Previsão de Sobrevivência ao Carcinoma Hepatocelular (HCC)

Descrição do Projeto

Este projeto foi desenvolvido como parte da disciplina "Elementos de Inteligência Artificial e Ciência de Dados" do curso de Licenciatura em Inteligência Artificial e Ciência de Dados. O objetivo é desenvolver um pipeline completo de ciência de dados para prever a sobrevivência de pacientes com Carcinoma Hepatocelular (HCC) um ano após o diagnóstico.

Estrutura do Projeto

O projeto está dividido nas seguintes etapas principais:

- Exploração de Dados: Análise exploratória dos dados, incluindo exame dos tipos de características, distribuição das classes, valores por atributo e identificação de inconsistências nos dados

- Pré-processamento de Dados: Processamento de características (imputação de valores ausentes, transformação e escalonamento de dados) e engenharia de características.

- Modelagem de Dados (Aprendizagem Supervisionada): Seleção de algoritmos de aprendizagem supervisionada, definição de conjuntos de treino e teste, e avaliação do desempenho dos modelos.

  Algoritmos utilizados:
  - Árvores de Decisão
  - K-Nearest Neighbors (KNN)
  - Random Forest
  - Gradient Boosting
  - Multi-Layer Perceptron (MLP)
  - Regressão Logística
  - Stacking Classifier
  - Support Vector Classifier (SVC)

- Avaliação dos Dados: Comparação dos resultados de classificação utilizando métricas de avaliação padrão (matriz de correlação, ROC/AUC, precisão, recall, acurácia).

- Interpretação dos Resultados: Extração de insights significativos dos resultados obtidos, explicação do comportamento dos modelos e recomendações para análises futuras.

Instruções para Execução

Pré-requisitos

Certifique-se de ter o Python 3 instalado. As bibliotecas necessárias podem ser instaladas usando o ficheiro requirements.txt.

Instalação

1. Clone o repositório do projeto:
   git clone <trabalho_HCD-2.ipynb >

1. cd < **Trabalho_HCD** >

2. Instale as dependências: pip install -r requirements.txt

Execução do Projeto

1. Navegue até ao diretório do projeto e abra o Jupyter Notebook
2. No Jupyter Notebook, abra o ficheiro trabalho_HCD-2.ipynb e, na segunda célula do código, coloque o caminho correto do dataset
   (data = pd.read_csv('/'colocar o caminho ' /hcc_dataset.csv')
3. Em seguida, carregue em "run all"

- Trabalho_HCD.ipynb: Jupyter Notebook contendo todo o pipeline de ciência de dados, desde a exploração de dados até à avaliação dos modelos.
- hcc_dataset.csv: Conjunto de dados utilizado no projeto.
- requirements.txt: Lista de bibliotecas Python necessárias para executar o projeto.

Dependências

O projeto utiliza as seguintes bibliotecas Python:

- pandas: Para manipulação e análise de dados.
- numpy: Para operações numéricas.
- seaborn: Para visualizações estatísticas avançadas.
- matplotlib: Para visualização de dados.
- scikit-learn: Para construção e avaliação dos modelos de aprendizagem supervisionada.
- imblearn: Para técnicas de oversampling.
- collections: Para operações com coleções.
- SMOTE: Para balanceamento de dados.

Como Utilizar o Programa

- Preparação: Certifique-se de que todas as dependências estão instaladas e que o conjunto de dados hcc_dataset.csv está disponível no diretório do projeto.

- Execução: Abra e execute o Jupyter Notebook trabalho_HCD.ipynb célula por célula, seguindo as instruções e observações fornecidas no notebook.
- Interpretação: Analise as visualizações e métricas de desempenho geradas para entender os resultados dos modelos.

Autoras

Catarina Abrantes
Liliana Silva
Mariana Fonseca