Predição de Diabetes Utilizando Modelos de Aprendizado de Máquina com o Dataset Pima Indians

Jonnas Pedro Beserra Gonçalves[1], João Vitor Farias de Souza[2], Cauã Rocha da Silva[1], Luciano de Souza Cabral[1]

[1]Curso Técnico de Desenvolvimento de Sistemas – Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE - Campus Jaboatão dos Guararapes)
Caixa Postal 54080-000 – Jaboatão dos Guararapes – PE – Brazil

jpbg@discente.ifpe.edu.br

jvfs16@discente.ifpe.edu.br

crs19@discente.ifpe.edu.br

luciano.cabral@jaboatao.ifpe.edu.br

Abstract: *This paper presents the development of the DiabPredict project, an application designed to predict the risk of diabetes using machine learning techniques. The project was developed as part of the Technical Course in Systems Development at IFPE – Jaboatão dos Guararapes Campus. The system aims to assist users and healthcare professionals in preventive monitoring of diabetes-related factors, promoting awareness and the use of accessible technologies for health support. The conception process, tools, methodology, and preliminary results are discussed.*

Keywords: *Diabetes prediction; Machine Learning; Digital Health; Web Applications.*

Resumo: *Este artigo apresenta o desenvolvimento do projeto DiabPredict, uma aplicação voltada à previsão de risco de diabetes utilizando técnicas de aprendizado de máquina. O trabalho foi realizado como parte das atividades do curso Técnico em Desenvolvimento de Sistemas do IFPE – Campus Jaboatão dos Guararapes. O sistema busca auxiliar usuários e profissionais de saúde no monitoramento preventivo de fatores associados à diabetes, promovendo a conscientização e o uso de tecnologias acessíveis para apoio à saúde. São descritos o processo de concepção, as ferramentas utilizadas, a metodologia de implementação e os resultados obtidos até o momento.*

Palavras-chave: *Previsão de diabetes; Aprendizado de Máquina; Saúde Digital; Aplicações Web.*

1. Introdução

O diabetes mellitus é uma das doenças crônicas mais incidentes no mundo, caracterizada por níveis elevados de glicose no sangue devido à deficiência na produção ou na utilização da insulina. A ausência de diagnóstico precoce agrava o quadro clínico, aumentando o risco de complicações cardiovasculares, renais e neurológicas. De acordo com dados da Organização Mundial da Saúde (OMS), o número de pessoas com diabetes tem crescido de forma significativa nas últimas décadas, configurando um problema global de saúde pública.

Com o avanço das tecnologias digitais e da Inteligência Artificial (IA), tornou-se possível aplicar métodos de aprendizado de máquina (Machine Learning) na área médica para identificar padrões complexos em dados clínicos. Essas técnicas podem auxiliar na previsão de doenças, permitindo diagnósticos mais rápidos e assertivos, além de favorecer a tomada de decisões por parte dos profissionais de saúde.

Neste contexto, o presente trabalho apresenta o DiabPredict, um sistema desenvolvido em linguagem Python que utiliza modelos supervisionados de aprendizado de máquina para estimar a probabilidade de um indivíduo desenvolver diabetes. A base de dados utilizada é o Pima Indians Diabetes Dataset, amplamente reconhecida na literatura científica por conter informações clínicas relevantes. O projeto foi estruturado em três etapas principais: análise e pré-processamento dos dados, treinamento e validação dos modelos, e implementação de uma interface interativa com o uso do Streamlit.

2. Metodologia

2.1 Base de Dados

A metodologia deste estudo foi estruturada em quatro etapas principais: coleta de dados, pré-processamento, modelagem e desenvolvimento da aplicação web. A base de dados utilizada foi a Pima Indians Diabetes, disponibilizada pelo UCI Machine Learning Repository, contendo informações de 768 pacientes do sexo feminino de origem indígena Pima. Entre as variáveis presentes, destacam-se glicose, pressão arterial, índice de massa corporal (IMC), idade e número de gestações. O objetivo principal do estudo foi prever a variável Outcome, que indica a presença (1) ou ausência (0) de diabetes.

2.2 Pré-processamento de Dados

Inicialmente, foi realizado o tratamento dos dados, com a verificação de valores ausentes e outliers. As variáveis foram normalizadas utilizando a técnica Standard Scaler para garantir que os modelos apresentaram maior estabilidade durante o treinamento.

2.3 Modelagem de Machine Learning

Em seguida, foram testados diferentes algoritmos de aprendizado de máquina, incluindo um classificador baseado em rede neural totalmente conectada, otimizado por algoritmo genético (Neuroevolution) e Random Forest. A divisão dos dados foi realizada em 80% para treino e 20% para teste, utilizando validação cruzada (cross-validation) para evitar sobreajuste (overfitting). As métricas utilizadas para avaliação dos modelos foram acurácia, precisão, revocação (recall) e F1-score.

2.4 Desenvolvimento da Aplicação Web

Após a etapa de modelagem, foi desenvolvido o protótipo da aplicação web com o framework Streamlit, permitindo a interação do usuário com o modelo treinado. O sistema recebe os dados clínicos como entrada e retorna a probabilidade estimada de o paciente apresentar diabetes.

3. Resultados e Discussão

3.1 Desempenho dos Modelos

Os experimentos realizados indicaram que os modelos de aprendizado de máquina apresentaram desempenhos consistentes. O modelo de Neuroevolution destacou-se com uma acurácia média de aproximadamente 82%, seguido pelo Random Forest, que apresentou 98% nos treinos mas apenas 73% nos testes.

Tabela 1. Comparação das métricas dos modelos de aprendizado de máquina.

Modelo	Acurácia treinos	Acurácia testes	Recall	F1-Score
Random Forest	**0.98**	0.73	0.68	**0.71**
Neuroevolution	0.83	**0.81**	**0.79**	**0.80**

3.2 Avaliação da Interface

A análise das métricas mostrou que o Neuroevolution apresentou o melhor equilíbrio entre precisão e revocação, evidenciando sua robustez para dados clínicos, o que é crucial para minimizar riscos de falsos negativos em aplicações médicas.

A interface desenvolvida com o Streamlit apresentou resultados satisfatórios em termos de usabilidade e clareza. O sistema foi projetado para oferecer uma experiência simples, permitindo ao usuário inserir valores clínicos e receber, de forma imediata, uma estimativa visual do risco de diabetes. Dessa forma, o DiabPredict se apresenta como uma ferramenta de apoio à tomada de decisão, podendo ser utilizada tanto em contextos educacionais quanto em iniciativas de triagem preventiva.

3.3 Limitações e Observações

Apesar dos resultados promissores obtidos pelo DiabPredict, algumas limitações precisam ser consideradas. A base de dados utilizada é relativamente pequena e restrita a pacientes do sexo feminino de origem Pima, o que pode comprometer a generalização dos modelos para outras populações. Além disso, possíveis vieses presentes nos dados podem afetar a performance dos algoritmos, principalmente em casos de distribuição desigual das variáveis clínicas. Para estudos futuros, recomenda-se o uso de bases de dados mais amplas e diversificadas, assim como a aplicação de técnicas de balanceamento e validação externa, visando aumentar a robustez e a confiabilidade do sistema.

4. Conclusão

O desenvolvimento do DiabPredict evidenciou o potencial do aprendizado de máquina na área da saúde, especialmente no diagnóstico precoce do diabetes tipo 2. O projeto demonstrou que técnicas supervisionadas, como Neuroevolution e Random Forest, podem alcançar bons níveis de acurácia com um conjunto de dados relativamente pequeno.

Além disso, a integração do modelo preditivo a uma interface interativa reforça a aplicabilidade prática da solução, tornando-a acessível a profissionais da área médica e a pacientes em geral. Como perspectivas futuras, propõe-se ampliar a base de dados, incorporar novos atributos clínicos e testar redes neurais artificiais para aprimorar o desempenho do modelo.

O DiabPredict representa, portanto, uma contribuição significativa para o uso de tecnologias acessíveis em prol da saúde preventiva, aproximando a Inteligência Artificial do cotidiano clínico e educativo.

Referências

American Diabetes Association (2023). Standards of Medical Care in Diabetes–2023. Diabetes Care, 46(Supplement_1), S1–S154.

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Pima Indians Diabetes Database. University of California, Irvine. Disponível em: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

Han, J., Kamber, M., & Pei, J. (2022). Data Mining: Concepts and Techniques (4th ed.). Elsevier.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Streamlit Inc. (2022). Streamlit Documentation. Disponível em: https://docs.streamlit.io

World Health Organization (2023). Diabetes fact sheet. Geneva: WHO.