# Classificação de indivíduos com diabetes conforme estilo de vida e saúde

Caio Macedo Alves[1]*; Julio Cesar Ruivo Costa[2]

[1] Pós-Graduando do MBA em Data Science e Analytics. R. Alecrim, 246 – Vila Kosmos; 21220-050 Rio de Janeiro, Rio de Janeiro, Brasil
[2] DoutorMestre em ciências com ênfase em modelagem molecular. R. da Reitoria, 373 – Butantã; 05508-220 São Paulo, São Paulo, Brasil
*autor correspondente: caiomacedoalves@hotmail.com

.

# Classificação de indivíduos com diabetes conforme estilo de vida e saúde

## Resumo

Este trabalho aplicou técnicas de aprendizado de máquina em uma base de dados do "Centers for Disease Control and Prevention" (CDC), contendo informações sobre saúde e estilo de vida de indivíduos diabéticos e não diabéticos. O objetivo foi identificar um classificador eficiente e compreender os indicadores de saúde mais relevantes para a presença da doença. Após testes com diferentes algoritmos, o modelo selecionado apresentou métricas de desempenho satisfatórias, incluindo acurácia e sensibilidade. Utilizando valores SHAP, foi possível avaliar o impacto de variáveis como pressão alta, índice de massa corporal (IMC) e colesterol alto no risco de diabetes, demonstrando a relevância desses fatores na prática médica.

**Palavras Chave:** Aprendizado de Máquina Supervisionado, Modelagem Estatística, Diabetes

.

## Introdução

A diabetes é uma doença caracterizada pela elevação da glicose no sangue, se manifestando por diversos sintomas como perda de peso, visão turva, polifagia, polidipsia, poliúria, entre outros. A hiperglicemia está relacionada com a falência de órgãos tais como olhos, rins, coração e vasos sanguíneos. Com os malefícios da doença, torna-se imprescindível o diagnóstico precoce para minimizar os efeitos. De acordo com os autores, temos a diabetes do tipo 1 e 2. O primeiro ocorre quando há a destruição das células produtoras de insulina pelos anticorpos do indivíduo, consequência de um defeito no sistema imunológico, tornando-se uma doença autoimune. Já a do tipo 2 é resultado da resistência à insulina e da deficiência na secreção da mesma (Gross et al.,2002).

Segundo Zimmet (2017) a diabetes, em especial a do tipo 2, é a maior epidemia global de todos os tempos. Isso se dá pelo frequente aumento de casos e, segundo o autor, as estimativas dos casos globais são constantemente subestimadas pela Federação Internacional de Diabetes (FDI), órgão responsável pela disponibilização dos dados de diabetes no mundo. De acordo com a FDI, no mundo há cerca de 415 milhões de diabéticos, número que já ultrapassa o total populacional nos Estados Unidos, que conta com 320 milhões de pessoas residindo no país (Zimmet, 2017). Ou seja, o número de diabéticos ultrapassa o número de residentes de diversos países, indicando que uma considerável parcela da sociedade sofre com os malefícios da doença.

De acordo com Abbot e Barbosa (2015), algumas epidemias podem ser combatidas de forma eficaz através da Tecnologia da Informação e Comunicação (TIC) em países em que profissionais da saúde são escassos e a telefonia celular é abundante. Por exemplo, a Libéria possui cerca de 4 milhões de habitantes e 200 médicos, e pela escassez de médicos perante o tamanho da população, o país pode não estar preparado em casos de alta demanda médica (Abbot e Barbosa, 2015). Por isso as TIC's auxiliam os habitantes a terem uma melhor compreensão das doenças, ajudam a prevenir a disseminação e como cuidar dos doentes. Além de situações de emergência, as TIC's são fundamentais no monitoramento, vigilância de doenças e gestão da cadeia de abastecimento.

Com a necessidade de diagnóstico precoce da doença aliado com o avanço tecnológico, novas técnicas de classificação surgem. Uma delas é a aplicação de modelos de "Machine Learning" para detectar indivíduos que possuem ou não uma determinada condição com base em uma série de informações relacionadas à saúde e ao estilo de vida. O artigo de Sposito et al. (2022) relata muito bem esse ponto, onde a autora consegue explicar de forma resumida o que é Aprendizado de Máquina, e as principais características de alguns modelos utilizados na medicina como Árvores de Decisão, Floresta Aleatória, SVM ("Support Vector

.

Machine"), Redes Bayesianas e outras técnicas. Os autores também relatam a aplicabilidade desses modelos em diversas áreas da medicina devido a alta capacidade dos modelos em detectar padrões e portanto realizar diagnósticos. Também podem ser utilizadas em diversas áreas, como no campo da dermatologia, onde um modelo de "Deep Learning" conseguiu classificar lesões de pele como benignas ou malignas com resultados similares a 21 dermatologistas. Em psicologia, um modelo de "Machine Learning" conseguiu reduzir os critérios de diagnósticos de 29 para 8 obtendo 100% de acurácia ao detectar pessoas com autismo. Vale ressaltar que os modelos são uma ferramenta que auxilia os profissionais da saúde a conseguir um diagnóstico precoce e focar os esforços nas condições clínicas do paciente. Com avanço da tecnologia permitindo a digitalização de prontuários médicos, junto com informações de exames de imagens e laborais, obtém-se um grande volume de dados que podem ser utilizados pelo modelo.

O artigo de Garcia et al. (2020) também ressalta a utilização de Inteligência Artificial na medicina, em especial "Deep Learning". Além de detectar o diagnóstico da doença, modelos de "Deep Learning" que utilizam imagens podem também auxiliar na vigilância epidemiológica, uma vez que conseguem também identificar insetos vetores de doenças. Como exemplo, a doença de Chagas que é transmitida pelo inseto conhecido como barbeiro onde há cerca de 152 subfamílias do inseto. É necessários anos de prática e estudo entre os especialistas da área para identificar corretamente o inseto vetor, e com um modelo de Deep Learning bem treinado a identificação passa a ser em instantes. Os autores enfatizam que a ferramenta de Inteligência Artificial não substitui o médico ou especialista, mas auxilia na confirmação e guia do diagnóstico clínico.

