Sistema de Predição de Doenças Cardiovasculares com XGBoost

Integrante

João Paulo Ladeia Santana

Resumo

Este projeto propõe o desenvolvimento de um sistema de predição de doenças cardiovasculares utilizando o algoritmo **XGBoost (Extreme Gradient Boosting)**, reconhecido por sua eficiência em problemas de classificação. O estudo será conduzido a partir de um dataset público contendo informações clínicas e demográficas de pacientes, com o objetivo de identificar fatores de risco e prever a probabilidade de desenvolvimento de doenças cardiovasculares. O trabalho insere-se no campo da Inteligência Artificial aplicada à saúde, com potencial de impacto positivo no apoio ao diagnóstico e na prevenção.

Introdução

Contextualização

As doenças cardiovasculares são a principal causa de mortalidade global, responsáveis por aproximadamente 17,9 milhões de mortes anuais segundo a Organização Mundial da Saúde (WHO, 2023). Fatores como hipertensão, tabagismo, diabetes, colesterol elevado e obesidade são conhecidos como determinantes relevantes para o risco cardiovascular (Yusuf et al., 2020). Nesse contexto, a Inteligência Artificial (IA) tem se mostrado uma aliada importante, permitindo a criação de modelos preditivos que auxiliam na identificação precoce de pacientes em risco.

Justificativa

A aplicação de algoritmos de aprendizado de máquina na saúde vem crescendo devido à sua capacidade de lidar com grandes volumes de dados e encontrar padrões complexos. O **XGBoost**, em especial, destaca-se pela robustez, velocidade de

processamento e performance em benchmarks de classificação (Chen; Guestrin, 2016). Sua aplicação na área médica pode contribuir para diagnósticos mais precisos e estratégias preventivas.

Objetivo

O objetivo deste projeto é construir um modelo preditivo para detecção de risco cardiovascular, utilizando **XGBoost** e avaliando seu desempenho em métricas como acurácia, precisão, recall, f1-score e AUC-ROC. Pretende-se identificar quais variáveis possuem maior relevância para a predição, fornecendo insights interpretáveis para profissionais de saúde.

Opção do Projeto

O projeto se enquadra na **opção Framework**, empregando bibliotecas de aprendizado de máquina em Python, como XGBoost, para solução de um problema de classificação na área da saúde.

Descrição do Problema

O problema a ser enfrentado é a dificuldade em identificar precocemente indivíduos em risco de desenvolver doenças cardiovasculares. Muitas vezes, sintomas são negligenciados ou surgem tardiamente, comprometendo a eficácia do tratamento. Um modelo preditivo pode auxiliar no **rastreamento preventivo** em larga escala, utilizando dados simples de prontuários médicos (idade, sexo, pressão arterial, colesterol, glicemia, IMC, hábitos de vida).

Aspectos Éticos e Responsabilidade

O uso de IA em saúde levanta questões éticas fundamentais:

- **Privacidade e confidencialidade**: os dados devem ser anonimizados e tratados de forma a proteger a identidade dos pacientes (LGPD, 2018).
- **Transparência**: modelos preditivos precisam fornecer explicabilidade para profissionais de saúde, evitando a sensação de “caixa-preta”.
- **Justiça e não discriminação**: o modelo deve evitar vieses relacionados a gênero, idade ou etnia, garantindo previsões equitativas.

- **Responsabilidade compartilhada**: o sistema é um apoio à decisão, não substituindo a análise médica.

Segundo Topol (2019), a IA deve ser utilizada de forma complementar, sempre sob supervisão profissional, para garantir que avanços tecnológicos não comprometam a autonomia e a ética médica.

Dataset e Análise Exploratória

Dataset

Será utilizado o **Cardiovascular Disease Dataset** (Kaggle, Cole Welkins, 2022), que contém milhares de registros de pacientes com variáveis clínicas e histórico de saúde.
Principais variáveis: idade, sexo, colesterol, glicemia, pressão arterial sistólica e diastólica, IMC, tabagismo, prática de atividade física, histórico de doenças.
Variável alvo: presença (1) ou ausência (0) de doença cardiovascular.

Análise Exploratória

A análise exploratória será realizada em Python (Pandas, Seaborn, Matplotlib), incluindo:

- Estatísticas descritivas (média, desvio-padrão, mediana).
- Distribuição da variável alvo (balanceamento de classes).
- Identificação de valores ausentes e outliers.
- Correlação entre variáveis (mapa de calor).
- Visualização de variáveis mais relevantes para risco cardiovascular.

Preparação dos Dados

- Tratamento de valores faltantes.
- Normalização de variáveis numéricas (ex.: idade, colesterol).
- Codificação de variáveis categóricas (ex.: sexo).
- Divisão em treino (70%) e teste (30%).

Metodologia e Resultados Esperados

Metodologia

1. limpeza do dataset.
2. Análise exploratória com gráficos e estatísticas.
3. Preparação dos dados (normalização, codificação, divisão treino/teste).
4. Treinamento do modelo com **XGBoost Classifier**.
5. Avaliação por métricas de classificação (Accuracy, Precision, Recall, F1-score, AUC-ROC).
6. Interpretação de importância das variáveis (feature importance).

Resultados Esperados

- Acurácia prevista acima de 75%.
- Identificação dos fatores mais relevantes para risco cardiovascular.
- Geração de um sistema preditivo interpretável, com potencial aplicação como ferramenta de apoio clínico.

Referências

- CHEN, T.; GUESTRIN, C. *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD, 2016. DOI: 10.1145/2939672.2939785.
- WORLD HEALTH ORGANIZATION (WHO). *Cardiovascular diseases (CVDs)*. 2023. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Acesso em: 19 set. 2025.
- YUSUF, S. et al. *Modifiable risk factors, cardiovascular disease, and mortality in 155,722 individuals from 21 high-income, middle-income, and low-income countries (PURE)*. Lancet, v. 395, p. 795–808, 2020. DOI: 10.1016/S0140-6736(19)32008-2.
- TOPOL, E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books, 2019.
- BRASIL. *Lei Geral de Proteção de Dados Pessoais (LGPD)*. Lei nº 13.709, de 14 de agosto de 2018.
- KAGGLE. *Cardiovascular Disease Dataset*. Cole Welkins. Disponível em: https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease