Mestrado em Engenharia Informática

Introdução a Aprendizagem Automática

09 de Janeiro 2025

Previsão de Risco de Doenças Cardiovasculares

Paulo Francisco Pinto (128962)

Vitor Barbosa (105248)

Resumo

Este projeto aplica a metodologia CRISP-DM para desenvolver um modelo preditivo capaz de identificar o risco de doenças cardiovasculares, utilizando dados provenientes de um sistema de vigilância de saúde pública dos Estados Unidos. A análise inclui o tratamento de dados ausentes, normalização de variáveis, codificação de atributos categóricos e a implementação de algoritmos supervisionados e não supervisionados, como Decision Trees, k-Means e Random Forest. O modelo final apresentou resultados robustos, com destaque para uma precisão de 87% obtida pela Árvore de Decisão, além de padrões relevantes identificados pelos algoritmos não supervisionados. Apesar de desafios como a presença de dados enviesados e a natureza autodeclarada do dataset, o estudo demonstra o potencial do uso de aprendizagem automática na área da saúde pública. Este trabalho contribui para a priorização de recursos e o desenvolvimento de políticas preventivas, posicionando-se como um exemplo de como tecnologias de machine learning podem ser aplicadas em problemas de impacto social significativo.

Palavras-chave: Doenças Cardiovasculares, Machine Learning, CRISP-DM, Previsão de Risco, Saúde Pública

Índice

1. Introdução

As doenças cardiovasculares representam uma das principais causas de mortalidade no mundo, com elevado impacto socioeconómico. Com o avanço das tecnologias de aprendizagem automática, torna-se possível identificar padrões em grandes conjuntos de dados, permitindo prever indivíduos em risco com maior precisão e eficiência. Este trabalho aplica a metodologia CRISP-DM para desenvolver um modelo preditivo robusto, capaz de identificar o risco de doenças cardiovasculares, utilizando um conjunto de dados provenientes de um sistema de vigilância de saúde pública dos Estados Unidos. Este estudo tem como objetivo fornecer suporte para intervenções preventivas e políticas de saúde pública mais eficazes.

2. Revisão da Literatura

O uso de aprendizagem automática na previsão de doenças cardiovasculares tem sido amplamente explorado em estudos recentes. Lupague et al. (2023) propuseram um modelo preditivo baseado em múltiplos fatores de risco, incluindo características demográficas e clínicas, demonstrando resultados significativos em termos de precisão. Pedregosa et al. (2011) destacaram a aplicabilidade do scikit-learn como ferramenta poderosa para a implementação de algoritmos de aprendizagem automática. Este trabalho expande estas abordagens ao avaliar estratégias específicas de tratamento de dados ausentes, codificação de variáveis e comparação entre diferentes algoritmos supervisionados e não supervisionados.

3. Metodologia

3.1. Entendimento do Negócio

O objetivo principal deste estudo é identificar fatores de risco significativos e desenvolver um modelo preditivo para auxiliar profissionais de saúde na identificação precoce de indivíduos em risco. A previsão eficiente pode apoiar intervenções personalizadas, reduzir custos associados ao tratamento de estágios avançados de doenças e priorizar recursos em saúde pública.

3.2. Entendimento dos Dados

O conjunto de dados utilizado é proveniente do sistema de vigilância BRFSS, que reúne informações sobre fatores demográficos, comportamentais e clínicos de indivíduos. Este dataset contém 19 atributos, incluindo variáveis como idade, IMC, hábitos de consumo alimentar e condições de saúde. A variável-alvo é binária, indicando a presença ou ausência de doenças cardiovasculares. Realizámos uma análise exploratória para compreender a estrutura dos dados, identificar valores ausentes e visualizar relações entre variáveis.

