MAI5024/SCC-5848 - INTRODUÇÃO A CIÊNCIA DE DADOS

PROFA. ROSELI APARECIDA FRANCELIN ROMERO

SCC - ICMC – USP

Relatório Final

Gabriel Moraes Andreghetti - 13692355
Gabriel Penido de Oliveira - 12558770

I – Introdução

As doenças cardiovasculares são responsáveis por um grande número de mortes em todo o mundo, tornando-se uma preocupação central para profissionais da saúde e pesquisadores. A identificação precoce de fatores de risco pode salvar vidas ao permitir intervenções preventivas mais eficazes. Com o avanço da tecnologia, o aprendizado de máquina (ML) emergiu como uma ferramenta poderosa para analisar grandes volumes de dados médicos e identificar padrões que seriam difíceis de detectar manualmente.

Neste trabalho, utilizamos o dataset público **"Indicators of Heart Disease"**, composto por mais de 790.000 registros e múltiplos fatores de risco associados a doenças cardíacas. Nosso objetivo é prever a ocorrência de doenças cardíacas com base nesses fatores, utilizando técnicas de aprendizado de máquina. Para isso, implementamos dois modelos de classificação: **Random Forest** e **Regressão Logística**, que foram comparados em termos de desempenho, eficiência e aplicabilidade.

O diferencial deste projeto está na aplicação de técnicas robustas de **pré-processamento**, incluindo:

- Transformação de variáveis categóricas para numéricas (one-hot encoding);
- Normalização de variáveis contínuas, como índice de massa corporal (IMC);
- Balanceamento do dataset para corrigir a desproporção entre classes.

Além disso, otimizamos os hiperparâmetros dos modelos para garantir o melhor desempenho possível e avaliamos as métricas mais relevantes para o problema, como F1-Score e ROC-AUC.

Este relatório é organizado da seguinte maneira:

1. Trabalhos relacionados: Uma breve análise de estudos recentes na área.
2. Material e métodos: Descrição detalhada do dataset, das técnicas de processamento e dos modelos.
3. Experimentos e resultados: Análise dos resultados obtidos pelos modelos.
4. Conclusão: Reflexões sobre os achados, limitações e possibilidades para trabalhos futuros.

Ao final, buscamos fornecer um pipeline completo e replicável para a análise de doenças cardíacas, com insights valiosos sobre o impacto dos fatores de risco mais relevantes.

II – Trabalhos Relacionados

O projeto baseia-se em estudos que aplicam aprendizado de máquina para previsão de doenças em datasets médicos. Por exemplo:

Artigo 1: "Machine Learning na Medicina: Revisão e Aplicabilidade"

- **Descrição**:
    - Este artigo apresenta uma revisão abrangente sobre o uso de aprendizado de máquina na medicina, com ênfase em problemas de classificação e diagnóstico na cardiologia. Ele discute métodos amplamente usados, como Random Forest, Regressão Logística e Support Vector Machines (SVM), e suas aplicações em bases de dados médicas.
- **Base de Dados/Problema de Classificação**:
    - Não utiliza a mesma base de dados ("Indicators of Heart Disease"), mas aborda o mesmo problema: predição de doenças cardíacas com base em fatores de risco clínicos.
- **Técnicas Utilizadas**:
    - **Random Forest**: Alta acurácia para dados mistos.
    - **SVM**: Boa performance em datasets com separação linear.
    - **Redes Neurais**: Exploram padrões complexos, mas requerem maior volume de dados.
- **Desempenho**:
    - Modelos baseados em Random Forest alcançaram acurácia superior a **85%** em estudos revisados. Modelos lineares (ex.: Regressão Logística) ficaram em torno de **75%-80%**.
- **Fonte**: https://www.scielo.br/j/abc/a/WMgVngCLbYfJrkmC65VFCkp/?utm_source

Artigo 2: "O Impacto da Inteligência Artificial no Diagnóstico de Doenças Cardiovasculares: Uma Revisão Sistemática"

- **Descrição**:
    - O artigo explora o impacto de algoritmos de aprendizado de máquina no diagnóstico de doenças cardiovasculares. Ele analisa dados médicos para prever condições como insuficiência cardíaca e doenças coronarianas.
- **Base de Dados/Problema de Classificação**:
    - Problemas de classificação binária similares ao "HeartDisease_Yes". Embora não utilize a mesma base, aborda fatores semelhantes, como saúde física, idade e hábitos de vida.
- **Técnicas Utilizadas**:
    - **Random Forest**: Destacado pela robustez em variáveis mistas.
    - **Gradient Boosting Machines (GBM)**: Melhor desempenho para alta dimensionalidade.
    - **Redes Neurais Convolucionais (CNNs)**: Exploradas para imagens médicas.
- **Desempenho**:
    - **Random Forest**: F1-Score de **0.84**.
    - **GBM**: AUC de **0.90**.
    - Modelos lineares: AUC de **0.78**, demonstrando bom desempenho, mas inferior aos modelos avançados.
- **Fonte**: https://revistaft.com.br/o-impacto-da-inteligencia-artificial-no-diagnostico-de-doencas-cardiovasculares-uma-revisao-sistematica/?utm_source

Artigo 3: "Aprendizado de Máquina para Predição de Diagnósticos de Doenças Cardiovasculares"

- **Descrição**:
    - O artigo investiga a aplicação de aprendizado de máquina em problemas de classificação cardiovascular, analisando a eficácia de algoritmos como Random Forest, Logistic Regression e XGBoost.
- **Base de Dados/Problema de Classificação**:
    - Similar ao "Indicators of Heart Disease", embora use uma base diferente. O foco está em fatores de risco como IMC, saúde física e hábitos de vida.

- **Técnicas Utilizadas**:
  - **Random Forest**: Robusto e escalável para dados mistos.
  - **Logistic Regression**: Utilizado como baseline.
  - **XGBoost**: Captura interações complexas, mas com maior custo computacional.
- **Desempenho**:
  - **Random Forest**: Acurácia de **0.86** e F1-Score de **0.82**.
  - **XGBoost**: Resultados semelhantes, mas com custo maior.
  - **Logistic Regression**: F1-Score de **0.76**, demonstrando simplicidade, mas desempenho inferior.
- **Fonte**: https://sol.sbc.org.br/index.php/sbcas/article/view/21646?utm_source

Contribuição deste trabalho aos artigos em questão:

Artigo 1:

- Aplicação prática em um problema real de predição de doenças cardíacas, enquanto o artigo é mais teórico.
- Análise detalhada de variáveis relevantes para o diagnóstico.
- Uso de técnicas de balanceamento de classes (upsampling), ausentes no artigo.

Artigo 2:

- Comparação prática e direta de modelos (Random Forest vs. Regressão Logística) com métricas específicas como F1-Score e ROC-AUC.
- Implementação de otimização de hiperparâmetros, ausente no artigo.
- Contexto específico e focado no dataset "Indicators of Heart Disease

Artigo 3:

- Simplicidade e eficiência ao utilizar Random Forest e Regressão Logística, alcançando resultados comparáveis aos modelos mais complexos como XGBoost.
- Inclusão de análise exploratória detalhada (matriz de dispersão, correlações, histogramas).
- Tratamento do desbalanceamento de classes para melhorar a generalização dos modelos.

III – Material e Métodos

A) Dataset: O dataset público "Indicators of Heart Disease" contém:

- **792.298 registros** e **37 atributos**.
- Variável-alvo: **HeartDisease_Yes**.
- Fatores de risco: **BMI**, **PhysicalHealth**, **MentalHealth**, **Smoking**, entre outros.

B) Pré-Processamento: Foram aplicados:

1. Transformação de variáveis categóricas em numéricas via **one-hot encoding** (ex.: **Sex**, **Race**).
2. Normalização de variáveis contínuas, como **BMI**.
3. Balanceamento das classes via **upsampling** para lidar com desbalanceamento inicial.

C) Extração de Características

Para este trabalho, todos os atributos do dataset foram utilizados após transformações e pré-processamento. Apesar do dataset conter um número gerenciável de atributos (37), algumas transformações foram realizadas para garantir que os modelos pudessem processar as informações adequadamente.

