TECH CHALLENGE – FASE 04

PREVISÃO DO NÍVEL DE OBESIDADE

Bruna Alves de Amorim

Matheus Cesar do Amaral

Donizeti Carlos dos Santos Junior

Yuri Tierno Popic

1. INTRODUÇÃO

A obesidade é uma condição médica caracterizada pelo acúmulo excessivo de gordura corporal, estando associada a diversos riscos à saúde, como doenças cardiovasculares, diabetes mellitus tipo 2, hipertensão arterial e outras comorbidades. Trata-se de um problema de saúde pública crescente em escala global, afetando indivíduos de diferentes faixas etárias e contextos socioeconômicos. Suas causas são multifatoriais, envolvendo fatores genéticos, comportamentais, ambientais e relacionados ao estilo de vida. Nesse cenário, o uso de técnicas de Machine Learning pode contribuir de forma significativa como ferramenta de apoio à tomada de decisão clínica, auxiliando profissionais de saúde na identificação e classificação do nível de obesidade de pacientes.

Este projeto foi desenvolvido no contexto do Tech Challenge, com o objetivo de construir um sistema preditivo capaz de estimar o nível de obesidade de indivíduos a partir de dados demográficos, antropométricos e comportamentais, utilizando algoritmos de aprendizado supervisionado e disponibilizando os resultados por meio de uma aplicação interativa.

2. BASE DE DADOS E DESCRIÇÃO DAS VARIÁVEIS

Para o desenvolvimento do modelo preditivo, foi utilizada a base de dados disponibilizada no arquivo *obesity.csv*. O conjunto de dados reúne informações relacionadas ao perfil dos indivíduos, incluindo gênero, idade, altura, peso e histórico familiar de sobrepeso, além de variáveis associadas aos hábitos alimentares, consumo de água, prática de atividade física, uso de dispositivos tecnológicos, tabagismo, consumo de álcool e meio de transporte utilizado.

A variável alvo do problema corresponde ao nível de obesidade, classificado em diferentes categorias, caracterizando o problema como uma tarefa de classificação multiclasse supervisionada. A diversidade de atributos presentes na base permite capturar diferentes dimensões do comportamento e do estilo de vida dos indivíduos, tornando o conjunto de dados adequado para a aplicação de modelos de Machine Learning.

3. PREPARAÇÃO DOS DADOS E ENGENHARIA DE ATRIBUTOS

A etapa inicial do projeto consistiu na preparação e transformação dos dados, incluindo a leitura da base, análise da consistência dos registros e adequação dos tipos de variáveis para o processo de modelagem. As variáveis categóricas foram tratadas por meio de técnicas de codificação, permitindo sua utilização pelos algoritmos de aprendizado de máquina.

Em seguida, foi realizada a etapa de feature engineering, fundamental para enriquecer o conjunto de dados com informações relevantes. Destaca-se, nesse processo, o cálculo do Índice de Massa Corporal (IMC), obtido a partir do peso e da altura dos indivíduos. O IMC é amplamente utilizado na área da saúde como indicador do estado nutricional e apresentou papel central tanto no treinamento do modelo quanto nas análises exploratórias e na aplicação preditiva. Essa variável foi incorporada explicitamente ao conjunto de atributos utilizados pelo modelo.

4. MODELAGEM E AVALIAÇÃO DOS ALGORITMOS

Com os dados devidamente preparados, o conjunto foi dividido em dados de treino e teste, possibilitando a avaliação do desempenho dos modelos em dados não utilizados durante o treinamento. Foram treinados e avaliados dois algoritmos de aprendizado supervisionado: Regressão Logística e Random Forest.

A Regressão Logística foi utilizada como modelo inicial de referência, por se tratar de um algoritmo amplamente empregado em problemas de classificação. Esse modelo apresentou uma acurácia de **87,32%**, indicando desempenho consistente na identificação dos níveis de obesidade a partir das variáveis disponíveis.

Posteriormente, foi treinado o modelo de Random Forest, que apresentou desempenho superior em relação à Regressão Logística. O Random Forest alcançou uma acurácia de **98,8%**, demonstrando maior capacidade de capturar relações não lineares e interações complexas entre os atributos. Com base nessa comparação objetiva de desempenho, o Random Forest foi definido como o modelo final da solução e salvo para utilização no ambiente de produção.

5. SISTEMA PREDITIVO E DEPLOY DA APLICAÇÃO

Após a definição do modelo final, foi desenvolvido um sistema preditivo utilizando o framework Streamlit. A aplicação carrega o modelo treinado e permite a realização de previsões em tempo real, a partir da inserção dos dados do paciente por meio de uma interface interativa. O sistema calcula automaticamente o IMC com base nos valores informados de peso e altura e utiliza todas as variáveis necessárias para gerar a previsão do nível de obesidade.

O resultado da predição é apresentado de forma clara, indicando a classe prevista e as probabilidades associadas a cada nível de obesidade, proporcionando maior transparência e suporte à interpretação dos resultados por parte da equipe médica. O modelo utilizado na aplicação é o Random Forest previamente treinado e validado.

6. ANÁLISE EXPLORATÓRIA E VISÃO ANALÍTICA

Além da funcionalidade preditiva, a aplicação desenvolvida incorpora uma visão analítica baseada na análise exploratória dos dados. Essa visão apresenta indicadores e gráficos que permitem compreender os principais padrões associados à obesidade na base de dados utilizada. São exibidas distribuições dos níveis de obesidade, análises por gênero, idade média, IMC médio, consumo de água e prática de atividade física.

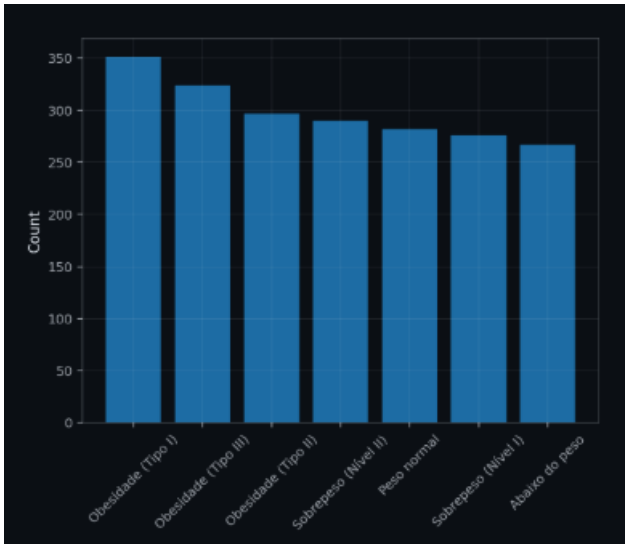
Também são disponibilizadas visualizações como histogramas, boxplots, gráficos de dispersão e matrizes de correlação entre variáveis numéricas, possibilitando uma análise mais aprofundada das relações entre os atributos. Essa abordagem oferece à equipe médica e aos gestores de saúde insights relevantes sobre os fatores associados à obesidade, contribuindo para uma visão mais estratégica do problema.

A análise exploratória foi conduzida por meio de um painel analítico integrado à aplicação, permitindo a visualização interativa dos dados utilizados no treinamento do modelo. O painel apresenta inicialmente indicadores agregados, nos quais se observa um total de 2.087 indivíduos após a aplicação de filtros, com predominância da classe Obesidade (Tipo I), representando

aproximadamente 16,8% da amostra. O IMC médio observado é de 29,77, valor consistente com a predominância de classes relacionadas ao sobrepeso e à obesidade. A idade média da amostra é de 24,4 anos, confirmando o perfil jovem da população analisada, enquanto o consumo médio de água (CH2O) situa-se em torno de 2,01, correspondente a um nível intermediário.

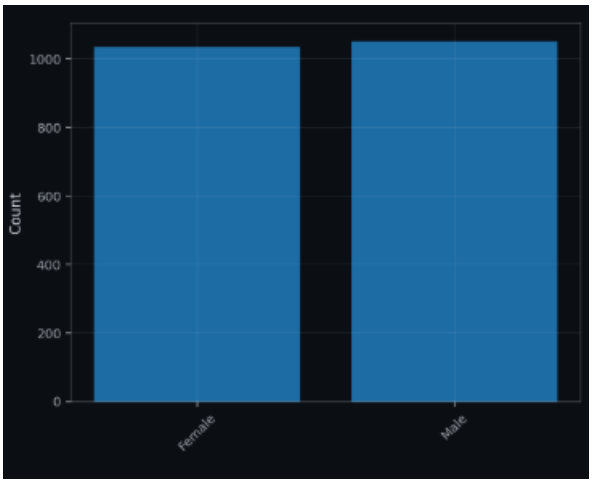
A distribuição dos indivíduos por classe de obesidade mostra que as categorias estão relativamente bem representadas, sem concentração extrema em uma única classe, conforme demonstrado no seguinte gráfico.

Distribuição por classe



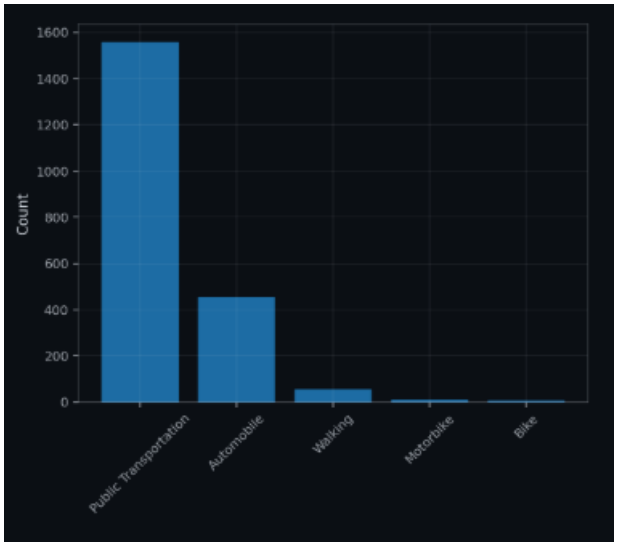
Esse equilíbrio contribui para a capacidade do modelo de diferenciar padrões entre os diferentes níveis de obesidade. A análise por gênero evidencia uma divisão bastante equilibrada entre indivíduos do sexo feminino e masculino, indicando que o gênero não atua como fator dominante isolado na classificação do nível de obesidade.

Distribuição por gênero

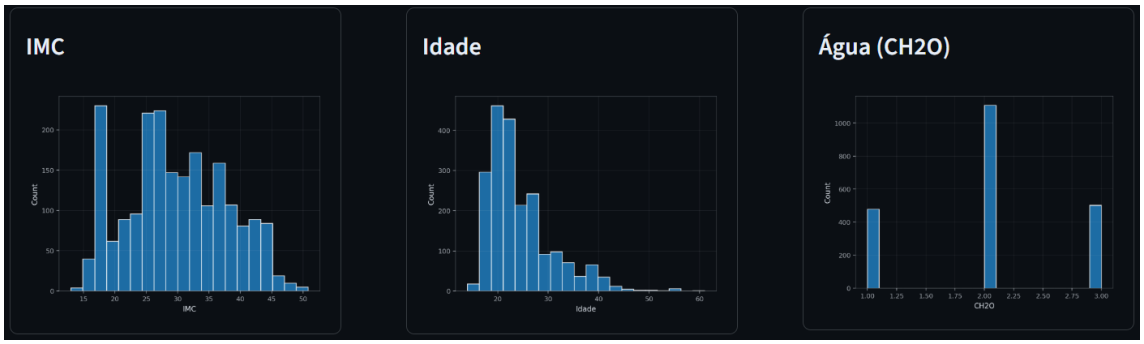


O gráfico de distribuição dos meios de transporte utilizados indica predominância do transporte público, seguido pelo uso de automóvel, enquanto modos ativos como caminhada e bicicleta apresentam menor representatividade. Esse padrão sugere comportamentos potencialmente mais sedentários em parte da população, o que é coerente com os níveis elevados de IMC observados em diversas classes.

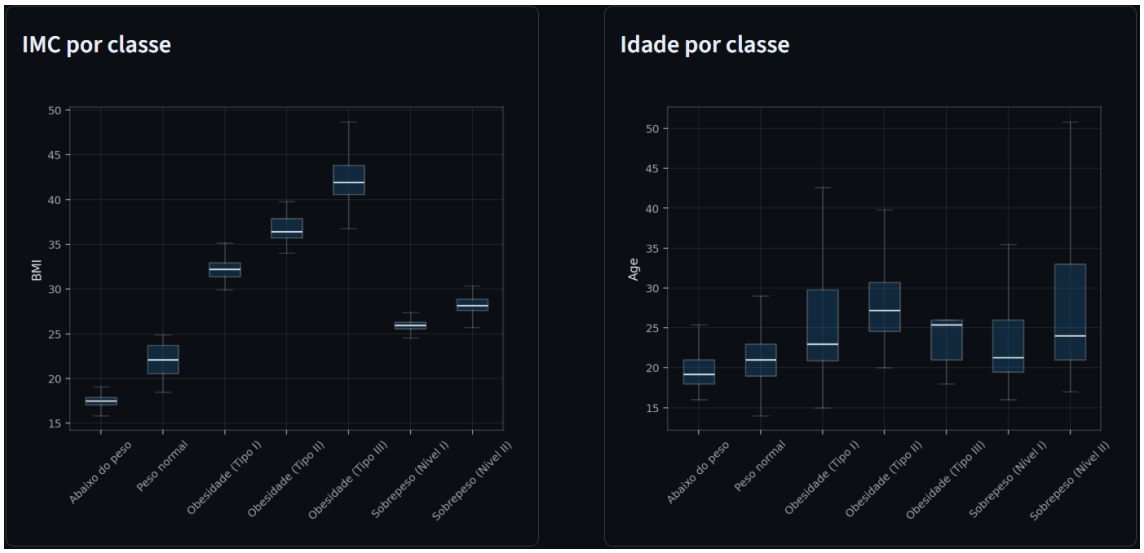
Transporte (estimado)



As distribuições individuais mostram que o IMC apresenta ampla dispersão, com concentração significativa entre valores associados ao sobrepeso e à obesidade. A distribuição da idade reforça a concentração em faixas etárias mais jovens, enquanto o consumo de água se concentra majoritariamente nos níveis intermediários da escala utilizada.



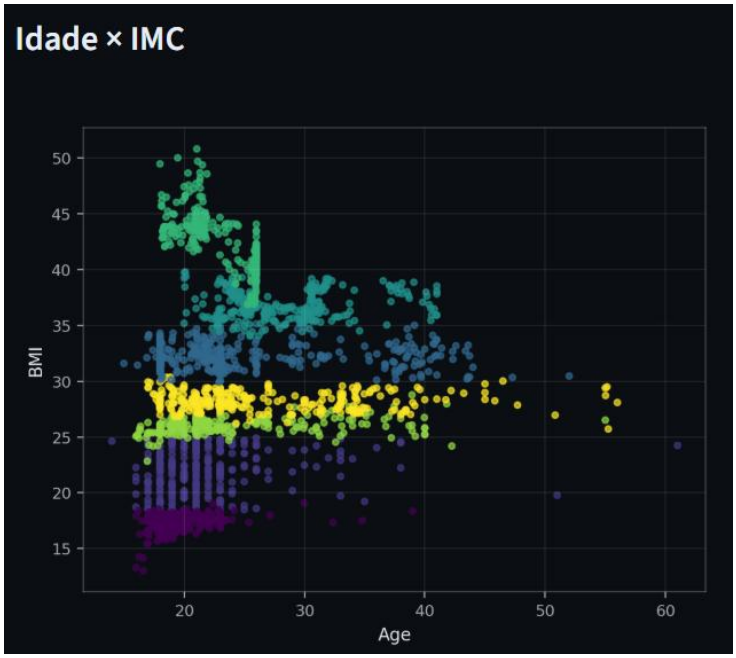
A análise do IMC por classe de obesidade, por meio do boxplot “IMC por classe” e “Idade por classe”, evidencia uma separação clara entre as categorias, com medianas progressivamente mais elevadas conforme o nível de obesidade aumenta. A classe Obesidade (Tipo III) apresenta os maiores valores de IMC, com pouca sobreposição em relação às classes de peso normal e abaixo do peso, reforçando o papel central do IMC como variável discriminante. A análise da idade por classe mostra sobreposição entre categorias, mas com tendência de idades medianas ligeiramente mais elevadas em classes mais severas de obesidade, indicando que a idade atua como fator complementar.



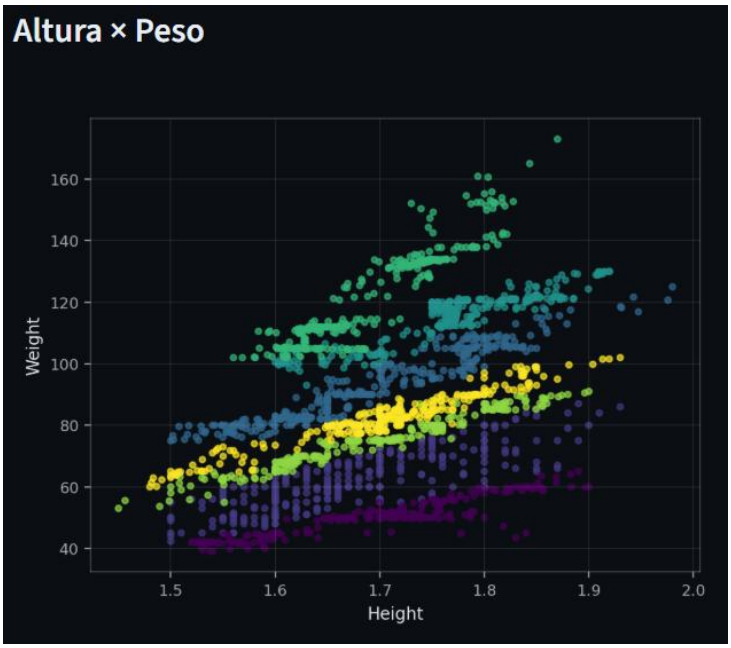
A matriz de correlação entre variáveis numéricas evidencia forte correlação positiva entre peso e IMC, bem como entre peso e altura, resultados esperados do ponto de vista antropométrico. Observa-se também correlação moderada entre idade e IMC, enquanto variáveis como prática de atividade física (FAF) e tempo de uso de dispositivos tecnológicos (TUE) apresentam correlações fracas, indicando relações mais complexas e não lineares. Esses achados ajudam a

explicar o desempenho superior do modelo Random Forest em relação à Regressão Logística.

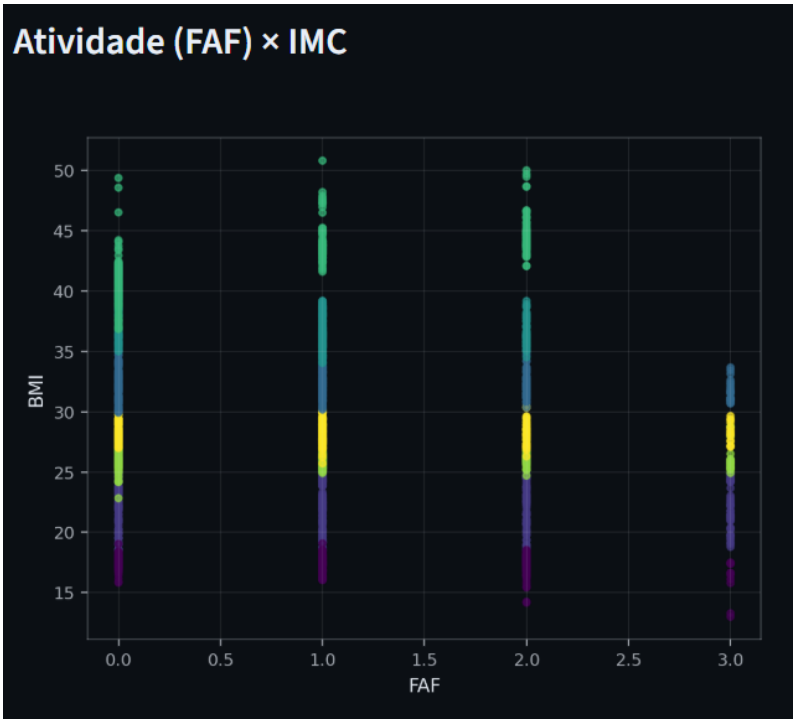
Os gráficos de dispersão reforçam essas conclusões. A relação entre idade e IMC mostra grande variabilidade dos valores de IMC em diferentes faixas etárias, indicando que a idade, isoladamente, não explica o nível de obesidade.



O gráfico de altura versus peso apresenta o padrão linear esperado, mas evidencia que indivíduos com alturas semelhantes podem apresentar pesos muito distintos, impactando diretamente o IMC.



Por fim, a relação entre prática de atividade física e IMC sugere que níveis mais elevados de atividade física estão associados, em média, a menores valores de IMC, embora com ampla dispersão, reforçando o caráter multifatorial da obesidade.



7. CONSIDERAÇÕES FINAIS

Os resultados obtidos ao longo do desenvolvimento do projeto demonstram que o uso de técnicas de Machine Learning é eficaz para a classificação do nível de

obesidade a partir de dados demográficos, antropométricos e comportamentais. A comparação entre os modelos treinados evidenciou diferenças relevantes de desempenho. A Regressão Logística, utilizada como modelo de referência, apresentou uma acurácia de 87,32%, indicando boa capacidade de identificar padrões nos dados, mesmo considerando sua abordagem linear. Esse resultado mostra que parte significativa da relação entre as variáveis e o nível de obesidade pode ser explicada de forma relativamente simples, especialmente por atributos diretamente relacionados à composição corporal, como peso, altura e IMC.

Entretanto, o modelo de Random Forest apresentou desempenho superior, alcançando uma acurácia de 98,8%. Esse ganho expressivo indica que o problema se beneficia de uma abordagem capaz de capturar relações não lineares e interações mais complexas entre os atributos, como a combinação entre hábitos alimentares, consumo de água, prática de atividade física, histórico familiar e tempo de exposição a comportamentos sedentários. A robustez do Random Forest torna o modelo particularmente adequado para o contexto do problema, no qual diferentes fatores atuam de forma conjunta na determinação do nível de obesidade.

A aplicação prática desses resultados foi viabilizada por meio do deploy do modelo em uma aplicação interativa desenvolvida com Streamlit. O sistema preditivo permite a realização de inferências em tempo real, apresentando não apenas a classe prevista, mas também as probabilidades associadas a cada nível de obesidade. Essa abordagem oferece maior transparência ao processo de decisão e possibilita que a equipe médica utilize o modelo como uma ferramenta de apoio, e não como substituto da avaliação clínica. A presença do IMC como variável calculada automaticamente e utilizada pelo modelo reforça a aderência do sistema à prática médica.

Complementarmente, a visão analítica construída a partir da análise exploratória dos dados possibilitou identificar padrões relevantes na base estudada, como diferenças nos níveis de obesidade associadas à idade, ao consumo de água, à prática de atividade física e ao comportamento alimentar. Essas análises fornecem subsídios importantes para uma compreensão mais ampla do perfil

dos indivíduos, permitindo que decisões clínicas e estratégias de prevenção sejam orientadas por evidências extraídas dos dados.

Com base nos resultados obtidos, recomenda-se que o sistema preditivo seja utilizado como uma ferramenta de suporte à triagem e ao acompanhamento de pacientes, especialmente em contextos nos quais há grande volume de atendimentos e necessidade de priorização de casos. A aplicação pode auxiliar na identificação precoce de indivíduos com maior risco de obesidade, permitindo intervenções mais rápidas e direcionadas. Além disso, a visão analítica pode ser utilizada pela gestão hospitalar e por equipes multidisciplinares como apoio à definição de programas de prevenção, educação alimentar e incentivo à prática de atividade física.

Como recomendações para evoluções futuras do projeto, destaca-se a possibilidade de ampliar a base de dados com novos registros, incorporar variáveis clínicas adicionais e realizar avaliações contínuas do desempenho do modelo à medida que novos dados forem coletados. Essas ações podem contribuir para manter a eficácia do sistema ao longo do tempo e ampliar seu potencial de aplicação no apoio à tomada de decisão em saúde.

REFERÊNCIAS

RIBEIRO, M. **Obesidade**. Disponível em: <https://drauziovarella.uol.com.br/doencas-e-sintomas/obesidade/>.

WORLD HEALTH ORGANIZATION. **Obesity and Overweight**. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.