UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS EXATAS

DEPARTAMENTO DE ESTATÍSTICA

Luciano Gustavo Martins Rocha

CONSTRUÇÃO DE UM MODELO DE CLASSIFICAÇÃO PARA PREVER A OCORRÊNCIA DE ACIDENTE DE TRÂNSITO COM VÍTIMAS NÃO FATAIS EM BELO HORIZONTE

Belo Horizonte

2025

Luciano Gustavo Martins Rocha

CONSTRUÇÃO DE UM MODELO DE CLASSIFICAÇÃO PARA PREVER A OCORRÊNCIA DE ACIDENTE DE TRÂNSITO COM VÍTIMAS NÃO FATAIS EM BELO HORIZONTE

Monografia de Especialização apresentada ao Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística Computacional Aplicada.

Orientador: Prof. Dr. Guilherme Lopes de Oliveira

Belo Horizonte

2025

Rocha, Luciano Gustavo Martins.

R672c Construção de um modelo de classificação para prever a ocorrência de acidente de trânsito com vítimas não fatais em Belo Horizonte [recurso eletrônico] / Luciano Gustavo Martins Rocha - 2025.

1 recurso online (38 f. il., color.) : pdf.

Orientador: Guilherme Lopes de Oliveira.

Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 36-38.

1. Estatística. 2. Análise de regressão logística. 3. Acidentes de trânsito - Controle preditivo. 4. Acidentes de trânsito - Belo Horizonte. I. Oliveira, Guilherme Lopes de. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação
Caixa Postal 702
31270-901 Belo Horizonte- MG – Brasil

Telefone (31) 3409-5923
Fax (31) 3499-5924
E-mail: pgest@ufmg.br
WEB: http://www.est.ufmg.br/posgrad/

ATA DO 344ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE LUCIANO GUSTAVO MARTINS ROCHA.

Aos sete dias do mês de abril de 2025, às 15:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística Computacional Aplicada, para julgar a apresentação do trabalho de fim de curso do aluno **Luciano Gustavo Martins Rocha**, intitulado: "*Construção de um modelo de classificação para prever a ocorrência de acidente de trânsito com vítimas não fatais em Belo Horizonte"*, como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Guilherme Lopes de Oliveira – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 04 de abril de 2025.

gov.br Documento assinado digitalmente
GUILHERME LOPES DE OLIVEIRA
Data: 07/04/2025 16:05:31-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Guilherme Lopes de Oliveira (orientador)
DECOM/CEFET-MG

gov.br Documento assinado digitalmente
GABRIELA OLIVEIRA
Data: 07/04/2025 17:14:12-0300
Verifique em https://validar.iti.gov.br

Prof. Dra. Gabriela Oliveira
IFMG

AGRADECIMENTOS

Agradeço a Deus por me conduzir até aqui, apesar das dificuldades que surgiram; ao meu querido pai e a minha querida mãe pela oportunidade da vida.

RESUMO

O trânsito das grandes metrópoles é um fenômeno complexo e dinâmico, que se altera ininterruptamente. Este trabalho tem o objetivo de aplicar o modelo de regressão logística para analisar os fatores que influenciam nas chances de ocorrência de acidentes de trânsito com vítimas não fatais, tendo como foco os diferenciais entre os fatores dias da semana, meses do ano, faixas de horário e tipos de logradouro nos quais o acidente ocorreu. Os dados utilizados referem-se aos acidentes ocorridos em 2023, na regional Centro-Sul do município de Belo Horizonte, Minas Gerais, Brasil. Os dados são provenientes do Sistema do Centro de Operações da Prefeitura Municipal de Belo Horizonte (SICOP-BH). Todas as variáveis mencionadas se mostraram estatisticamente significativas ao nível de 10% para explicar a ocorrência de acidente de trânsito com vítimas não fatais. Contudo, para as variáveis mês e faixa de horário, algumas categorias não se mostraram significativas. Por exemplo, os meses de fevereiro, março e abril, quando tomados em relação ao mês de janeiro, não trazem consigo nenhuma informação relevante. O classificador apresentou uma área sob a curva ROC de aproximadamente 80.0%, sugerindo uma capacidade razoável de classificar a ocorrência do fenômeno. Apesar de apresentar algumas limitações, a estratégia de modelagem aplicada nesse estudo evidencia aspectos quantitativos do fenômeno em questão e pode auxiliar na busca de uma administração pública cada vez mais baseada em evidências, visando a uma melhor gestão do trânsito e na correta alocação de recursos humanos e materiais da Administração Pública Municipal nessa área.

Palavras-chave: acidente de trânsito; regressão logística; classificador; Belo Horizonte.

ABSTRACT

The traffic in large metropolitan areas is a complex and dynamic phenomenon that changes continuously. This study aims to apply the logistic regression model to analyze the factors influencing the likelihood of traffic accidents with non-fatal victims, focusing on the differences between the factors: day of the week, month of the year, time range, and type of road where the accident occurred. The data used refer to accidents that occurred in 2023 in the Centro-Sul region of Belo Horizonte, Minas Gerais, Brazil. The data were obtained from the Operations Center System of the Belo Horizonte City Hall (SICOP-BH). All the mentioned variables were statistically significant at the 10% level in explaining the occurrence of traffic accidents with non-fatal victims. However, for the variables "month" and "time range," some categories were not significant. For example, the months of February, March, and April, when compared to January, did not provide any relevant information. The classifier achieved a ROC curve area of approximately 80.0%, suggesting a reasonable ability to classify the occurrence of the phenomenon. Despite some limitations, the modeling strategy applied in this study highlights quantitative aspects of the phenomenon and can contribute to the advancement of evidence-based public administration. This approach aims to improve traffic management and optimize the allocation of human and material resources within the municipal administration in this area.

Keywords: traffic accident; logistic regression; classifier; Belo Horizonte.

LISTA DE FIGURAS E TABELAS

SUMÁRIO

1. INTRODUÇÃO

O trânsito das grandes metrópoles constitui-se em um fenômeno de alta complexidade e envolve diversos fatores que vão desde a infraestrutura urbana até o comportamento de motoristas e pedestres, ensejando a atuação dos órgãos do poder público responsáveis pelo planejamento e fiscalização de trânsito na busca da incolumidade e da segurança pública de seus cidadãos. Neste contexto, os acidentes de trânsito são eventos relevantes de estudo devido aos seus efeitos tanto na saúde de motoristas e pedestres quanto na logística do tráfego como um todo.

De acordo com Rozestraten (1988, *apud* Zampiere, 2010), a maioria dos acidentes de trânsito tem o fator humano como principal fator, devido à conduta dos motoristas. Por outro lado, alguns acidentes de trânsito têm as condições naturais e/ou infraestrutura urbana e das vias como fatores de ocorrência.

Nesse sentido, com o intuito de nortear políticas públicas eficientes para a redução de acidentes de trânsito, o uso da estatística e da ciência de dados para o correto entendimento de fatores (variáveis) que sejam relevantes para explicar o fenômeno em questão faz-se necessário. Ademais, a construção de modelos de classificação capazes de prever situações adversas no trânsito urbano (ainda que com um certo grau de incerteza) auxilia na tomada de decisões estratégicas e a devida alocação dos recursos humanos e materiais (que são sabidamente escassos), o que corrobora com o princípio constitucional da Eficiência da Administração Pública.

De fato, estudos qualitativos e quantitativos têm utilizados métodos estatísticos para avaliar fatores ligados ao trânsito. Saraiva e dos Santos (2024), por exemplo, abordam a modelagem temporal de óbitos em acidentes de trânsito no Brasil, aplicando modelos de regressão e de *Machine Learning* com dados de 2011 a 2023. Também utilizando-se de informações sobre regionais, tipo e circunstâncias do acidente, além de informações relacionadas aos dias da semana e faixa de horário, Goulart (2018) apresentou uma modelagem espacial dos acidentes de trânsito ocorridos no município de Montes Claros, Minas Gerais.

Especialmente, Cunha (2019) apresentou um estudo de caso no município de Belo Horizonte, Minas Gerais, envolvendo dados de acidentes de trânsito de 2011 a 2015, onde foram analisados fatores como o perfil do motorista envolvido, tipo de veículo e severidade do acidente, além do dia da semana, horário e logradouro. A partir da análise descritiva dos dados,

verificou-se que a maior ocorrência de acidentes de trânsito acontece durante a semana, de segunda à quarta-feira, período da tarde, na região central da cidade, onde possivelmente há um volume de tráfego maior. A autora limitou-se, no entanto, à parte descritiva e qualitativa do fenômeno, não tendo aplicado alguma modelagem estatística robusta para, por exemplo, estimar as diferenças de risco de acidentes com base nos fatores considerados.

Dentro desse escopo, e considerando que acidentes de trânsito tendem a ocorrer mais comumente em locais, horários e épocas em que haja maior concentração de veículos e pedestres em circulação, é interessante e necessário quantificar de que forma isto ocorre.

Assim, os objetivos deste trabalho são, utilizando ferramentas de *Machine Learning*, ajustar um modelo de regressão logística, que auxiliará no entendimento dos fatores (variáveis) disponíveis no banco de dados (dia da semana, mês, horário do dia e tipo de logradouro) que são estatisticamente significativas para explicar o fenômeno '*acidente de trânsito com vítimas não fatais'* em um grande município do Brasil e, baseado nestes fatores (variáveis): **1)** estabelecer a probabilidade e a chance (*odds*) de este fenômeno ocorrer; **2)** ajustar um modelo de classificação, com intuito de predizer se o fenômeno ocorrerá ou não ocorrerá.

Mais especificamente, serão considerados dados do município de Belo Horizonte, Minas Gerais, em 2023. A escolha de aplicação para este ano considerou o intuito de realizar um estudo que seja mais contemporâneo, levando em conta o fato de que o trânsito é um fenômeno dinâmico e que apresenta alterações ao longo do tempo.

