

POSTECH

TECH CHALLENGE 4

Alberto de Franca Marchiori
Alef de Sousa Pereira
Leticia Lauria Lopes

Modelo Preditivo Interativo de Obesidade
Grupo 150

Pauta

O que este relatório aborda

- Objetivo do Projeto
- Aquisição e Limpeza dos Dados
- Engenharia de Atributos
- Análise Exploratória dos Dados
- Preparação e Treinamento de Modelo Preditivo
- Aplicação no Streamlit

Objetivo do Projeto

Criar um modelo de Machine Learning para auxiliar a previsão de obesidade com base nos hábitos declarados pela pessoa, por meio de aplicação interativa utilizando o Streamlit.

Para critérios mínimos de sucesso, buscamos uma acurácia mínima de 75% das previsões realizadas.

Aquisição e Limpeza dos Dados

Foi utilizada a base disponibilizada de pessoas do desafio contendo informações sobre os hábitos e classificação de obesidade de cada um.

Realizamos a limpeza e transformação da base com:

- Conversão dos campos binários de Gênero, Histórico Familiar, FAVC, SMOKE e SCC para valores de 0 e 1
- Conversão dos campos de frequência CAEC e CALC para valores de 0 a 3, sendo 0 equivalente a 'Nunca' e 3 equivalente a 'Sempre'
- Conversão do campo de meio de transporte principal MTRANS para valores equivalentes a baixa, média e alta intensidade física
- Conversão dos campos de quantidade AGE, FCVC, NCP, CH2O, FAF e TER para inteiro devido a erros de ponto flutuante contidos na base
- Conversão do Peso e Altura para duas casas decimais

Dicionário de dados:

- Gender: Gênero.
- Age: Idade.
- Height: Altura em metros.
- Weight: Peso em kgs.
- family_history: Algum membro da família sofreu ou sofre de excesso de peso?
- FAVC: Você come alimentos altamente calóricos com frequência?
- FCVC: Você costuma comer vegetais nas suas refeições?
- NCP: Quantas refeições principais você faz diariamente?
- CAEC: Você come alguma coisa entre as refeições?
- SMOKE: Você fuma?
- CH2O: Quanta água você bebe diariamente?
- SCC: Você monitora as calorias que ingere diariamente?
- FAF: Com que frequência você pratica atividade física?
- TER: Quanto tempo você usa dispositivos tecnológicos como celular, videogame, televisão, computador e outros?
- CALC: Com que frequência você bebe álcool?
- MTRANS: Qual meio de transporte você costuma usar?
- Obesity_level (coluna alvo): Nível de obesidade

Engenharia de Atributos

Criação de Variáveis

Classificação de Obesidade

Criamos três diferentes classificações para as métricas de obesidade para aplicação no modelo, sendo elas:

- Classificação em Dois Níveis: os níveis de Peso Insuficiente, Normal e Sobrepeso foram agrupadas como Não-Obesidade
- Classificação em Três Níveis: os Níveis de Peso Insuficiente e Normal foram agrupados como Não-Obesidade, e os de Sobrepeso e Obesidade foram agrupados em classes respectivas
- Classificação em Quatro Níveis: cada classificação foi agrupada individualmente

Definição do Target

O que o modelo deverá prever

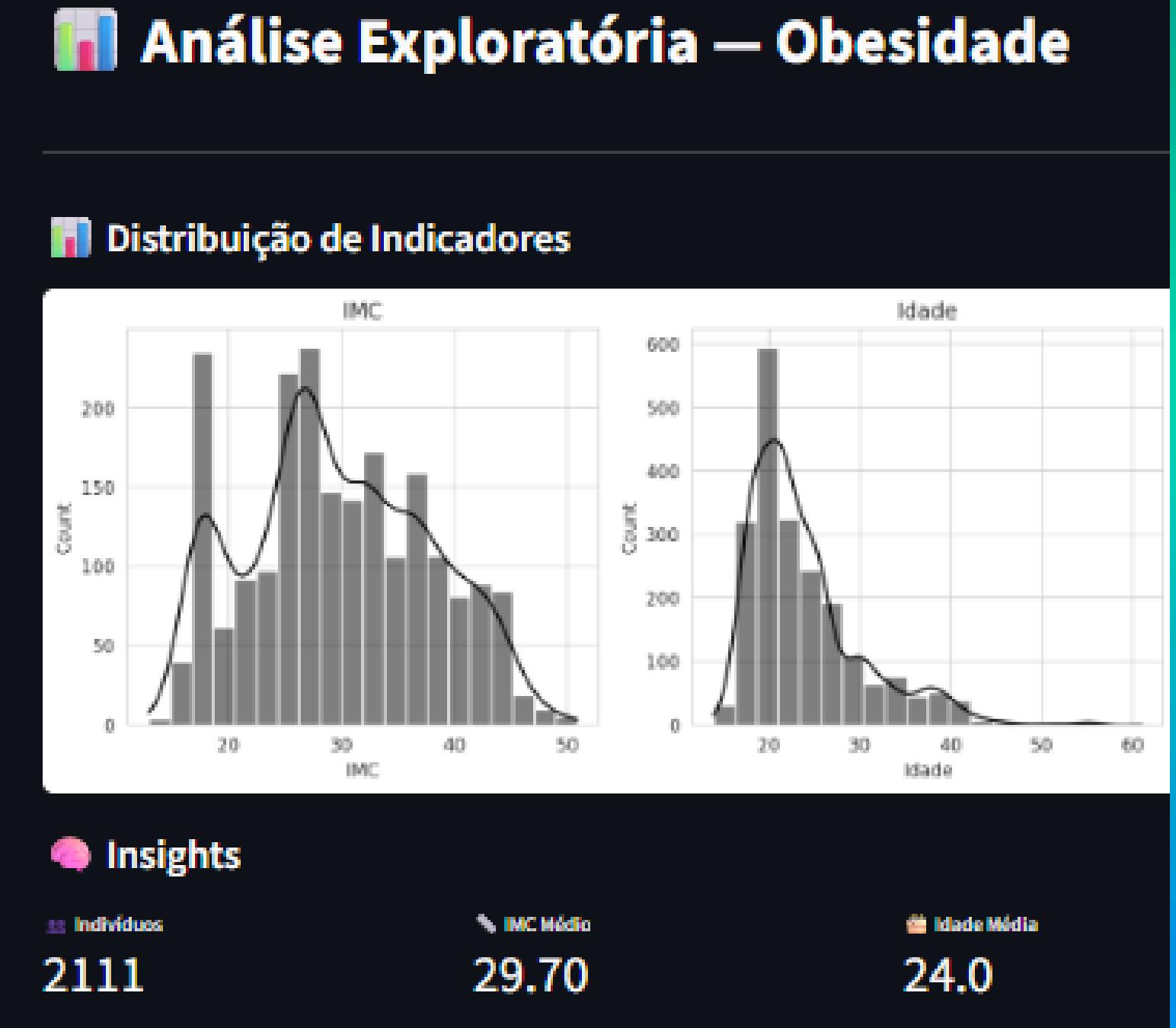
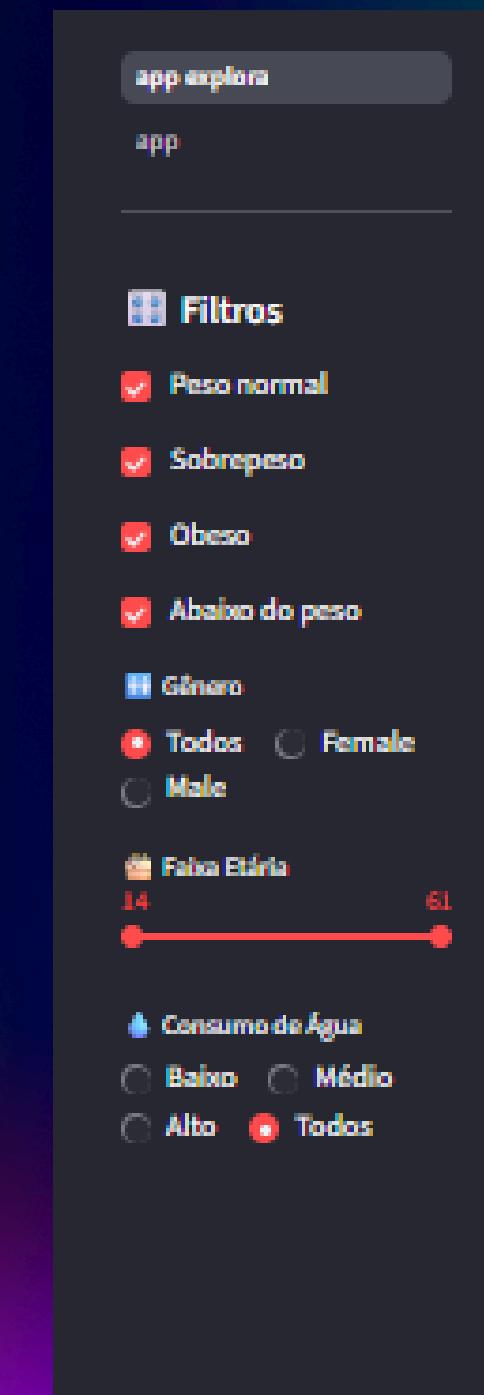
O target do modelo utilizado foi definido utilizando a classificação de obesidade criada.

Para evitar overfitting do modelo, os campos anteriores de Nível de Obesidade, Peso e Altura foram eliminados, pois estão fortemente relacionados ao cálculo do IMC e à classificação.

Análise Exploratória dos Dados

A aplicação dispõe de uma página interativa para filtragem de critérios para análise exploratória dos dados e suas métricas correspondentes.

Podemos ver, no exemplo, que se trata de uma base majoritariamente jovem, com IMC médio de sobrepeso, histórico familiar de obesidade e alto consumo calórico.



Preparação da Base para Previsão

Modelos testados:
Regressão Logística
Random Forest
XGBoost

A base foi dividida em treino e teste, utilizando a volumetria de 80% de dados para treino e 20% para teste, com estratificação.

Todos os atributos de hábito foram utilizados para previsão nos modelos, com a classificação em níveis criada sendo utilizada como Target.

Nos testes, foi observado que todas as divisões de classificação obtiveram bons resultados preditivos, com os resultados no modelo escolhido sendo de 91% para dois níveis, 84% para três e 82% para quatro.

Para interesse da aplicação, preferimos a utilização de três níveis de classificação, uma vez que identificar o sobre peso é ideal antes do desenvolvimento para obesidade.

Treinamento dos Modelos

	precision	recall	f1-score	support
0	0.77	0.78	0.77	112
1	0.39	0.18	0.25	116
2	0.69	0.90	0.78	195
accuracy			0.67	423
macro avg	0.62	0.62	0.60	423
weighted avg	0.63	0.67	0.63	423

	precision	recall	f1-score	support
0	0.85	0.88	0.86	112
1	0.75	0.77	0.76	116
2	0.91	0.88	0.89	195
accuracy			0.85	423
macro avg	0.84	0.84	0.84	423
weighted avg	0.85	0.85	0.85	423

	precision	recall	f1-score	support
0	0.85	0.82	0.84	112
1	0.73	0.80	0.76	116
2	0.91	0.87	0.89	195
accuracy			0.84	423
macro avg	0.83	0.83	0.83	423
weighted avg	0.84	0.84	0.84	423

Regressão

Testado - 63% de Acurácia

- Maior precisão para Peso Normal (77%) e Obesidade (78%) que para Sobrepeso (25%)
- Testado principalmente na configuração de classificação em dois níveis, onde teve acurácia de 73%, sendo mais indicado para previsões binárias

Random Forest

Testado - 85% de Acurácia

- Precisão equilibrada de 86% para Peso Normal, 76% para Sobrepeso e 89% para Obesidade
- Simples e estável para uso, bom para baseline de classificações

XGBoost

Escolhido - 84% de Acurácia

- Precisão equilibrada de 84% para Peso Normal, 76% para Sobrepeso e 89% para Obesidade
- Utilização de LOGLOSS para melhorar a consistência do modelo
- Considerado melhor para capturar padrões complexos entre hábitos

Aplicação no Streamlit

A aplicação publicada no Streamlit faz a chamada do modelo treinado utilizando dados de entrada do usuário para retornar a previsão de risco de obesidade.

Também oferecemos uma página interativa, como mencionado anteriormente (slide 6) para possibilitar a análise de insights da base utilizada para treino.

 **Sistema Preditivo de Risco de Obesidade**

Este aplicativo utiliza um modelo de Machine Learning (XGBoost) para prever o risco de obesidade de um paciente com base em fatores de estilo de vida.

Preencha os dados a esquerda e clique em Fazer Previsão de Risco



Dados do Paciente

Gênero: Feminina

Idade (anos): 31

Tem histórico familiar de Obesidade?: Não

Consume Fast food frequentemente?: Não

É Fumante?: Não

Você monitora as calorias que ingere diariamente?: Não

Come alimentos calóricos entre as refeições?: Nunca

Consume Álcool ?: Nunca

Consume Vegetais ?: Nunca

Passa muito tempo no celular?:

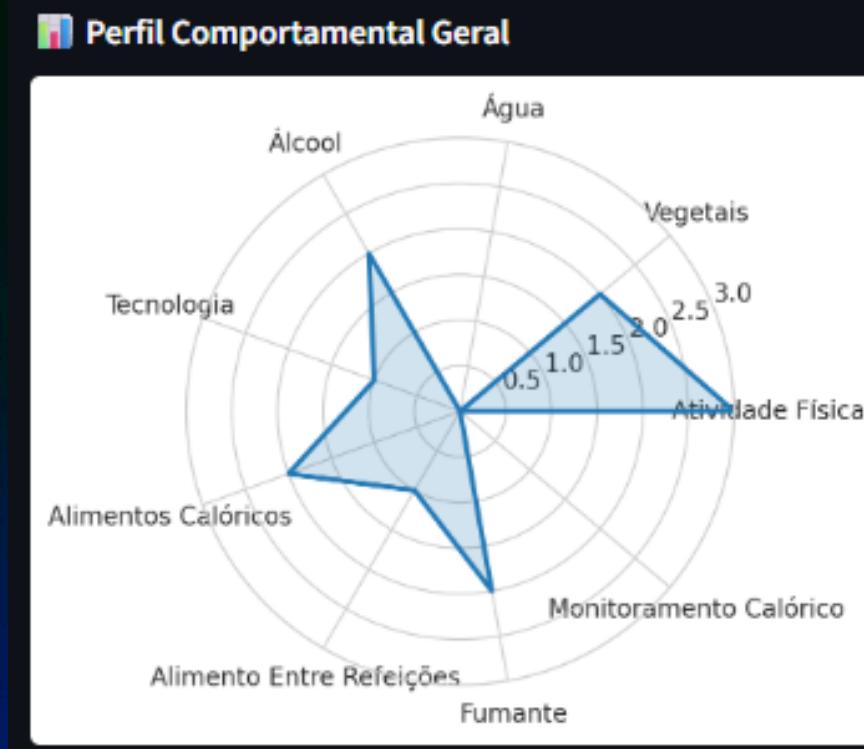
Fazer Previsão de Risco

 Voltar para
Análise
Exploratória

Previsão de risco do paciente

Além da previsão de risco, também retornamos os insights dos comportamentos prejudiciais, benéficos e visão geral de forma personalizada.

A classificação de obesidade retornada apresenta também o percentual de confiabilidade específico da previsão feita utilizando a probabilidade de classe da função predict_proba, comparando os hábitos preenchidos com a base utilizada

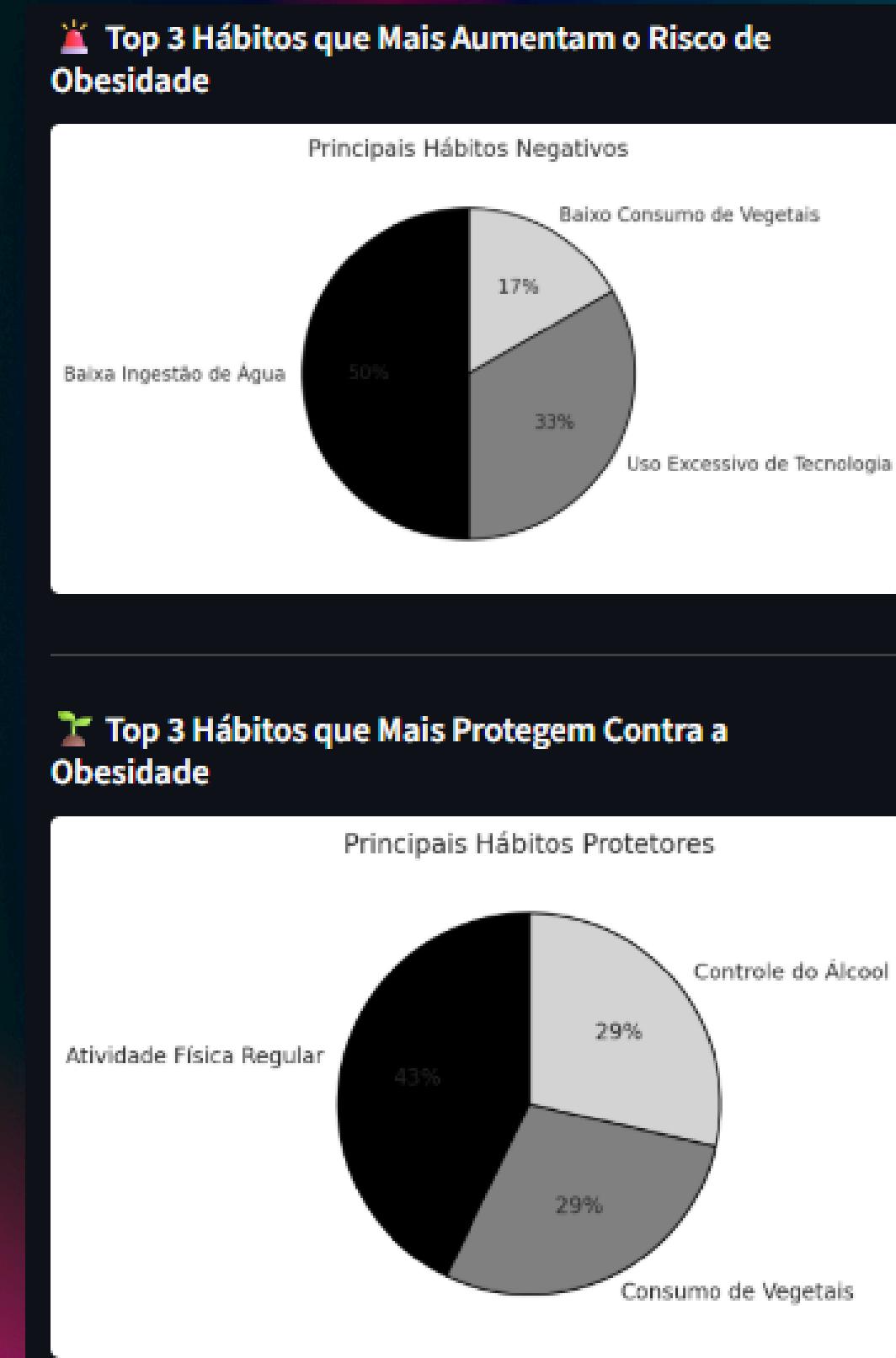


Resultado da Previsão para a Equipe Médica

Status de Risco: SOBREPESO / RISCO MÉDIO

O modelo prevê que o paciente está na categoria SOBREPESO / RISCO MÉDIO com 68.40% de confiança.

Risco MÉDIO: Recomenda-se monitoramento e ajuste de hábitos para evitar progressão para obesidade.



Obrigado!

Caso tenha alguma dúvida, entre em contato.

Alberto Marchiori

RM362799

Alef Pereira

RM362855

Leticia Lopes

RM362795