Descrição do Projeto

O objetivo deste projeto é desenvolver modelos de aprendizagem de máquina que se ajustem aos conjuntos de dados selecionados e fornecer uma solução para prever novas entradas.

O projeto deve ser realizado por grupos constituídos por dois estudante, Em caso de número impar de estudantes um grupo poderá ser constituído por três estudantes ou esse estudante poderá fazer o trabalho sozinho.

O grupo deve entregar um PowerPoint juntamente com o arquivo Jupyter Notebook criado para realizar o projecto.

O Jupyter Notebook deve descrever corretamente todas as etapas executadas. A apresentação deve começar por **descrever** o **trabalho** e **o conjunto de dados utilizado**, e continuar com as seguintes três secções para os tópicos seguintes, terminando com uma breve conclusão.

- [6,0Pomadas] Limpeza de Dados e Análise Exploratória
    - [4,0 Pontos] Carregue os dados em Python e analise se faltam dados, qual é o número e tipo de características e qual é o número de amostras. Apresente um gráfico com as distribuições (por exemplo, histograms) de duas características. Utilize uma estratégia para preencher os dados em falta, justificando as opções tomadas.
    - [2,0 Pontos] Verifique se é necessário normalizar ou padronizar os dados. Justifique a decisão.

- [8,0 Pontos] Método de Machine Learning Aplicado (por exemplo, Classificação, Regressão, Clustering, etc.)

  O desempenho do modelo implementado pelo grupo deve ser analisado no final.

    - [2,0 Pontos] Escolher o modelo de ML certo. Justifique a decisão.
    - [2,0 Pontos] Desenvolvimento do modelo
    - [2,0 Pontos] Melhorar a precisão dos modelos desenvolvidos. Descreva e justifique todas as decisões tomadas.
    - [2,0 Pontos] Apresentar os resultados de validação do modelo de aprendizagem automática desenvolvido. Por exemplo, medições de validação, matrizes de confusão e limites de decisão em formato de imagem (sugestão: http://www.tarekatwan.com/index.php/2017/12/how-to-plot-a-confusion-matrix-in-python/)

- [4,0 Pontos] Discussão dos Resultados

  Apresentar uma breve discussão dos resultados obtidos pelos diferentes modelos implementados, indicando e justificando o método que melhor se adequa aos dados apresentados (se aplicável).

- [2,0 Pontos] Apreciação global do projeto

O grupo deve escolher um tema do projeto, a partir da lista de tópicos apresentada abaixo, até **8.6.20223** e entregue via moodle até **23/06/2023** Cada tópico do projeto só pode ser selecionado por um grupo. A seleção é feita por ordem de chegada e deve ser registado no excel partilhado nesta pasta (elementos do grupo e tema do projeto)

Lista de conjuntos de dados

(escolher no excel partilhado no moodle)

1. Conjunto de dados Abalone (regressão)
   Mais informações e informações disponíveis em: https://archive.ics.uci.edu/ml/datasets/Abalone

   Conjunto de dados para ajudar a prever a idade do abalone a partir de medições físicas

2. Conjunto de dados de recomendação de cupão no veículo (Classification)
   Para mais informações e conjuntos de dados, consultar: https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation

   Conjunto de dados para estudar se uma pessoa aceitará um cupom recomendado a ela em diferentes cenários de condução.

3. Conjunto de dados do tipo de cobertura (classificação)
   Mais informações e conjunto de dados disponíveis em:
   https://archive.ics.uci.edu/ml/datasets/Covertype

4. Conjunto de dados de popularidade de notícias on-line (regressão)
   Para mais informações e conjuntos de dados, consultar:
   https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

   Este conjunto de dados resume um conjunto heterogêneo de recursos sobre artigos publicados pelo Mashable em um período de dois anos. O objetivo é prever o número de compartilhamentos nas redes sociais (popularidade).

5. Conjunto de dados do conjunto de dados do feijão seco (classificação)
   Mais informações e conjunto de dados disponíveis em:
   https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset

   Imagens de 13.611 grãos de 7 diferentes grãos secos registrados were tiradas com uma câmera de alta resolução. Um total de 16 características, 12 dimensões e 4 formas de forma, foram obtidas a partir dos grãos. O objetivo é aprender a classificar os tipos de feijão de acordo com suas características.

6. Conjunto de dados do VHC (classificação)
   Mais informações e conjunto de dados disponíveis em:
   https://archive.ics.uci.edu/ml/datasets/HCV+data

   O conjunto de dados contém valores laboratoriais dos dadores de sangue e dos doentes da hepatite C e valores demográficos como a idade. O atributo alvo para a classificação é a Categoria (dadores de sangue vs. Hepatite C (incluindo a sua evolução ("apenas" Hepatite C, Fibrose, Cirrose)).

7. Conjunto de dados de posturas manuais MoCap (clustering)
   Mais informações e conjunto de dados disponíveis em:
   https://archive.ics.uci.edu/ml/datasets/MoCap+Hand+Postures

Foram registados 5 tipos de posturas das mãos de 12 utilizadores utilizando marcadores ED não rotulados ligados aos dedos de uma luva num ambiente de captura de movimento. Devido à resolução e oclusão, valores ausentes são comuns.

Este conjunto de dados pode ser usado para uma variedade de tarefas, a mais óbvia das quais é o reconhecimento da postura através da classificação. Pode-se também tentar a identificação do usuário. Alternativamente, pode-se realizar clustering (restrito ou não) para descobrir distribuições de marcadores como uma tentativa de prever identidades de marcadores ou obter descrições estatísticas /visualizações das postagens.

8. Conjunto de dados do conjunto de dados de compartilhamento de bicicletas (regressão)

Para mais informações e conjuntos de dados, consultar:
https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Previsão da contagem de aluguel de bicicletas por hora ou dia com base nas configurações ambientais e sazonais

9. Conjunto de dados das classes de bioconcentração QSAR (regressão)

Para mais informações e conjuntos de dados, consultar:
https://archive.ics.uci.edu/ml/datasets/QSAR+Bioconcentration+classes+dataset

Conjunto de dados de bioconcentração manual (BCF, peixes) e classes mecanísticas para modelagem QSAR

--- Prever o Fator de Bioconcentração (FBC) em unidades logarítmicas (regressão)

10. Conjunto de Dados de Intenção de Compra de Compradores Online (Classificação, Clustering)

Mais informações e conjunto de dados disponíveis em:
https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

Das 12.330 sessões do conjunto de dados, 84,5% (10.422) foram amostras de classe negativa que não terminaram com compras , e o restante (1908) foram amostras de classe positiva terminando com compras.

11. Facebook Live Sellers na Tailândia Data set (Clustering)

Mais informações e conjuntos de dados disponíveis em:
https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand

Páginas no Facebook de 10 vendedores tailandeses de moda e cosméticos. Publicações de natureza diferente (vídeo, fotos, status e links). As métricas de engajamento consistem em comentários, compartilhamentos e reações.

12. Conjunto de dados de cogumelos (classificação)

Mais informações e conjunto de dados disponíveis em:
https://archive.ics.uci.edu/ml/datasets/Mushroom

Este conjunto de dados inclui descrições de amostras hipotéticas correspondentes a 23 espécies de cogumelos dourados da família Agaricus e Lepiota (pp. 500-525). Cada espécie é identificada como definitivamente comestível, definitivamente venenosa ou de comestibilidade desconhecida e não recomendada. Esta última classe foi combinada com a venenosa. O Guia afirma claramente que não existe uma regra simples para determinar a comestibilidade de um cogumelo, nenhuma regra como "folhetos três, que seja" para o Carvalho Venenoso e Ivy.

13. Conjunto de dados de mapeamento de crowdsourcing (classificação)

Mais informações e conjunto de dados disponíveis em:
https://archive.ics.uci.edu/ml/datasets/Crowdsourced+Mapping

Este conjunto de dados foi derivado de dados geoespaciais de duas fontes: 1) imagens de satélite da série temporal Landsat dos anos 2014-2015 e 2) polígonos georreferenciados crowdsourced com rótulos de cobertura da terra obtidos de OpenStreetMap. Os polígonos crowdsourced cobrem apenas uma pequena parte da área da imagem, e são usados para extrair dados de treinamento da imagem para classificar o resto da imagem. O principal desafio com o conjunto de dados é que tanto as imagens quanto os dados de crowdsourcing contêm ruído (devido à cobertura de nuvens nas imagens e rotulagem imprecisa/ digitalização de polígonos).

14. Conjunto de Dados de Qualidade do Vinho (Classificação, Regressão)

Mais informações e conjunto de dados disponíveis em:
https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Os dois conjuntos de dados estão relacionados com variantes tintas e brancas do vinho verde português. Para mais detalhes, consulte: [Web Link] ou a referência [Cortez et al., 2009]. Devido a questões de privacidade e logística, apenas estão disponíveis variáveis físico-químicas (entradas) e sensoriais (saída) (por exemplo, não existem dados sobre tipos de uva, marca de vinho , preço de venda do vinho, etc.).

Estes tasets podem ser vistos como tarefas de classificação ou regressão. As classes são ordenadas e não equilibradas (por exemplo, há muito mais vinhos normais do que excelentes ou pobres). Algoritmos de deteção de outlier podem ser usados para detetar os poucos vinhos excelentes ou pobres. Além disso, não temos certeza se todas as variáveis de entrada são relevantes. Portanto, pode ser interessante testar métodos de seleção de recursos.