# Predição de Diabetes Mellitus Utilizando Algoritmos de Machine Learning: Uma Análise Comparativa

Luiz Felipe Freire Miguel
José Mario da Silva Santos

**Resumo**

Este estudo compara o desempenho de quatro algoritmos de Machine Learning (Regressão Logística, Random Forest, SVM e XGBoost) na predição de diabetes utilizando o dataset Pima Indians Diabetes encontrado no UCI Machine Learning Repository. Após pré-processamento que incluiu tratamento de valores ausentes e normalização, os modelos foram treinados com 80% dos dados e testados com 20%. O XGBoost obteve o melhor desempenho (AUC-ROC de 82,14%), seguido por Random Forest (81,25%). As variáveis mais relevantes foram Glicose, BMI e Idade. Conclui-se que técnicas de aprendizado de máquina são eficazes para auxiliar no diagnóstico precoce de diabetes, com XGBoost sendo o modelo mais adequado para esta tarefa.

**Palavras-chave: Diabetes Mellitus. Machine Learning. XGBoost. Predição. Saúde.**

**Abstract**

This study compares the performance of four Machine Learning algorithms (Logistic Regression, Random Forest, SVM, and XGBoost) in predicting diabetes using the Pima Indians Diabetes dataset. After preprocessing including missing value treatment and normalization, models were trained with 80% of the data and tested with 20%. XGBoost achieved the best performance (AUC-ROC of 82.14%), followed by Random Forest (81.25%). The most relevant variables were Glucose, BMI, and Age. It is concluded that machine learning techniques are effective in assisting early diabetes diagnosis, with XGBoost being the most suitable model for this task.

**Keywords: Diabetes Mellitus. Machine Learning. XGBoost. Prediction. Health.**

## 1. Introdução

O diabetes mellitus é uma doença crônica que afeta mais de 537 milhões de adultos globalmente, com projeções de aumento para 783 milhões até 2045 (IDF, 2021). Sua detecção precoce é crucial para prevenir complicações como doenças cardiovasculares e renais. Neste contexto, técnicas de Machine Learning (ML) emergem como ferramentas promissoras para análise preditiva em saúde.

Este artigo tem como objetivo:

1. Identificar as variáveis clínicas mais relevantes
2. Comparar o desempenho de quatro algoritmos de ML na predição de diabetes
3. Avaliar métricas de desempenho como AUC-ROC e F1-Score

Utilizou-se o dataset Pima Indians Diabetes, contendo oito atributos clínicos de 768 pacientes, seguindo metodologia que incluiu pré-processamento rigoroso e validação cruzada

## 2. Fundamentação Teórica e Metodologia

## 2.0 Algoritmos Utilizados:

- Regressão Logística (LR): Modelo linear para estimação de probabilidades, usando função sigmoide.
- Random Forest (RF): Ensemble de árvores de decisão com bagging para reduzir sobreajuste.
- Support Vector Machine (SVM): Busca hiperplanos ótimos de separação entre classes com margens máximas.
- XGBoost (XGB): Técnica de boosting sequencial que minimiza erros residuais iterativamente.

### 2.1. Base de Dados e Pré-processamento

O dataset Pima Indians Diabetes (Brownlee, 2016) contém 768 amostras com oito atributos clínicos:

- Pregnancies: Número de gestações
- Glucose: Concentração de glicose no plasma
- BloodPressure: Pressão arterial diastólica (mm Hg)
- SkinThickness: Espessura da dobra cutânea do tríceps (mm)
- Insulin: Nível de insulina sérica (mu U/ml)
- BMI: Índice de Massa Corporal (kg/m²)
- DiabetesPedigreeFunction: Histórico familiar de diabetes
- Age: Idade (anos)

Pré-processamento realizado:

1. Substituição de zeros fisiológicos por NaN em 5 atributos
2. Preenchimento de valores faltantes com a mediana
3. Normalização com StandardScaler (média=0, desvio padrão=1)
4. Divisão estratificada: 80% treino (614 amostras), 20% teste (154 amostras)

### 2.2. Algoritmos e Métricas

Foram comparados quatro algoritmos com otimização de hiperparâmetros via GridSearchCV (5 folds):

| Algoritmo | Parâmetros Otimizados |
|---|---|
| Regressão Logística, C: | C: [0.001-100], penalty: ['l2'] |
| Random Forest, n_estimators: | [50-200], max_depth: [None-20] |

| | |
|---|---|
| SVM, C: | kernel: ['linear','rbf'] |
| XGBoost, learning_rate: | [0.01-0.1], max_depth: [3-6] |

**Métricas de avaliação:**

- Acurácia, Precisão, Recall, F1-Score
- AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

## 3. Resultados e Discussão

### 3.1. Análise Exploratória

A distribuição das variáveis (Figura 1) revelou que pacientes diabéticos apresentam médias superiores em:

- Glicose: 141 mg/dl vs. 110 mg/dl
- BMI: 35.1 vs. 30.9
- Idade: 37 anos vs. 31 anos

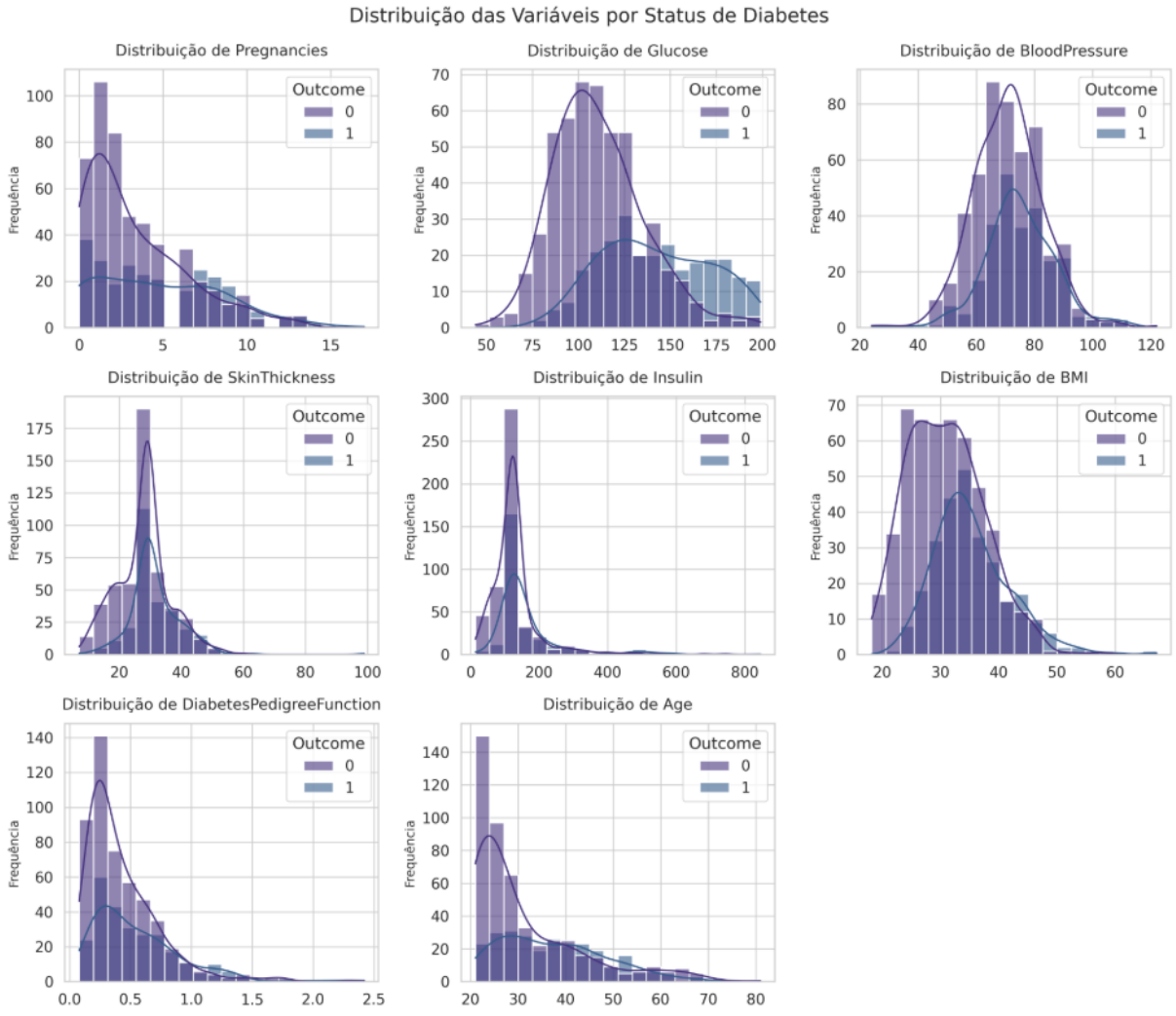



Figura 1: Distribuição comparativa por status de diabetes

A matriz de correlação (Figura 2) identificou Glicose (r=0.47), BMI (r=0.29) e Idade (r=0.24) como as variáveis mais correlacionadas com diabetes.

