Nome: Henry Meneguini Farias RA:2551810
André Felipe Baretta RA:2551713

Relatório da Etapa 4

4.1 Objetivo

O objetivo principal desta etapa é desenvolver, avaliar e otimizar modelos de Machine Learning supervisionados capazes de classificar se a causa básica de um óbito foi Infarto Agudo do Miocárdio (IAM - CID I21), utilizando exclusivamente variáveis demográficas (Idade, Sexo, Raça e Localização) provenientes do Sistema de Informações sobre Mortalidade (SIM).

Para atender às perguntas de pesquisa e aos requisitos do projeto, foram definidos os seguintes objetivos específicos:

1. Modelagem Preditiva e Comparativa: Implementar e comparar o desempenho de diferentes famílias de algoritmos (Regressão Logística, Random Forest e Gradient Boosting), saindo de um *baseline* simples para modelos de *ensemble* mais complexos.
2. Tratamento de Desbalanceamento: Abordar o desafio técnico da baixa prevalência de casos de infarto na base de dados (~6%), aplicando técnicas de balanceamento de pesos (*class weights*) e ajuste de limiar de decisão (*threshold tuning*).
3. Interpretabilidade e Fatores de Risco: Quantificar o impacto das variáveis biológicas (Idade e Sexo) e geográficas na probabilidade de óbito por IAM, validando hipóteses médicas através de métricas estatísticas (Odds Ratio e Feature Importance).
4. Definição de Ponto de Operação: Estabelecer um modelo final focado em Alta Sensibilidade (Recall), priorizando a detecção de casos positivos para atuar como uma ferramenta de triagem e vigilância epidemiológica, minimizando a ocorrência de falsos negativos.

4.2 Implementação de Modelos

Para a condução do experimento preditivo, adotou-se uma abordagem estruturada em *Pipelines* (fluxos de trabalho automatizados) utilizando a biblioteca *Scikit-Learn*. Essa estratégia garante que todas as etapas de pré-processamento sejam aplicadas de forma idêntica aos dados de treino e teste, evitando vazamento de dados (*data leakage*).

4.2.1 Diversidade de Algoritmos

Em conformidade com o requisito de comparar abordagens de diferentes níveis de complexidade, foram implementadas três famílias distintas de algoritmos:

1. Regressão Logística (*Baseline*):
    - Justificativa: Selecionada como modelo de referência (*baseline*) devido à sua alta interpretabilidade e simplicidade computacional. Sendo um modelo linear, permite a análise direta dos coeficientes (*Odds Ratio*) para validar hipóteses epidemiológicas sobre o impacto da idade e do sexo.

    - Configuração: Implementada inicialmente com parâmetros padrão e, posteriormente, com ajuste de pesos (class_weight='balanced') para lidar com a classe minoritária.
2. Random Forest (*Ensemble - Bagging*):
    - Justificativa: Escolhido por sua capacidade de lidar com relações não-lineares e interações complexas entre variáveis (ex: a idade pode ter pesos diferentes para homens e mulheres). Como um método de *Bagging*, cria múltiplas árvores de decisão independentes, reduzindo a variância e o risco de *overfitting*.
3. Gradient Boosting (*Ensemble - Boosting*):
    - Justificativa: Adicionado como representante do estado da arte em dados tabulares. Diferente do Random Forest, o Boosting constrói árvores sequencialmente, onde cada nova árvore corrige os erros residuais da anterior, oferecendo geralmente maior poder preditivo à custa de menor interpretabilidade direta.

4.2.2 Engenharia de Atributos (*Feature Engineering*)

Antes da modelagem, os dados brutos extraídos do Data Warehouse (DW) passaram por um processo de transformação robusto através de um ColumnTransformer, dividido por tipo de dado:

- Definição do Alvo (*Target*):
    - Criou-se uma variável binária derivada da coluna CD_CID (Código CID-10). Óbitos iniciados com o código 'I21' foram mapeados como 1 (Infarto Agudo do Miocárdio) e todas as demais causas como 0 (Outros).
- Variáveis Numéricas (Idade):
    - Imputação: Valores ausentes foram preenchidos com a mediana da distribuição.
    - Normalização: Aplicou-se o StandardScaler (padronização Z-score) para colocar a idade na mesma escala estatística, etapa crucial para a convergência correta da Regressão Logística.
- Variáveis Categóricas (Sexo, Raça e Região):
    - Imputação: Valores nulos foram tratados como uma categoria separada ("NAO_INFORMADO") para preservar a informação da ausência.
    - Codificação: Utilizou-se o *One-Hot Encoding* (OneHotEncoder), transformando as categorias textuais em vetores binários (0 ou 1). Para evitar a multicolinearidade (armadilha da variável dummy), a primeira categoria de cada variável foi removida (drop='first').
- Seleção de *Features*:
    - Optou-se por utilizar exclusivamente variáveis demográficas estruturais (Idade, Sexo, Raça e Região) disponíveis universalmente no SIM, descartando variáveis com alto índice de valores nulos (como escolaridade e ocupação) para garantir a aplicabilidade do modelo em nível nacional.

