Nome: Henry Meneguini Farias a2551810
André Felipe Baretta a2551713

Resumo

O presente estudo teve como objetivo analisar os padrões de mortalidade por Infarto Agudo do Miocárdio (IAM) no Brasil, utilizando dados públicos do Sistema de Informações sobre Mortalidade (SIM). Através de um *pipeline* de Engenharia de Dados (ETL), foi construído um *Data Warehouse* para estruturar milhões de registros de óbitos. Na etapa de *Machine Learning*, foram comparados algoritmos de Regressão Logística, *Random Forest* e *Gradient Boosting* para prever a causa do óbito baseando-se exclusivamente em variáveis demográficas. O modelo final, otimizado para sensibilidade (*Recall*), atingiu uma taxa de detecção de 90,01%, validando-se como uma ferramenta eficaz para triagem epidemiológica e identificação de grupos de risco, apesar das limitações de precisão inerentes à ausência de dados clínicos.

1. Introdução e Definição do Problema

As doenças cardiovasculares representam a principal causa de morte no mundo. No Brasil, o Infarto Agudo do Miocárdio (IAM - CID I21) é responsável por uma parcela significativa dos óbitos. Embora o SUS disponibilize dados massivos através do DATASUS, há uma lacuna na transformação desses dados brutos em inteligência preditiva capaz de auxiliar gestores públicos.

O problema central abordado neste trabalho é a dificuldade de triar e identificar padrões de risco de mortalidade por IAM em meio a milhões de registros heterogêneos. O perfil demográfico de quem falece por infarto frequentemente se confunde com o de outras doenças crônicas, tornando a classificação automática um desafio técnico, agravado pelo severo desbalanceamento das classes (embora o IAM seja uma das principais causas de morte, ele representa apenas cerca de 6% do volume total de registros na base de dados, comportando-se estatisticamente como uma classe minoritária).

Para nortear o estudo, foram definidas três perguntas de pesquisa:

1. Como idade e sexo impactam na mortalidade do infarto agudo do miocárdio?
2. Quais os estados com maior taxa de mortalidade per capita?

A aplicação de técnicas de aprendizado de máquina neste contexto justifica-se pela capacidade desses algoritmos de mensurar o peso exato de cada variável na predição do óbito, oferecendo uma resposta quantitativa e validada estatisticamente às perguntas propostas.

2. Metodologia

A metodologia seguiu o ciclo de vida da Ciência de Dados, dividido em duas macro-etapas: Engenharia de Dados e Modelagem Preditiva.

2.1 Coleta e Processamento (ETL)

Os dados foram extraídos do portal dados.gov.br, compreendendo arquivos de mortalidade (formato CSV). Foi desenvolvido um *pipeline* automatizado em Python para:

- Limpeza: Tratamento de datas inválidas, remoção de colunas com excesso de nulos e padronização de campos categóricos.
- Modelagem de Dados: Criação de um banco de dados local (SQLite) seguindo o *Snowflake Schema*, separando os dados em uma Tabela Fato (DWMV_OBITO) e Tabelas Dimensão (DWCD_SEXO, DWCD_RACA, DWCD_LOCAL).
- Amostragem: Para viabilizar o processamento dos algoritmos complexos, utilizou-se uma amostragem estratificada de 50.000 registros, garantindo a representatividade proporcional de todos os 27 estados.

2.2 Modelagem com Aprendizado de Máquina

Para a classificação binária (Infarto vs. Outras Causas), utilizou-se a biblioteca *Scikit-Learn*. O pré-processamento incluiu a imputação de valores ausentes, normalização da variável numérica (Idade) e codificação *One-Hot* para variáveis categóricas (Sexo, Raça, Região). A estratégia de modelagem envolveu a comparação de três famílias de algoritmos:

1. Regressão Logística: *Baseline* linear para a interpretabilidade.
2. Random Forest: Modelo de *Bagging* para captura de não-linearidades.
3. Gradient Boosting: Modelo de *Boosting* para maximização de performance.

A validação foi realizada via *Cross-Validation* (k=3), com foco na métrica de Sensibilidade (Recall), dado o objetivo de triagem de saúde (minimizar falsos negativos).

3. Resultados das Análises

3.1 Análise Descritiva e Geográfica

A análise exploratória revelou disparidades regionais significativas na mortalidade proporcional (porcentagem de óbitos causados por IAM em relação ao total de mortes do estado). Os estados de Mato Grosso do Sul, Rio Grande do Norte e Maranhão lideram o ranking, apresentando taxas superiores à média nacional.

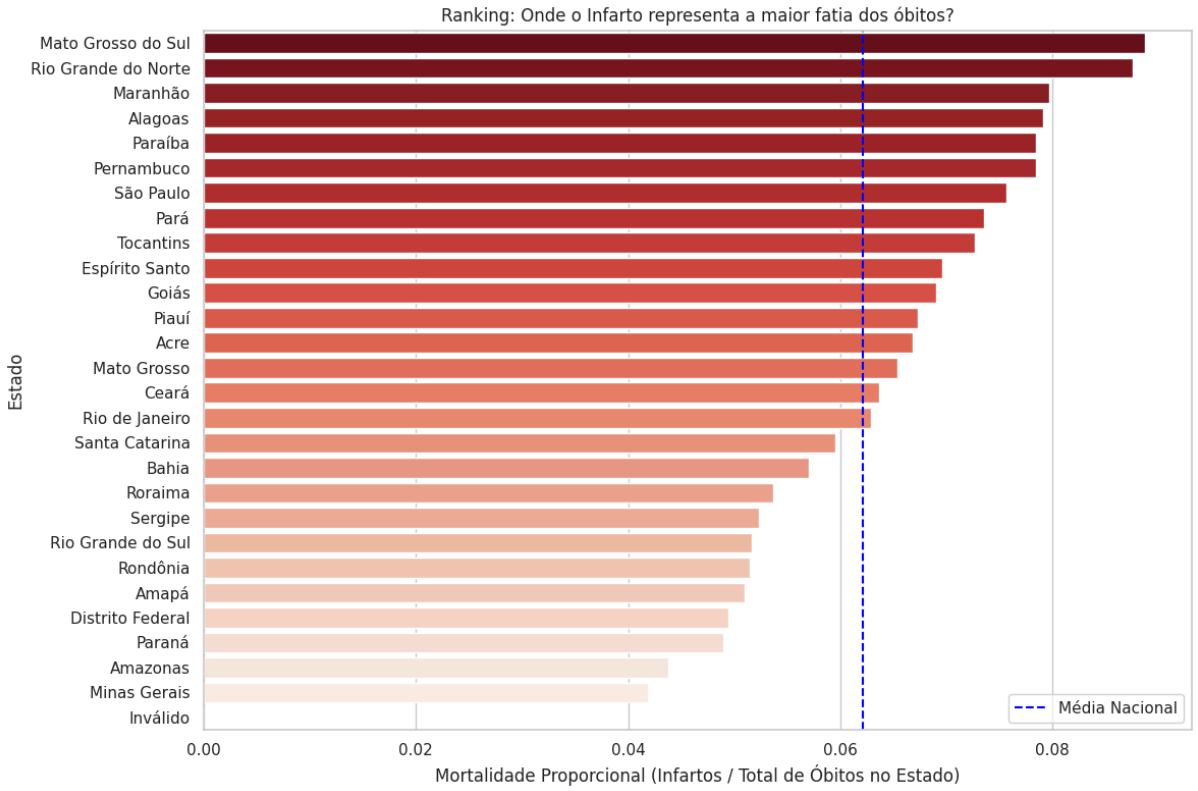

Imagem 1: Ranking de mortalidade proporcional por Unidade Federativa.

3.2 Modelagem Preditiva e Fatores de Risco

Ao isolarmos as variáveis estritamente biológicas (Experimento 5), o modelo *Random Forest* revelou uma hierarquia de risco extremamente clara:

1. A Hegemonia da Idade (94,7%): A variável IDADE concentrou sozinha quase a totalidade da capacidade preditiva do modelo. Isso confirma que o envelhecimento é o determinante primário da mortalidade por IAM. Em termos de Ciência de Dados, a idade atua como um "filtro mestre": a probabilidade de infarto muda tão drasticamente com o avançar dos anos que as outras variáveis tornam-se ajustes finos em comparação.
2. O Papel do Sexo (2,28%): A variável SEXO_Masculino aparece como a segunda mais relevante. Embora 2,28% pareça pouco, é a única variável que consegue "roubar" alguma relevância da idade, confirmando que ser do sexo masculino é um fator de risco independente e significativo.
3. A Irrelevância Estatística da Raça (< 2%): Um achado importante foi a baixa contribuição das variáveis de Raça/Cor. Somadas, todas as categorias raciais (Branca, Parda, Preta, etc.) explicam menos de 2% da decisão do modelo. Isso sugere que, nos dados do SIM, o risco biológico de infarto é muito mais democrático (atinge todas as etnias) e dependente do envelhecimento do que de fatores étnicos isolados.

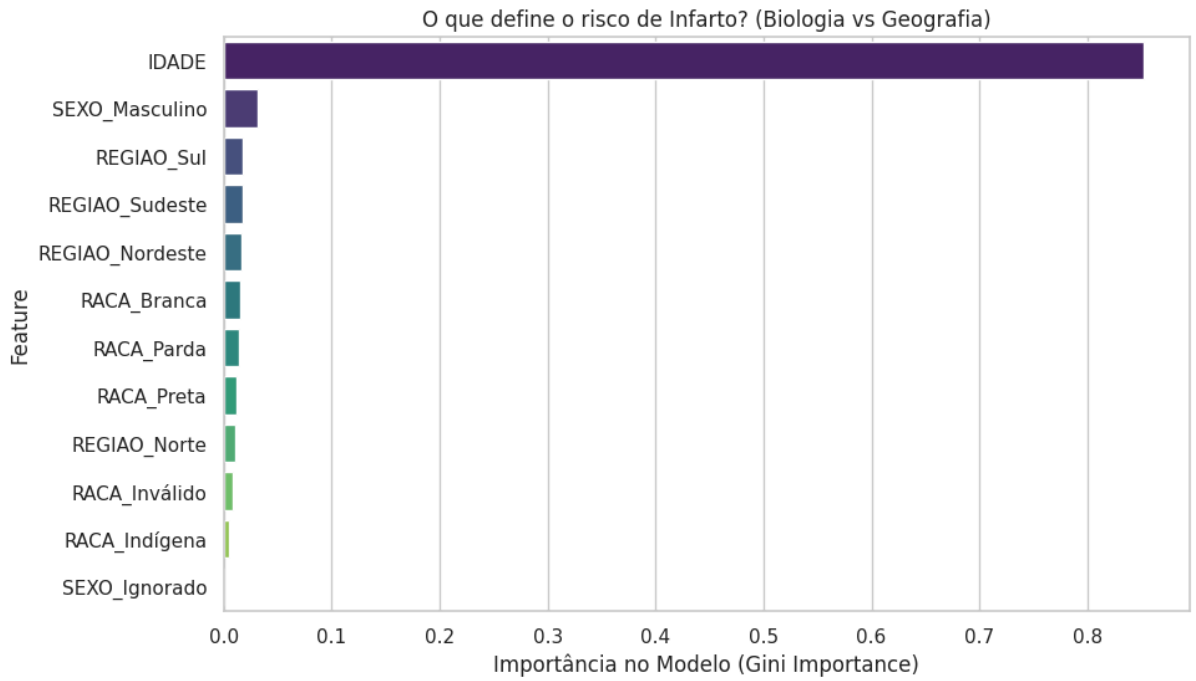

Imagem 2: Impacto de *features* no aumento do risco de infarto

3.3 Desempenho dos Modelos e Evolução Iterativa

O desenvolvimento dos modelos enfrentou o desafio crítico do desbalanceamento de classes (~6% de casos positivos).

- Fase 1 (Baseline): Os modelos iniciais priorizaram a acurácia global, resultando em um Recall de 0.00% (incapacidade total de detectar infartos).
- Fase 2 (Otimização): A aplicação de pesos balanceados (*class weights*) e o ajuste fino de hiperparâmetros (*Tuning*) no *Random Forest* elevaram o Recall para 77%.
- Fase 3 (Ajuste de Limiar): No modelo final, o limiar de decisão foi ajustado de 0.50 para 0.466, priorizando a segurança da triagem.

