RELATÓRIO DO TRABALHO FINAL DE CIÊNCIA DE DADOS

RAFAEL AUGUSTO DE SOUZA - RA 2134756, LUIZ GUILHERME GERON MANFRIM COELHO - RA 2134624

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ – CAMPO MOURÃO

CIÊNCIA DA COMPUTAÇÃO – OPT004

rafaelaugusto@alunos.utfpr.edu.br, luizguilhermecoelho@alunos.utfpr.edu.br

1. Resumo

Sinistros de trânsito representam uma das principais causas de morte e lesões, com impactos sociais, econômicos e de saúde pública significativos. A fim de encontrar os fatores que estão mais associados à ocorrência de sinistros graves nas rodovias federais e saber se seria possível prever a gravidade deles com base nas características de sua ocorrência, analisamos um dataset de sinistros em rodovias federais de 2007, com uso de estatística e modelos de IA. Elaboramos 3 hipóteses: fatores ambientais e temporais influenciam a gravidade do sinistro, o número de veículos envolvidos e o tipo de pista são preditores significativos da gravidade e modelos de ensemble apresentam desempenho superior na predição da gravidade do acidente. Todas foram aceitas.

2. Problema e Perguntas de Pesquisa

No Brasil, os sinistros de trânsito representam uma das principais causas de morte e lesões, com impactos sociais, econômicos e de saúde pública significativos. Analisar dados de sinistros em rodovias federais permite compreender os fatores associados à ocorrência e gravidade dos acidentes, fornecendo percepções e evidências para políticas de prevenção e melhoria da segurança rodoviária.

Neste trabalho, perguntamos quais fatores estão mais associados à ocorrência de sinistros graves nas rodovias federais em 2007 e se seria possível prever a gravidade deles com base nas características de sua ocorrência.

Elaboramos 3 hipóteses: fatores ambientais e temporais influenciam a gravidade do sinistro, o número de veículos envolvidos e o tipo de pista são preditores significativos da gravidade e modelos de ensemble apresentam desempenho superior na predição da gravidade do acidente.

3. Metodologia e Limitações

O dataset usado foi de sinistros em rodovias federais em 2007, apresentando 127675 instâncias e 26 colunas (incluindo parâmetros da condição do tempo e da pista, dados temporais e de localização).

Inicialmente verificamos se havia a presença de dados faltantes, inconsistentes, outliers ou fora do padrão. Encontramos alguns dados ausentes, mas nenhuma inconsistência, despadronização ou outlier significativo. Duas colunas apresentavam classes muito amplas (causa_acidente possuía a classe “Outras” e condicao_metereologica a classe “Ignorada”), que foram tratadas como valor ausente. Removemos todas as instâncias com algum valor ausente, o que reduziu o dataset para 79811 instâncias. O tipo das colunas br e km foi alterado para numérico.

Esses dados foram complementados com a população de cada município, extraída do dataset do IBGE com a estimativa da população em 2006. Foi necessária uma padronização dos nomes de cidade deste dataset para o casamento de padrão com os nomes no dataset de sinistros. apenas 172 instâncias do dataset de sinistros não tiveram correspondência e ficaram sem a informação da população.

Os dados foram passados para formato Tidy, com a alteração do tipo do campo data_inversa para datetime e a inclusão das colunas mes, fim_semana, dia e hora.

Usando DuckDB fizemos algumas análises com consultas SQL. Também geramos análises univariada, bivariada, multivariada e testes de hipóteses.

Além das análises estatísticas clássicas, empregamos métodos de aprendizado de máquina para construir modelos preditivos de gravidade dos sinistros. Inicialmente, definimos uma variável-alvo binária denominada grave, que assume valor 1 quando há qualquer vítima não ilesa (isto é, presença de mortos, feridos leves ou feridos graves) e 0 caso contrário. A partir disso, removemos do conjunto de atributos todas as variáveis diretamente relacionadas às vítimas (como número de mortos, feridos, feridos graves, ilesos e classificação oficial do acidente), bem como variáveis de identificação e localização excessivamente específicas, a fim de evitar vazamento de informação e garantir que a predição se baseasse apenas em características contextuais e estruturais do evento. Sobre esse conjunto de dados limpo e enriquecido com atributos temporais (mês, dia, indicador de fim de semana e faixa horária), treinamos modelos de Regressão Logística, Random Forest e MLP e a combinação destes, avaliados em um esquema de divisão estratificada em conjuntos de treino e teste.

Como principal limitação podemos destacar o uso de dados apenas de 2007, o que pode dificultar a visualização de tendências mais gerais com casos atípicos deste ano. Nesse dataset não haviam dados georreferenciados, que exporiam com maior precisão o local de cada sinistro.

4. Resultados das Análises

Inicialmente, foram feitas análises com consultas SQL. Verificamos o número difio de sinistros com média móvel de 7 dias, suavizando a visualização dos casos diários para evidenciar períodos com maior quantidade de casos. Podemos notar esses períodos no início e fim do ano, além de 4 picos ao longo do restante do ano. Ver a figura 1.

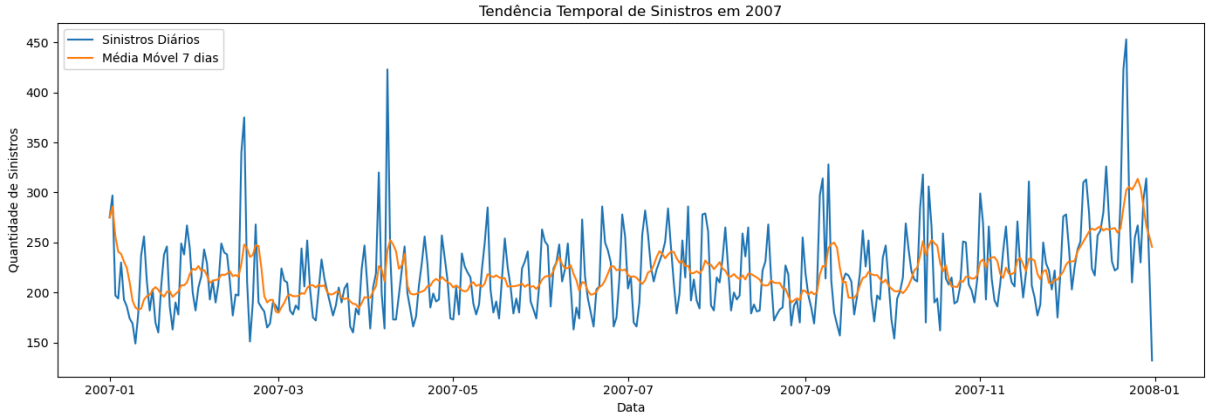

Figura 1: Sinistros com média móvel de 7 dias.

Consultamos a média de mortos por tipo de sinistro, demonstrando a mortalidade de cada um. Os tipos mais letais foram colisão frontal, atropelamento e colisão com bicicleta. Resultados exibidos na figura 2.

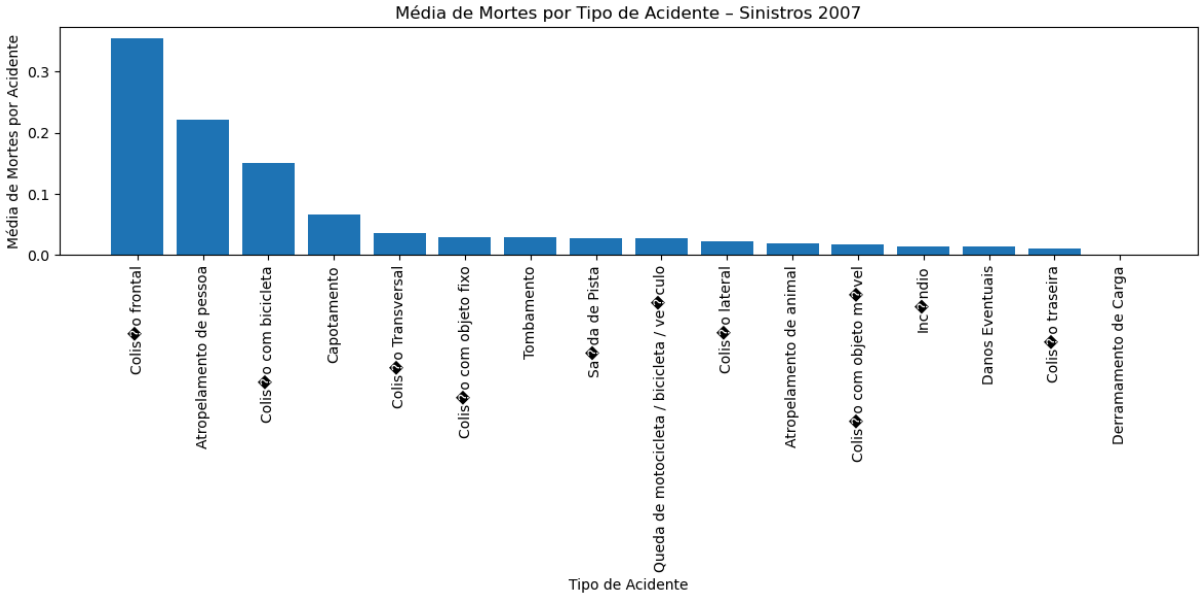

Figura 2: Média de mortes por tipo de sinistro.

Buscamos as causas de ontem em 80% do número de mortos. Foram encontradas 5 causas: falta de atenção, velocidade incompatível, ultrapassagem indevida, dormindo e desobediência à sinalização. Figura 3.

causa_acidente	mortos	acumulado	perc_acumulado
Falta de aten��o	1436.0	1436.0	0.425
Velocidade incompat�vel	513.0	1949.0	0.577
Ultrapassagem indevida	494.0	2443.0	0.723
Dormindo	214.0	2657.0	0.786
Desobedi�ncia � sinaliza��o	209.0	2866.0	0.848
Ingest�o de �lcool	172.0	3038.0	0.899
Defeito mec�nico em ve�culo	117.0	3155.0	0.933
Animais na Pista	96.0	3251.0	0.962
N�o guardar dist�ncia de seguran�a	65.0	3316.0	0.981
Defeito na via	64.0	3380.0	1.000

Figura 3: Acumulado da porcentagem de mortos por tipo de sinistro.

Verificamos os dias com número anormal de sinistros usando Z-Score com acima de 2.5. 6 dias apareceram nessa busca. Todos estão próximos de datas comemorativas, como o Natal (dias 21 e 22), o Carnaval (16 e 17 de fevereiro), a Independência (9 de setembro) e a Páscoa (8 de abril). Figura 4.

	dia	total_sinistros	z_score
0	2007-12-22	453	5.442866
1	2007-12-21	423	4.746075
2	2007-04-08	423	4.746075
3	2007-02-17	375	3.631208
4	2007-02-16	340	2.818284
5	2007-09-09	328	2.539567

Figura 4: Dias com número de sinistros anormal.

Por fim, vimos a relação entre número de veículos e a gravidade dos sinistros. Os sinistros com maior número de pessoas afetadas (feridas, feridas graves ou mortas) se dão com o maior número de veículos. As linhas de média de feridos e feridos graves demonstram alguma oscilação. Figura 5.

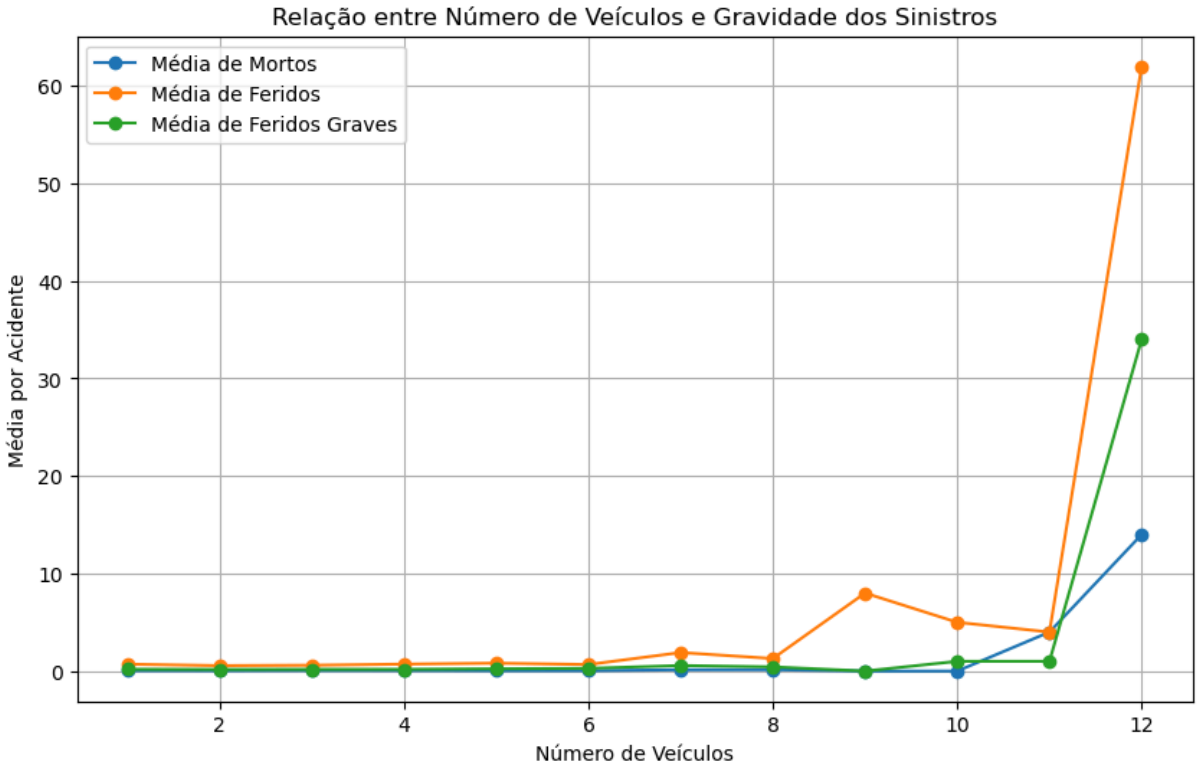

Figura 5: Relação entre número de veículos e gravidade dos sinistros.

Além dessas consultas, realizamos análises univariada, bivariada, multivariada e testes de hipóteses. A univariada foi feita com as colunas mortos, feridos, feridos_graves, veículos, pessoas e hora. O número de feridos e mortos em sua grande maioria é zero, sendo normalmente 1 o número de pessoas envolvidas. Na hora, observamos um crescimento do periodo da madrugada até o fim da tarde e início da noite (entre 15 e 20 horas) e então uma queda. A maior frequência do número de veículos envolvidos é 2, mas podemos notar uma quantidade expressiva de sinistros com apenas 1. Ver figura 6.

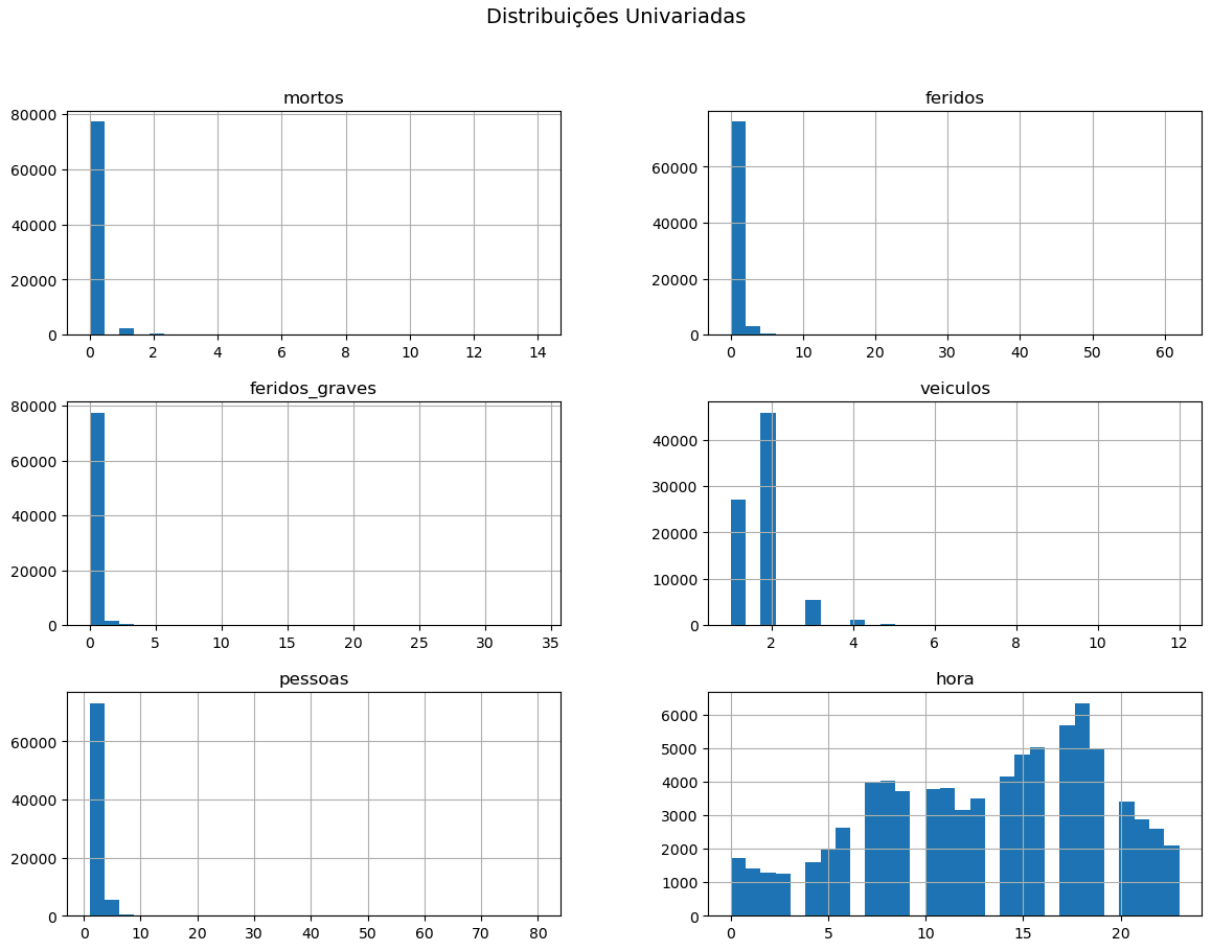

Figura 6: Distribuições univariadas.

Para a análise bivariada usamos um heatmap com a correlação entre as colunas mortos, feridos, feridos_graves, veiculos, pessoas e hora. Verificamos correlações altas entre variáveis numéricas que se somam, como feridos_graves com feridos e pessoas com feridos. Identificamos uma correlação significativa entre pessoas e veiculos.

Na análise multivariada usamos a técnica PCA com as colunas veiculos, mortos, feridos, hora e pop_2006. Os pontos estão concentrados em apenas uma região, indicando que são parecidos em relação às colunas usadas. Existem alguns pontos que fogem dessa região, outliers. A escala maior de PC1 indica que ele explora melhor a variação dos dados. Figura 7.

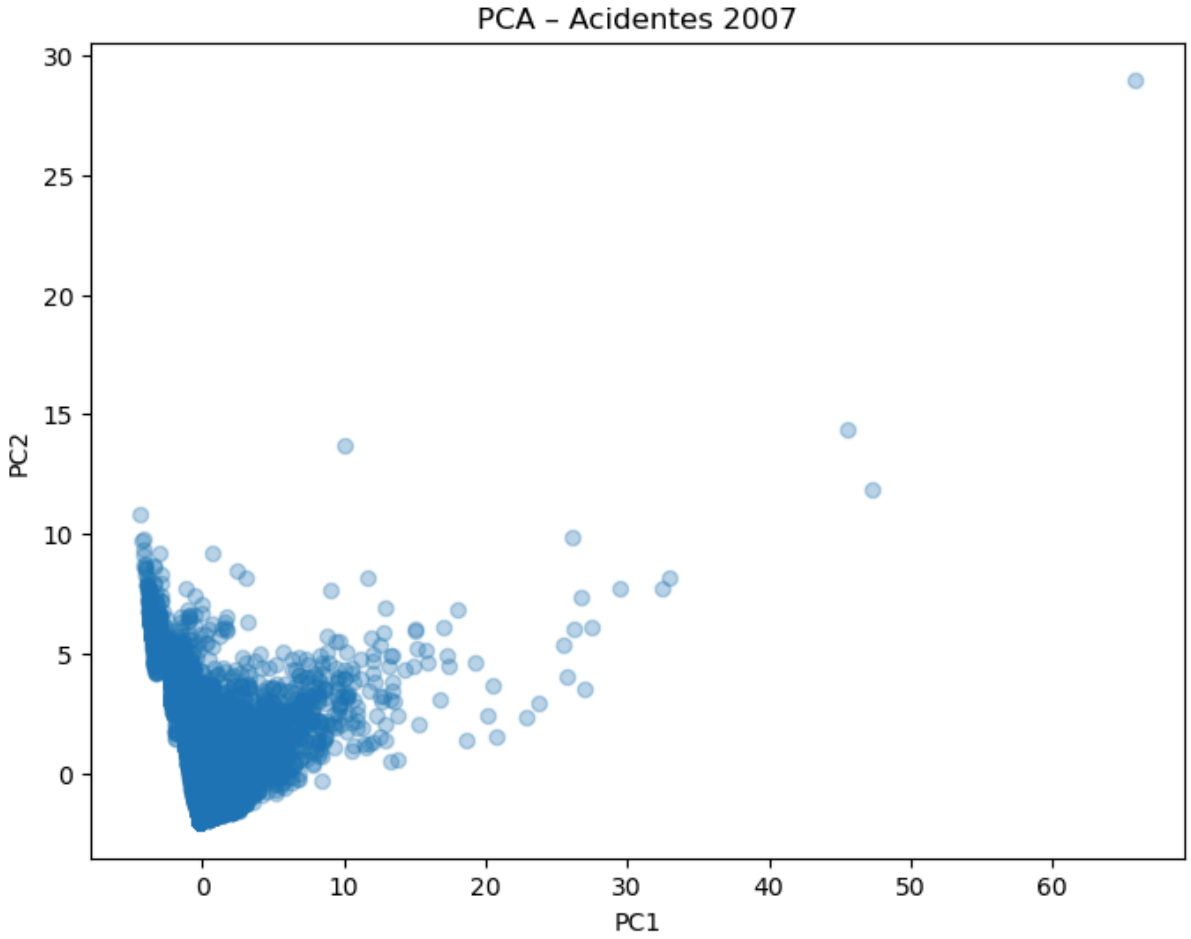

Figura 7: PCA.

Testamos a primeira hipótese (fatores ambientais e temporais influenciam a gravidade do sinistro) com a análise de variância para o número de mortos e a fase do dia, que resultou em p-value de 3.8233613031366624e-47, e com o teste qui-quadrado para a condição meteorológica e a gravidade, p-value de 3.5852705469560646e-17. Em ambos os casos rejeitamos fortemente a hipótese nula e aceitamos H1.

A segunda hipótese (o número de veículos envolvidos e o tipo de pista são preditores significativos da gravidade) foi testada com um modelo de regressão logística que chegou ao valor p de 0.002 para o tipo de pista múltipla, 0.000 para o tipo de pista simples e 0.045 para o número de veículos. Todas as variáveis testadas foram significativas, logo, aceitamos a hipótese.

Em relação à terceira hipótese, os resultados obtidos (figura 8) indicam que técnicas de ensemble superam modelos individuais de forma consistente. Embora o Random Forest tenha apresentado melhor desempenho dentre os classificadores isolados, com F1-score e recall superiores à Regressão Logística e ao MLP, foi a combinação dos três modelos que produziu o melhor resultado geral. O ensemble demonstrou maior estabilidade preditiva, melhor capacidade de generalização e melhor compromisso entre detecção de acidentes graves e controle de falsos positivos. Dessa forma, a terceira hipótese é considerada aceita, uma vez que a solução baseada em Ensemble apresentou desempenho superior e mais consistente do que qualquer modelo empregado sozinho.

	Modelo	F1	Acurácia	Precisão	Recall
0	Regressão Logística	0.382192	0.640105	0.576948	0.285737
1	Random Forest	0.433106	0.640857	0.562404	0.352147
2	LR Reduzido	0.364231	0.634780	0.565910	0.268532
3	RF Reduzido	0.435332	0.630771	0.538516	0.365332
4	MLP Completo	0.415441	0.649001	0.591503	0.320148
5	MLP Reduzido	0.426936	0.648688	0.585646	0.335906
6	Ensemble RF+MLP (Reduzido)	0.537313	0.634906	0.530657	0.544139
7	Ensemble LR+RF+MLP (Reduzido)	0.544674	0.634780	0.529537	0.560701
8	Ensemble LR+RF+MLP (Reduzido, tunado)	0.595025	0.566560	0.467789	0.817334

Figura 8: Resultados dos modelos.

Os resultados obtidos ao longo das análises exploratórias e da modelagem com IA convergem para um mesmo conjunto de fatores explicativos. As consultas SQL, as estatísticas descritivas e os testes de hipóteses indicaram que as características temporais (fase do dia e proximidade de feriados), o número de veículos e determinadas causas oficiais do acidente estão associadas à maior severidade dos sinistros. Por sua vez, os modelos de aprendizado de máquina reforçaram esse padrão ao atribuir importância a variáveis como causa, tipo de pista e número de veículos envolvidos.

Do ponto de vista das perguntas de pesquisa, a combinação de EDA e modelos de IA permitiu responder de forma afirmativa à primeira e à segunda hipótese. A gravidade dos sinistros não se distribui de maneira aleatória, mas é fortemente influenciada por fatores comportamentais, estruturais e temporais. Além disso, os modelos treinados alcançaram desempenho suficiente para demonstrar que é possível prever, com grau razoável de acerto, se um acidente tende a ser grave utilizando apenas informações observáveis no momento da ocorrência, confirmando a terceira hipótese.

5. Discussão dos Resultados

Os resultados confirmam que os sinistros de trânsito nas rodovias federais brasileiras em 2007 seguem padrões consistentes e não ocorrem de forma aleatória. A análise temporal evidenciou maior concentração de acidentes em períodos festivos, indicando forte relação entre volume de tráfego e risco de ocorrência. Da mesma forma, certos tipos de sinistro, como colisões frontais e atropelamentos, apresentaram maior letalidade, indicando que nem todos os acidentes possuem o mesmo potencial de gravidade.

A análise de Pareto mostrou que a maior parte das mortes está associada a um conjunto reduzido de causas, falta de atenção, velocidade incompatível, ultrapassagem indevida, dormindo e desobediência à sinalização. Esses achados reforçam a necessidade de políticas públicas mais direcionadas, em substituição a campanhas genéricas.

Os resultados dos modelos de aprendizado de máquina reforçam as conclusões das análises exploratórias ao identificar como mais relevantes as mesmas variáveis observadas na EDA, como tipo de pista, número de veículos envolvidos e fatores comportamentais. A solução baseada em ensemble apresentou o melhor desempenho, superando os modelos individuais e evidenciando que abordagens híbridas são mais adequadas para lidar com a complexidade dos dados de trânsito.

Apesar dos bons resultados, reconhece-se que a ausência de variáveis como volume real de tráfego, velocidade aferida e dados georreferenciados mais precisos limita o potencial preditivo dos modelos. Ainda assim, os resultados demonstram que é viável prever a gravidade dos sinistros com base apenas em informações contextuais.

6. Recomendações Práticas

Com base nos resultados obtidos, recomenda-se que ações preventivas em segurança viária sejam direcionadas a fatores de risco claramente identificados. Em particular, é fundamental priorizar fiscalizações em rodovias de pista simples, onde foi observado maior número de sinistros, bem como intensificar operações de controle em períodos críticos como feriados prolongados e meses de maior fluxo veicular.

Programas educativos devem ser focados nos principais comportamentos de risco identificados, como ignorar a sinalização, excesso de velocidade, ultrapassagens indevidas e desatenção. Além disso, os achados indicam a necessidade de investimentos estruturais, como melhorias de sinalização, manutenção viária e duplicação de trechos críticos.

Do ponto de vista tecnológico, recomenda-se a adoção gradual de sistemas de apoio à decisão baseados em dados, integrando análises estatísticas e modelos de aprendizado de máquina para priorização de trechos e horários de maior risco.

7. Trabalhos Futuros

Como continuidade deste estudo, recomenda-se a ampliação da base de dados para incluir múltiplos anos, permitindo a análise de tendências de longo prazo e a construção de modelos mais generalizáveis. A incorporação de novas variáveis, como volume de tráfego, velocidade média, dados meteorológicos de maior resolução e informações georreferenciadas, tende a aprimorar significativamente o desempenho preditivo.

Do ponto de vista metodológico, sugere-se a experimentação de modelos mais avançados, como técnicas de boosting, calibração de probabilidades e métodos de explicabilidade, visando maior interpretabilidade e confiabilidade das predições.

8. Referências

Dataset principal: Sinistros de Trânsito Agrupados Por Ocorrência em 2007 (https://dados.gov.br/dados/conjuntos-dados/sinistros-de-transito-agrupados-por-ocorrencia).

Dataset complementar: estimativa da população no ano de 2006 (https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?edicao=17283&t=downloads).

Material e aulas do professor doutor em Ciência da Computação Eduardo Henrique Monteiro Pena ministradas na UTFPR-CM.

Chat GPT (dúvidas técnicas).