



Modelo de predição de Diabetes

Matheus Afonso Leite da Silva, Jonas Tavares Lewis, Renan Nunes dos Santos,
Gabriel Narvaes de Macedo

Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

{10409312, 10403619, 10409432, 10409681}@mackenzie.br

Resumo. O projeto tem como objetivo desenvolver um modelo de Inteligência Artificial utilizando técnicas de Machine Learning para prever a ocorrência de diabetes em pacientes, utilizando a base de dados Pima Indians Diabetes do Kaggle. O estudo será conduzido por meio da aplicação de algoritmos de classificação supervisionada. Será realizada uma análise exploratória dos dados, seguida do pré-processamento adequado, a fim de desenvolver um modelo preditivo capaz de apoiar a tomada de decisão na saúde, auxiliando profissionais na detecção de pacientes com maior risco de desenvolver diabetes.

Palavras-chave: Inteligência Artificial; Machine Learning; Classificação; Previsão de Diabetes; Saúde

Abstract. The project aims to develop an Artificial Intelligence model using Machine Learning techniques to predict the occurrence of diabetes in patients, using Kaggle's Pima Indians Diabetes database. The study will be conducted through the application of supervised classification algorithms. An exploratory analysis of the data will be carried out, followed by appropriate pre-processing, in order to develop a predictive model capable of supporting decision-making in healthcare, assisting professionals in detecting patients at greater risk of developing diabetes.

Keywords: Artificial intelligence; Machine Learning; Classification; Diabetes Forecast; Health

1. Introdução

1.1 Contextualização

A diabetes é uma das doenças crônicas mais comuns no mundo e tem crescido bastante nos últimos anos. Essa condição pode trazer várias complicações para a saúde das pessoas, principalmente quando não é diagnosticada ou tratada de forma adequada. Por isso, a detecção precoce é muito importante para ajudar médicos e pacientes a tomarem decisões mais rápidas e eficientes.

Segundo a Organização Mundial da Saúde (OMS), a prevalência global de diabetes em adultos passou de 7% em 1990 para 14% em 2022. Estima-se que cerca de 589 milhões de adultos com idade entre 20 e 79 anos vivam com diabetes atualmente, número que deve subir para aproximadamente 853 milhões até 2050. (INTERNATIONAL DIABETES FEDERATION, 2023)

Além disso, os impactos vão além dos aspectos de saúde individuais, visto que há custos econômicos relevantes para sistemas de saúde e para a sociedade. Por exemplo, nos Estados Unidos, pessoas com diagnóstico de diabetes têm gastos médicos em média 2,6 vezes maiores do que os que não têm a doença (ZHU et al., 2023).

1.2 Justificativa

O aumento dos casos de diabetes em todo o mundo reforça a necessidade de ferramentas que auxiliem na detecção precoce da doença. Quanto mais cedo um paciente é identificado com risco de desenvolver diabetes, maiores são as chances de receber acompanhamento adequado e não desenvolver a doença. Nesse sentido, métodos computacionais que utilizam dados podem apoiar profissionais de saúde na tomada de decisão.

A Inteligência Artificial tem se mostrado uma ferramenta importante na área da saúde, pois permite analisar grandes quantidades de dados e encontrar padrões que muitas vezes não são percebidos por humanos. Dessa forma, é possível criar modelos que auxiliem na previsão de doenças, como a diabetes, a partir das informações dos pacientes.

Dessa forma, o uso de Machine Learning para previsão de diabetes representa não apenas uma aplicação prática dos conceitos estudados na disciplina, mas também uma contribuição para a área da saúde, ao oferecer suporte adicional a médicos e instituições no processo de diagnóstico e prevenção.

1.3 Objetivo

O objetivo do projeto é desenvolver um modelo de Machine Learning capaz de prever a ocorrência de diabetes em pacientes a partir da base de dados Pima Indians Diabetes, disponível no Kaggle. Para isso, serão aplicados algoritmos de classificação supervisionada.

1.4 Opção do projeto

A opção escolhida para o desenvolvimento do projeto foi a Opção Framework, que consiste em empregar bibliotecas e ferramentas de Machine Learning em Python para a resolução de um problema de classificação. Para isso, será utilizado o framework scikit-learn e pandas, aplicando diferentes algoritmos supervisionados sobre o conjunto de dados selecionado.

2. Descrição do Problema

A diabetes é uma doença que se não diagnosticada a tempo, pode causar diversas complicações para a saúde de uma pessoa e nem sempre os sintomas aparecem de forma clara, o que dificulta a identificação dela. Isso gera a necessidade de métodos que analisam variáveis clínicas e indicam a probabilidade de um paciente desenvolver a doença.

O problema a ser resolvido é justamente a previsão da diabetes em pacientes, utilizando informações clínicas básicas, como níveis de glicose, índice de massa corporal, pressão arterial e idade. A proposta é usar algoritmos de classificação supervisionada para criar um modelo que auxilie no diagnóstico, apoiando os profissionais da saúde.

3. Aspectos Éticos do uso da IA

Decisões tomadas com base em modelos preditivos podem impactar diretamente a vida das pessoas, por isso é fundamental que a IA seja usada como ferramenta de apoio, e não como substituto da avaliação médica, pois erros de classificação podem levar a diagnósticos equivocados. Outro aspecto importante é a privacidade e proteção dos dados dos pacientes, já que as informações clínicas são sensíveis e devem ser tratadas de forma segura, garantindo anonimização quando necessário e evitando qualquer tipo de uso inadequado.

Além disso, se o conjunto de dados não representar bem diferentes perfis de pacientes, o modelo pode acabar dando resultados que não funcionam para todos. Por isso, é papel do desenvolvedor analisar se os dados possuem qualidade e tentar reduzir esses problemas para que a solução seja mais confiável.

4. Descrição detalhada do Dataset

O dataset utilizado neste projeto é o *Pima Indians Diabetes Database*, disponível na plataforma *Kaggle*. Ele foi coletado pelo *National Institute of Diabetes and Digestive and Kidney Diseases* e contém informações clínicas de mulheres de pelo menos 21 anos do grupo indígena Pima, da região do Arizona, nos Estados Unidos.

A base possui 768 registros, cada um representando um paciente, e conta com 8 variáveis independentes: número de gestações, nível de glicose, pressão arterial, espessura da pele, nível de insulina, índice de massa corporal (IMC), histórico familiar

de diabetes (*Diabetes Pedigree Function*) e idade. A variável dependente é binária, indicando se o paciente tem (1) ou não tem (0) diabetes.

Como os dados disponibilizados já são anônimos, não é necessário realizar nenhuma anonimização adicional.

A primeira etapa do trabalho foi carregar o dataset de diabetes utilizando a biblioteca Pandas. Em seguida, foi feita uma análise exploratória inicial para conhecer melhor a base. Para isso, foram consultadas as informações gerais do dataset (quantidade de linhas, colunas e tipos de dados) e também foram calculadas as estatísticas descritivas, como média, mediana, valores mínimos e máximos.

Outro ponto analisado foram os valores ausentes e inválidos. Algumas variáveis como glicose, pressão arterial, espessura da pele, insulina e IMC aparecem com valor igual a zero, o que não faz sentido. Esses valores foram tratados da seguinte forma: primeiro, os zeros foram substituídos por NaN (valores nulos), e depois os valores ausentes foram preenchidos usando a mediana de cada coluna. Esse método foi escolhido porque a mediana é menos afetada por valores extremos, deixando o dataset mais consistente.

6. Metodologia

A metodologia do projeto seguiu quatro etapas principais. Primeiro, foi feita uma análise exploratória para entender o dataset e identificar valores incorretos, como zeros em variáveis que não poderiam ser zero. Esses valores foram substituídos por NaN e preenchidos com a mediana de cada coluna.

Depois, os dados foram normalizados para que todas as variáveis tivessem a mesma escala, o que ajuda no desempenho dos modelos. Em seguida, o dataset foi dividido em grupos de treino com 80% da base e teste com 20% da base.

Os algoritmos aplicados foram KNN e Regressão Logística, ambos treinados com os dados tratados. Por fim, o desempenho foi avaliado por meio das métricas de acurácia e precisão.

7. Resultados obtidos

Os dois modelos apresentaram resultados dentro do esperado. A Regressão Logística teve acurácia de 70%, com bom equilíbrio entre as classes. O KNN também teve desempenho próximo, mas foi um pouco mais sensível ao valor de k e à normalização, apresentando acurácia de 0.7532 e precisão de 0.6600, sendo o modelo com melhor desempenho geral. Já a Regressão Logística apresentou acurácia de 0.7013 e precisão de 0.5870.

Em ambos os modelos, a classe negativa (sem diabetes) teve mais acertos, o que é comum no dataset por ele ser desbalanceado. Mesmo assim, os modelos conseguiram identificar uma quantidade relevante de casos positivos, mostrando utilidade prática para triagem de pacientes.

8. Conclusão

O projeto teve como objetivo desenvolver um modelo capaz de prever a ocorrência de diabetes utilizando dados clínicos básicos. Com as etapas realizadas e os testes aplicados, foi possível alcançar resultados satisfatórios. O modelo KNN obteve o melhor desempenho, enquanto a Regressão Logística também apresentou resultados consistentes, embora um pouco inferiores.

Com isso, é possível concluir que os resultados esperados foram alcançados. Os modelos construídos demonstraram capacidade de auxiliar na identificação de pacientes com risco de desenvolver diabetes, contribuindo como uma ferramenta de apoio na área da saúde.

9. Endereço GitHub e Youtube

###

10. Referências bibliográficas e Bibliografia

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). Urgent action needed as global diabetes cases increase four-fold over past decades. 2024. Disponível em: <https://www.who.int/news/item/13-11-2024-urgent-action-needed-as-global-diabetes-cases-increase-four-fold-over-past-decades>. Acesso em: 20 set. 2025.

INTERNATIONAL DIABETES FEDERATION (IDF). IDF Diabetes Atlas. 10. ed. 2023. Disponível em: <https://diabetesatlas.org>. Acesso em: 20 set. 2025.

ZHU, Y. et al. Economic costs of diabetes in the U.S. in 2022. Diabetes Care, v. 46, n. 1, p. 50-60, 2023. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/37909353/>. Acesso em: 20 set. 2025.

KAGGLE. Pima Indians Diabetes Database. Disponível em: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>. Acesso em: 20 set. 2025.