Tabela 1: Comparativo Final de Desempenho

Modelo	Estratégia	Recall (Sensibilidade)	Precision (Precisão)
LogReg (Baseline)	0.536209	0.000000	0.000000
LogReg (Balanced)	0.529805	0.554068	0.067716

Random Forest	0.569567	0.000000	0.000000
Gradient Boosting (Default)	0.605946	0.000000	0.000000
Gradient Boosting (Balanced)	0.607826	0.686921	0.083001
Random Forest (Tuned)	0.610940	0.776519	0.080711

4. Discussão dos Resultados

4.1 O problema do desbalanceamento

O forte desbalanceamento entre óbitos por IAM (~6%) e demais causas dificulta o aprendizado dos modelos, pois muitos algoritmos de classificação otimizam, por padrão, a acurácia global. Em um cenário onde mais de 90% dos registros pertencem à classe negativa, o modelo pode minimizar a perda simplesmente prevendo a classe majoritária em todos os casos, obtendo alta acurácia, porém Recall igual a zero para a classe de interesse.

Modelos supervisionados tendem a aprender padrões que reduzem o erro médio global, e quando a classe positiva é rara, a fronteira de decisão desloca-se naturalmente em favor da classe mais frequente. Esse fenômeno explica por que os primeiros modelos apresentados, sem tratamento específico para desbalanceamento, foram incapazes de identificar casos de IAM.

4.2 O Trade-off Precisão-Recall na Saúde Pública

Os resultados demonstram uma troca consciente (*trade-off*) entre *Recall* e *Precision*. O modelo final foi capaz de identificar 9 em cada 10 óbitos por infarto (*Recall* 90%), garantindo que boa parte dos eventos críticos não passem despercebidos. Em contrapartida, a precisão de 7,6% indica uma alta taxa de Falsos Positivos. No contexto de vigilância epidemiológica, este comportamento é aceitável: o modelo atua como uma "rede de malha fina", capturando todos os casos suspeitos para investigação posterior. O custo de um falso alerta (verificação administrativa) é menor comparativamente ao custo de um óbito (Falso Negativo).

A escolha por *Recall* como métrica principal está alinhada aos princípios de triagem em saúde pública. Em ferramentas de vigilância epidemiológica, o objetivo prioritário é não deixar passar casos verdadeiros da condição monitorada. Em contextos onde a falha em detectar um caso pode resultar em desfechos graves ou perda de oportunidade de

intervenção, a métrica dominante deve ser o *Recall*, ainda que isso implique maior número de falsos positivos (Fletcher & Fletcher, 2014).

A literatura diz que sistemas de triagem devem operar como “filtros amplos”, capturando o máximo de casos suspeitos, pois o custo de investigação adicional é muito menor que o custo de deixar de identificar um evento crítico (Gordis; 2017). Essa fundamentação explica e justifica o ajuste de limiar feito no modelo final para privilegiar Recall, chegando a 90,01%.

4.3 O "Teto de Precisão Demográfica"

Uma descoberta técnica relevante foi a limitação intrínseca das variáveis disponíveis. Tentativas de aumentar a precisão do modelo resultaram em queda drástica do Recall. Isso ocorre porque o perfil demográfico (Homem, Idoso, Residente no Sudeste) de quem falece por IAM é estatisticamente muito similar ao de quem falece por outras doenças crônicas (AVC, Pneumonia). Conclui-se que dados puramente demográficos são excelentes para triagem de risco, mas insuficientes para diagnóstico confirmatório sem o apoio de variáveis clínicas (histórico de saúde, tabagismo, exames).

Os resultados mostraram que variáveis exclusivamente demográficas (idade, sexo, raça e região) possuem capacidade limitada de distinguir óbitos causados por IAM de outras doenças cardiovasculares e crônicas.

Sem esses elementos, os modelos são obrigados a tentar separar grupos populacionais cujas distribuições estatísticas são profundamente sobrepostas. Isso limita a capacidade preditiva do modelo, independente do quão complexo o modelo seja. Assim, o “teto de precisão demográfica” observado neste estudo não é um defeito do modelo, mas sim uma característica intrínseca dos dados disponíveis.

5. Recomendações e Trabalhos Futuros

Com base nos resultados, recomenda-se a utilização deste modelo como ferramenta de gestão e vigilância, e não como ferramenta de diagnóstico clínico individual. O algoritmo desenvolvido é capaz de processar milhões de registros em segundos, gerando "mapas de calor" de risco que podem orientar a alocação de recursos preventivos em municípios com taxas de mortalidade proporcionais anômalas.

Para trabalhos futuros, sugere-se:

1. Dados Climáticos: Integração com dados de temperatura, visto que extremos climáticos podem ser gatilhos para IAM.
2. Variáveis de Causa Associada: Utilizar as linhas de "causas secundárias" do atestado de óbito (comorbidades como Diabetes/Hipertensão) para tentar romper o "teto de precisão" identificado neste estudo.
3. Modelos especifistas: Testar outros modelos que lidem melhor com classes minoritárias, IAM representava apenas 6% do nosso *dataset*, um modelo que lide melhor com esse caso pode conseguir precisão mais alta.

6. Conclusão

O presente estudo demonstrou que é possível transformar dados brutos do Sistema de Informações sobre Mortalidade (SIM) em conhecimento estruturado e acionável por meio de técnicas de Engenharia de Dados e Machine Learning. A partir da construção de um Data Warehouse e da aplicação de modelos supervisionados, foi possível analisar padrões nacionais de mortalidade por Infarto Agudo do Miocárdio e propor um classificador capaz de identificar casos prováveis com alta sensibilidade mas baixa precisão.

A evolução do desempenho, partindo de um cenário inicial de Recall nulo para um modelo final capaz de atingir 90,01% de Sensibilidade, evidencia o impacto das estratégias de balanceamento, ajuste de limiar e escolha criteriosa de métricas orientadas à triagem epidemiológica. Como discutido na literatura (Gordis, 2017; Fletcher & Fletcher, 2014), a priorização do Recall sobre a *Precision* é não apenas metodologicamente adequada, como clinicamente necessária em sistemas cujo objetivo primário é não deixar casos graves passarem despercebidos.

Os resultados também revelaram uma limitação estrutural dos dados demográficos utilizados: a ausência de variáveis clínicas, comportamentais e laboratoriais estabelece um "teto de precisão" que impede a diferenciação fina entre IAM e outras causas crônicas, o que reforça que o modelo proposto deve ser compreendido como uma ferramenta de triagem populacional, e não de diagnóstico individual.

Apesar dessas limitações, o estudo oferece contribuições práticas relevantes. O *pipeline* desenvolvido possibilita processar milhões de registros de forma padronizada, permitindo identificar anomalias regionais, estimar grupos de maior vulnerabilidade e apoiar políticas públicas de prevenção e alocação de recursos em saúde. Além disso, o modelo pode ser integrado a sistemas de vigilância para geração automática de alertas epidemiológicos.

Para trabalhos futuros, recomenda-se o enriquecimento das bases por meio da vinculação com variáveis climáticas, comorbidades e causas secundárias de óbito, permitindo testar se essas informações rompem o teto de precisão observado. Abordagens mais avançadas, como redes neurais profundas ou modelos probabilísticos hierárquicos, também podem ser investigadas em bases mais ricas.

Em síntese, o estudo confirma a viabilidade de usar dados públicos e modelos preditivos para ampliar a compreensão da mortalidade cardiovascular no Brasil, oferecendo uma base quantitativa sólida para a tomada de decisão em saúde pública e abrindo caminho para sistemas automatizados de triagem epidemiológica de alta escala.

Referências

GORDIS, Leon. *Epidemiologia*. 5. ed. Rio de Janeiro: Elsevier, 2017.

FLETCHER, Robert; FLETCHER, Suzanne. *Epidemiologia Clínica: Elementos Essenciais*. 5. ed. Artmed, 2014.