Licenciatura em Ciência de Dados Aplicada
Aprendizagem Automática
2023/2024

Trabalho prático
Análise de Dados usando Aprendizagem Automática

Este trabalho propõe uma investigação aprofundada no campo da exploração e análise de dados utilizando técnicas de aprendizagem automática. O objetivo principal é capacitar os alunos a aplicar metodologias adequadas para explorar e compreender conjuntos de dados, realizar pré-processamento, aplicar técnicas de análise não supervisionada e avaliar o desempenho de diferentes algoritmos de aprendizagem automática supervisionada.

Para tal, deverá ser criado um Jupyter Notebook, estruturado em secções, que englobe os passos da análise realizada e explique de forma sucinta os procedimentos realizados e as decisões tomadas ao longo da análise (3 primeiras etapas). Os alunos deverão também criar uma aplicação simples que permita usar os modelos gerados para fazer novas previsões.

Adicionalmente, os grupos farão uma apresentação intermédia do trabalho realizado. Para tal deveram usar o Jupyter Notebook do trabalho.

Os trabalhos podem ser realizados para diversos tipos de problemas (classificação binária, classificação multiclasse e regressão), sendo da responsabilidade do grupo a escolha do problema e do conjunto de dados correspondente (podem encontrar algumas sugestões de onde encontrar datasets no final do enunciado).

Os grupos de trabalho devem ser compostos por 2 elementos (com excepção de um grupo que poderá ter 3 elementos).

Genericamente, ao longo do trabalho, devem ser realizadas as seguintes etapas:

1. **Exploração Inicial e Pré-processamento:**
    a. Revisão da documentação disponível sobre o conjunto de dados.
    b. Análise exploratória do conjunto de dados.
    c. Pré-processamento dos dados, incluindo tratamento de valores omissos e possivelmente geração e seleção de atributos.
    d. Descrição das características dos dados e justificação das escolhas de pré-processamento.
    e. Inclusão de gráficos exploratórios iniciais que ilustrem as principais características dos dados.

2. **Análise Não Supervisionada**
    a. Uso de técnicas de redução de dimensionalidade.
    b. Análise dos resultados obtidos para as técnicas de redução de dimensionalidade e da visualização de dados.
    c. Uso de métodos de clustering.
    d. Análise dos resultados obtidos a partir dos algoritmos de clustering.

3. **Aprendizagem automática supervisionada:**
    a. Comparação da performance de diversos modelos/algoritmos de aprendizagem automática.
    b. Uso de métodos de ensemble.
    c. Cálculo de métricas de erro e utilização de métodos de estimação de erro adequados.
    d. Optimização de hiperparametros.
    e. Escolha do melhor modelo alcançado e interpretação dos resultados quando possível.
    f. Análise crítica dos resultados obtidos nesta etapa.

4. **Desenvolvimento de uma aplicação simples:**
    a. Desenvolvimento de uma aplicação simples que permita usar os modelos treinados para fazer novas previsões
    b. A aplicação deve ser intuitiva e fácil de usar
    c. Recomenda-se o uso da framework streamlit (opcional)

Datas importantes:

- Escolha definitiva dos dados e dos grupos de trabalho:
    - 28 de Março de 2024
    - Até à data os datasets escolhidos deverão ser discutidos com o docente para aprovação
- Apresentação: 16 de Maio
- Submissão do trabalho: 30 de Maio

Avaliação:

A avaliação dos trabalhos será feita recorrendo a:

- Apresentação intermédia: 20 %
- Notebook final: 60%
- Aplicação: 20%

O docente reserva se o direito de justificadamente poder atribuir classificações distintas aos vários elementos de cada grupo.

Alguns recursos para escolher datasets:

- UCI Machine Learning Repository: https://archive.ics.uci.edu/datasets
- Kaggle datasets: https://www.kaggle.com/datasets
- Awesome public datasets repository: https://github.com/awesomedata/awesome-public-datasets/tree/master