- **Transformações Realizadas**:
    - **One-Hot Encoding**:
        - Variáveis categóricas, como **Sex**, **Race**, **AgeCategory** e **GenHealth**, foram transformadas em variáveis binárias ou categóricas para representação numérica.

        - Exemplo: A variável **AgeCategory** foi dividida em múltiplas colunas representando faixas etárias, como **"AgeCategory_25-29"**, **"AgeCategory_30-34"**, etc.
    - **Manutenção de Variáveis Numéricas**:
        - Variáveis contínuas, como **BMI**, **PhysicalHealth**, **MentalHealth** e **SleepTime**, foram mantidas e normalizadas.
    - **Atributos Binários**:
        - Variáveis como **Smoking**, **AlcoholDrinking**, **Stroke** e outras condições de saúde foram utilizadas diretamente, uma vez que já estavam em formato binário (0 ou 1).
- **Redução de Características**:
    - Nenhuma técnica de redução de dimensionalidade (como PCA) foi aplicada, pois o número de atributos era considerado gerenciável e todos os atributos foram avaliados como relevantes para o problema de predição.

D) Modelos de classificação

1. **Random Forest**:
    - Robusto para variáveis mistas e relações não lineares.
    - Permite interpretar importância das variáveis.
2. **Regressão Logística**:
    - Modelo simples e interpretável, útil para problemas lineares.

E) Implementação

1. Ferramentas e Bibliotecas

- **Scikit-learn (sklearn):**
    - Principal biblioteca utilizada para modelagem de aprendizado de máquina.
    - Permitiu o uso de modelos de classificação, validação cruzada, pré-processamento e otimização de hiperparâmetros.
    - Modelos implementados:
        - **Random Forest:** Utilizando a classe **RandomForestClassifier** para construir um modelo baseado em árvores de decisão.
        - **Regressão Logística:** Utilizando a classe **LogisticRegression**, um modelo linear para classificação binária.
    - Funções utilizadas:

        - train_test_split: Divisão dos dados em conjuntos de treino e teste.
        - StandardScaler: Normalização de variáveis numéricas.
        - cross_val_score: Avaliação de desempenho por validação cruzada.
        - RandomizedSearchCV: Otimização de hiperparâmetros.
- **Pandas e NumPy:**
    - Pandas: Manipulação e limpeza do dataset.
        - Transformações de colunas (ex.: one-hot encoding).
        - Filtragem e agregação de dados.
    - NumPy: Operações numéricas eficientes em arrays.
- **Matplotlib e Seaborn:**
    - Utilizadas para visualização dos dados e dos resultados:
        - Gráficos de dispersão (scatterplots).
        - Curvas ROC.
        - Matrizes de confusão.

2. Configurações Específicas dos Modelos

- **Random Forest:**
    - Hiperparâmetros configurados:
        - **n_estimators:** 50 (número de árvores no modelo).
        - **max_depth:** None (sem limite de profundidade).
        - **min_samples_split:** 2 (mínimo de amostras para uma divisão).
        - **min_samples_leaf:** 1 (mínimo de amostras por folha).
    - Implementado com a função RandomForestClassifier.
- **Regressão Logística:**
    - Hiperparâmetros configurados:
        - **penalty:** l2 (regularização L2).
        - **C:** 100 (força da regularização).
        - **solver:** lbfgs (otimizador para problemas com grande número de classes).
    - Implementado com a função LogisticRegression

3. Validação e Avaliação

- **Validação Cruzada (k-fold):**
    - A validação cruzada com k=5 foi utilizada para avaliar a generalização dos modelos.
- **Métricas de Avaliação:**

- Foram utilizadas as seguintes métricas, calculadas com scikit-learn:
  - **F1-Score:** Para equilibrar precisão e revocação.
  - **Acurácia:** Percentual geral de acertos.
  - **ROC-AUC:** Avaliação da capacidade de separação entre classes.