No banco de dados coletado, os acidentes de trânsito são subdivididos em três classes: **1)** acidente de trânsito sem vítimas; **2)** acidente de trânsito com vítimas não fatais; **3)** acidente de trânsito com vítimas fatias. A classe "*acidente de trânsito com vítimas não fatais*" é disparadamente a classe de maior frequência. Portando, apenas esta classe de ocorrências será objeto de estudo. Além disso, considerando que o trânsito de cada regional no município de Belo Horizonte (e mesmo de cada bairro dentro de uma mesma regional) apresenta características muito específicas, com o intuito de realizar um estudo que seja mais local e homogêneo, apenas as ocorrências da regional Centro-Sul serão consideradas, sendo esta a regional com maior fluxo diário de pessoas e veículos.

A organização da sequência deste trabalho está feita de forma que: no Capítulo 2 são apresentados os dados e métodos estatísticos aplicados; no Capítulo 3, são apresentados os resultados da análise descritiva e da modelagem dos dados; e no Capítulo 4 são feitas as discussões e considerações finais.

2. MATERIAIS E MÉTODOS

2.1 Definição da Base de Dados

Os dados são provenientes do Sistema do Centro de Operações da Prefeitura Municipal de Belo Horizonte (SICOP-BH)[1] e referem-se aos acidentes de trânsito com vítimas não fatais ocorridos em Belo Horizonte, na regional Centro-Sul, em 2023.

O SICOP-BH registra apenas as ocorrências ditas INTEGRADAS, as quais caracterizam-se por ter sua resolução realizada por mais de uma das instituições que compõem o Centro de Operações da Prefeitura Municipal de Belo Horizonte (COP-BH).

Para melhor entendimento do que se trata uma ocorrência integrada, considere o seguinte exemplo hipotético: "*um motorista, alcoolizado colide em um poste de luz. Como consequência da colisão, o motorista fica gravemente ferido e o poste de luz cai, arrebentando os fios de eletricidade e deixando-os expostos*". No atendimento e resolução do acidente mencionado nesse exemplo, seria necessária a atuação de mais de uma instituição pública: **1)** a Companhia Energética local, para assegurar o desligamento dos fios que porventura estivessem energizados; **2)** o SAMU ou o Corpo de Bombeiros Militar, para prestar socorro ao motorista ferido; **3)** a Guarda Civil Municipal de Belo Horizonte ou a Polícia Militar de Minas Gerais, para autuar o motorista pelo crime de dirigir embriagado; **4)** a Gerência de Manutenção, para realizar reparos estruturais que porventura fossem necessários.

Portanto, visto que o SICOP-BH não registra ocorrências que não sejam integradas, o banco de dados que será utilizado nesse trabalho pode ser considerado como sendo um "recorte" do fenômeno em estudo. Em princípio, este banco de dados contém 1.530 observações (cada observação representa um acidente de trânsito com vítimas não fatais ocorridos em Belo Horizonte, na regional Centro-Sul, em 2023) e 4 variáveis categóricas, mencionadas na Tabela 1 a seguir.

Tabela 1 - Descrição das variáveis contidas no banco de dados

NOME DA VARIÁVEL	DESCRIÇÃO	CATEGORIAS
DIA_SEMANA	Refere-se ao dia da semana em que o acidente de trânsito com vítimas não fatais ocorreu	DOMINGO, SEGUNDA-FEIRA, TERÇA-FEIRA, QUARTA-FEIRA, QUINTA-FEIRA, SEXTA-FEIRA, SÁBADO
MÊS	Refere-se ao mês em que o acidente de trânsito com vítimas não fatais ocorreu	JANEIRO, FEVEREIRO, MARÇO, ABRIL, MAIO, JUNHO, JULHO, AGOSTO, SETEMBRO, OUTUBRO, NOVEMBRO, DEZEMBRO
FAIXA_HORARIO	Refere-se a faixa de horário em que o acidente de trânsito com vítimas não fatais ocorreu	ENTRE 0H E 02H59, ENTRE 3H E 05H59, ENTRE 6H E 08H59, ENTRE 9H E 11H59, ENTRE 12H E 14H59, ENTRE 15H E 17H59, ENTRE 18H E 20H59, ENTRE 21H E 23H59
LOGRADOURO	Refere-se ao tipo de logradouro em que o acidente de trânsito com vítimas não fatais ocoreu	AVE (Avenida), ROD (Rodovia), RUA (Rua)
VALOR_LÓGICO	Refere-se à variável resposta	Assume 1, caso um acidente de trânsito com vítimas não fatais tenha ocorrido; assume 0, caso contrário

[1] O SICOP-BH é um sistema de uso interno e restrito.

2.1.1 Definição da Variável Resposta Binária

No estudo e uso das técnicas de regressão logística, a ocorrência do fenômeno que estiver sendo objeto de estudo é denominada por ***evento*** ou ***sucesso*** (a não ocorrência do fenômeno que estiver sendo objeto de estudo é denominada por ***não evento*** ou ***fracasso*** ou ***falha***).

Nesta monografia, a ocorrência de um "*acidente de trânsito com vítimas não fatais ocorridos em Belo Horizonte, na regional Centro-Sul, em 2023*" será denominada ***evento***; a não ocorrência será denominada ***não evento***.

Sistemas operacionais como o SICOP-BH ou o Registro de Eventos de Defesa Social (sistema que formaliza ocorrências de fatos policiais das Polícias Civil e Militar do Estado de Minas Gerais) registram apenas os ***eventos*** (ou seja, o sistema respectivo apenas realiza um registro quando há a ocorrência de, por exemplo, um acidente de trânsito ou de um furto de celular). Porém, a regressão logística exige que o banco de dados também contenha os ***não eventos***, para que ele seja capaz de "entender" ou "aprender" quais são os fatores (variáveis) que são de fato importantes para explicar o ***evento*** e o ***não evento***.

O problema de se utilizar um banco de dados proveniente de sistemas operacionais, que naturalmente registrarão apenas os ***eventos***, é que o algoritmo entenderá que, submetido a certos fatores (variáveis), o ***evento*** é certo, ou seja, há uma probabilidade de 100% de ocorrer.

Para corrigir o problema em questão, um banco de dados conveniente foi "criado" a partir dos dados provenientes do SICOP-BH, utilizando o *software* R (R CORE TEAM, 2024). O "novo" banco de dados, o qual será denominado por Banco de Dados para Modelagem, foi construído de acordo com os seguintes passos: **Passo 1)** para cada um dos 365 dias do ano, 24 observações (linhas do banco de dados) foram criadas. Essas 24 observações são a combinação de 8 faixas de horários e 3 tipos de logradouros. Para melhor entendimento, considere a Tabela 2 abaixo. **Passo 2)** o valor lógico "1" foi atribuído àquelas observações que contivessem as variáveis para as quais havia registro, no SICOP, de acidente de trânsito com vítimas não fatias (***evento*** = 1). Ao contrário, o valor lógico "0" foi atribuído àquelas observações que contivessem as variáveis para as quais não havia registro, no SICOP, de acidente de trânsito com vítimas não fatais (***não evento*** = 0).

Tabela 2 - Filtro do banco de dados

DIA	DIA_SEMANA	MES	FAIXA_HORARIO	LOGRADOURO	REGIONAL	VALOR_LOGICO
7	TERÇA-FEIRA	MARÇO	ENTRE 0H E 02H59	AVE	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 3H E 05H59	AVE	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 6H E 08H59	AVE	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 9H E 11H59	AVE	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 12H E 14H59	AVE	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 15H E 17H59	AVE	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 18H E 20H59	AVE	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 21H E 23H59	AVE	CENTRO-SUL	1
7	TERÇA-FEIRA	MARÇO	ENTRE 0H E 02H59	ROD	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 3H E 05H59	ROD	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 6H E 08H59	ROD	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 9H E 11H59	ROD	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 12H E 14H59	ROD	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 15H E 17H59	ROD	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 18H E 20H59	ROD	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 21H E 23H59	ROD	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 0H E 02H59	RUA	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 3H E 05H59	RUA	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 6H E 08H59	RUA	CENTRO-SUL	1
7	TERÇA-FEIRA	MARÇO	ENTRE 9H E 11H59	RUA	CENTRO-SUL	1
7	TERÇA-FEIRA	MARÇO	ENTRE 12H E 14H59	RUA	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 15H E 17H59	RUA	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 18H E 20H59	RUA	CENTRO-SUL	0
7	TERÇA-FEIRA	MARÇO	ENTRE 21H E 23H59	RUA	CENTRO-SUL	0

Para este banco de dados, para alguns dias do ano, há mais de 24 observações. Esse fato aconteceu porque há combinações (faixas de horários e tipos de logradouros) nas quais mais de um acidente trânsito com vítimas não fatais ocorreu. Por exemplo, o dia 15/03/2023 apresenta 25 observações, já que na faixa de horário entre 09:00 e 11:59, em uma avenida da regional Centro sul, houve 02 acidentes de trânsito com vítimas não fatais.

Ao final dos dois passos mencionados, visto que há combinações nas quais mais de um acidente trânsito com vítimas não fatais aconteceu, o Banco de Dados para Modelagem apresentou 216 observações a mais do que se esperaria para a combinação entre 24 observações por dia do ano e 365 dias no ano (24 x 365 = 8.760 observações esperadas). O que se considerou para a modelagem, portanto, é um banco de dados composto por 8.976 observações, sendo 1.530 ***eventos*** e 7.446 ***não eventos***[2].

[2] O Banco de Dados para Modelagem apresenta as seguintes proporções: 17,0% de ***eventos*** e 83,0% de ***não eventos***.

2.1.2 Separação da Base de Treino e de Teste

Finalmente, com intuito de testar a qualidade do modelo de classificação que será ajustado, o Banco de Dados para Modelagem foi dividido em duas partes: 1) **banco de treinamento**, composto por 70% das observações (6.283 observações) e 2) **banco de teste**, composto por 30% das observações (2.693 observações).

2.2 Regressão Logística

A Regressão Logística é uma técnica estatística utilizada para determinar a probabilidade média de ocorrência do fenômeno que está sendo objeto de estudo - representado pela variável dependente Y (CORDEIRO; DEMÉTRIO e MORAL, 2024).

Diferentemente das técnicas de regressão linear, onde a variável dependente Y se apresenta na forma quantitativa e um dos objetivos é modelar a relação quantitativa entre as variáveis, na técnica de regressão logística a variável dependente Y se apresenta na forma qualitativa (categórica) e um dos objetivos é modelar a probabilidade média de o fenômeno representado por Y ocorrer, com base no comportamento de variáveis explicativas X. Quando a variável dependente Y apresenta apenas duas categorias, a regressão é dita Regressão Logística Binomial; quando a variável dependente Y apresenta mais de duas categorias, a regressão é dita Regressão Logística Multinomial (MONTGOMERY; RUNGER, 2016).

Segundo FÁVERO e BELFIORE (2017), no caso da Regressão Logística Binomial, a expressão geral da probabilidade média estimada de o fenômeno em estudo ocorrer, para cada observação do banco de dados, é dada pela seguinte expressão:

$$\mathrm{P}(Yi = 1|Xi) = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+ \ldots +\beta_k X_{ki}}}{1 + e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+ \ldots +\beta_k X_{ki}}}$$

onde $P(Yi = 1 \mid Xi)$ representa a probabilidade de o fenômeno $\mathrm{Y_i}$ ocorrer dado que certo valor da variável explicativa $\mathrm{X_i}$ ocorreu; $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$ representa o *logito*; β_0 representa a constante (intercepto); $\beta_1, \beta_2, \ldots, \beta_k$ representam os parâmetros estimados para cada variável explicativa X e o subscrito i representa cada observação da amostra (i = 1, 2, 3, ... , n, em que n é o tamanho da amostra).

Além disto, destacam que os conceitos de probabilidade e de chance (*odds*) são distintos. A chance (*odds*) representa a relação existente entre a probabilidade de um fenômeno ocorrer e a probabilidade de ele não ocorrer. Matematicamente, a chance (*odds*) é dada pela seguinte expressão:

$$Chance\ (odds) = \frac{p}{(1-p)}$$

onde p = a probabilidade de o fenômeno ocorrer e $(1 - p)$ = a probabilidade de o fenômeno não ocorrer. Logo, a chance (*odds*) é dada pela exponencial do *logito*, de acordo com a seguinte expressão:

$$Chance\ (odds) = e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+ \ldots +\beta_k X_{ki}}$$

Todavia, no estudo da regressão logística, p não é conhecida *a priori* e deve, portanto, ser estimada. Isto é feito diretamente através das estimativas de máxima verossimilhança para os parâmetros $\beta_0,\ \beta_1, \ldots, \beta_p$.

Teste de hipóteses apropriados são utilizados para a análise da significância desses parâmetros, sendo a hipóteses nula $Ho: \beta\_j = 0$ para j = 1, ..., p. A decisão é tomada com base em um certo nível de significância[3], que representa a probabilidade máxima aceita de se cometer o **erro tipo I** (rejeitar a hipótese nula, sendo ela verdadeira) ao ser comparado com o p-valor (ou *p-value*) do teste, que é a probabilidade de se observar um resultado tão extremo (ou mais) quanto o resultado obtido nos dados, supondo que a hipótese nula seja verdadeira. Detalhes sobre os testes específicos no caso da regressão logística podem ser vistos, por exemplo, em Cordeiro, Demétrio e Moral (2024).

Nas aplicações de técnicas de *Machine Learning* para classificação, há técnicas cujo foco é exclusivamente a predição; mas também há técnicas que apresentam um foco inferencial, além da predição. A regressão logística e uma destas técnicas que, a depender dos interesses do pesquisador, pode ser utilizada para fazer inferência ou para fazer predição ou para ambas as finalidades. Nesta monografia, a regressão logística será utilizada para fazer inferência e predição.

2.3 Avaliação da Capacidade Preditiva do Modelo Logístico

2.3.1 Conceitos e Definições

Curva ROC (***Receiver Operating Characteristic***) - A Curva ROC é um gráfico que avalia o desempenho de um modelo de classificação binário para diferentes valores de *cut-off*. Ela

[3] Nesta monografia, o Nível de Significância que será utilizado é de 10%.

compara a taxa de verdadeiros positivos (sensitividade) com a taxa de falsos positivos (1 - especificidade) (CORDEIRO; DEMÉTRIO e MORAL, 2024).

Área Abaixo da Curva (AUC - *Area Under the Curve*) - A AUC é um índice número que mede a qualidade da Curva ROC. Ela indica o quão bem o modelo classifica corretamente os ***eventos*** e os ***não eventos*** em um problema de classificação binária (CORDEIRO; DEMÉTRIO e MORAL, 2024).

Sensitividade - A sensitividade (ou *Recall*, ou *True Positive Rate* - TPR) mede a capacidade de o modelo identificar corretamente os ***eventos*** (casos positivos) (CORDEIRO; DEMÉTRIO e MORAL, 2024).

$$Sensitividade = \frac{VP}{VP + FN}$$

onde: VP = Verdadeiros Positivos (observações que são ***eventos*** e o modelo classifica corretamente como ***eventos***) e FN = Falsos Negativo (observações que são ***eventos*** e o modelo classifica incorretamente como ***não eventos***).

Especificidade - A especificidade (ou *True Negative Rate* - TNR) mede a capacidade de o modelo identificar corretamente os ***não eventos*** (casos negativos) (CORDEIRO; DEMÉTRIO e MORAL, 2024).

$$Especificidade = \frac{VN}{VN + FP}$$

onde: VN = Verdadeiros Negativos (observações que são ***não eventos*** e o modelo classifica corretamente como ***não eventos***) e FP = Falsos Positivos (observações que são ***não eventos*** e o modelo classifica incorretamente como ***eventos***).

Acurácia - A acurácia é uma métrica utilizada para avaliar o desempenho de modelos de classificação. Ela mede a proporção de previsões corretas em relação ao total de previsões feitas (CORDEIRO; DEMÉTRIO e MORAL, 2024).

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN}$$

onde: VP = Verdadeiros Positivos; VN = Verdadeiros Negativos; FP = Falsos Positivos e FN = Falsos Negativo.

Cut-off* -** O *cut-off* (também denominado por limiar ou *threshold*) é um valor entre 0 e1 usado para converter as probabilidades preditas pelo modelo em classes. Ele define a partir de qual probabilidade uma observação será classificada como ***evento (1) ou ***não evento*** (0) (CORDEIRO; DEMÉTRIO e MORAL, 2024).

Matriz de Confusão - A Matriz de Confusão é uma tabela que resume o desempenho de um modelo de classificação comparando as previsões do modelo com os valores reais (observados). Ela é amplamente usada para avaliar modelos de classificação (CORDEIRO; DEMÉTRIO e MORAL, 2024).

2.3.2 Métricas de Avaliação

Segundo CORDEIRO, DEMÉTRIO e MORAL, 2024), um modelo de classificação é tanto melhor quanto mais próximo de 1 a AUC estiver. Ao contrário, um modelo de classificação é tanto pior quanto mais próximo de 0.5 a AUC estiver. Uma AUC igual a 1 significa que o modelo é capaz de classificar corretamente todas as observações do banco de dados[4]. Ao contrário, uma AUC igual a 0.5 indica que o modelo é completamente aleatório e realiza classificações ao acaso.

A Curva ROC é construída a partir dos seguintes passos: **Passo 1)** tomando um número suficientemente grande de *cut-offs* (valores entre 0 e 1 a partir do qual as probabilidades estimadas irão indicar um ***evento***), o algoritmo calcula a sensitividade e a especificidade para um *cut-off* específico e plota esse ponto no gráfico - no eixo y plota-se a Sensitividade e no eixo x plota-se a diferença entre 1 e a Especificidade (1 - Especificidade)[5]. **Passo 2**) a partir da origem (especificidade = 1, sensitividade = 0), o gráfico é construído considerando um *cut-off* muito próximo à maior probabilidade média de o ***evento*** ocorrer, estimada pela função *predict* do pacote *stats*. O processo se repete para diversos *cut-offs*, até que o gráfico atinja o ponto de especificidade = 0 e sensitividade = 1. Neste ponto, o gráfico é construído considerando um *cut-off* muito próximo à menor probabilidade média de o ***evento*** ocorrer, estimada pela função *predict* do pacote *stats*.

A interpretação básica da Curva ROC, em especial nos pontos P_1 (especificidade = 1, sensitividade = 0) e P_2 (especificidade = 0, sensitividade = 1) é: em P_1, o *cut-off* é tão alto que,

[4] Na prática, uma AUC = 1 só seria possível se não houvesse observações que se sobrepusessem umas às outras (o que é praticamente impossível de acontecer). Nesse caso, haveria o *cut-off* para o qual sensitividade = 1 e especificidade = 1. Essa AUC é denominada por AUC Teórica.

[5] O eixo x é construído de forma invertida (começando com a Especificidade = 1 e terminando com a Especificidade = 0).

se for eleito como sendo o *cut-off* a ser utilizado no modelo, esse modelo errará todos os ***eventos*** e acertará todos os ***não eventos***; em P_2, o *cut-off* é tão baixo que, se for eleito como sendo o *cut-off* a ser utilizado no modelo, esse modelo acertará todos os ***eventos*** e errará todos os ***não eventos***.

Dito de outra forma: em P_1, o *cut-off* é tão alto que, se for eleito como sendo o *cut-off* a ser utilizado no modelo, esse modelo classificará, indevidamente, todos os ***eventos*** como sendo ***não eventos*** (gerando Falsos Negativos ou Falsos ***Não Eventos***); em P_2, o *cut-off* é tão baixo que, se for eleito como sendo o *cut-off* a ser utilizado no modelo, esse modelo classificará, indevidamente, todos os ***não eventos*** como sendo ***eventos*** (gerando Falsos Positivos ou Falsos ***Eventos***).

Eleger o *cut-off* ideal é uma etapa complexa no processo de construção do classificador e depende fundamentalmente do ramo de atividade em questão e das finalidades do pesquisador. Para esta etapa, existem alguns critérios amplamente difundidos para a eleição deste *cut-off*. Um destes critérios determina que o *cut-off* ideal seria 50% (ou seja, se a probabilidade média estimada de o ***evento*** ocorrer for igual ou superior a 50%, o classificador deveria classificá-lo como ***evento***; caso contrário, deveria classificá-lo como ***não evento***). Outro critério determina que o *cut-off* ideal seria aquele para o qual a sensitividade é igual a especificidade. Por fim, um terceiro critério determina que o *cut-off* ideal seria aquele que maximiza a acurácia do classificador.

Em bancos de dados muito desbalanceados, os quais caracterizam-se por ter um número muito maior de ***não eventos*** do que de ***eventos*** (ou vice e versa), utilizar, por exemplo, o critério do *cut-off* que maximiza a acurácia do classificador conduzirá, inevitavelmente, ao favorecimento da categoria majoritária.

Em bancos de dados muito desbalanceados, um critério interessante para eleger o *cut-off* ideal pode ser obtido a partir do seguinte questionamento: seja $P_{máx}$ o ponto na Curva ROC Teórica, no qual especificidade = 1 e sensitividade =1; ou seja, $P_{máx}$ (1, 1). A questão é "qual é o ponto na Curva ROC ajustada para o qual a distância até $P_{máx}$ é mínima ?" A resposta a esse questionamento conduz ao ponto onde estaremos o mais próximo possível do ponto que maximiza a AUC. A AUC máxima é obtida quando a especificidade = 1 e sensitividade = 1; ou seja, em $P_{máx}$. Logo, encontrar o ponto mais próximo de $P_{máx}$ equivale a ter encontrado o ponto

ideal, e consequentemente o *cut-off*, que maximiza a especificidade e a sensitividade do banco de dados que está sendo objeto de estudo[6].

O cálculo desta distância, denominada Distância Euclidiana, é obtido de acordo com a seguinte expressão:

$$D_{PajustPmáx} = \sqrt{\left\{\left(1 - X_{Pajust}\right)^2 + \left(1 - Y_{Pajust}\right)^2\right\}}$$

onde P_{ajust} = um ponto qualquer na Curva ROC ajustada; $D_{PajustPmáx}$ = a distância Euclidiana entre P_{ajust} e $P_{máx}$; X_{Pajust} = a coordenada X do P_{ajust} e Y_{Pajust} = a coordenada Y do P_{ajust}.

Tendo sido eleito o *cut-off* ideal, o modelo de classificação, quando ajustado, classificará como ***evento*** todas as observações cuja probabilidade média estimada for igual ou superior a este *cut-off* e classificará como ***não evento*** todas as observações cuja probabilidade média estimada for inferior a ele. Vale destacar que, como Assunção, Izbicki e Prates (2024) argumentam, no intuito de melhorar a previsão em conjuntos de dados desbalanceados, ajustar o *cut-off* do classificador pode produzir resultados semelhantes às técnicas usuais de balanceamento de dados, como *oversampling* ou *subsampling* (KAUR et al., 2019; CHAWLA et al., 2022).

Por fim, com o auxílio da matriz de confusão, um resumo comparando as previsões obtidas pelo modelo com os valores reais (observados) pode ser implementado. Na Tabela 3 abaixo, um exemplo de uma Matriz de Confusão.

Tabela 3 - Matriz de Confusão

		VALORES OBSERVADOS	
		0	**1**
VALORES PREDITOS	**0**	VN	FN
	1	FP	VP

[6] Para uma abordagem mais detalhada sobre este tema, recomenda-se a leitura de (Gabriel O. Assunção, Rafael Izbicki e Marcos O. Prates. *Is Augmentation Effective in Improving Prediction in Imbalanced Datasets* ? *Journal of Data Science*, 2024)

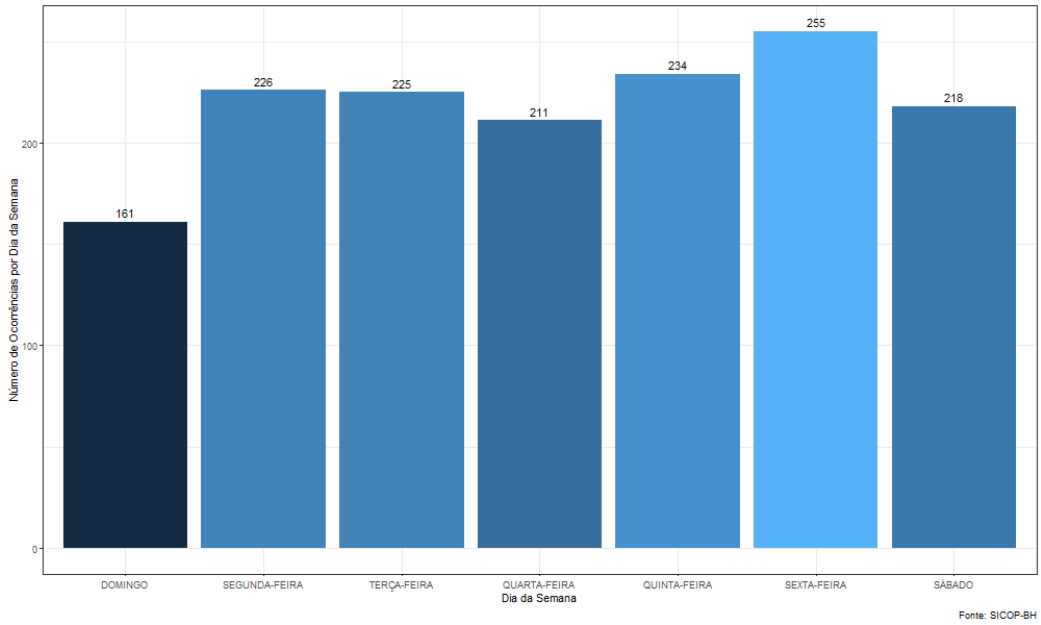
3. RESULTADOS

3.1 Análise Descritiva dos Dados

Abaixo, nas Figuras de 1 a 4, a distribuição de frequência dos acidentes de trânsito com vítimas não fatais ocorridos em Belo Horizonte, na regional Centro-Sul, em 2023.

A média das ocorrências por DIA DA SEMANA é de 218.60, com o desvio-padrão igual a 28.96 e assimetria igual a -1.0265 (sugerindo uma distribuição com assimetria acentuada). Em geral, ocorrem menos acidentes aos domingos do que nos demais dias (Figura 1).

Figura 1 - Distribuição dos Acidentes de Trânsito com Vítimas não Fatais ocorridos na Regional Centro-Sul, por DIA DA SEMANA, em 2023



A média das ocorrências por MÊS é de 127.50, com o desvio-padrão igual a 17.30 e assimetria igual a -0.1050 (sugerindo uma distribuição relativamente simétrica). Em geral, ocorrem menos acidentes três primeiros meses do ano, período típico de férias escolares e com menor circulação de veículos (Figura 2).

A média das ocorrências por FAIXA DE HORÁRIO é de 191.25, com o desvio-padrão igual a 97.26 e assimetria igual a -0.5824 (sugerindo uma distribuição com assimetria moderada). Em geral, ocorrem mais acidentes no período do dia entre 15h e 20h59 (Figura 3).

Figura 2 - Distribuição dos Acidentes de Trânsito com Vítimas não Fatais ocorridos na Regional Centro-Sul, por MÊS, em 2023

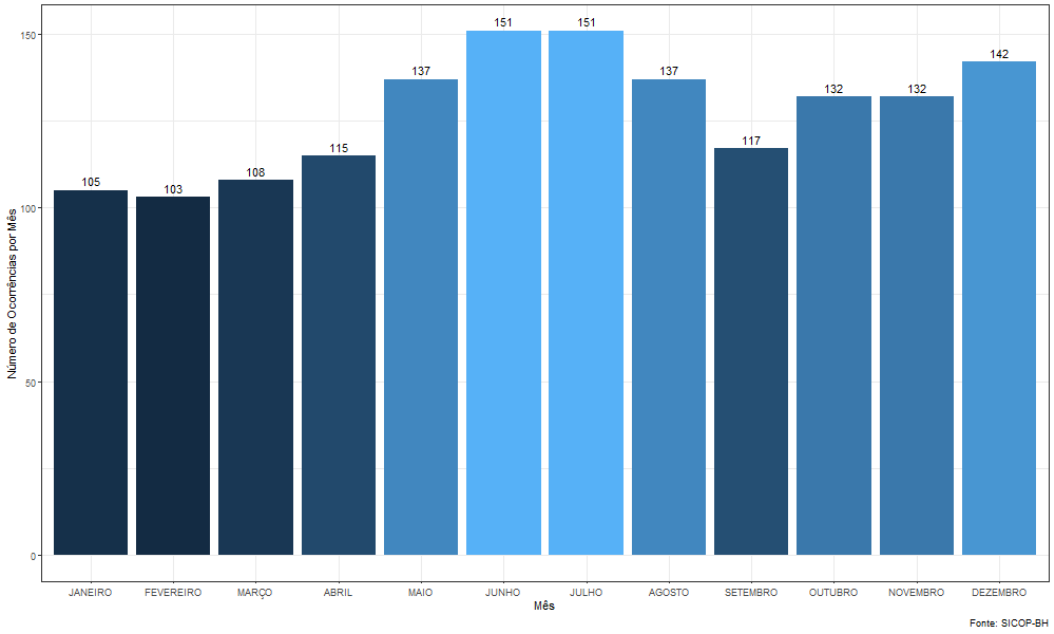
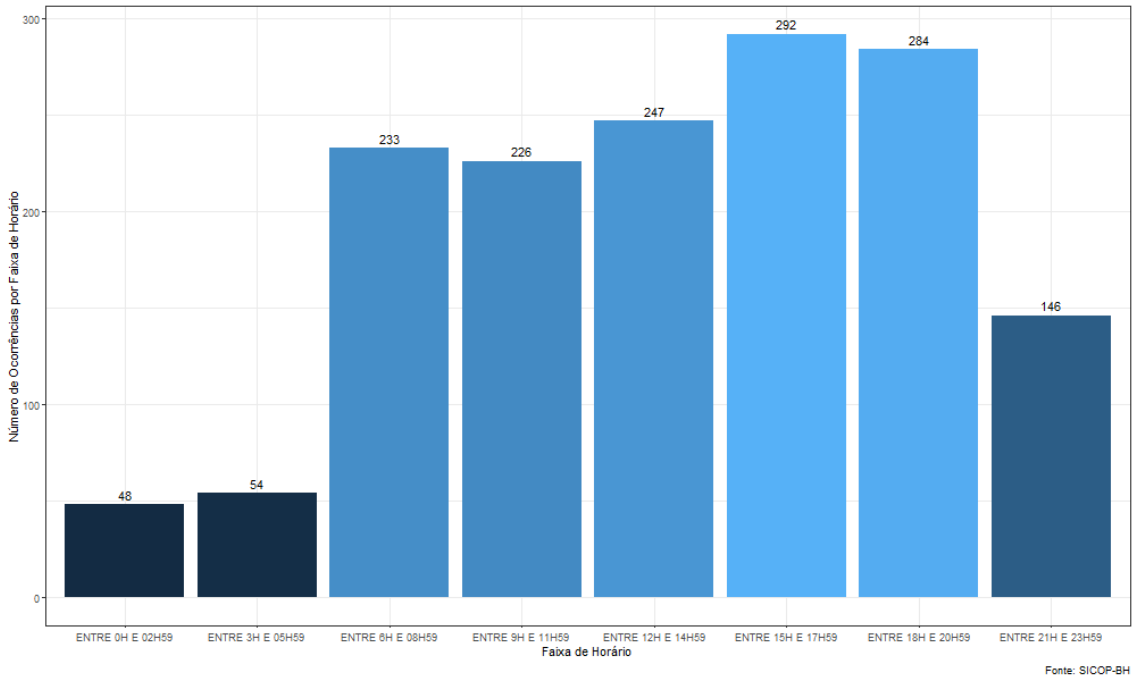
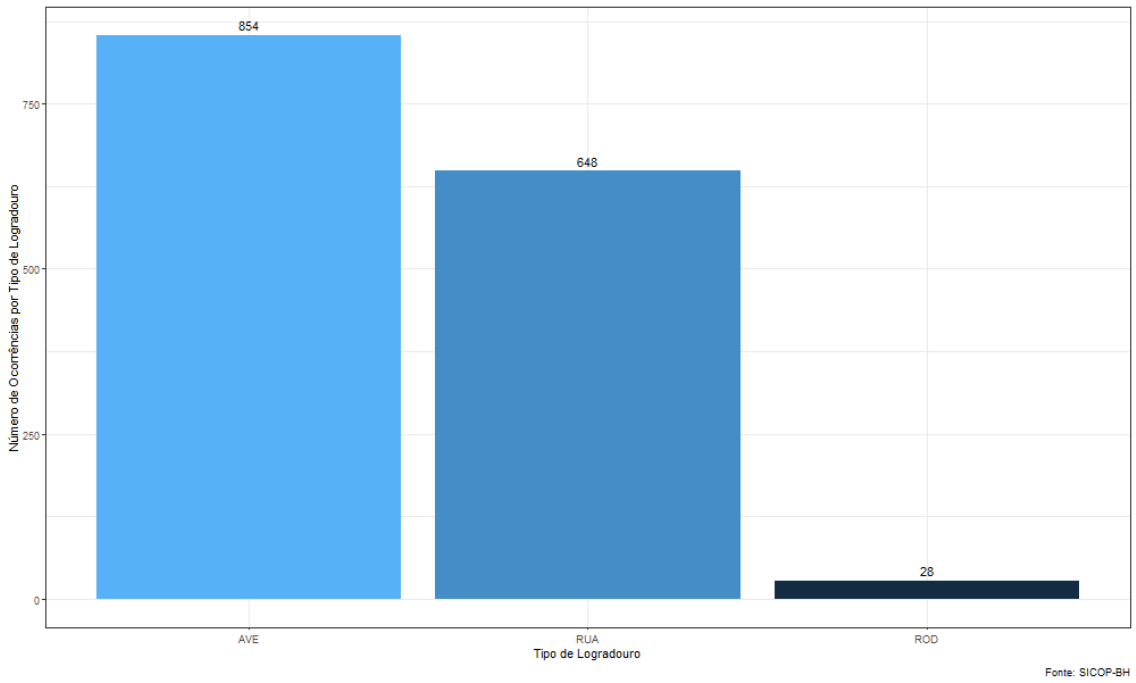


Figura 3 - Distribuição dos Acidentes de Trânsito com Vítimas não Fatais ocorridos na Regional Centro-Sul, por FAIXA DE HORÁRIO, em 2023



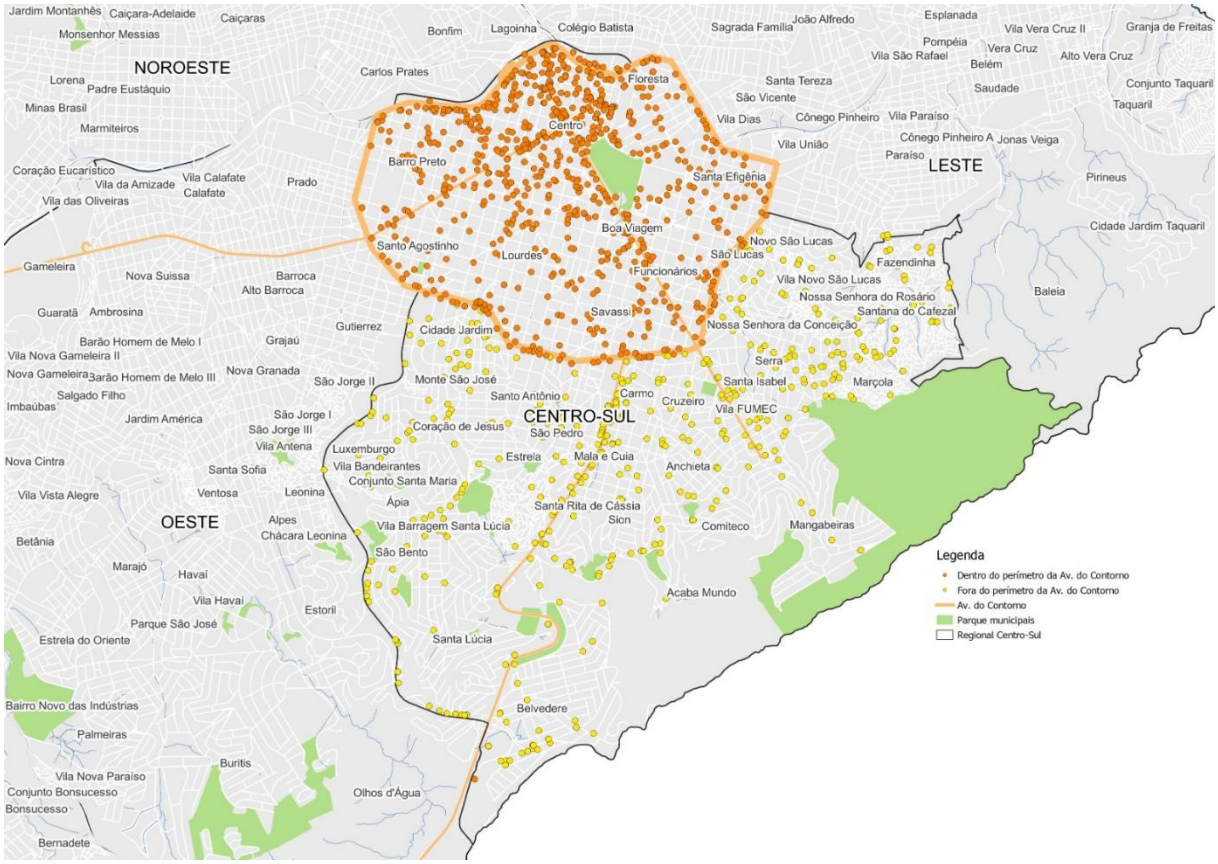
A média das ocorrências por TIPO DE LOGRADOURO é de 191.25, com o desvio-padrão igual a 97.26 e assimetria igual a -0.5824 (sugerindo uma distribuição com assimetria moderada). Há uma frequência muito baixa de acidentes em rodovias no espaço considerado no estudo (Figura 4)

Figura 4 - Distribuição dos Acidentes de Trânsito com Vítimas não Fatais ocorridos na Regional Centro-Sul, por TIPO DE LOGRADOURO, em 2023



A Figura 5 apresenta a distribuição espacial dos acidentes de trânsito com vítimas não fatais ocorridos na regional Centro-Sul de Belo Horizonte em 2023. A delimitação dos acidentes ocorridos dentro e fora do perímetro da Avenida do Contorno, importante via de tráfego em Belo Horizonte, tem caráter meramente exemplificativo e os dados considerados nesse estudo se referem a toda a área da regional.

Figura 5 - Distribuição Espacial dos Acidentes de Trânsito com Vítimas não Fatais ocorridos na Regional Centro-Sul, em 2023



3.2 Análise e Interpretação do Modelo Logístico

Toda a análise foi realizada por meio da plataforma RStudio do *software* R. Utilizando a função *glm* do pacote *stats*, o modelo de regressão logística foi ajustado no banco de treinamento, gerando a saída mencionada na Tabela 4, de onde algumas informações importantes podem ser obtidas:

1) com relação à variável DIA_SEMANA, tendo sido o domingo eleito como categoria de referência, todos os demais dias da semana se mostraram estatisticamente significativos para explicar o ***evento*** (p valor < 0.10). Os coeficientes positivos observados na coluna "Estimativa" nos remetem a ideia de que a probabilidade de o ***evento*** ocorrer em qualquer dia da semana é maior do que a probabilidade de ele ocorrer nos domingos.

2) com relação à variável MES, tendo sido o mês de janeiro eleito como categoria de referência, os meses de maio a dezembro se mostraram estatisticamente significativos para explicar o ***evento*** (p valor < 0.10). Os coeficientes positivos observados na coluna "Estimativa"

destes meses nos remetem a ideia de que a probabilidade de o ***evento*** ocorrer nestes meses é maior do que a probabilidade de ele ocorrer em janeiro. Já os meses fevereiro, março e abril não se mostraram estatisticamente significativos para explicar o ***evento*** (p valor > 0.10), sugerindo que, nestes meses, a probabilidade de o ***evento*** ocorrer é a mesma de ocorrer em janeiro.

3) com relação à variável FAIXA_HORÁRIO, tendo sido a faixa de horário entre12:00 e 14:59 eleita como categoria de referência, as faixas entre 00:00 e 02:59, 03:00 e 05:59, 15:00 e 17:59 e 21:00 e 23:59 se mostraram estatisticamente significativas para explicar o ***evento*** (p valor < 0.10). Os coeficientes negativos observado na coluna "Estimativa" para as faixas de horários entre 00:00 e 02:59, 03:00 e 05:59 e 21:00 e 23:59 nos remetem a ideia de que a probabilidade de o ***evento*** ocorrer nestas faixas de horários é menor do que a probabilidade de ele ocorrer entre 12:00 e 14:59 (o contrário é verdadeiro para a faixa de horário entre 15:00 e 17:59). As faixas de horários entre 06:00 e 08:59, 09:00 e 11:59 e 18:00 às 20:59 não se mostraram estatisticamente significativas para explicar o ***evento*** (p valor > 0.10), sugerindo que, nestas faixas de horários, a probabilidade de o ***evento*** ocorrer é a mesma de ocorrer entre 12:00 e 14:59.

4) com relação à variável LOGRADOURO, tendo sido o tipo de logradouro RUA eleito como categoria de referência, todos os demais tipos de logradouro se mostraram estatisticamente significativos para explicar o ***evento*** (p valor < 0.10). Ademais, o coeficiente positivo observado na coluna "Estimativa" para o tipo de logradouro AVENIDA nos remete a ideia de que a probabilidade de o ***evento*** ocorrer em uma avenida da regional Centro-sul é maior do que a probabilidade de ele ocorrer em uma rua da regional Centro-Sul (o contrário é verdadeiro para o tipo de logradouro RODOVIA).

Tabela 4 - Saída do modelo de regressão logística ajustado no banco de treinamento

Nome da Categoria	Estimativa	P value	*Odds*
Intercepto	-1,680	0.000	-
SEGUNDA-FEIRA	0.387	0.007	1,473
TERÇA-FEIRA	0.495	0.001	1,640
QUARTA-FEIRA	0.337	0.023	1,401
QUINTA-FEIRA	0.466	0.001	1,594
SEXTA-FEIRA	0.571	0.000	1,770
SÁBADO	0.295	0.047	1,343
FEVEREIRO	0.160	**0.415**	-
MARÇO	0.170	**0.381**	-
ABRIL	0.297	**0.120**	-
MAIO	0.396	0.035	1,486
JUNHO	0.605	0.001	1,831
JULHO	0.543	0.003	1,721
AGOSTO	0.456	0.015	1,578
SETEMBRO	0.385	0.044	1,470
OUTUBRO	0.415	0.029	1,514
NOVEMBRO	0.348	0.064	1,416
DEZEMBRO	0.385	0.043	1,470
ENTRE 0H E 02H59	-1,828	0.000	0,161
ENTRE 3H E 05H59	-1,830	0.000	0,160
ENTRE 6H E 08H59	-0.143	**0.279**	-
ENTRE 9H E 11H59	-0.144	**0.283**	-
ENTRE 15H E 17H59	0.246	0.054	1,279
ENTRE 18H E 20H59	0.119	**0.359**	-
ENTRE 21H E 23H59	-0.636	0.000	0,529
AVE	0.318	0.000	1,374
ROD	-3,436	0.000	0,032

Ainda com base na Tabela 4 (coluna *Odds*), algumas informações importantes podem ser obtidas:

1) com relação à variável DIA_SEMANA, tendo sido o domingo eleito como categoria de referência, **1.1)** a chance de o ***evento*** ocorrer em uma segunda-feira é multiplicada pelo fator 1.473, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em uma segunda-feira é, em média, 47,3% maior do que em um domingo; **1.2)** a chance de o ***evento*** ocorrer em uma terça-feira é multiplicada pelo fator 1.640, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em uma terça-feira é, em média, 64,0% maior do que em um domingo; **1.3)** a chance de o ***evento*** ocorrer em uma quarta-feira é multiplicada pelo fator 1.401, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em uma quarta-feira é, em média, 40,1% maior do que em um domingo; **1.4)** a chance de o ***evento*** ocorrer em uma quinta-feira é multiplicada pelo fator 1.594, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em uma quinta-feira é, em média, 59,4% maior do que em um domingo; **1.5)** a chance de o ***evento*** ocorrer em uma sexta-feira é multiplicada pelo fator 1.770, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em uma sexta-feira é, em média, 77,0% maior do que em um domingo; **1.6)** a chance de o ***evento*** ocorrer em um sábado é multiplicada pelo fator 1.343, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em um sábado é, em média, 34,3% maior do que em um domingo.

2) com relação à variável MES, tendo sido o mês de janeiro eleito como categoria de referência, **2.1)** a chance de o ***evento*** ocorrer em maio é multiplicada pelo fator 1.486, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em maio é , em média, 48,6% maior do que em janeiro; **2.2)** a chance de o ***evento*** ocorrer em junho é multiplicada pelo fator 1.831, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em junho é , em média, 83,1% maior do que em janeiro; **2.3)** a chance de o ***evento*** ocorrer em julho é multiplicada pelo fator 1.721, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em julho é , em média, 72,1% maior do que em janeiro; **2.4)** a chance de o ***evento*** ocorrer em agosto é multiplicada pelo fator 1.578, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em agosto é , em média, 57,8% maior do que em janeiro; **2.5)** a chance de o ***evento*** ocorrer em setembro é multiplicada pelo fator 1.470, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em setembro é , em média, 47,0% maior do que em janeiro; **2.6)** a chance de o ***evento*** ocorrer em outubro é multiplicada pelo fator 1.514, ou seja, mantidas as demais condições constantes, a

chance de o ***evento*** ocorrer em outubro é , em média, 51,4% maior do que em janeiro; **2.7)** a chance de o ***evento*** ocorrer em novembro é multiplicada pelo fator 1.416, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em novembro é , em média, 41,6% maior do que em janeiro; **2.8)** a chance de o ***evento*** ocorrer em dezembro é multiplicada pelo fator 1.470, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em dezembro é , em média, 47,0% maior do que em janeiro.

3) com relação à variável FAIXA_HORÁRIO, tendo sido a faixa de horário entre 12:00 e 14:59 eleita como categoria de referência, **3.1)** a chance de o ***evento*** ocorrer entre 00:00 e 02:59 é multiplicada pelo fator 0.161, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer entre 00:00 e 02:59 é , em média, 83,9% menor do entre 12:00 e 14:59; **3.2)** a chance de o ***evento*** ocorrer entre 03:00 e 05:59 é multiplicada pelo fator 0.160, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer entre 03:00 e 05:59 é, em média, 84,0% menor do entre 12:00 e 14:59; **3.3)** a chance de o ***evento*** ocorrer entre 15:00 e 17:59 é multiplicada pelo fator 1.279, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer entre 15:00 e 17:59 é , em média, 27,9% maior do entre 12:00 e 14:59; **3.4)** a chance de o ***evento*** ocorrer entre 21:00 e 23:59 é multiplicada pelo fator 0.529, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer entre 21:00 e 23:59 é , em média, 47,1% menor do entre 12:00 e 14:59.

4) com relação à variável LOGRADOURO, tendo sido o tipo de logradouro RUA eleito como categoria de referência, **4.1)** a chance de o ***evento*** ocorrer em uma avenida da regional Centro-Sul é multiplicada pelo fator 1.374, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em uma avenida da regional Centro-Sul é , em média, 37,4% maior do que em uma rua; **4.2)** a chance de o ***evento*** ocorrer em uma rodovia da regional Centro-Sul é multiplicada pelo fator 0.032, ou seja, mantidas as demais condições constantes, a chance de o ***evento*** ocorrer em uma rodovia da regional Centro-Sul é , em média, 96,8% menor do que em uma rua.

Na realidade, estas são algumas poucas análises quantitativas que podem ser realizadas a partir Tabela 4 (coluna *Odds*). Ao todo, 384 análises de *odds* distintas podem ser realizadas com a combinação de todas as categorias que se mostraram estatisticamente significativas de

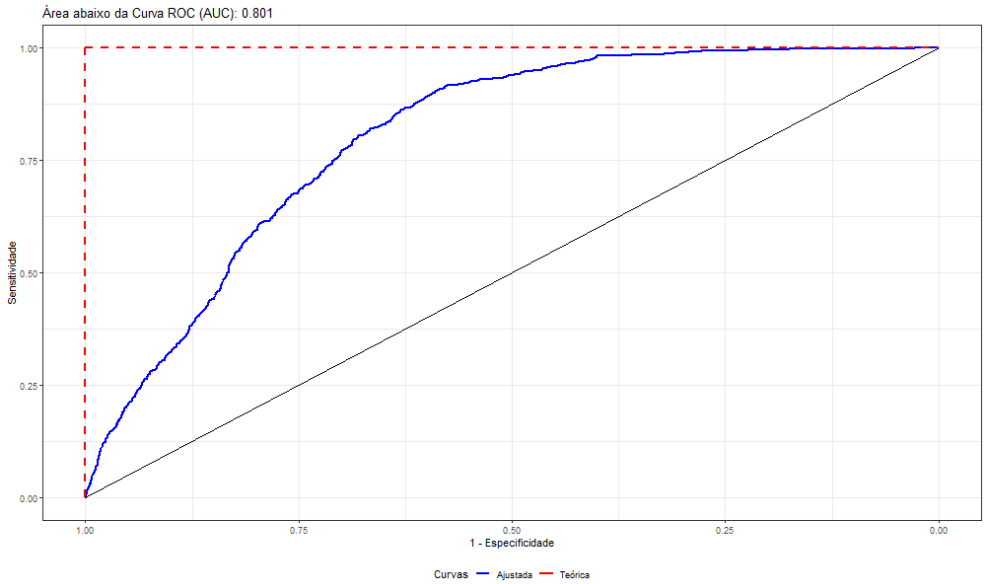
cada uma das 04 variáveis (6 dias da semana x 8 meses do ano x 4 faixas de horários x 2 tipos de logradouro = 384)[7].

Considerando que as categorias de referência são o domingo, mês de janeiro, faixa de horário entre 12:00 e 14:59 e logradouro do tipo RUA, um pesquisador poderia estar interessado em estudar, por exemplo, enquanto, em média, aumenta ou diminui a chance de o ***evento*** ocorrer em uma sexta-feira, do mês de maio, entre 15:00 e 17:59 em uma avenida da regional Centro-Sul ou em uma terça-feira, do mês de agosto, entre 00:00 e 02:59 em uma rodovia da regional Centro-Sul; e assim sucessivamente.

3.3 Análise da Capacidade Preditiva do Modelo Logístico

Utilizando a função *roc* do pacote *pROC*, a Curva ROC foi construída. A Figura 6 abaixo mostra o resultado. A AUC obtida a partir da Curva ROC ajustada no banco de treinamento do Banco de Dados para Modelagem foi: AUC = 0.801; e o *cut-off* ideal encontrado, após a realização do procedimento mencionado na seção 2.3.2, foi: *cut-off* = 0.2346 (23.46%). Ou seja, esse é o *cut-off* que maximiza a sensitividade e a especificidade do Banco de Dados para Modelagem.

Figura 6 - Curva ROC Ajustada no banco de treinamento e Curva Teórica



[7] Neste cálculo não foram computados: a) 01 dia da semana, porque foi eleito como categoria de referência; b) 04 meses, sendo 01 mês porque foi eleito como categoria de referência e 03 meses porque não se mostraram estatisticamente significativos; c) 04 faixas de horários, sendo uma faixa de horário porque foi eleita como categoria de referência e 03 faixas de horários porque não se mostraram estatisticamente significativas; d) 01 tipo de logradouro, porque foi eleito como categoria de referência.

Desta forma, o modelo de classificação classificou como ***evento*** todas as observações cuja probabilidade média estimada foi igual ou superior a 0.2346 (23.46%) e classificou como ***não evento*** todas as observações cuja probabilidade média estimada foi inferior a 0.2346 (23.46%).

Após o modelo de regressão logística ter sido devidamente ajustado, o *cut-off* ideal ter sido eleito e os valores preditos terem sido obtidos com base neste *cut-off*, a função *confusionMatrix* do pacote *caret* foi utilizada para gerar a Matriz de Confusão do Banco de Treinamento. A Tabela 5 abaixo mostra os resultados.

Tabela 5 - Matriz de Confusão Ajustada no Banco de Treinamento

		VALORES OBSERVADOS	
		0	**1**
VALORES PREDITOS	**0**	3.595	212
	1	1.651	825

Baseado na Tabela 5, os valores da acurácia, da sensitividade e da especificidade podem ser obtidos, conforme a Tabela 6 abaixo:

Tabela 6 - Acurácia, Sensitividade e Especificidade obtidas no Banco de Treinamento

Acurácia	Sensitividade	Especificidade
70.35%	79.56%	68.53%

Com relação à acurácia, o resultado observado indica que o modelo de classificação ajustado é capaz de prever corretamente 70.35% e incorretamente 29.65% das observações. Dito de outra forma: em 70.35% das vezes, o modelo disse que o ***evento*** aconteceria e ele de fato aconteceu ou o modelo disse que o ***evento*** não aconteceria e ele de fato não aconteceu; em 29.65% das vezes, o modelo disse que o ***evento*** aconteceria, mas ele não aconteceu ou o modelo disse que o ***evento*** não aconteceria, mas ele aconteceu.

Com relação à sensitividade e à especificidade, os resultados observados indicam que o modelo de classificação ajustado é capaz de prever corretamente 79.56% dos ***eventos*** e 68.53% dos ***não eventos***. Dito de outra forma: em 79.56% das vezes, o modelo vai dizer que o ***evento*** acontecerá e ele de fato acontecerá (VP); em 20.44% das vezes, o modelo vai dizer que o ***evento*** não acontecerá, mas ele acontecerá (FN); em 68.53% das vezes, o modelo vai dizer que o ***evento***

não acontecerá e ele de fato não acontecerá (VN); em 31.47% das vezes, o modelo vai dizer que o ***evento*** acontecerá, mas ele não acontecerá (FP).

Utilizando a função *confusionMatrix* do pacote *caret*, a Matriz de Confusão do Banco de Teste foi construída. A Tabela 7 abaixo mostra os resultados e fornece subsídio para o cálculo dos valores da acurácia, da sensitividade e da especificidade, conforme a Tabela 8 abaixo. Comparando as Tabelas 6 e 8, verifica-se uma aderência entre as métricas respectivas no banco de dados de treinamento e de teste e uma boa capacidade de generalização do modelo.

Tabela 7 - Matriz de Confusão Ajustada no Banco de Teste

		VALORES OBSERVADOS	
		0	**1**
VALORES PREDITOS	**0**	1.505	108
	1	695	385

Tabela 8 - Acurácia, Sensitividade e Especificidade obtidas no Banco de Teste

Acurácia	Sensitividade	Especificidade
70.18%	78.09%	68.41%

4. DISCUSSÃO

A escolha do domingo como sendo a categoria de referência da variável DIA_SEMANA e a escolha do mês de janeiro como sendo a categoria de referência da variável MES se deu com o intuito de criar a noção de como o ***evento*** se comporta em relação a um dia em que a grande maioria da sociedade está em gozo de descanso semanal e em relação a um mês em que escolas, universidades e alguns ramos de atividade estão de férias, respectivamente. Desta forma é possível inferir, baseado na Tabela 4 (coluna *Odds*), de forma quantitativa (e não mais apenas de forma intuitiva), que o fluxo de veículos e pedestres é, de fato, um fator preponderante na ocorrência de acidentes de trânsito.

A escolha do tipo de logradouro RUA como sendo a categoria de referência da variável LOGRADOURO se deu com o intuito de criar a noção de como o ***evento*** se comporta em relação a uma via em que, em tese, a velocidade permitida é menor do que a velocidade permitida em uma avenida ou em uma rodovia (onde a velocidade permitida para o tráfego de veículos é a maior dentre os três tipos de logradouro considerados). A hipótese inicial era que, para um mesmo mês, mesmo dia da semana e mesma faixa de horário, a ocorrência do ***evento*** seria tanto maior quanto maior fosse a velocidade permitida para a via em questão. Está hipótese foi refutada visto que, baseado na Tabela 4 (coluna *Odds*), mantidas as demais variáveis constantes, a chance de o ***evento*** ocorrer em uma rodovia da regional Centro-Sul é , em média, 96,8% menor do que em uma rua. Aliás, neste contexto, visto o número baixo de ocorrências observadas em rodovias, para trabalhos futuros talvez seja mais interessante considerar uma rodovia que esteja contida dentro dos limites do município de Belo Horizonte como sendo uma avenida.

Ao contrário da linha de raciocínio utilizada para demais variáveis, o critério utilizado para escolher a faixa de horário que estaria na categoria de referência não foi “do menor para o maior fluxo de veículos e pedestres”. Quando a faixa de horário entre 00:00 e 02:59 ou entre 03:00 e 05:59 foram tomadas como referências, todas as demais faixas de horários se mostraram estatisticamente significativas. Porém, o estudo não se mostrou interessante nas faixas de horários em que sabidamente há um fluxo intenso de veículos e pedestres (06:00 às 19:00 aproximadamente). Desta forma, a escolha da faixa de horário entre 12:00 e 14:59 como sendo a categoria de referência da variável FAIXA_HORARIO se deu com o intuito de criar a seguinte noção: dentre as faixas de horários sabidamente complexas, de que forma o ***evento*** se comporta em relação àquela faixa de horário que nem está contida no horário de “pico” diurno nem está contida no horário de “pico” vespertino/noturno? O resultado obtido, talvez contraintuitivo,

mostra que, no “pico” diurno (assim considerado a faixa de horário entre 06:00 e 08:59), o ***evento*** se comporta de forma similar ao da faixa de horário de referência; também contraintuitivo, o mesmo ocorre com os ***eventos*** ocorridos no “pico” noturno (assim considerado a faixa de horário entre 18:00 e 20:59).

Neste contexto, para trabalhos futuros, talvez seja mais interessante considerar as faixas de horários entre 00:00 e 02:59 e 03:00 e 05:59 como sendo uma única faixa de horários (00:00 às 05:59) ou fracionar, “de hora em hora”, as faixas de horários consideradas como “pico” diurno e “pico” noturno, para se ter uma ideia mais precisa de qual é a faixa de horário que pode ser realmente considerada “pico” e se ter uma ideia de como o ***evento*** se comporta nesta faixa de horário fracionada. Desta forma, é possível otimizar os esforços dos órgãos competentes e a alocação de recursos humanos e materiais naquilo que de fato importa.

O p-valor apresentado nos meses de fevereiro, março e abril estão dispostos de forma decrescente, evidenciando que, em fevereiro e março o ***evento*** ainda é similar ao que acontece em janeiro; mas, a partir de abril, o ***evento*** começa a se alterar, apesar de ainda não apresentar significância estatística.

No geral, o modelo de classificação ajustado apresentou uma capacidade preditiva razoável (AUC = 0.801, Sensitividade = 79.56% e Especificidade = 68.53%). Este resultado era um resultado esperado, visto a quantidade reduzida de variáveis explicativas utilizadas. Segundo Fávero e Belfiore (2017), a melhor forma de melhorar a capacidade preditiva de um modelo de classificação é a inserção, neste modelo, de variáveis que sejam estatisticamente significativas para explicar o ***evento***.

Neste contexto, surge uma crítica a esta monografia, visto se tratar de um estudo relativamente limitado. Na prática, é sabido que acidentes de trânsito têm relação direita com o fluxo de veículos e pedestres, ou seja, locais, datas e horários em que haja mais fluxo de veículos e pedestres em circulação tendem a apresentar maior número de ocorrências desta natureza. Esta monografia não trouxe informações referentes ao fluxo de veículos e pedestres de forma direta (o que talvez exija a utilização de *softwares* ou aplicativos caros para sua obtenção), mas trouxe informações que, indiretamente, conduzem a uma boa noção dele. Todavia, tentar explicar e predizer um fenômeno da magnitude do trânsito de uma grande metrópole, como é o caso de Belo Horizonte, com apenas 4 variáveis, talvez não seja o mais prudente a ser feito (apesar de que, com apenas essas 4 variáveis, grande parte do ***evento*** já foi explicado). Muitas outras variáveis poderiam ser inseridas no modelo para que, aquelas que se mostrassem

estatisticamente significativas, pudessem auxiliar no entendimento e predição do ***evento***. A seguir, alguns exemplos destas variáveis: **1)** "o ***evento*** se comporta de forma diferente em dias de chuva ?"; **2)** caso positivo, "de que forma o ***evento*** se comporta em função dos diferentes volumes de chuva ?"; **3)** "o ***evento*** se comporta de forma diferente em dias de *shows*, festas públicas ou nas proximidades de locais onde haja concentração de bares ?"; **4)** caso positivo, "de que forma o ***evento*** se comporta em função dos diferentes públicos (tamanho do público, faixa etário do púbico, etc.) ?"; **5)** "o ***evento*** se comporta de forma diferente em ruas e avenidas dotadas de radar ou outro meio de vigilância eletrônica ?"; **6)** caso positivo, "a qual distância média esse radar deve estar para reduzir a ocorrência do ***evento*** ?"; além de diversas outras variáveis que poderiam ser estudadas para os fins mencionados.

Todavia, o SICOP-BH não registra estas informações, limitando a qualidade do estudo. Seria necessária, portanto, a integração de alguns órgãos e sistemas da Prefeitura de Belo Horizonte, com a consequente cessão de dados, para a melhoria deste modelo e para a implementação de muitos outros modelos que podem ser desenvolvidos com a regressão logística e outras técnicas de *Machine Learning*.

Outra possível limitação do estudo diz respeito à forma de definição da variável resposta, visto que ela não considera o montante de acidentes observados em cada combinação de dia, logradouro e horário. Tem-se apenas a indicação de ocorrência ou não de um acidente dentro de um determinado estrato. Modelos apropriados para dados de contagem, como Poisson e Binomial Negativo, poderiam ser aplicados para estimar a taxa de incidência do evento em cada estrato. Contudo, a forma de definição binária adotada permite mensurar o quanto a mudança em um dos efeitos aumenta a probabilidade de se observar um acidente e a interpretação prática das *odds ratios* geradas pela análise realizada neste estudo pode ser de grande valia para a tomada de decisão por parte dos gestores.

5. CONCLUSÃO

O trânsito das grandes metrópoles é um fenômeno complexo e dinâmico, que se altera ininterruptamente. Este estudo não pretende, portanto, exaurir a discussão em torno deste assunto.

Apesar de nesta monografia não ter sido utilizada informações referentes ao fluxo de veículos e pedestres de forma direta, as 4 variáveis utilizadas conduzem, indiretamente, a uma boa noção dele e, sozinhas, foram capazes de explicar grande parte do ***evento*** em questão. Neste sentido, trabalhos como este, pela sua "facilidade" e baixo custo de implementação (em relação a, por exemplo, o custo de *softwares* e aplicativos que controlem o fluxo via satélite), podem utilizados na gestão do trânsito e na correta alocação de recursos humanos e materiais da Administração Pública Municipal.

Para estudos futuros, algumas propostas se destacam: **1)** outros modelos de *Machine Learning*, tais como *Random Forest*, *XGBoost* e *Support Vector Machine*, por exemplo, podem ser testados com o intuito de melhorar a capacidade de prever o ***evento***, além de pesquisar e utilizar outras variáveis explicativas que sejam relevantes para explicá-lo; **2)** desenvolver um algoritmo/aplicativo (utilizando o pacote *shiny* do software R) capaz de calcular, automaticamente, as probabilidades e chances de o ***evento*** ocorrer, além de ser capaz de classificar, também automaticamente, novas observações; **3)** com o uso das técnicas da Estatística Espacial, utilizar o georreferenciamento dos ***eventos*** e de outras fatores de interesse para entender se há correlação espacial entre eles **4)** por fim, com o auxílio e a cooperação mútua de outros órgãos da Administração Pública Municipal, Estadual e Federal, implementar as técnicas de *Machine Learning* mencionadas (e outras que porventura se façam necessárias) para realizar estudos de outros fenômenos de interesse coletivo, tais como fenômenos relacionados a chuva, a crimes (furto, roubo, feminicídio, homicídios, etc.), a desordens públicas (morador em situação de rua, etc.).

Apesar das limitações, a estratégia de modelagem aplicada nesse estudo pode auxiliar na persecução de uma administração pública cada vez mais baseada em evidências, em prol de serviços públicos eficientes e de qualidade para Belo Horizonte.

6. REFERÊNCIAS

ACHIM, Zeileis, TORSTEN, Hothorn (2002). *Diagnostic Checking in Regression Relationships*.

ASSUNÇÃO, Gabriel O.; IZBICKI, Rafael e PRATES, Marcos O. *Is Augmentation Effective in Improving Prediction in Imbalanced Datasets* ? *Journal of Data Science*, 2024, 1-16, DOI 10.6339/24-JDS1154.

CHAWLA, N.V.; BOWYER, K. W.; HALL, L.O. e KEGELMEYER, W. P. (2002). SMOTE: *Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, 16: 321–357.*

CORDEIRO, G.M; DEMÉTRIO, C.G.B e MORAL, R.A. Modelos Lineares Generalizados e Aplicações**.** Blucher, 2024 - 1ª Edição, Coleção Projeto Fisher.

CUNHA, Izabella Bauer de Assis. Modelagem da Informação para Cidades Inteligentes: Aplicação em Acidentes de Trânsito em Belo Horizonte, 2019. Disponível em https://repositorio.ufmg.br/bitstream/1843/31524/1/Disserta%C3%A7%C3%A3o_PPGGOC_Izabella%20Bauer_Modelagem%20da%20Informa%C3%A7%C3%A3o%20para%20Cidades%20Inteligentes.pdf. Acessado em 25 de março de 2025.

FÁVERO, Luiz Paulo e BELFIORE, Patrícia. Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel, SPSS e Stata. 1ª Edição - Rio de Janeiro: Editora Elsevier, 2017.

GOULART, Wagner Santos. Modelagem Estatística dos Acidentes de Trânsito na Cidade de Montes Claros/MG, 2018. Disponível em https://repo.ppgmcs.com.br/wp-content/uploads/tainacan-items/17/222/MODELAGEM-ESTATISTICA-DOS-ACIDENTES-

DE-TRANSITO-NA-CIDADE-DE-MONTES-CLAROS-MG.pdf. Acessado em 25 de março de 2025.

KAUR, H; PANNU, H. S. e MALHI, A. K. (2019). *A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Computing Surveys (CSUR)*, 52(4): 1–36.

KUHN, M. (2008). *Building Predictive Models in R Using the caret Package. Journal of Statistical Software*, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05.

LONG, J. A. (2022). _jtools: *Analysis and Presentation of Social Scientific* Data_. R package version 2.2.0, https://cran.r-project.org/package=jtools.

MONTGOMERY, Douglas C.; RUNGER George C. Estatística Aplicada e Probabilidade para Engenheiros. 6ª Edição - Rio de Janeiro: Editora LTC, 2016.

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/.

SARAIVA, João Pedro Melani e DOS SANTOS, Pedro Augusto Borges. Modelo Preditivo de Óbitos no Trânsito Brasileiro, 2024. Disponível em https://www.onsv.org.br/pdi/dados/modelo-preditivo-de-obitos-no-transito-brasileiro. Acessado em 25 de março de 2025.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: *an open-source package for R and S+ to analyze and compare ROC curves*. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/.