Este trabalho tem como objetivo colaborar com os estudos epidemiológicos utilizando modelos classificatórios de "Machine Learning" para entender melhor a relação da diabetes com o estilo de vida e indicadores de saúde. Além disso, selecionar o classificador que possui melhores métricas de desempenho para obter uma maior assertividade no diagnóstico da doença.

.

## Material e Métodos

### 1.0 Base de Dados

Para realizar este trabalho, foi utilizado uma base de dados pública disponibilizada no Kaggle denominada “Diabetes Health Indicator Dataset”. Esse é um conjunto de dados de uma pesquisa americana feita pelo “Centers for Disease Control and Prevention” (CDC), onde anualmente é realizada uma pesquisa telefônica com diversas perguntas em que há a coleta de informações relacionadas à saúde e estilo de vida de cada paciente, tendo também a indicação da presença ou não de diabetes em cada um dos entrevistados.

A base de dados é da pesquisa feita no ano de 2015, contendo respostas de 70.692 pessoas (números de linhas) e 22 indicadores (número de colunas) considerando a variável resposta que representa o diagnóstico de diabetes em formato binário. As variáveis explicativas que foram utilizadas são de caráter pessoal, demográfico e comportamental (estilo de vida) de cada pessoa. O dicionário completo das variáveis está na Tabela 1 abaixo:

Tabela 1. Descrição das variáveis

| Nome da variável | Tipo da variável | Descrição da Variável |
|---|---|---|
| Diabetes_binary | Binário | 0 = sem diabetes<br>1 = pré-diabetes ou diabetes |
| HighBP | Binário | 0 = não tem pressão alta<br>1 = tem pressão alta |
| HighChol | Binário | 0 = não tem colesterol alto<br>1 = tem colesterol alto |
| CholCheck | Binário | 0 = não checkou o colesterol em 5 anos<br>1 = checkou o colesterol em 5 anos |
| BMI | Numérico | Índice de Massa Corporal |
| Smoker | Binário | Você já fumou 100 cigarros na sua vida? 0 = não 1 = sim |
| Stroke | Binário | Já teve AVC? 0 = não 1 = sim |
| HeartDiseaseorAttack | Binário | Já teve doença coronariana ou infarto do miocárdio? 0 = não, 1 = sim |
| PhysActivity | Binário | Fez atividade física nos últimos 30 dias sem incluir o trabalho. 0 = não, 1 = sim |
| Fruits | Binário | Consome frutas 1 ou mais vezes por dia? 0 = não, 1 = sim |

.

| | | |
|---|---|---|
| Veggies | Binário | Consome vegetais 1 ou mais vezes por dia? 0 = não, 1 = sim |
| HvyAlcoholConsump | Binário | Tem consumido bebida alcoolica? (Homens mais de 14 drinks por semana e mulheres mais de 7)? 0 = não, 1 = sim |
| AnyHealthcare | Binário | Tem algum tipo de planos/seguro saúde? 0 = não, 1 = sim |
| NoDocbcCost | Binário | Nos últimos 12 meses você precisou de uma consulta médicas e não conseguiu custear? 0 = não, 1 = sim |
| GenHlth | Categórico | No geral, sua saúde é? 1 = excelente, 2 = muito boa, 3 = boa, 4 = razoável, 5 = ruim |
| MentHlth | Inteiro | Considerando saúde emocional que inclui estresse, depressão e problemas emocionais, por quantos dias nos últimos 30 sua saúde emocional não foi boa? 1 – 30 |
| PhysHlth | Inteiro | Considerando saúde física que inclui doenças e lesões físicas, por quantos dias nos últimos 30 sua saúde física não foi boa? 1 – 30 |
| DiffWalk | Binário | Você tem muita dificuldade para andar ou subir escadas? 0 = não, 1 = sim |
| Sex | Binário | 0 = feminino, 1 = masculino |
| Age | Categórico | Escala de 1 a 13 |
| Education | Categórico | 1 = nunca frequentou a escola ou apenas jardim de infância, 2 = 1ª a 8ª serie (ensino fundamental), 3 = 9ª a 11ª serie (algum ensino médio), 4 = ensino médio completo, 5 = cursando faculdade, 6 = faculdade completa |
| Income | Categórico | Escala de 1 a 8 |

Fonte: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

**2.0 Pré-processamento**

Como a base de dados possui variáveis do tipo categórica e numérica, foram realizados os tratamentos adequados a cada tipo.Para as variáveis numéricas (BMI, MentHlth e PhysHlth) foi feita a normalização Z-Score para padronizar a escala das variáveis. Esse método de padronização consiste em transformar os dados para que tenham média zero e

.

desvio padrão um, sendo bastante utilizado em métodos de aprendizado de máquinas supervisionado. Assim, a fórmula do Z-Score é descrita como:

$$Z = \frac{X - \mu}{\sigma}, \quad (1)$$

sendo Z o valor padronizado, X a observação, $\mu$ e $\sigma$ as respectivas médias e desvio padrão da variável numérica.

Já para as variáveis categóricas com mais de duas categorias (GenHlth,Age, Education e Income) foi realizado o processo para transformá-las em variáveis “dummys”. Essa técnica consiste em transformar todas as variáveis categóricas em binárias 0 ou 1. Por exemplo, supondo que uma variável categórica contenha 3 possíveis valores, cria-se então 3 variáveis binárias para representar os 3 valores. Ao realizar esse processo é importante se atentar ao problema de multicolinearidade do modelo, uma vez que uma das variáveis binárias dummy pode ser escrita em forma de combinação linear das outras variáveis. Devido a isso, a categoria 1 foi excluída para as quatro variáveis que passaram pelo processo de dummy.

Após a realização do tratamento adequado para cada tipo de variável, o dataset foi separado em 80% treino e 20% teste, através de amostragem aleatória mantendo a proporção da variável target. Para essa parte do estudo foram utilizados os pacotes “scikit-learn” e “pandas” do software python.

## 3.0 Classificadores

O aprendizado de máquina supervisionado possui como característica os rótulos da variável resposta (“target”) previamente conhecidos. Assim é possível aplicar um classificador/regressor que se ajusta aos dados e posteriormente realiza predições com uma base de interesse. De acordo com Domingos (2012), existem vários princípios fundamentais que orientam a prática de aprendizado de dados supervisionado. Entre eles, destacam-se a importância da seleção das variáveis explicativas (“features”) e a necessidade de evitar o “overfitting”. O autor ressalta que a eficácia de um modelo depende da qualidade dos dados e “features” selecionadas, e não apenas do tipo de algoritmo selecionado. Este trabalho foi feito com classificadores, uma vez que a variável resposta do modelo é uma variável categórica binária que indica a presença ou não da diabetes em determinado indivíduo. Foi utilizado cinco classificadores diferentes: “Gradiente Boosting”, “AdaBoost”, “XGBoost”, “Random Forest” e “Regressão Logística”.

Como demonstrado por Friedman (2001) e Chen & Guestrin (2016), no método “Boosting” há a construção sequencial de modelos em que cada modelo subsequente corrige os erros dos modelos anteriores, com o objetivo de reduzir o viés e melhorar a precisão do modelo. O “Adaboost” consiste em ajustar iterativamente os pesos das instâncias de treinamento,

focando nos mais difíceis de classificar (Freund & Schapire, 1997) e o “XGBoost” é uma otimização do “Gradient Boosting” que inclui técnicas avançadas de regularização para evitar “overfitting”.

Já no método “Bagging” há o treinamento de múltiplos modelos independentes em paralelo usando diferentes subconjuntos dos dados, através de amostragem aleatória com reposição. Esse método reduz a variância do modelo e melhora a estabilidade das previsões combinando múltiplos modelos independentes. O “RandomForest” é um modelo clássico do tipo “Bagging” e utiliza agregação de árvores de decisões para melhorar as predições.

A regressão logística que também foi utilizada neste trabalho consiste em um modelo clássico estatístico que utiliza a função logística para modelar a probabilidade de um evento binário ocorrer ou não com base nas variáveis explicativas, sendo bastante utilizada devido a sua eficácia e fácil interpretação. Vale ressaltar que neste método não há a criação de múltiplos modelos como nos anteriores, e sim a criação de um único modelo (Hosmer, 2013).

## 4.0 Métricas de Desempenho

Para avaliar a qualidade dos modelos e poder compará-los, foram calculadas algumas métricas de desempenho. Cada uma delas fornece uma perspectiva diferente em relação à performance dos modelos, sendo necessário analisá-las em conjunto para chegar a uma conclusão completa. As métricas utilizadas foram: Acurácia, Sensibilidade, Especificidade, “Precision”, “F1-Score” e a Área Sob a Curva ROC (AUC).

De início foi calculado a matriz de confusão, ferramenta essencial para avaliação de desempenho do modelo. Ela é uma tabela que resume as previsões feitas por um modelo em relação aos valores reais conhecidos. Seus elementos são:

- Verdadeiros Positivos (TP): Número de casos em que o modelo previu corretamente a classe positiva.
- Verdadeiros Negativos (TN): Número de casos em que o modelo previu corretamente a classe negativa.
- Falsos Positivos (FP): Número de casos em que o modelo previu erroneamente a classe positiva.
- Falsos Negativos (FN): Número de casos em que o modelo previu erroneamente a classe negativa.

Com a matriz de confusão calculada, é possível obter as métricas de desempenho. A acurácia representa a proporção de previsões certas e é importante para ter uma visão geral do desempenho do modelo. Sua fórmula é:

.

$$Acurácia = \frac{TN+TP}{TN+TP+FN+FP} \quad (2)$$

É importante avaliar a acurácia em conjunto com outras duas métricas: sensibilidade (“recall”) e especificidade. A primeira mede a capacidade do modelo em identificar corretamente os casos positivos, e a segunda de identificar corretamente os casos negativos (Powers, 2011). A sensibilidade acaba sendo bastante utilizada em contexto de diagnóstico de doenças, já que o objetivo é identificar todos os indivíduos com a condição clínica e realizar o tratamento imediato. As fórmulas das métricas são:

$$Sensibilidade = \frac{TP}{TP+FN} \quad (3)$$

$$Especificidade = \frac{TN}{TN+FP} \quad (4)$$

A precisão (“precision”) é a proporção de verdadeiros positivos sobre o total de positivos indicados pelo modelo, sendo crucial em aplicações onde falsos positivos são caros/perigosos (Saito & Rehmsmeier, 2015). O “F1-Score” é a média harmônica da sensibilidade e precisão, balanceando essas duas métricas. As fórmulas dessas métricas são:

$$Precisão = \frac{TP}{TP+FP} \quad (5)$$

$$F1 = 2 . \frac{Precisão\ .\ Sensibilidade}{Precisão+Sensibilidade} \quad (6)$$

A Curva ROC é uma ferramenta visual que mostra a relação entre a taxa de verdadeiros positivos, ajudando a balancear entre sensibilidade e especificidade (Fawcett, 2006). A Área Sob a Curva (AUC) é uma medida agregada de desempenho do modelo em todos os limiares de classificação.

## 5.0 Tunning

Após a escolha do melhor classificador, é possível extrair um pouco mais de desempenho do modelo através do “tunning”. Essa etapa consiste em otimizar os hiperparâmetros, como taxa de aprendizado, profundidade máxima da árvore e número de estimadores, com o objetivo de encontrar combinações que obtêm o melhor desempenho em termos de métricas como acurácia, sensibilidade e as outras citadas na seção anterior.

Foi realizada a técnica de Grid Search utilizando a biblioteca do scikit-learn em python, focando em maximizar a sensibilidade. Os classificadores que passaram pela tunning possuem diferentes parâmetros testados. Para o Gradiente Boosting e o XGBoost foi testado

número de estimadores (100, 200 e 300), taxa de aprendizado (0,01, 0,1 e 0,2) e profundidade máxima das árvores (3, 5, 7). Para o Adaboost foi testado o número de estimadores (50, 100, 200) e taxa de aprendizado (0,01, 0,1, 0,5 e 1). Por fim, para o Random Forest foi testado número de estimadores (100, 200 e 300) e profundidade máxima da árvore (10, 20 e 30).

### 6.0 Impacto das Variáveis

Foi realizada uma análise de impacto das features na previsão do modelo para entender o comportamento do modelo através do método “SHapley Additive exPlanations” (SHAP), que oferece uma interpretação consistente na contribuição de cada variável explicativa na predição. A técnica é baseada na teoria dos jogos e tem como objetivo distribuir de forma justa a recompensa (importância) entre um grupo de jogadores (“features”). O método consiste em calcular a contribuição marginal de cada variável explicativa para a previsão considerando todas as combinações possíveis, o que caracteriza ser uma abordagem consistente (Lundberg & Lee, 2017). Com essa técnica é possível ter uma análise mais robusta e detalhada à nível global (quais features são mais importantes) e local (impacto das features em uma previsão específica).

## Resultados e Discussão

Inicialmente foi realizada uma breve análise exploratória na base de dados com o objetivo de ter um melhor entendimento sobre as variáveis trabalhadas. O dataset possui 70.692 observações e 22 variáveis incluindo a variável “target”. Esse tamanho de amostra é representativo para a população alvo que são os habitantes dos Estados Unidos.

Foi verificado que o dataset está balanceado, ou seja, 50% das observações da variável respostas são um (contém diabetes) e os outros 50% são zero (não contém diabetes). É uma prática comum treinar modelos de machine learning com datasets balanceados para evitar enviesamento do modelo. Quando o treinamento é realizado com datasets desbalanceados, o modelo pode tender a classificar majoritariamente na classe de maior frequência, buscando maximizar a acurácia. No entanto, isso pode mascarar o verdadeiro desempenho do modelo em relação às classes menos representadas. Esse viés geralmente pode ser detectado através da análise de métricas como sensibilidade e especificidade, onde uma métrica pode ser muito alta e a outra muito baixa, indicando desempenho desigual entre as classes.

A Figura 1 abaixo apresenta os histogramas das variáveis consideradas. Por mais que o histograma seja indicado para variáveis numéricas contínuas, foi utilizado o método apenas para ter um breve resumo das distribuições de probabilidade das variáveis explicativas utilizadas no treinamento.

.

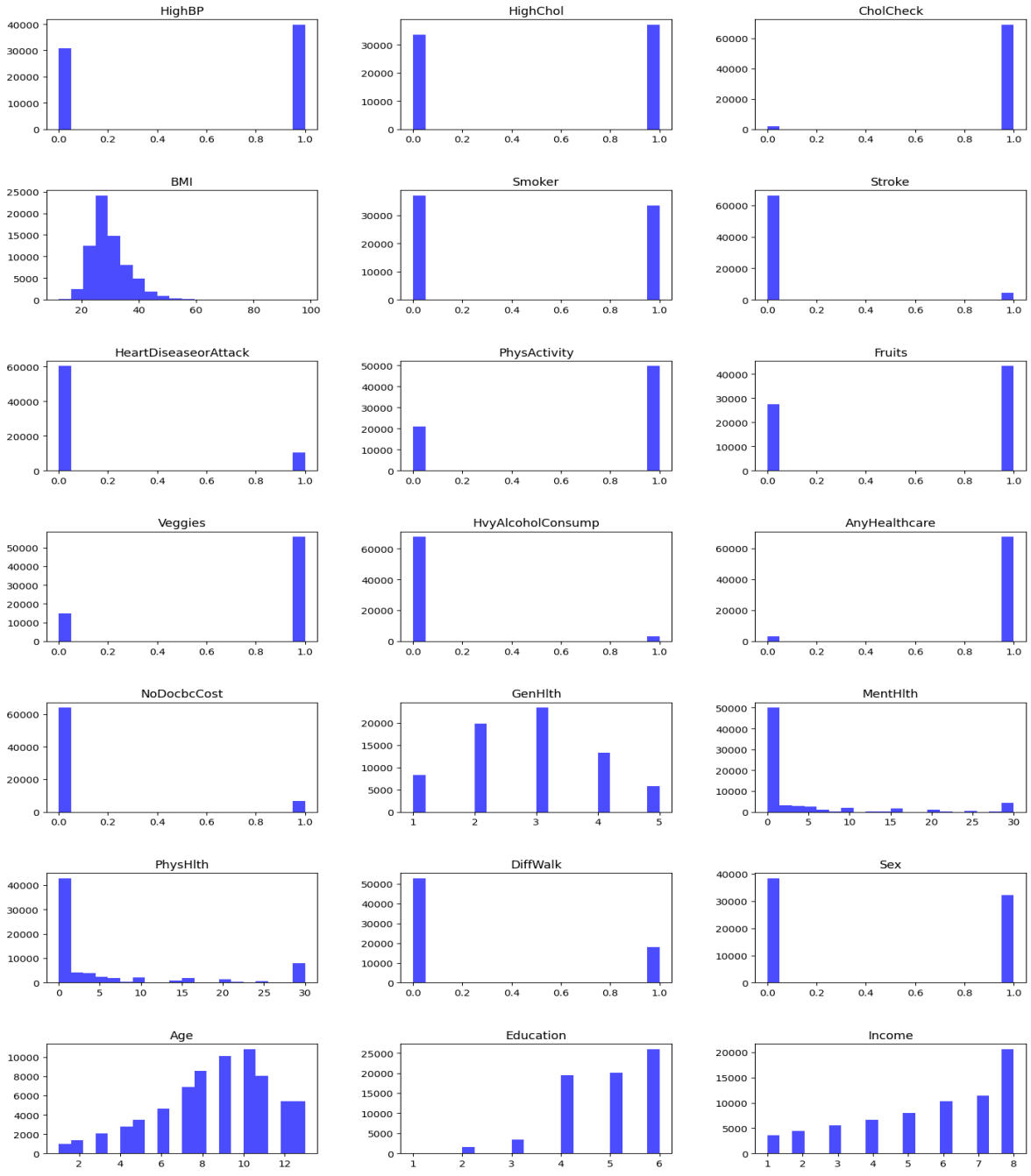


Figura 1 Histograma das covariáveis do modelo

Percebe-se que a maioria das variáveis são binárias (contendo o valor um ou zero), apenas três variáveis são numéricas (MentHlth, PshysHlth e BMI) e quatro variáveis categóricas possuem mais de duas categorias (Age, Income, Education e GenHlth). Essas quatro passaram pelo processo de dummyzação (transformar cada categoria em uma variável binária), onde foram excluídas as variáveis GenHlth_1, Education_1, Income_1, Age_1 para evitar o problema de multicolinearidade.