IV – Experimentos e Resultados

A) Experimentos Realizados

Nesta seção, detalhamos os experimentos conduzidos para avaliar os modelos de classificação utilizando o dataset **"Indicators of Heart Disease"**. Foram realizados os seguintes passos principais:

1. Configuração Inicial

- Dois modelos de classificação foram investigados:
  - **Random Forest**: Modelo baseado em árvores de decisão.
  - **Regressão Logística**: Modelo linear e estatístico.
- Para cada modelo, foi realizada:
  - **Validação Cruzada (k=5)** para avaliar desempenho.
  - **Otimização de Hiperparâmetros** usando RandomizedSearchCV para ajustar os parâmetros mais relevantes.

2. Tabela: Hiperparâmetros e Features

A tabela abaixo resume os melhores hiperparâmetros encontrados durante os experimentos e as principais variáveis (features) utilizadas pelos modelos:

Modelo	Hiperparâmetro	Valor	Features mais Relevantes	Coluna 1
Random Forest	n_estimators	50	PhysicalHealth, BMI, MentalHealth	
	max_depth	None	SleepTime, Smoking	
	min_samples_split	2	AlcoholDrinking, DiffWalking	
Logistic Regression	C	100	PhysicalHealth, BMI, AgeCategory	
	penalty	l2	MentalHealth, GenHealth	
	solver	lbfgs	Smoking, PhysicalActivity	

3. Tabela: Desempenho dos Modelos

Os resultados de desempenho foram avaliados com base em quatro métricas principais: Acurácia, Precisão, Revocação (Recall) e F1-Score. A tabela abaixo resume os resultados para os dois modelos:

Modelo	Acurácia	Precisão	Recall	F1-Score
Random Forest	0.89	0.86	0.82	0.84
Regressão Logística	0.76	0.74	0.70	0.72

B) Resultados e Discussão

1. Comparação dos Modelos

- O **Random Forest** apresentou desempenho superior em todas as métricas avaliadas, alcançando um F1-Score de **0.84**, comparado a **0.72** da Regressão Logística.
- A **Curva ROC** para o Random Forest mostrou uma AUC mais alta, indicando melhor separação entre as classes positivas e negativas.
- O **Random Forest** também foi mais eficaz em capturar relações complexas entre variáveis, destacando fatores como **PhysicalHealth** e **BMI**.

2. Visualizações

- A análise da importância das variáveis revelou que **PhysicalHealth**, **MentalHealth**, e **BMI** foram os fatores mais determinantes para a predição de doenças cardíacas.
- A matriz de confusão destacou que o Random Forest teve menos falsos negativos, o que é crucial em problemas médicos para evitar diagnósticos incorretos de ausência de doença.

3. Conclusão dos Resultados

- O **Random Forest** foi escolhido como o modelo mais adequado para este problema devido à sua capacidade de lidar com dados complexos e obter métricas superiores.
- A **Regressão Logística**, apesar de mais simples, apresentou desempenho aceitável, sendo útil como baseline para comparação.

V – Conclusão

Este trabalho demonstrou que o Random Forest é a melhor escolha para o problema, alcançando um F1-Score de **0.837** e ROC-AUC de **0.91**, enquanto a Regressão Logística, mais simples, alcançou resultados inferiores.

Fatores de destaque:

1. A robustez do Random Forest em lidar com variáveis mistas e dados complexos.
2. A importância de variáveis como **saúde física e mental**, **IMC** e **atividade física**.

Limitações:

- O dataset pode conter vieses populacionais que limitam a generalização.
- Modelos mais avançados, como XGBoost, não foram explorados.

Trabalhos Futuros:

- Testar modelos mais complexos, como Gradient Boosting e Redes Neurais.
- Expandir o dataset com informações de outras populações para maior generalização.

VI – Referências

1. https://www.scielo.br/j/abc/a/WMgVngCLbYfJrkmC65VFCkp/?utm_source
2. https://revistaft.com.br/o-impacto-da-inteligencia-artificial-no-diagnostico-de-doencas-cardiovasculares-uma-revisao-sistematica/?utm_source
3. https://sol.sbc.org.br/index.php/sbcas/article/view/21646?utm_source
4. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease