FCD 2023 Trabalho Prático 1

November 4, 2024

Enunciado

O objectivo do trabalho prático é transformar um conjunto de dados com potenciais inconsistências e problemas num conjunto de dados coerente.

Para tal, será necessário utilizar técnicas presentes nos módulos referentes às bibliotecas NumPy e Pandas.

O produto final deverá ser um conjunto de dados devidamente limpo, pronto para que sejam aplicadas as mais variadas técnicas de visualização e machine learning (que será a segunda parte do trabalho).

Metodologia

Devem começar por fazer download do vosso conjunto de dados que estará disponível no Moodle dentro da área referente ao trabalho prático.

Cada ficheiro comprimido terá dois ficheiros no seu conteúdo: um é o dataset (termina em .csv) e o outro é um ficheiro com detalhes sobre o dataset.

Importante: Na última linha do ficheiro details.txt terão um link para o Kaggle (plataforma de onde retirei o conjunto de dados). Nesse link encontrarão também informação que poderá ser útil para vocês.

De seguida devem começar por analisar o conjunto de dados percebendo quais as variáveis disponíveis, quais os valores que cada uma das variáveis toma, etc.

Ao analisarem o conjunto de dados deverão referir todas as inconsistências e problemas que encontraram e também qual foi a forma escolhida para os corrigir.

No final, devem produzir um conjunto de dados sem essas inconsistências.

Devem também descrever todo o conjunto de dados e cada um dos campos do dataset. Isto é, devem descrever qual a finalidade, quais os valores possíveis, entre outras coisas que achem relevantes, para cada uma das variáveis.

Entrega

Devem ser entregues os seguintes componentes:

- Dataset devidamente tratado após o processamento.
- Um ficheiro Jupyter Notebook que deverá conter todo o código utilizado e também deverá servir de relatório, indicando todas as decisões, problemas e formas de os corrigir que foram sucedendo na realização do trabalho. Também no relatório deve constar uma descrição exaustiva do dataset. Essencialmente, o relatório deve descrever o vosso conjunto de dados com todas as métricas que considerem relevantes assim como explicar o processo de limpeza efectuado, caso seja necessário.

Os trabalhos serão apresentados individualmente. Para tal apresentação pedia que preparassem uma apresentação de no máximo 5 minutos para me explicarem o que fizeram (Podem preparar slides se preferirem, mas não é estritamente necessário).

Noto que um peso muito grande da nota no trabalho será dada pela vossa capacidade de justificar as vossas decisões: tanto no relatório como, especialmente, na apresentação onde poderão ser confrontados com algumas escolhas que fizeram.

Podem utilizar os fóruns do Moodle caso vão tendo dúvidas em algum aspecto do trabalho. Assim, tanto eu como os colegas podemos ajudar!

Data limite para entrega do trabalho: **18 de Novembro de 2024**

Data de apresentação do trabalho: **19 de Novembro de 2024 até 22 de Novembro de 2024 em hora a combinar com os alunos**