3.3. Preparação dos Dados

Durante a preparação dos dados, implementámos:

- **Tratamento de valores ausentes:** Utilizámos a mediana para substituir valores ausentes em variáveis numéricas.
- **Codificação de variáveis categóricas:** Aplicámos **Label Encoding** em variáveis como "Sexo", "Consumo de Álcool" e "Histórico de Fumo".
- **Verificação do dataset tratado:** Validámos o dataset para assegurar a consistência e ausência de valores ausentes.

3.4. Modelagem

Testámos algoritmos supervisionados, como Decision Trees, Multi-layer Perceptron e k-NN, e não supervisionados, como k-Means e DBSCAN. A modelagem foi realizada com base nas seguintes etapas:

- Separação dos dados em conjuntos de treino (80%) e teste (20%).
- Codificação de variáveis categóricas utilizando One-Hot Encoding, quando necessário.
- Avaliação utilizando métricas como precisão, recall e F1-score.

3.5. Avaliação

Os modelos foram avaliados utilizando métricas predefinidas. A Árvore de Decisão demonstrou o melhor desempenho com precisão de 87%, enquanto o k-Means foi eficaz em identificar padrões em subgrupos específicos.

3.6. Implantação

O pipeline desenvolvido foi concebido para ser reutilizável em diferentes contextos. O modelo foi salvo em formato joblib para facilitar a integração em sistemas reais de suporte à decisão.

4. Descrição do Dataset e Preparação

O dataset original contém 308.854 registos e 19 atributos. Durante a preparação, valores ausentes foram tratados utilizando a mediana, e variáveis categóricas foram codificadas para valores numéricos. A análise exploratória revelou correlações significativas entre variáveis como idade, IMC e risco de doenças cardiovasculares.

5. Experimentos e Resultados

Realizámos uma série de experimentos para avaliar diferentes abordagens de modelagem. Os principais resultados incluem:

- **Decision Trees:** Apresentaram o melhor desempenho global com uma precisão de 87% e F1-score elevado.
- **k-Means:** Destacou-se na análise não supervisionada, identificando dois clusters principais associados a comportamentos de risco.

6. Discussão e Limitações

Os resultados obtidos demonstraram a eficácia de modelos supervisionados na previsão de risco. No entanto, desafios foram identificados, como:

- A presença de vieses nos dados, devido à natureza autorreportada do sistema de vigilância.
- A dificuldade em lidar com variáveis categóricas com múltiplos níveis.
- A necessidade de algoritmos mais avançados para lidar com características complexas nos dados.

7. Conclusões e Trabalhos Futuros

Este estudo demonstra o potencial do uso de aprendizagem automática na saúde pública, especialmente na previsão de doenças cardiovasculares. Trabalhos futuros podem explorar:

- A inclusão de novos atributos, como genéticos ou ambientais.
- A integração do modelo em sistemas de suporte à decisão em tempo real.

8. Referências

- Lupague et al. (2023). Integrated machine learning model for comprehensive heart disease risk assessment based on multi-dimensional health factors. European Journal of Computer Science and Information Technology, 11(3):44–58.
- Pedregosa et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

4. Anexos

- **Notebook:**

  "IAAprojeto_CRISPDM - Vitor_Barbosa_105248_Francisco_Pinto_128962"

  Contém todo o código efetuado para resolução do problema apresentado no projeto, assim como a aplicação da metodologia CRISP-DM.

- **Enunciado:**

  "IAA_Project_2025.pdf"

- **Dados:**

  Pasta estruturada por: *"ZipComCsv > CVD_cleaned.csv"*

- **Exercícios Anteriores:**

  "IAAex1 - Vitor_Barbosa_105248_Francisco_Pinto_128962.ipynb"
  "IAAex2 - Vitor_Barbosa_105248_Francisco_Pinto_128962.ipynb"
  "IAAex3 - Vitor_Barbosa_105248_Francisco_Pinto_128962.ipynb"
  "IAAex4 - Vitor_Barbosa_105248_Francisco_Pinto_128962.ipynb"

  Exercícios já previamente entregues em aula, que serviram como referência para aplicação da resolução deste projeto final.