Comparing Explainability Frameworks for Accident Severity Prediction

Diego Souza Rodrigues[1], **Rodolfo Ayala Lopes**[2], **Rodrigo Silva** [1]

[1]Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil

[2]Confederação Nacional das Cooperativas do Sicoob
Brasília - DF - Brasil

[3]Departamento de Computação, Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil

diego.rodrigues@aluno.ufop.edu.br

Abstract. *Predicting traffic crash severity from administrative databases is limited by the absence of behavioral variables, making it more important to identify which variables are consistently associated with severity across explainability methods than to maximize model accuracy. This paper compares SHapley Additive exPlanations (SHAP) and Average Marginal Effects (AMEs) across different classifiers on a dataset of Brazilian federal highway crashes (2017–2025). Results reveal a stable core of explanatory variables, stronger agreement among SHAP-based combinations, and more heterogeneous, low-to-moderate agreement between SHAP and AMEs.*

Resumo. *A predição da gravidade de sinistros com bases administrativas é limitada pela ausência de variáveis comportamentais, tornando mais relevante identificar quais variáveis são consistentemente associadas à gravidade entre métodos de explicabilidade do que maximizar a acurácia. Este artigo compara SHapley Additive exPlanations (SHAP) e Efeitos Marginais Médios (AMEs) em diferentes classificadores na base de sinistros em rodovias federais brasileiras (2017–2025). Os resultados revelam um núcleo estável de variáveis, concordância mais forte entre combinações baseadas em SHAP, e concordância mais heterogênea, de baixa a moderada, entre SHAP e AMEs.*

1. Introdução

Os sinistros de trânsito representam um grave problema de saúde pública no Brasil [Carvalho 2020], e as rodovias federais concentram parcela significativa dos eventos [Polícia Rodoviária Federal 2026]. As variáveis disponíveis na base de dados da Polícia Rodoviária Federal (PRF), porém, são majoritariamente contextuais e inferidas pós-facto, não capturando diretamente os mecanismos causadores [Bucsuházy et al. 2020], o que limita os ganhos de acurácia independentemente do modelo [Mannering et al. 2020]. A questão relevante desloca-se, portanto, para a *interpretabilidade*: quais variáveis são consistentemente associadas à gravidade entre diferentes paradigmas de modelagem? Fatores estáveis entre métodos distintos são mais informativos do que associações de um único modelo, pois reduzem o risco de que os resultados reflitam a especificação escolhida em vez de padrões estruturais [Molnar et al. 2020].

Para abordar essa questão, empregamos três classificadores e dois frameworks de explicabilidade, quantificando a concordância entre os rankings das quatro combinações framework-classificador via Spearman, Kendall, Rank-Biased Overlap (RBO) e Jaccard.

Assim, este trabalho parte da hipótese de que há um núcleo de variáveis explicativas cuja relevância é preservada independentemente do classificador e do framework de explicabilidade. As perguntas de pesquisa atacadas mais diretamente são: $(Q1)$ AME e SHAP concordam para o modelo de Regressão Logística nesta base de sinistros de trânsito? $(Q2)$ O SHAP é estável quando aplicado aos diferentes modelos para previsão de gravidade de sinistros de trânsito?

Para o gestor de tráfego, essas perguntas definem onde aplicar recursos limitados de fiscalização, sinalização e manutenção: priorizar um fator cuja relevância depende do modelo ou método de explicação escolhido é investir em possível artefato de especificação, não em algo estrutural. Variáveis relevantes entre métodos distintos oferecem sinais mais confiáveis para essas intervenções.

A literatura existente aplica SHAP ou AMEs de forma isolada, cada estudo interpreta um único modelo com um único explicador [Dong et al. 2022, Chen et al. 2025, Franceschi et al. 2022], sem investigar se as narrativas de importância produzidas por paradigmas distintos de fato convergem. Comparações sistemáticas entre métodos de explicabilidade são escassas no domínio de segurança viária e, quando existem, restringem-se a pares modelo–explicador sem quantificação formal da concordância entre rankings [Saarela and Jauhiainen 2021, Swamy et al. 2022]. O presente trabalho preenche essa lacuna ao confrontar dois frameworks e três classificadores sobre a mesma base, com métricas de concordância complementares e intervalos de confiança por bootstrap.

As principais contribuições são: (i) evidência empírica de um núcleo estável de variáveis, cuja relevância se mantém entre frameworks e paradigmas de modelagem distintos, fornecendo base mais confiável para priorização de intervenções; (ii) evidência de que a divergência entre AME e SHAP não se explica apenas pela forma de decomposição; (iii) evidência de efeito heterogêneo da fiscalização eletrônica, com inversão de sinal entre métodos, que recomenda análise espacialmente desagregada antes de qualquer decisão de alocação de radares; e (iv) um protocolo de comparação transferível a outros domínios onde múltiplos explicadores são aplicados ao mesmo problema.

1.1. Trabalhos Relacionados

A aplicação de aprendizado de máquina à predição de severidade de sinistros já é um campo consolidado na literatura. A revisão sistemática apresentada em [Wen et al. 2021] cobrindo 74 estudos entre 2014 e 2025 mostra que modelos baseados em gradient boosting e ensembles superam consistentemente modelos lineares em acurácia, mas a AUC típica em bases administrativas raramente passa de 0,75 — um indício de que a própria estrutura dos dados impõe um teto de desempenho, independentemente do algoritmo escolhido. No Brasil, [Franceschi et al. 2022] aplicam regressão logística multinomial aos registros da PRF e apontam tipo de pista, condições meteorológicas e velocidade como os preditores mais relevantes de gravidade, em linha com a literatura internacional [Zhang et al. 2021]; um estudo mais recente, testando sete classificadores em rodovias federais do Rio de Janeiro (2020–2023), confirma que nem o ajuste fino de hiperparâmetros consegue levar a acurácia além de 73–75% [Mafort and Kappel 2025].

Nesse cenário o interesse desloca-se para a interpretabilidade, e é nesse ponto que diferentes estudos convergem para o SHAP, combinado a LightGBM, XGBoost e Random Forest em bases dos EUA, Reino Unido, China e Índia [Lundberg and Lee 2017, Lundberg et al. 2020, Dong et al. 2022, Chen et al. 2025, Wang and Serre 2025, Khanum et al. 2025]. Velocidade, tipo de pista e condições meteorológicas emergem repetidamente como fatores dominantes — um padrão semelhante ao núcleo estável identificado no presente trabalho.

O ponto em comum entre esses estudos, no entanto, é também sua principal limitação: cada um interpreta um único modelo, com um único método de explicabilidade, sem investigar sistematicamente se as narrativas de importância produzidas por paradigmas distintos de fato convergem.

Na área da saúde, em [Saarela and Jauhiainen 2021] compara-se a importância de variáveis e concluem que combinar múltiplos explainers produz resultados mais confiáveis do que qualquer método isolado; [Swamy et al. 2022] avaliam SHAP, LIME e outros métodos via Spearman e documentam concordância moderada a alta para variáveis dominantes e baixa para variáveis periféricas. A instabilidade de rankings de importância em métodos baseados em amostragem foi documentada por [Hameed and Roth 2024], que recomenda ICs por bootstrap. Esse conjunto de evidências, contudo, permanece restrito à área da saúde. Até onde é de nosso conhecimento, nenhum estudo em segurança viária confrontou sistematicamente AME e SHAP em múltiplos classificadores sobre a mesma base, com quantificação formal da concordância entre rankings.

2. Base de Dados e Engenharia de Atributos

A base primária é composta pelos registros abertos da PRF [Polícia Rodoviária Federal 2026] (2017–2025), com 4.069.582 registros brutos e 37 atributos. Após concatenação e remoção de `pesid` (identificador da pessoa envolvida) duplicados, e de 1.403 registros com nulos em `sentido_via` e `frequencia`, obteve-se 631.157 registros: 548.510 não graves (86,9%) e 82.647 graves (13,1%).

A base foi enriquecida com atributos derivados: limites de velocidade (`vel_max`) conforme o Código de Trânsito Brasileiro [Brasil 1997]; distância ao radar mais próximo (`distancia_radar_m`), calculada sobre dados do DNIT [Departamento Nacional de Infraestrutura de Transportes (DNIT) 2024]; três binárias derivadas de `tracado_via` (`reta`, `declive`, `intersecao_de_vias`); três atributos de condição meteorológica — `condicao_ignorada` (sem informação), `pista_molhada` (chuva, garoa, granizo) e `visibilidade_ruim` (nevoeiro, precipitação); a frequência histórica de sinistros por km (`frequencia`); `uf_br_km`, identificando segmentos via concatenação de UF, rodovia e quilômetro; e `feriado`, obtido via biblioteca Holidays, expressando a proximidade em dias ao feriado mais próximo (0 = data do próprio feriado).

O pipeline de codificação tem três etapas: (1) **cíclicas** — `turno` (P=4), `dia_semana` (P=7) e `mes` (P=12) por sin/cos; (2) **alta cardinalidade** — Target Encoding com suavização bayesiana (smoothing=10) [Micci-Barreca 2001] em `uf_br_km`; (3) **numéricas** — MinMaxScaler $[0, 1]$ ajustado apenas no treino para `vel_max`, `distancia_radar_m` e `frequencia`.

As variáveis selecionadas cobrem geometria e infraestrutura (`tipo_pista`,

reta, declive, intersecao_de_vias), condições operacionais (pista_molhada, visibilidade_ruim, condicao_ignorada), temporais cíclicas (turno, dia_semana, mes), contexto espacial e fiscalização (uf_br_km, frequencia, vel_max, distancia_radar_m) e temporal não cíclica (feriado).

As demais variáveis foram descartadas por serem identificadores sem valor semântico, estarem associadas à ocorrência (não ao contexto) do sinistro, terem sido substituídas por atributos derivados mais representativos, apresentarem redundância ou multicolinearidade [Dormann et al. 2013], corresponderem a elementos de baixa frequência com risco de coeficientes instáveis [Harrell et al. 2001], ou introduzirem dependência espacial não modelada, no caso de coordenadas geográficas [Anselin 1988].

3. Metodologia

Três classificadores, cobrindo paradigmas de modelagem distintos, foram utilizados. A Regressão Logística [Cox 1958] serve como referência paramétrica clássica e é o único modelo no qual os Efeitos Marginais Médios (AMEs) são analiticamente bem definidos, permitindo contraste direto com o SHAP na mesma escala de probabilidade. O Random Forest [Breiman 2001], um ensemble de árvores por bagging que captura não linearidades sem pressupor forma funcional, representa um paradigma não paramétrico de referência. O CatBoost [Prokhorenkova et al. 2018] aplica gradient boosting com tratamento nativo de variáveis categóricas via ordered boosting, escolhido pela presença de variáveis de alta cardinalidade na base, nas quais esse tratamento tende a superar codificações genéricas.

Todos os modelos foram implementados em Python[1]. Especificamente, foram configurados: (i) uma Regressão Logística estimada por Máxima Verossimilhança via algoritmo IRLS (*Iteratively Reweighted Least Squares*), com distribuição binomial e máximo de 500 iterações, e pesos inversamente proporcionais à frequência de cada classe (peso 0: 0,5745; peso 1: 3,857); (ii) um Random Forest com 300 estimadores, profundidade máxima 8, mínimo de 50 amostras por folha e ponderação balanceada de classes; (iii) um CatBoost com 300 iterações, profundidade 4, taxa de aprendizado de 0,05 e peso da classe positiva $\approx 6{,}71$. A ponderação de classes foi preferida ao *oversampling* por SMOTE porque preserva a distribuição empírica das covariáveis, propriedade necessária para que AMEs e SHAP reflitam a estrutura real dos dados [Mannering et al. 2020].

A variável-alvo binária é `grave = 1` quando o sinistro envolve ao menos uma vítima fatal ou gravemente ferida. O conjunto analisado contém exclusivamente registros de sinistros, de modo que AMEs e SHAP expressam a variação média em $P(\texttt{grave} = 1)$ condicional à ocorrência do sinistro. Todos os modelos usaram divisão temporal: treino 2017–2024 (558.847 registros) e teste 2025 (72.310 registros) [Shmueli 2010].

Para interpretar esses modelos, o trabalho emprega dois frameworks de explicabilidade: valores SHAP, aplicáveis aos três classificadores, e Efeitos Marginais Médios (AMEs) [Cameron 2005], restritos à Regressão Logística, pois modelos de árvore produzem funções constantes por partes, nas quais derivadas parciais são mal definidas [Hastie et al. 2009].

Dois agregados globais são extraídos para cada variável k [Lundberg and Lee 2017]: $E[\phi_k]$ (efeito médio com sinal) e $E[|\phi_k|]$ (importância

[1]Python 3.x; `statsmodels` para Regressão Logística e `scikit-learn`/`CatBoost` para os demais.

absoluta para ranqueamento). Para variáveis cíclicas, reporta-se $E\left[\sqrt{\phi_{\sin}^2 + \phi_{\cos}^2}\right]$. O ranqueamento é baseado em $E[|\phi_k|]$, que evita cancelamentos espúrios. Um exemplo concreto: `uf_br_km` no RF tem $E[\phi_k] \approx -0{,}002$ (quase zero) mas $E[|\phi_k|] = 0{,}111$ (1º no ranking), pois o sinal médio reflete cancelamento de efeitos positivos e negativos de diferentes segmentos.

Para o *CatBoost*, usou-se um explicador baseado na estrutura da árvore, configurado para escala de probabilidade e amostragem (*background*: 500 observações; teste: 2.000). Para *Random Forest* e Regressão Logística, adotou-se um explicador baseado em amostragem com *wrapper* explícito para saída em probabilidade (*background*: 100 centroides via K-means; teste: 2.000) — um explicador linear foi descartado por operar em *log-odds*. Essa escolha não é arbitrária: por padrão, o explicador baseado em árvore pode retornar valores em log-odds, e a documentação do pacote `shap`[2] adverte que sua saída não corresponde necessariamente à probabilidade prevista [SHAP developers 2026]; o explicador por amostragem evita essa ambiguidade por construção. A propriedade de eficiência ($\phi_0 + \sum_k \phi_k(i) = \hat{P}(i)$) foi verificada em todos os modelos.

Os AMEs foram calculados via Máxima Verossimilhança: $AME_k = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{P}(y_i=1)}{\partial x_{ki}}$, com significância pelo método delta ($\alpha = 0{,}05$) [Cameron 2005]. Mesmo operando na mesma escala de probabilidade, $E[\phi_k]$ e AME_k são quantidades matematicamente distintas: o AME é a derivada parcial média (estimador analítico global), enquanto o SHAP é a média das contribuições aditivas locais relativas a um valor de referência [Shapley et al. 1953, Lundberg and Lee 2017]. Concordâncias empíricas constituem evidência de estabilidade de efeitos; divergências refletem não linearidades, interações e heterogeneidade [Shmueli 2010, Hastie et al. 2009]. Três classes de métricas são aplicadas aos seis pares de combinações framework-classificador possíveis de comparação. Em todas elas, S e T denotam os dois rankings (ou conjuntos ordenados) de $m = 17$ variáveis obtidos pelos dois métodos de explicabilidade sendo comparados.

Correlações com ICs por bootstrap. Spearman ρ [Spearman 1987] e Kendall τ [Kendall 1938] sobre o ranking completo das $m = 17$ variáveis, com ICs de 95% por bootstrap ($n_{\text{boot}} = 5.000$ reamostras) — mais robusto que aproximações assintóticas para $m = 17$ [Efron and Tibshirani 1994].

RBO Extrapolado (RBO_{EXT}). [Webber et al. 2010] Medida de similaridade entre duas listas ordenadas (rankings) S e T, adequada para comparar rankings incompletos e não necessariamente conjuntos, dando peso maior às posições superiores. Seja $S_{:d}$ o prefixo de tamanho d da lista S (analogamente para $T_{:d}$), e seja $X_d = |S_{:d} \cap T_{:d}|$ o tamanho da interseção entre os prefixos top-d de S e T. O parâmetro $p \in (0, 1)$ controla a taxa de decaimento do peso atribuído a cada profundidade d: quanto menor p, maior a concentração do peso nas posições do topo do ranking. O RBO extrapolado, calculado a partir de uma avaliação até a profundidade k, é definido como

$$\text{RBO}_{\text{EXT}}(S, T, p, k) = \frac{X_k}{k} \cdot p^k \; + \; \frac{1-p}{p} \sum_{d=1}^{k} \frac{X_d}{d} \cdot p^d$$

No presente trabalho, utiliza-se $p = 0{,}9$ e $k = m = 17$ variáveis.

[2]Pacote `shap` para Python, usado para todos os valores SHAP deste trabalho.

Jaccard no top-k. $J_k(S,T) = |S_k \cap T_k|/|S_k \cup T_k|$ para $k \in \{5, 10\}$, onde S_k e T_k denotam os top-k subconjuntos de S e T, capturando respectivamente o núcleo crítico e o conjunto ampliado de variáveis relevantes, independentemente da ordenação interna.

4. Resultados

Em relação ao desempenho preditivo, os três modelos apresentam perfil semelhante (Tabela 1): acurácia em torno de 63–64%, recall para a classe grave entre 41% e 46%, e AUC $\approx$ 0,58. A acurácia obtida é inferior ao baseline trivial da classe majoritária (86,9%) como consequência direta da ponderação de classes (*class weighting*) para aumentar a sensibilidade à classe minoritária; portanto, não deve ser interpretada como a métrica operativa deste estudo, sendo complementada pela acurácia balanceada e pela PR-AUC (Tabela 1). O AUC próximo de 0,58 confirma o teto de desempenho imposto pela natureza dos dados administrativos [Mannering et al. 2020].

Diante desse poder discriminativo limitado, cabe ressaltar que SHAP e AME quantificam a contribuição de cada variável para a saída do *modelo*, não para o fenômeno subjacente [Molnar et al. 2020]; com a maior parte da variância não explicada pelas covariáveis, os rankings refletem o sinal parcial captado pelo modelo, e não uma estrutura causal robusta. Isso não invalida o exercício comparativo proposto: o objetivo central é avaliar a convergência entre métodos de explicação, não estabelecer determinantes causais da gravidade. A concordância entre SHAP e AME no núcleo estável (`uf_br_km`, `vel_max`, `frequencia`) é, nesse sentido, informativa por ser robusta à escolha do método, ainda que modesta em magnitude; já os rankings periféricos devem ser lidos como hipóteses para investigação futura, não como evidência definitiva.

A Tabela 2 consolida as quatro combinações framework-classificador. Um núcleo estável de duas variáveis emerge: `uf_br_km` no top-2 em todos os frameworks (1º no AME, no SHAP-Logístico e no SHAP-RF; 2º no SHAP-CatBoost) e `vel_max` entre 1º e 3º; `frequencia` no top-3 no AME, no SHAP-Logístico e no SHAP-RF, recuando ao 4º apenas no SHAP-CatBoost por diluição de correlação com `uf_br_km`. As principais divergências são: `condicao_ignorada` (4º no AME, mas 14º–16º nos SHAPs — diagnóstico da decomposição aditiva); `turno` (top-9 em todos, fase dominante 323° identificada pelo AME); `distancia_radar_m` (negativo no AME e próximo de zero no SHAP-Logístico, mas positivo nos SHAPs de ensemble — efeito heterogêneo, Seção 5); `tipo_pista` (3º–5º nos SHAPs, 13º no AME — interações não capturadas pela derivada parcial); `visibilidade_ruim` (significativa na Regressão Logística, $p = 0{,}044$, sinal negativo em todos); e `feriado` (irrelevante em todos, $p = 0{,}962$).

Tabela 1. Métricas de desempenho no conjunto de teste (2025). [1]Métricas para a classe grave. AUC $\approx$ 0,58 nos três modelos.

Modelo	Acurácia	Precisão[1]	Recall[1]	F1[1]
Regressão Logística	0,6397	0,1743	0,4159	0,2457
Random Forest	0,6270	0,1802	0,4634	0,2595
CatBoost	0,6375	0,1852	0,4616	0,2644

A Tabela 3 apresenta os testes de concordância para os seis pares, calculados sobre o ranking das 17 variáveis (cíclicas agregadas via $\sqrt{\sin^2 + \cos^2}$).

Tabela 2. Comparação de rankings entre as quatro combinações framework-classificador. Cíclicas: $|\mathrm{AME}| = \sqrt{\mathrm{AME}_{\sin}^2 + \mathrm{AME}_{\cos}^2}$; $E[|\phi_k|] = E[\sqrt{\phi_{\sin}^2 + \phi_{\cos}^2}]$. **(ns) =** $p > 0{,}05$**. Ordenado pelo AME.**

Variável	AME — Logística				SHAP — Logística			SHAP — RF			SHAP — CatBoost		
	AME	|AME|	p	Rk	$E[\phi_k]$	$E[\|\phi_k\|]$	Rk	$E[\phi_k]$	$E[\|\phi_k\|]$	Rk	$E[\phi_k]$	$E[\|\phi_k\|]$	Rk
uf_br_km	+3,901	3,901	<.001	1	+0,003	0,148	1	−0,002	0,111	1	−0,015	0,140	2
frequencia	+0,116	0,116	<.001	2	+0,001	0,009	3	+0,003	0,016	3	−0,002	0,023	4
vel_max	+0,067	0,067	<.001	3	+0,002	0,021	2	+0,070	0,070	2	+0,164	0,164	1
condicao_ignorada	+0,057	0,057	<.001	4	+0,001	0,001	14	+0,000	0,000	16	+0,000	0,000	16
pista_molhada	−0,037	0,037	<.001	5	−0,004	0,004	6	−0,002	0,002	9	−0,002	0,002	11
turno (cícl.)	—	0,036	<.001	6	—	0,026	9	—	0,016	6	—	0,023	7
declive	+0,031	0,031	<.001	7	+0,004	0,004	7	+0,001	0,001	12	+0,006	0,006	9
intersecao_de_vias	−0,019	0,019	<.001	8	−0,002	0,002	12	−0,000	0,000	15	−0,001	0,001	13
dia_semana (cícl.)	—	0,015	<.001	9	—	0,011	15	—	0,005	8	—	0,006	10
uso_solo	+0,014	0,014	<.001	10	+0,007	0,007	4	−0,000	0,005	7	−0,018	0,019	5
visibilidade_ruim	−0,013	0,013	.044	11	−0,002	0,002	11	−0,002	0,002	10	−0,002	0,002	12
reta	−0,013	0,013	<.001	12	+0,004	0,004	8	+0,001	0,001	11	+0,006	0,006	8
tipo_pista	+0,008	0,008	<.001	13	−0,003	0,005	5	+0,001	0,009	4	−0,022	0,024	3
distancia_radar_m	−0,007	0,007	.009	14	−0,000	0,001	13	+0,007	0,008	5	+0,011	0,014	6
mes (cícl.)	—	0,007	<.001	15	—	0,005	16	—	0,001	14	—	0,003	15
sentido_via	−0,004	0,004	.001	16	−0,002	0,002	10	−0,000	0,000	17	−0,000	0,000	17
feriado	≈ 0	≈ 0	.962	17 (ns)	≈ 0	≈ 0	17	−0,000	0,000	13	+0,000	0,001	14

Tabela 3. Concordância de rankings. IC 95% por bootstrap ($n = 5.000$**). RBO com** $p = 0{,}9$**.** J_5/J_{10}**: Jaccard no top-5/10.**

Par	ρ	IC 95%	τ	IC 95%	RBO	J_5/J_{10}
SHAP vs AME						
SHAP-RF vs AME	0,475	[−0,125; 0,845]	0,338	[−0,122; 0,705]	0,793	0,43/0,54
SHAP-Cat vs AME	0,466	[−0,144; 0,825]	0,309	[−0,147; 0,669]	0,655	0,43/0,54
SHAP-Logit vs AME	0,610	[0,096; 0,896]	0,471	[0,057; 0,785]	0,809	0,43/0,54
SHAP vs SHAP						
RF vs CatBoost	**0,944**	[0,792; 0,985]	**0,824**	[0,667; 0,951]	**0,820**	0,67/0,67
RF vs Logit	0,684	[0,257; 0,911]	0,515	[0,171; 0,794]	0,869	0,67/0,54
CatBoost vs Logit	0,779	[0,428; 0,920]	0,574	[0,277; 0,789]	0,784	1,00/0,67

Para os pares SHAP vs AME, a concordância global é heterogênea ($\rho \in [0{,}47; 0{,}61]$), variando de baixa (SHAP-CatBoost e SHAP-RF) a moderada (SHAP-Logístico), com ICs amplos que incluem zero em dois pares — reflexo do pequeno $n = 17$ e das divergências estruturais. O RBO, porém, permanece relativamente elevado ($\geq 0{,}65$), indicando que as discordâncias se concentram mais no meio e no fundo do ranking do que no topo. $J_5 = 0{,}43$ para todos os três pares — reflexo sobretudo de `condicao_ignorada` (Δ_{rank} entre 10 e 12) e `tipo_pista` (Δ entre 8 e 10). No top-10, $J_{10} = 0{,}54$ nos três pares (Figura 1).

Para SHAP vs SHAP, a concordância é sistematicamente superior à observada entre SHAP e AME. RF vs CatBoost apresenta a maior concordância do conjunto ($\rho = 0{,}944$; $\tau = 0{,}824$; RBO$= 0{,}820$; $J_{10} = 0{,}67$). Os pares envolvendo o SHAP-Logístico são mais heterogêneos: CatBoost vs Logit atinge $\rho = 0{,}779$, enquanto RF vs Logit cai para $\rho = 0{,}684$, com maiores discrepâncias concentradas em `distancia_radar_m` (Δ de 7–8) e `dia_semana` (Δ de 5–7). Nos dois pares envolvendo o SHAP-Logístico, o padrão RBO $>\rho$ se repete (RF vs Logit: RBO$= 0{,}869 > \rho = 0{,}684$; CatBoost vs Logit: RBO$= 0{,}784 > \rho = 0{,}779$), sugerindo que as divergências residuais são majoritariamente periféricas. Já em RF vs CatBoost — o par de maior concordância global — ocorre o

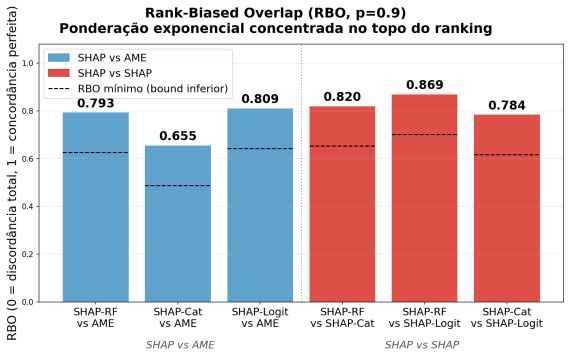


Figura 1. Rank-Biased Overlap (RBO, $p = 0{,}9$) para os seis pares de frameworks, com o RBO mínimo (sem extrapolação) indicado pela linha tracejada.

inverso (RBO$= 0{,}820 < \rho = 0{,}944$): com o ranking já quase idêntico ponta a ponta, não há margem para o RBO, que pondera o topo, superar um ρ próximo do máximo. A Figura 2 ilustra essa concordância por variável: `condicao_ignorada` é a que mais se afasta da diagonal nos três pares SHAP vs AME, enquanto os pares SHAP vs SHAP se concentram próximos à linha de identidade.

5. Discussão

A inclusão do SHAP-Logístico permite distinguir dois tipos de divergência. Quando AME e SHAP-Logístico concordam mas divergem dos ensembles — como em `pista_molhada` (sinal negativo nos quatro frameworks, com magnitude bem menor nos SHAPs) — a diferença é atribuível a não linearidades capturadas pelos modelos de árvore, ainda que sem inversão de sinal neste caso específico. Quando AME diverge do SHAP-Logístico para o *mesmo* modelo — como em `condicao_ignorada` (4º no AME, 14º no SHAP-Logístico) — a diferença decorre da forma como a decomposição aditiva trata variáveis com efeito mediado por correlações no background.

O caso de `distancia_radar_m` é o mais explicativo: AME negativo ($-0{,}0074$, consistente com radares instalados em segmentos de alto risco estrutural) versus SHAP-RF ($+0{,}0074$) e SHAP-CatBoost ($+0{,}0106$) positivos, com SHAP-Logístico próximo de zero ($-0{,}0003$). Sinal de que o efeito da fiscalização eletrônica é mediado por geometria, velocidade e histórico local [Wilson et al. 2010], e sensível ao método de explicação.

A estabilidade de `uf_br_km` no top-2 em todas as quatro combinações e de `vel_max` no top-2 em três delas fornece base robusta: gestão dinâmica de velocidade e manutenção prioritária em trechos de alta sinistralidade têm suporte convergente [Aarts and Van Schagen 2006]. A relevância de `turno` (fase $323°$) justifica fiscalização na transição noite-madrugada. A não significância de `feriado` deve ser interpretada com cautela, dado o desempenho preditivo moderado dos modelos, e não questiona operações de fiscalização focadas em feriados — cuja justificativa (maior movimentação de veículos e exposição ao risco) é independente desse resultado. Para `distancia_radar_m`, a divergência de sinais recomenda análises espacialmente desagregadas. O $J_{10} = 0{,}54$ em todos os pares SHAP vs AME demonstra que o conjunto ampliado de variáveis relevantes é parcialmente compartilhado, ainda que a sobreposição não seja tão extensa quanto entre pares SHAP vs SHAP.

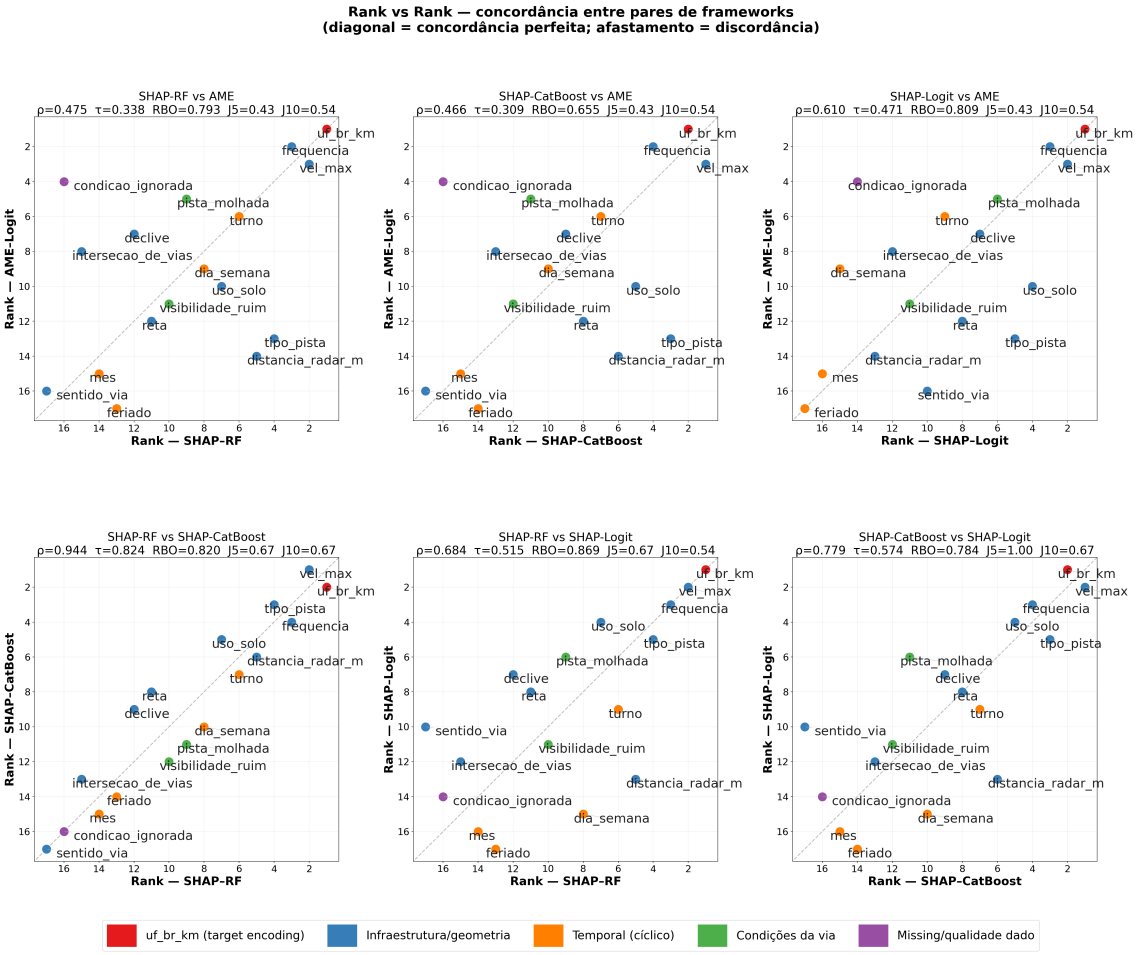


Figura 2. Rank vs. rank para os seis pares de frameworks (diagonal = concordância perfeita; cores indicam a categoria da variável).

6. Conclusão

Quanto à concordância entre AME e SHAP para a Regressão Logística (Q1), os dois frameworks apresentam concordância moderada mesmo operando sobre o mesmo modelo ($\rho = 0{,}610$; $J_5 = 0{,}43$) — na verdade, a maior concordância entre os três pares SHAP-vs-AME testados. O par com menor concordância do conjunto (SHAP-CatBoost vs AME, $\rho = 0{,}466$) envolve dois *modelos diferentes*, não o mesmo modelo, o que qualifica a hipótese inicial: a forma de decomposição (aditiva vs. derivada parcial) não é, isoladamente, o fator dominante da divergência — o classificador subjacente também importa, e nem sempre menos do que o framework de explicação. O caso mais expressivo permanece `condicao_ignorada`, deslocada 10 posições entre os dois frameworks para a mesma Regressão Logística.

Quanto à estabilidade do SHAP entre diferentes classificadores (Q2), a concordância é sistematicamente superior à observada entre SHAP e AME, mas não uniforme: quase perfeita entre os ensembles ($\rho = 0{,}944$; RBO$= 0{,}820$) e moderada a alta nos pares com o SHAP-Logístico ($\rho \in [0{,}68; 0{,}78]$). O padrão RBO $>$ ρ se mantém na maioria dos pares, indicando que boa parte das divergências residuais é periférica, embora de forma menos acentuada do que inicialmente observado.

Ambas as respostas qualificam parcialmente a hipótese central: `uf_br_km` e `vel_max` compõem um núcleo estável no topo do ranking em todos os frameworks e classificadores, com `frequencia` no top-4 em todos eles. A dominância de `uf_br_km`, contudo, é um sinal de heterogeneidade espacial não explicada — ela captura, de forma agregada, tudo aquilo que não foi medido em cada trecho: geometria local, histórico de manutenção, padrão de uso. Identificar quais atributos físicos e operacionais concretos sustentam essa relevância é, portanto, a agenda natural para trabalhos futuros, tanto pela incorporação de variáveis de infraestrutura mais granulares quanto pela análise espacialmente desagregada do efeito da fiscalização eletrônica, cujos sinais conflitantes em `distancia_radar_m` reforçam essa necessidade.

O presente trabalho possui limitações que merecem registro. O número reduzido de atributos ($n = 17$) produz intervalos de confiança amplos para as estatísticas de ranking, podendo limitar a significância estatística de comparações individuais. O teto de desempenho preditivo (AUC $\approx$ 0,58) restringe a leitura das importâncias como reflexo do fenômeno real e não apenas do sinal parcial apresentado pelo modelo. Os resultados derivam de uma única base e agência (PRF), o que limita a generalização a outros contextos sobre segurança viária. Por fim, efeitos de codificação podem influenciar a magnitude e o ranking dos efeitos estimados, de forma distinta entre AME e SHAP.

Agradecimentos

Os autores agradecem ao CNPq, à FAPEMIG e à PROPPI/UFOP pelo suporte financeiro ao projeto de pesquisa.

Referências

Aarts, L. and Van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, 38(2):215–224.

Anselin, L. (1988). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.

Brasil (1997). Código de trânsito brasileiro (lei nº 9.503, de 23 de setembro de 1997). Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9503.htm. Acesso em: 2 out. 2024.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bucsuházy, K., Matuchová, E., Zvala, R., Moravcová, P., Kostíková, M., and Mikulec, R. (2020). Human factors contributing to the road traffic accident occurrence. *Transportation research procedia*, 45:555–561.

Cameron, A. (2005). Microeconometrics: methods and applications. *Cambridge University*.

Carvalho, C. H. R. d. (2020). Custos dos acidentes de trânsito no brasil: Estimativa simplificada com base na atualização das pesquisas do ipea sobre custos de acidentes nos aglomerados urbanos e rodovias.

Chen, J., Liu, P., Wang, S., Zheng, N., and Guo, X. (2025). Prediction and interpretation of crash severity using machine learning based on imbalanced traffic crash data. *Journal of Safety Research*, 93:185–199.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

Departamento Nacional de Infraestrutura de Transportes (DNIT) (2024). Controle de velocidade — dados abertos dnit. `https://servicos.dnit.gov.br/dadosabertos/dataset/controle-de-velocidade`. Acesso em: 02 out. 2024.

Dong, S., Khattak, A., Ullah, I., Zhou, J., and Hussain, A. (2022). Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with shapley additive explanations. *International journal of environmental research and public health*, 19(5):2925.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Franceschi, L., Kaesemodel, L., de Vargas, V. d. C. C., Konrath, A. C., Nakamura, L. R., Ramires, T. G., Barreto, C. B. M., and Valente, A. M. (2022). Factors related to highway crash severity in brazil through a multinomial logistic regression model. *Transportes*, 30(1):2566–2566.

Hameed, I. and Roth, A. (2024). Statistical significance of feature importance rankings. *arXiv preprint arXiv:2401.15800*.

Harrell, F. E. et al. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Khanum, H., Garg, A., Faheem, M. I., and Kulkarni, R. (2025). A methodological framework for road accident severity prediction for indian highways using machine learning models. *MethodsX*, page 103728.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mafort, E. V. and Kappel, M. A. (2025). 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*, 23:244–252.

Mannering, F., Bhat, C. R., Shankar, V., and Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic methods in accident research*, 25:100113.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD explorations newsletter*, 3(1):27–32.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B. (2020). Pitfalls to avoid when interpreting machine learning models.

Polícia Rodoviária Federal (2026). Dados abertos da prf. `https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf`. Acesso em: jan. 2026.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Saarela, M. and Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, (2).

SHAP developers (2026). shap.treeexplainer — SHAP latest documentation. `https://shap.readthedocs.io/en/latest/generated/shap.TreeExplainer.html`. Accessed: 2026-08-14.

Shapley, L. S. et al. (1953). A value for n-person games.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, pages 289–310.

Spearman, C. (1987). The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

Swamy, V., Radmehr, B., Krco, N., Marras, M., and Käser, T. (2022). Evaluating the explainers: black-box explainable machine learning for student success prediction in moocs. *arXiv preprint arXiv:2207.00551*.

Wang, C. and Serre, T. (2025). A hybrid approach to investigating factors associated with crash injury severity: Integrating interpretable machine learning with logit model. *Applied Sciences*, 15(19):10417.

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Wen, X., Xie, Y., Jiang, L., Pu, Z., and Ge, T. (2021). Applications of machine learning methods in traffic crash severity modelling: current status and future directions. *Transport reviews*, 41(6):855–879.

Wilson, C., Willis, C., Hendrikz, J. K., Le Brocque, R., and Bellamy, N. (2010). Speed cameras for the prevention of road traffic injuries and deaths. *Cochrane database of systematic reviews*, (11).

Zhang, X., Wen, H., Yamamoto, T., and Zeng, Q. (2021). Investigating hazardous factors affecting freeway crash injury severity incorporating real-time weather data: Using a bayesian multinomial logit model with conditional autoregressive priors. *Journal of safety research*, 76:248–255.