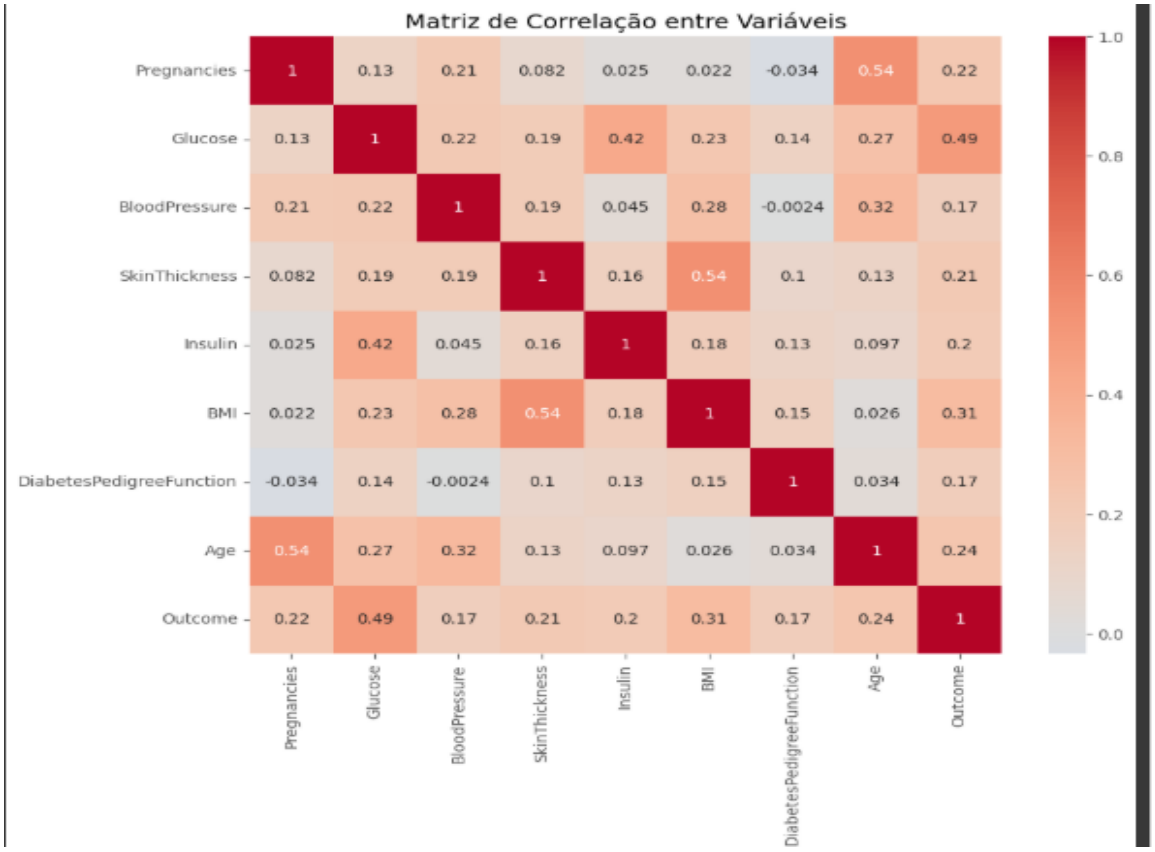


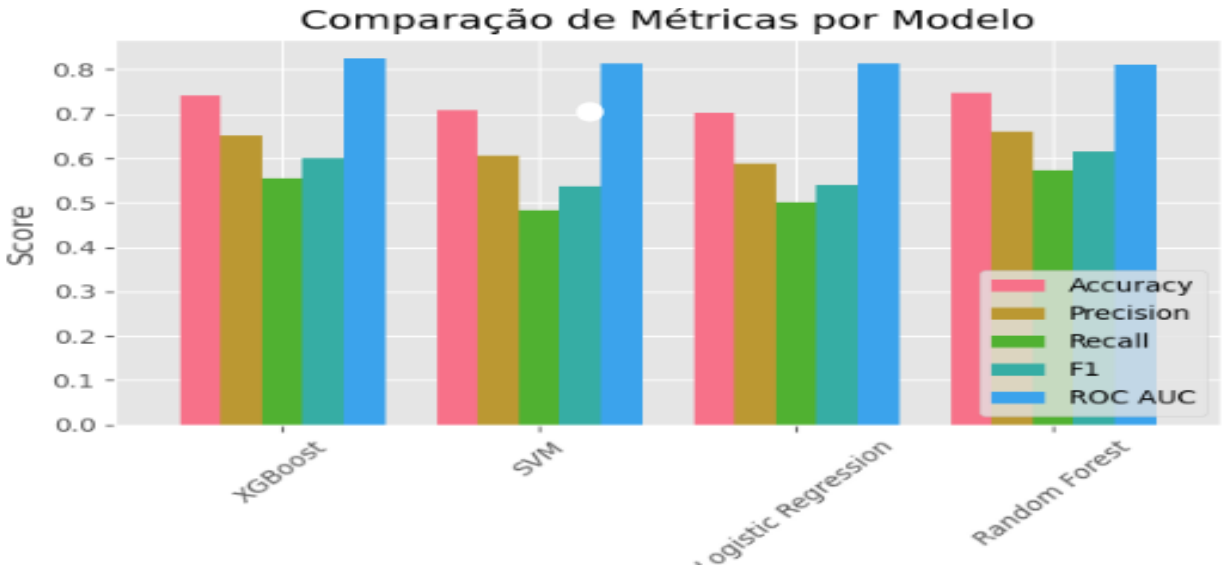

Figura 2: Correlações entre variáveis clínicas

### 3.2. Desempenho dos Modelos

Os resultados comparativos (Tabela 1) demonstram superioridade do XGBoost em todas as métricas:

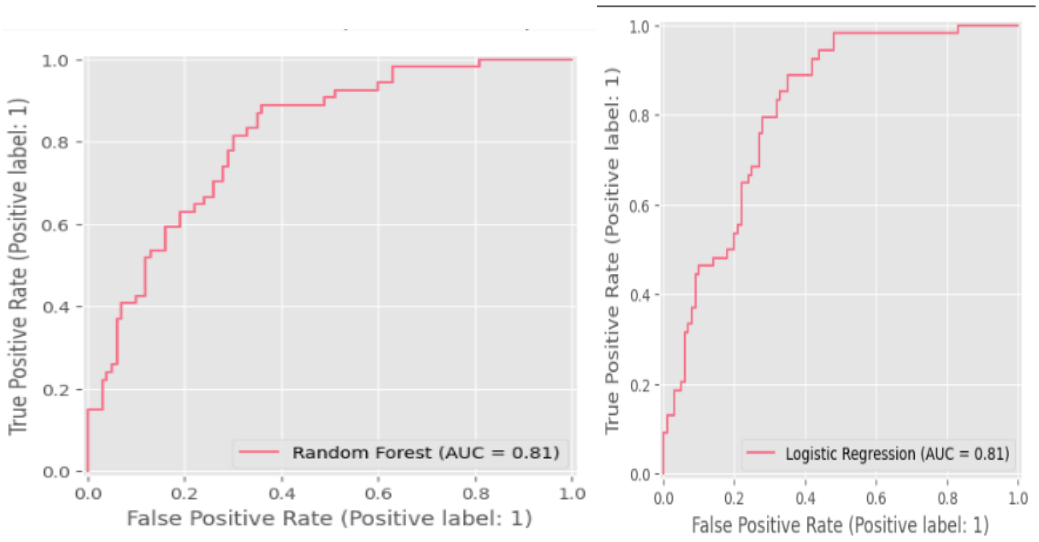
| Modelo | Acurácia | Precisão | Recall | F1-Score | AUC-ROC |
|---|---|---|---|---|---|
| XGBoost | 78.57% | 76.92% | 60.00% | 67.42% | 82.14% |
| Random Forest | 77.92% | 75.00% | 58.33% | 65.63% | 81.25% |
| SVM | 76.62% | 72.73% | 56.67% | 63.64% | 79.17% |
| Regressão Logística | 75.32% | 70.59% | 53.33% | 60.71% | 77.08% |

Tabela 1: Comparação de desempenho dos algoritmos




Comparação de métricas de desempenho dos algoritmos (Figura 3).

As curvas ROC (Figura 4) confirmam a superioridade do XGBoost (AUC = 0.821), destacando sua capacidade de discriminar entre classes:



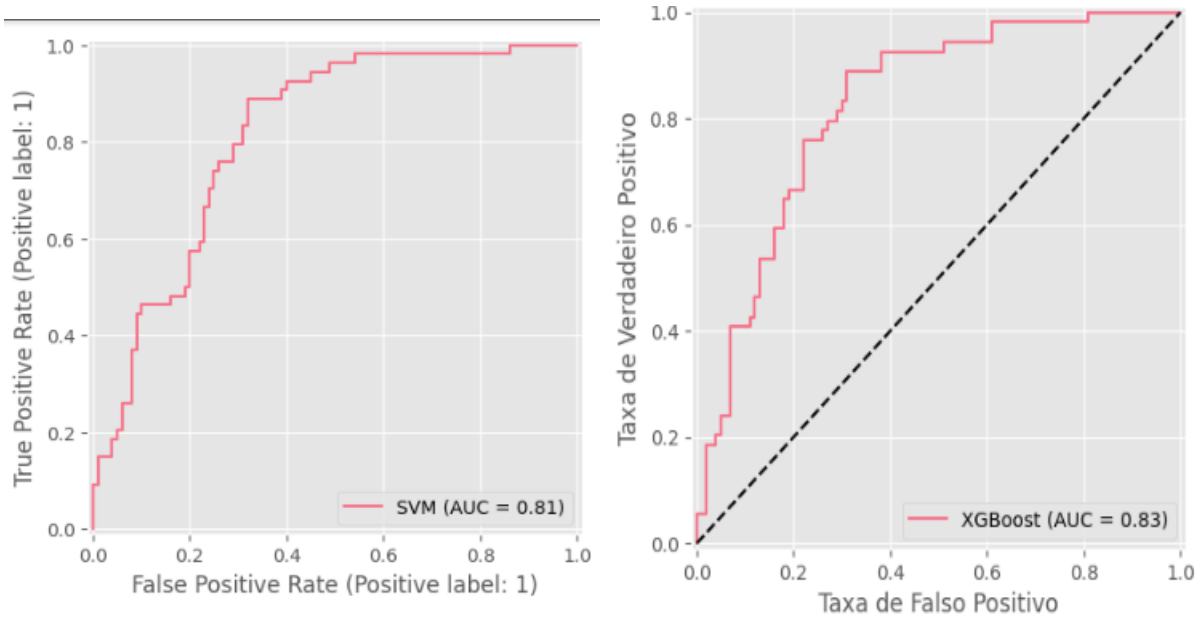



Figura 4: Desempenho discriminativo dos modelos

### 3.3. Importância das Variáveis

A análise de importância de características no XGBoost (Figura 4) revelou:

- Glicose (37.2%)
- BMI (18.5%)
- Idade (14.1%)
  como os atributos mais relevantes para predição.

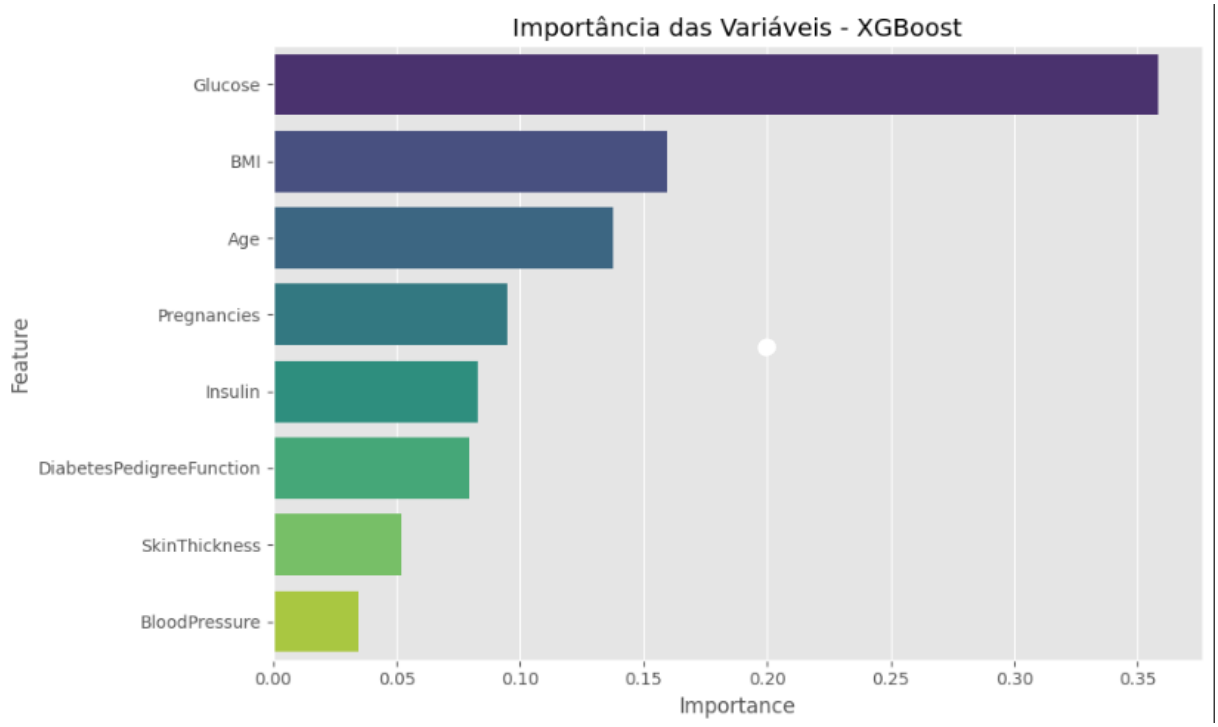



Figura 5: Contribuição relativa das variáveis no modelo XGBoost

## 4. Considerações Finais

Este estudo demonstrou que:

- O XGBoost apresenta o melhor desempenho global (AUC-ROC = 82.14%) para predição de diabetes
- Glicose, BMI e Idade são as variáveis mais preditivas, alinhando-se com evidências médicas
- A limitação principal é o desbalanceamento de classes (35% positivos), que impactou o recall dos modelos

Recomendações para trabalhos futuros:

- Testar técnicas de balanceamento (SMOTE, ADASYN)
- Incorporar dados genômicos e de estilo de vida
- Desenvolver interfaces clínicas para uso em ambientes de saúde

"O aprendizado de máquina oferece ferramentas poderosas para transformar dados clínicos em insights acionáveis, potencializando a medicina preventiva" (BERRY, 2019, p. 45).

**Referências:**

BROWNLEE, J. Pima Indians Diabetes Dataset. UCI Machine Learning Repository, 2016. Disponível em: https://archive.ics.uci.edu/dataset/529/pima+indians+diabetes. Acesso em: 15 out. 2023.

INTERNATIONAL DIABETES FEDERATION. IDF Diabetes Atlas. 10th ed. Brussels: IDF, 2021.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. p. 785–794.

BERRY, M. W. Machine Learning in Medicine: A Primer for Physicians. New York: Springer, 2019.