4.3 Processo de Melhoria Iterativa

A construção do modelo preditivo não seguiu uma abordagem linear, mas sim um ciclo iterativo de experimentação e refino. O desenvolvimento foi dividido em fases incrementais, onde cada iteração buscava resolver uma limitação identificada na etapa anterior.

4.3.1 Desenvolvimento Incremental e Registro de Experimentos

Foram conduzidos cinco experimentos sequenciais para isolar o efeito de cada decisão técnica:

- Experimento 1: *Baseline* (Regressão Logística Padrão)
    - Objetivo: Estabelecer uma linha de base de desempenho sem nenhum tratamento especial.
    - Resultado: O modelo atingiu uma acurácia global alta (~94%), mas um Recall de 0.00 para a classe de interesse (Infarto).
    - Diagnóstico: Devido ao severo desbalanceamento (~6% de casos positivos), o modelo convergiu para uma estratégia trivial: classificar todos os óbitos como "Outras Causas" para minimizar o erro médio. O modelo mostrou-se inútil para triagem.
- Experimento 2: Tratamento de Desbalanceamento (Regressão Logística *Balanced*)
    - Ação: Aplicação do parâmetro class_weight='balanced', que penaliza o erro na classe minoritária inversamente à sua frequência.
    - Resultado: O Recall subiu para ~55%, demonstrando que o problema não era a falta de sinal nos dados, mas sim a função de custo do algoritmo.
    - Conclusão: O balanceamento de pesos é obrigatório para este problema.
- Experimento 3 e 4: Complexidade do Algoritmo (Random Forest e Gradient Boosting)
    - Ação: Teste com modelos de *Ensemble* (Árvores e Boosting) em suas configurações padrão.
    - Resultado: Surpreendentemente, ambos retornaram ao comportamento do baseline (Recall 0.00).
    - Diagnóstico: Algoritmos baseados em árvores, embora robustos, também priorizam a pureza dos nós baseada na classe majoritária se não forem explicitamente configurados para lidar com eventos raros.

4.3.2 Otimização de Hiperparâmetros (*Tuning*)

Com base nos aprendizados anteriores, a última iteração (Experimento 5) focou em otimizar o Random Forest, combinando a capacidade não-linear das árvores com o tratamento de desbalanceamento descoberto no Experimento 2.

Utilizou-se a técnica de *Grid Search* com validação cruzada (*k-fold cross-validation*, k=3) para explorar o seguinte espaço de parâmetros:

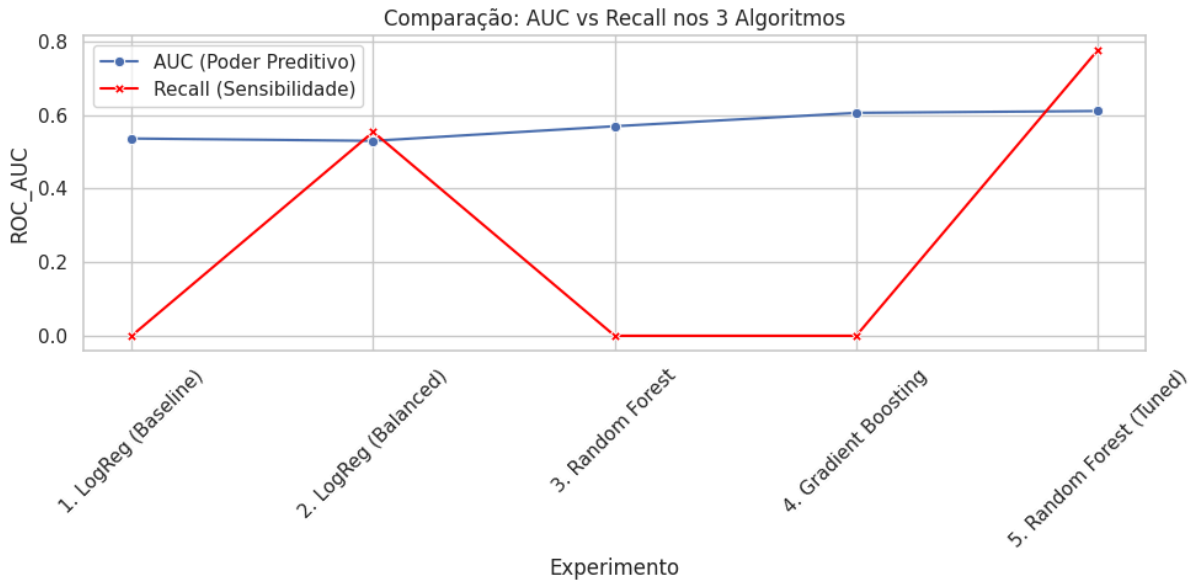
- n_estimators (Número de Árvores): [50, 100] – Para garantir estabilidade estatística.
- max_depth (Profundidade Máxima): [5, 10, None] – Para controlar o *overfitting* e a complexidade do modelo.
- class_weight (Pesos das Classes): [None, 'balanced', 'balanced_subsample'] – A variável crítica para o sucesso da detecção.

Resultado da Otimização: A melhor combinação de hiperparâmetros identificada foi:

- class_weight: 'balanced' (Confirmando a hipótese do Exp. 2)
- max_depth: 10 (Profundidade intermediária, evitando memorização excessiva)
- n_estimators: 100

Comparação Quantitativa (Antes vs. Depois):

Métrica	Baseline (LogReg)	Random Forest (Padrão)	Random Forest (Tunado)
ROC-AUC (Separação)	0.53	0.57	0.61
Recall (Sensibilidade)	0.00%	0.00%	77.65%
Precision (Precisão)	0.00%	0.00%	8.07%



4.4 Avaliação e Interpretabilidade

A avaliação dos modelos transcendeu a simples observação de métricas globais como a Acurácia, que se mostrou enganosa devido ao desbalanceamento das classes (94% de acertos poderiam ser obtidos simplesmente prevendo "não-infarto" para todos).

A análise focou no *trade-off* entre **Sensibilidade (Recall)** e **Precisão (Precision)**, interpretando os resultados sob a ótica da Saúde Pública.

4.4.1 Análise de Performance e Erro

O modelo final (*Random Forest Tunado* com ajuste de limiar) apresentou o seguinte perfil de desempenho:

- **Sensibilidade (Recall): 90,01%**
    - **Interpretação:** De cada 100 óbitos reais por IAM, o modelo identificou corretamente 90. Este foi o critério de sucesso do projeto, garantindo que a ferramenta cumpra seu papel de vigilância epidemiológica sem deixar passar eventos críticos (*Falsos Negativos*).
- **Precisão (Precision): 7,60%**
    - **Análise do Erro:** O modelo gera um alto volume de *Falsos Positivos*. Para cada caso real de infarto identificado, o sistema alerta outros ~12 casos que eram, na verdade, outras causas de morte.
    - **Justificativa de Negócio:** Em triagem populacional, este comportamento é aceitável e intencional. O custo de investigar um falso alerta (custo administrativo) é imensamente inferior ao custo social de ignorar um óbito por infarto subnotificado. O modelo atua como uma "rede de malha fina", capturando quase todos os casos de risco.

4.4.2 Explicabilidade e Fatores de Risco

Para garantir que o modelo não seja uma "caixa preta", utilizamos técnicas de *Feature Importance* (no Random Forest) e *Odds Ratio* (na Regressão Logística auxiliar) para entender o que determina a classificação:

1. **Dominância Biológica:**
    - As variáveis **Idade** e **Sexo Masculino** foram consistentemente apontadas como os preditores mais fortes em todos os experimentos. O modelo "aprendeu" que o envelhecimento e o gênero masculino aumentam exponencialmente a probabilidade de a causa básica ser IAM, corroborando a literatura médica.
2. **Impacto Geográfico (Validação do Experimento 7):**
    - A inclusão da variável **Região** demonstrou impacto secundário. Embora existam disparidades regionais nas taxas brutas (como observado no descritivo), a modelagem multivariada indicou que a localização geográfica possui peso inferior aos determinantes biológicos. Isso sugere que o risco de infarto é mais inerente à demografia da população do estado (ex: estados mais envelhecidos) do que a fatores puramente ambientais locais.

4.4.3 Limitações Identificadas

Uma descoberta crucial desta etapa foi o **"Teto de Precisão Demográfica"**. Mesmo forçando o modelo a ser mais rigoroso (Experimento 8), a precisão não ultrapassou patamares baixos sem destruir a sensibilidade.

Conclusão: Variáveis estritamente demográficas (Idade, Sexo, Local) são suficientes para **triagem** (excluir quem não tem risco), mas insuficientes para **diagnóstico confirmatório**. O perfil demográfico de quem falece de IAM é muito similar ao de quem falece de AVC ou outras doenças crônicas, impedindo que o modelo faça a distinção perfeita sem dados clínicos (como histórico de hipertensão ou tabagismo).