.

Foi analisada a correlação entre as variáveis numéricas, apresentadas na Figura 2. A análise mostra que as variáveis possuem correlação baixa entre si, indicando que não há a necessidade de exclusão de mais variáveis:

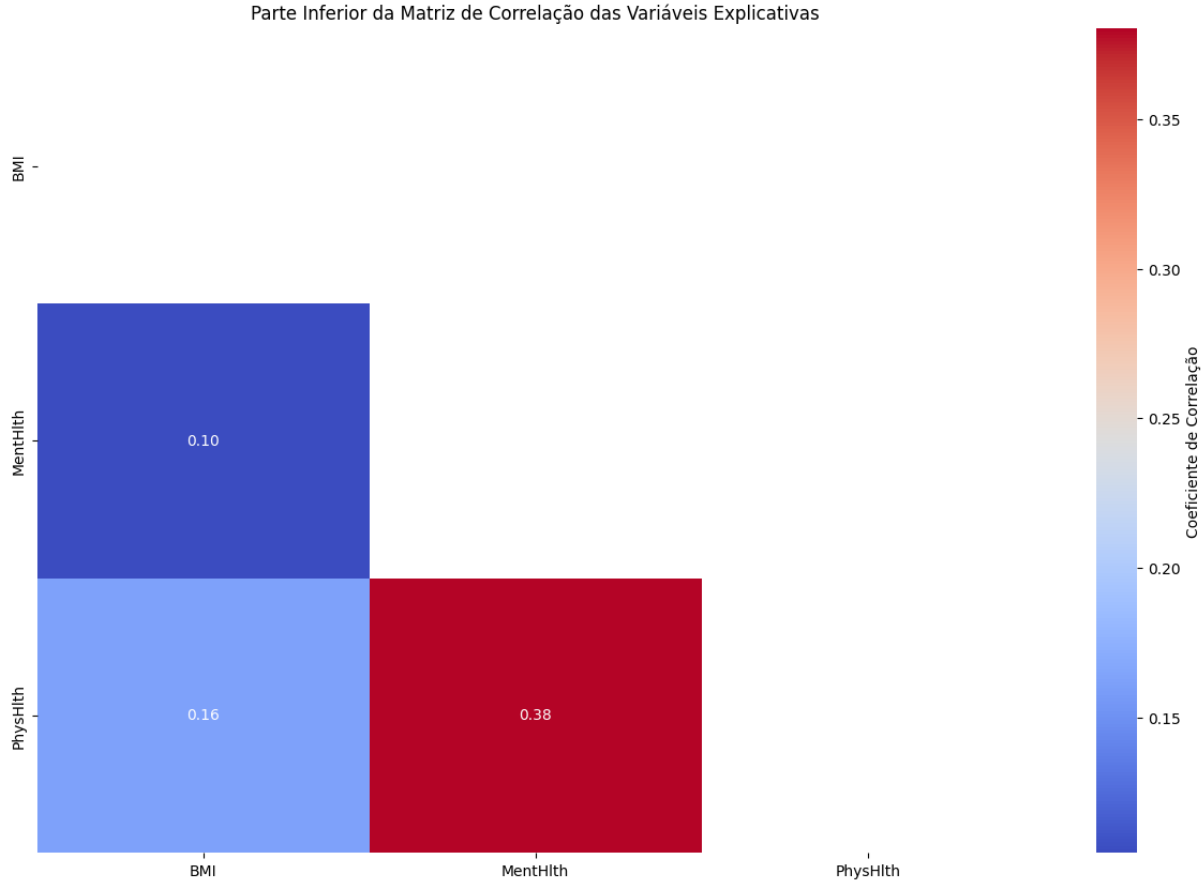


Figura 2 Parte inferior da matriz de correlação das variáveis numéricas

Como última parte do pré-processamento, as variáveis numéricas foram padronizadas através o Z-Score especificado na fórmula (1), e assim separadas em treino e teste, com 20% da base para teste utilizando o train test split da biblioteca “scikit-learn”.

O treinamento dos diferentes classificadores foi realizado e as respectivas métricas de desempenho se encontram na Tabela 2 abaixo. Vale ressaltar que foi considerado o cutoff de 0,5 na matriz de confusão para realizar os cálculos das métricas.

Tabela 2: Métricas de desempenho dos classificadores

| Classificador | Acurácia | Sensibilidade | Especificidade | Precisão | F-Score | AUC |
|---|---|---|---|---|---|---|
| GBM | 74,95% | 77,92% | 72,00% | 73,00% | 75,62% | 82,65% |
| XGBoost | 74,84% | 78,89% | 70,81% | 73,00% | 75,76% | 82,23% |
| Adaboost | 74,70% | 75,96% | 73,44% | 74,00% | 74,96% | 82,54% |

.

| | | | | | | |
|---|---|---|---|---|---|---|
| RF | 73,44% | 77,67% | 69,23% | 72,00% | 74,46% | 80,60% |
| Logística | 75,03% | 77,75% | 72,32% | 74,00% | 75,64% | 82,68% |

Também é possível avaliar os modelos através do gráfico com as respectivas Curvas ROC e AUC dos classificadores. Percebe-se que a performance foi bem parecida, com a Regressão Logística obtendo os melhores números.

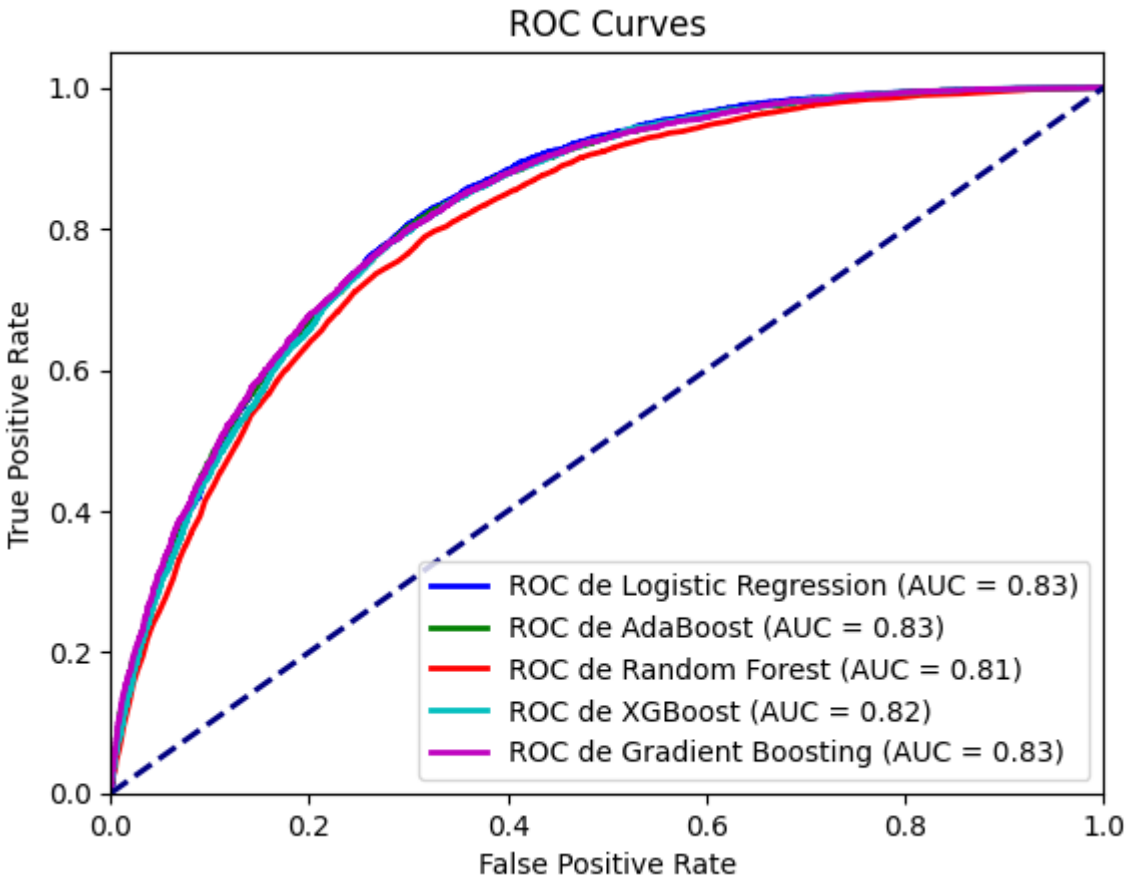


Figura 3: Curva ROC dos classificadores

Foi realizada a etapa de “tunning” do modelo para extrair um pouco mais de performance dos modelos. Na parte de Materiais e Métodos foi especificado os hiperparâmetros testados, e nesta parte há a melhor combinação de hiperparâmetros com as respectivas métricas de desempenho. O foco foi na maximização da Sensibilidade, métrica bastante acompanhada em análises de diagnóstico, pois mede a capacidade do modelo em classificar corretamente os indivíduos que possuem diabetes. Assim, após a etapa de tunning os modelos tunados tiveram os seguintes desempenho:

.

Tabela 3: Métricas de desempenho após etapa de tunning

| Classificador | Acurácia | Sensibilidade | Especificidade | Precisão | F-Score | AUC |
|---|---|---|---|---|---|---|
| GBM | 75,52% | 79,30% | 71,76% | 74,00% | 76,36% | 83,02% |
| XGBoost | 75,35% | 79,30% | 71,43% | 73,00% | 76,24% | 82,95% |
| Adaboost | 75,13% | 77,25% | 73,03% | 74,00% | 75,60% | 82,83% |
| RF | 74,46% | 79,04% | 69,90% | 72,00% | 75,52% | 82,12% |

Considerando a Tabela 3 acima, percebe-se que o Gradiente Boosting Machine (GBM) obteve melhor desempenho final, inclusive na métrica alvo que é a sensibilidade. Vale ressaltar que esse modelo contém os hiperparâmetros: taxa de aprendizado igual a 0,2, profundidade máxima da árvore igual a 3 e número de estimadores igual a 300.

Para a última etapa do projeto, foi realizado o impacto das variáveis explicativas na predição do modelo através do método SHAP. A Figura 4 abaixo mostra a importância de cada feature no modelo, calculada pelos valores SHAP e estão ordenados do mais ao menos relevante. Percebe-se que o HighBP é a “feature” mais importante, com maior impacto no modelo, indicando que indivíduos com pressão alta têm maior risco de ter diabetes. A segunda “feature” mais importante foi o BMI, indicando que variações no BMI têm grande influência nas previsões de diabetes. Outras variáveis também se mostraram importantes, como quem declara que sua saúde no geral está boa, razoável ou ruim (GenHlth_3.0, GenHlth_4.0 e GenHlth_5.0) e a variável indicadora de colesterol alto (HighCol). Variáveis como HvyAlcoholConsump, Age_3.0 e MentHlth têm um impacto médio menor nas predições, indicando que contribuem menos para a predição do modelo.

.

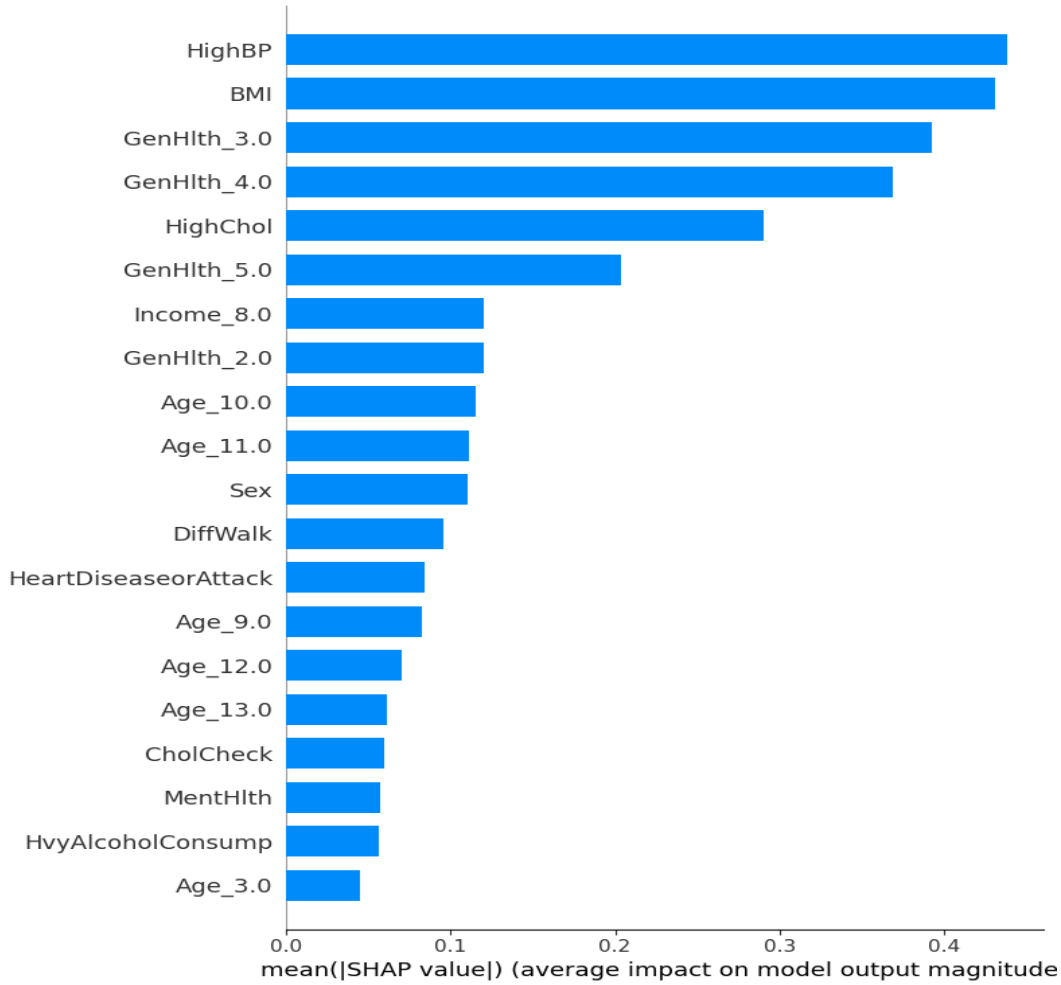


Figura 4: Importância das features através do SHAP values

A Figura 5 abaixo mostra a distribuição de valores SHAP para cada “feature”. As cores indicam o valor de cada variável, onde vermelho representa valores altos e azul valores baixos. Com a coloração e dispersão dos pontos é possível entender a relação entre os valores das features e o impacto nas previsões do modelo. Como a base de dados possui muitas variáveis binárias (0 ou 1) o método sintetiza 1 como vermelho e 0 como azul.

Percebe-se que para indivíduos com pressão alta (altos valores de HighBP) tem impacto positivo na predição, aumentando a chance do indivíduo ser classificado como diabético, enquanto não ter pressão alta diminui essa chance. Para o BMI há um comportamento semelhante, valores altos tendem a aumentar as probabilidades de diabetes (valores positivos de SHAP), enquanto valores mais baixos reduzem a probabilidade. Além da pressão alta, indivíduos com colesterol alto também tem chance aumentada de ter diabetes. Uma variável de renda Income_8.0 indica que quem tem uma menor probabilidade de serem preditos como diabéticos.

Os indicadores de percepção pessoal de saúde (GenHlth_3.0, GenHlth_4.0 e GenHlth_5.0) também contribuem para a diabetes, onde quem tem a percepção de que a saúde está boa, razoável ou ruim aumentam as chances de ter diabetes. A variável sexo indica que ser do sexo masculino aumenta a chance de ser caracterizado como diabético, enquanto ser do sexo feminino diminui essa chance.

.

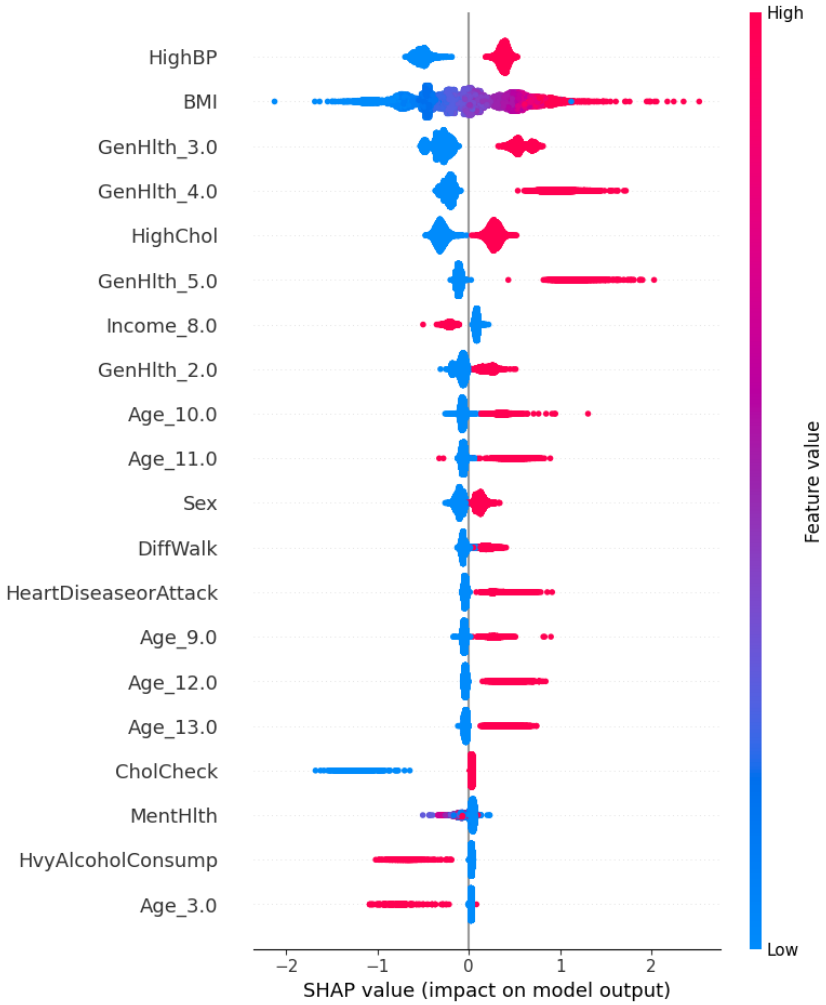


Figura 5: Impacto das covariáveis nas previsões do modelo

**Conclusão(ões) ou Considerações Finais**

Este trabalho demonstrou a eficácia de técnicas de aprendizado de máquina na classificação de indivíduos com diabetes, utilizando uma base de dados pública do CDC. O objetivo foi selecionar o melhor classificador com base em métricas de desempenho, como acurácia e sensibilidade, e avaliar o impacto de indicadores de saúde. Os modelos aplicados, como Gradiente Boosting, XGBoost, Adaboost, Random Forest e Regressão Logística, apresentaram resultados semelhantes, com o Gradiente Boosting obtendo uma sensibilidade de 79,3% após o ajuste de hiperparâmetros. A análise dos valores SHAP revelou que variáveis como pressão alta, índice de massa corporal (IMC) e colesterol alto foram os principais fatores de risco associados à presença de diabetes.

.

Entretanto, algumas limitações deste estudo devem ser destacadas. A base de dados utilizada foi balanceada, o que pode influenciar o desempenho do modelo em ambientes reais, onde a proporção de indivíduos com e sem diabetes é geralmente desbalanceada. Essa condição pode levar a uma superestimação da capacidade preditiva, especialmente em métricas como acurácia. Além disso, o conjunto de variáveis disponível pode não abranger outros fatores importantes que afetam o diagnóstico, como aspectos genéticos e socioeconômicos. Recomenda-se que futuras pesquisas considerem essas limitações e testem os modelos em amostras mais representativas.

Para trabalhos futuros, sugere-se explorar diferentes abordagens metodológicas. Testes com outras técnicas de aprendizado de máquina, como redes neurais profundas e algoritmos de ensemble avançados, podem contribuir para aumentar a precisão dos diagnósticos. Além disso, ajustes em parametrizações, como diferentes estratégias de balanceamento de classes e variações nos hiperparâmetros, podem ser investigados. A integração de dados adicionais, como exames laboratoriais e histórico médico detalhado, também pode fornecer uma análise mais robusta, auxiliando na identificação de novos fatores de risco e no desenvolvimento de estratégias mais eficazes de prevenção e tratamento da diabetes.

## Agradecimento

Primeiramente gostaria de agradecer aos meus pais, Vera Lúcia e Fernando Alves, por terem dedicado suas vidas à minha, e por me ensinar a valorizar a educação. Agradeço especialmente também a minha namorada, Carolina Muylaert, que me acompanhou e incentivou nessa jornada, sendo sempre uma influência positiva em minha vida. Agradeço também a todos os funcionários da USP-ESALQ por se dedicarem e conseguir entregar uma ensino de extrema qualidade. Por fim, sou extremamente grato à Universidade Federal Fluminense, lugar onde me graduei em estatística que me proporcionou uma base enorme.

## Referências

Gross, J.L.; Silveiro, S.P.; Camargo, J.L.; Reichelt, A.J.; Azevedo, M.J. de. (2002). Diabetes melito: diagnóstico, classificação e avaliação do controle glicêmico. *Arquivos Brasileiros de Endocrinologia & Metabologia, 46*(1), 16–26. https://doi.org/10.1590/S0004-27302002000100004

.

Zimmet, P.Z. Diabetes and its drivers: the largest epidemic in human history?. *Clin Diabetes Endocrinol* 3, 1 (2017). https://doi.org/10.1186/s40842-016-0039-3

Abbott, P. A., & Barbosa, S. F. F. (2015). Usando tecnologia da informação e mobilização social para combater doenças. *Acta Paulista de Enfermagem, 28*(1), 1. https://doi.org/10.1590/1982-0194201500001

Paixão, G. M. de M., Santos, B. C., Araujo, R. M. de, Ribeiro, M. H., Moraes, J. L. de, & Ribeiro, A. L. (2022). Machine learning na medicina: Revisão e aplicabilidade. *Arquivos Brasileiros de Cardiologia, 118*(1), 95–102. https://doi.org/10.36660/abc.20200596

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87. https://doi.org/10.1145/2347736.2347755

Friedman, Jerome. (2000). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 29.10.1214/aos/1013203451.

Chen, T.; Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Freund, Y.; Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

Hosmer Jr., D.W.; Lemeshow, S.; Sturdivant, R.X. (2013). *Applied logistic regression*. John Wiley & Sons, Inc. ISBN: 9780470582473. DOI: 10.1002/9781118548387.

Powers, David. (2008). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Mach. Learn. Technol.. 2.

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015 Mar 4;10(3):e0118432. doi: 10.1371/journal.pone.0118432. PMID: 25738806; PMCID: PMC4349800.

.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Lundberg, S.M.; Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

**Apêndice ou Anexo**