Artificial Intelligence and Supervised Learning

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Inteligência Artificial (IART)

GRUPO A1_111
João Cordeiro (202205682)
Luciano Ferreira (202208158)
Tomás Telmo (202206091)

Predict Students' Dropout/ Academic Success

Características do data set: Conjunto de dados com variáveis numéricas e categóricas, tratado com pré-processamento adequado.
Resultados e Comparaçoes: Performance comparativa dos três algoritmos implementados.
Análises Adicionais: A análise de importância das variáveis e das suas devidas importancias

Objetivo do trabalho:

- Desenvolver modelos de aprendizagem supervisionada para prever a situação acadêmica dos estudantes (graduado, desistente ou em curso**)**
- Identificar fatores determinantes que influenciam o desempenho acadêmico

Definição do Problema

Tipo de problema

– Classificação multiclasse (3 classes possíveis):

- **Dropout** (Desistiu)
- **Enrolled** (Ainda matriculado)
- **Graduate** (Graduado)

Variável-Alvo

– Coluna **Target** do dataset; representa o estado final do aluno.

- Será necessário codificar esta variável para utilizar para utilizar em modelos de ML (ex: Dropout = 0, Enrolled = 1, Graduate = 2)

Relevância do Estudo

- Permite detectar estudantes em risco de forma antecipada.
- Apoia decisões pedagógicas e a gestão de recursos nas instituições de ensino.
- Tem aplicabilidade real e impacto direto na melhoria dos resultados académicos.

Metodologia

Pré-processamento dos dados (limpeza, normalização, codificação de variáveis)

- Divisão treino/teste
- Treinamento com modelos como:
  - Random Forest
  - Árvore de Decisão
  - k-NN
- Avaliação com métricas como:
  - Precisão: Relevância das previsões positivas.
  - *Recall:* Cobertura dos verdadeiros positivos.
  - *F1-Score*: Média harmônica entre precisão e recall.
  - Relatório de Classificação : Resumo das métricas por classe.
- Visualização de Desempenho
  - Matriz de Confusão
  - Gráfico Comparativo de Modelos

Trabalho Relacionado

Estudos Relevantes

- **Krüger et al. (2023)**: Aplicaram técnicas de aprendizagem automática explicáveis para prever o abandono escolar, destacando a importância de dados atualizados e ricos para modelos eficazes. (Fonte 1)
- **Nature (2025):** Utilizaram o algoritmo CatBoost em logs de atividade do Moodle para prever o abandono escolar, demonstrando a eficácia de modelos baseados em interações em plataformas de e-learning. (Fonte 2)
- **ScienceDirect (2024):** Compararam modelos de aprendizagem automática, como árvores de decisão e SVM, utilizando dados de transcrições e demográficos para prever o abandono escolar. (Fonte 3)
- **MDPI (2023):** Implementaram modelos de aprendizagem automática para prever desempenho académico e abandono, integrando ferramentas de análise de aprendizagem para apoiar equipas de tutoria. (Fonte 4)

Principais Conclusões

- Modelos como Random Forest, SVM e CatBoost são frequentemente utilizados devido à sua capacidade de lidar com dados complexos e desequilibrados.
- A integração de dados de plataformas de e-learning, como o Moodle, melhora a precisão das previsões.
- A explicabilidade dos modelos é crucial para a adoção por parte das instituições educativas.

- **Fonte 1:** https://www.sciencedirect.com/science/article/abs/pii/S0957417423014355?
- **Fonte 2:** https://www.nature.com/articles/s41598-025-93918-1?
- **Fonte 3:** https://www.sciencedirect.com/science/article/pii/S0160791X24000228?
- **Fonte 4:** https://www.mdpi.com/2306-5729/7/11/146?

Ferramentas Utilizadas e Algoritmos de Classificaçao

Algoritmos de Classificação

- Árvore de Decisão - Útil para entender a importância das variáveis.
- Random Forest - Conjunto de múltiplas árvores de decisão
- k-NN - Classificação baseada na proximidade entre os dados.

Técnicas adicionais

- GridSearchCV: Ajuste de hiperparâmetros para selecionar a melhor configuração de cada modelo.
- MinMaxScaler: Técnicas de normalização de variáveis para melhorar o desempenho dos modelos.
- PCA : Utilizada para redução de dimensionalidade e análise exploratória.

Ferramentas

- Python
- Jupyter Notebook: Ambiente interativo para desenvolvimento.
- Pandas & NumPy: Manipulação e análise de dados.
- Seaborn & Matplotlib: Visualização de dados e resultados.
- Scikit-learn: Biblioteca para machine learning
  - Pré-processamento dos dados
  - Treinamento e avaliação dos modelos
  - Métricas de desempenho

Trabalho Realizado até a data

Importação de Bibliotecas

- pandas, numpy, seaborn, matplotlib, sklearn, warning

Carregamento e Pré-processamento de Dados

- Leitura dos dados CSV
- Tratamento de valores ausentes.
- Codificação de variáveis categóricas e do alvo
- Normalização dos dados numéricos

Análise Exploratória

- Heatmap de correlações.
- Gráfico de taxa de abandono por curso.
- Análise dos atributos mais correlacionados.

Separação das Features e Target

- X com as variáveis independentes.
- Y com a variável-alvo.

Redução de Dimensionalidade com PCA

- Aplicado PCA para combinar atributos correlacionados

Divisão em conjuntos de Treino e Teste

- train_test_split com 20% dos dados para teste.

Treino e Otimização de Modelos

- Modelos utilizados:
- Árvore de Decisão
- Random Forest
- K-Nearest Neighbors (KNN)
- Utilização de GridSearchCV para otimizar hiperparâmetros.

Avaliação de Desempenho dos Modelos

- Métricas: Acurácia, Precisão, Revocação, F1 Score.
- Random Forest teve o melhor desempenho (~75,7%).

Curva de Aprendizado

- Avaliação de overfitting.

Conclusões

- Random Forest foi identificado como o melhor modelo.
- PCA ajudou a simplificar o conjunto de dados com boa retenção de variância.