):** a modelagem explícita da variância condicional dos resíduos por meio de modelos da família GARCH (BOLLERSLEV, 1986), em resposta à heterocedasticidade identificada no diagnóstico do SARIMA;
- **Trechos críticos na escala do quilômetro:** a evolução da detecção de trechos críticos para a escala do quilômetro, aproveitando o georreferenciamento aprimorado para localizar *hotspots* com maior precisão.

Dos achados emergem elementos relevantes para a segurança viária, cabendo às autoridades competentes, e não a esta pesquisa, avaliar e definir as ações a partir deles. A concentração da letalidade nas colisões frontais associadas à ultrapassagem indevida e ao transitar na contramão, sobretudo em pistas simples, expõe um padrão de risco que merece atenção. No mesmo sentido, a maior letalidade observada na madrugada e nos fins de semana, bem como a relevância dos atropelamentos noturnos, evidenciam períodos e situações de risco elevado. A forte concentração geográfica da mortalidade nas BR-101 e BR-116, e nos trechos críticos identificados pelo escore de anomalia, constitui outro achado relevante, ao apontar pontos da malha rodoviária federal nos quais os óbitos se concentram de forma desproporcional. Por fim, a previsão mensal fornecida pelo modelo SARIMA disponibiliza estimativas de curto prazo sobre novas ocorrências de acidentes, informação que pode subsidiar avaliações de planejamento por parte das autoridades competentes.

Quanto ao uso dos produtos desenvolvidos, a base tratada e o painel interativo constituem um instrumento de diagnóstico contínuo, capaz de apoiar o monitoramento da acidentalidade e a avaliação do efeito de intervenções, da visão nacional ao diagnóstico de um trecho específico. Cabe ressaltar que os dados de origem, oriundos do DATATRAN, já são públicos; os artefatos elaborados nesta pesquisa, a saber, a base consolidada, o *pipeline* reprodutível e o painel, permanecem, neste momento, no âmbito acadêmico, sendo

a sua disponibilização pública uma evolução natural do trabalho. Conclui-se, assim, que a engenharia e a mineração de dados aplicadas a bases públicas de acidentes constituem um caminho efetivo para converter grandes volumes de registros brutos em conhecimento interpretável e acionável, contribuindo para que as autoridades competentes e a sociedade identifiquem tendências, reconheçam pontos críticos e fundamentem estratégias de prevenção nas rodovias federais brasileiras.

Bibliografia

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, v. 57, n. 1, p. 289–300, 1995.

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, v. 31, n. 3, p. 307–327, 1986.

BOX, G. E. P. et al. *Time Series Analysis: Forecasting and Control*. 5. ed. Hoboken: Wiley, 2015.

Confederação Nacional do Transporte. *Pesquisa CNT de Rodovias 2024: Relatório Gerencial*. Brasília, 2024. Disponível em: ⟨https://cnt.org.br/pesquisa-cnt-rodovias-2024⟩. Acesso em: 27 maio 2026.

COSTA, J. de J.; BERNARDINI, F. C.; FILHO, J. V. A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. *AtoZ: novas práticas em informação e conhecimento*, v. 3, n. 2, p. 139–157, 2014.

CUNHA, I. B. de A. *Modelagem da informação para cidades inteligentes: aplicação em acidentes de trânsito de Belo Horizonte*. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Belo Horizonte, 2019.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, v. 74, n. 366, p. 427–431, 1979.

ELVIK, R. et al. *The Handbook of Road Safety Measures*. 2. ed. Bingley: Emerald Group Publishing, 2009.

ENGLE, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, v. 50, n. 4, p. 987–1007, 1982.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, n. 3, p. 37–54, 1996.

FüHR, G. K.; INACIO, E. C. Predição de acidentes de trânsito em santa catarina: avaliando o impacto do pré-processamento dos dados no desempenho dos modelos. *Revista Brasileira de Iniciação Científica*, v. 13, 2026.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3. ed. Boston: Morgan Kaufmann, 2011.

HAYKIN, S. *Neural Networks and Learning Machines*. 3. ed. Upper Saddle River: Prentice Hall, 2009.

INMON, W. H. *Building the Data Warehouse*. 4. ed. Indianapolis: Wiley, 2005.

Instituto de Pesquisa Econômica Aplicada. *Balanço da 1ª década de ação pela segurança no trânsito no Brasil e perspectivas para a 2ª década*. Brasília, 2023. Disponível em: ⟨https://repositorio.ipea.gov.br/⟩. Acesso em: 27 maio 2026.

JARQUE, C. M.; BERA, A. K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, v. 6, n. 3, p. 255–259, 1980.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. [S.l.], 2007.

KWIATKOWSKI, D. et al. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, v. 54, n. 1–3, p. 159–178, 1992.

LAKATOS, E. M.; MARCONI, M. de A. *Fundamentos de Metodologia Científica*. 9. ed. São Paulo: Atlas, 2021.

LIMA, J. P. da S. *GeoJSON das rodovias federais do Brasil (simplificado via Mapshaper)*. 2022. GitHub Gist. Dados originais do Banco de Informações de Transportes (BIT). Disponível em: ⟨https://gist.github.com/jaumpedro214/8524d59a2e7791c504eda0783975bcb0⟩. Acesso em: 31 maio 2026.

LIMA, J. P. da S. *Traffic Accidents on BRs: arquitetura data lakehouse para acidentes em rodovias federais*. 2023. Repositório GitHub. Disponível em: ⟨https://github.com/jaumpedro214/traffic-accidents-br-data-project⟩. Acesso em: 31 maio 2026.

MAFORT, E. V.; KAPPEL, M. A. A. Técnicas de aprendizado de máquina para predição de gravidade de acidentes em rodovias do estado do rio de janeiro. *Revista Eletrônica de Iniciação Científica em Computação (REIC)*, v. 23, n. 1, p. 244–252, 2025.

MARTINS, I. E. S.; ANDRADE, M. H. S. Aplicação de técnicas de aprendizado de máquina na análise de ocorrências de trânsito de belo horizonte-mg. *Journal of Innovation and Science: research and application (JOINS)*, v. 1, n. 1, 2021.

MCKINNEY, W. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 51–56.

Ministério da Infraestrutura. *BIT Mapas: Banco de Informações de Transportes*. 2025. Portal gov.br. Disponível em: ⟨...⟩. Acesso em: 1 fevereiro 2026.

PAGE, M. J. et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, v. 372, n. n71, 2021.

PAULA, D. A. de. Estado, sociedade civil e hegemonia do rodoviarismo no brasil. *Revista Brasileira de História da Ciência*, Rio de Janeiro, v. 3, n. 2, p. 142–156, 2010.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

Polícia Rodoviária Federal. *Dados Abertos — Acidentes (DATATRAN)*. 2025. ⟨https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf⟩. Acesso em: 16 jun. 2026.

RASCHKA, S. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of Open Source Software*, v. 3, n. 24, p. 638, 2018.

SANTOS, A.; SANTOS, E. Análise da evolução da mortalidade por sinistros de trânsito no brasil: O impacto assistencial e socioeconômico no sistema Único de saúde. *Estudos de Saúde Pública e Mobilidade (Compilação de Dados Umane/Estadão)*, v. 12, n. 2, p. 45–58, 2026.

SEABOLD, S.; PERKTOLD, J. statsmodels: Econometric and statistical modeling with python. In: *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 92–96.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining - Mineração de Dados*. Rio de Janeiro: Ciência Moderna, 2009.

WARE, C. *Information Visualization: Perception for Design*. 3. ed. Waltham: Morgan Kaufmann, 2012.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. ed. Boston: Morgan Kaufmann, 2011.