Universidade do Minho
Departamento de Informática
Mestrado [Integrado] em Engenharia Informática
Mestrado em Matemática e Computação

Dados e Aprendizagem Automática
1º Ano, 1º Semestre
Ano letivo 2023/2024

Trabalho Prático de Grupo
Outubro, 2023

Tema

Conceção e otimização de modelos de *Machine Learning*.

Objetivos de Aprendizagem

Com a realização deste trabalho prático pretende-se sensibilizar e motivar os alunos para a conceção e desenvolvimento de um projeto de *Machine Learning* utilizando, entre outros, os modelos de aprendizagem abordados ao longo do semestre.

Enunciado

A energia solar é uma das principais fontes de energias renováveis, desempenhando não só um papel fundamental na transição para fontes de energia limpa e renovável, mas também na promoção da sustentabilidade ambiental. Para além de ser crucial otimizar o uso da energia solar, a relação entre o gasto e a produção energética é essencial para permitir um planeamento eficaz do consumo energético e a integração harmoniosa de sistemas de energia solar em redes elétricas existentes.

Este trabalho prático consiste no desenvolvimento de modelos de *Machine Learning* capazes de prever, com precisão, a quantidade de energia elétrica, em kWh, gerada por painéis solares e injetada na rede elétrica, a cada hora do dia, tendo por base uma grande diversidade de atributos, que vão desde dados meteorológicos e informações geográficas, a históricos de gasto e produção energética elétrica. Este é um problema de previsão de energia com impacto significativo na eficiência energética, mas também na redução das emissões de gases com efeito estufa e na promoção da sustentabilidade. Com isto em consideração, foi colecionado um *dataset* que contém dados referentes à produção energética de determinados painéis solares na cidade de Braga (o *dataset* cobre um período que vai desde setembro de 2021 até abril de 2023).

Este enunciado prático engloba 2 TAREFAS.

TAREFA DATASET GRUPO:

É para fazer as 2 tarefas ao mesmo tempo

- Consultar, analisar e selecionar um *dataset* de entre os que estão acessíveis a partir de fontes como, por exemplo, o ***Google Dataset Search*** ou ***Kaggle***; O objetivo é cada grupo ter um dataset diferente
- Explorar, analisar e preparar o *dataset* selecionado, procurando extrair conhecimento relevante no contexto do problema em questão; Não escolher datasets como acertar em euromilhões - não há padrões e as bolas são aleatórias.
- Conceção e otimização de múltiplos modelos de *Machine Learning*;
- Obtenção e análise crítica de resultados.

O dataset de teste é um dataset mais curto do que o que vai ser usado para calcular a classificação final. É o dataset completo daquilo que vamos estar a usar, o mais pequeno. Os resultados que nós vemos no ranking enquanto vamos fazendo as submissões do teste são só para aquele conjunto pequeno de dados. -> O que é q isto significa? Que quando os stores usarem o nosso modelo no dataset completo para saber a classificação, os resultados que obtiamos com o dataset mais pequeno vai mudar. Se nós fizermos um modelo muito bom, com tipo 98% para o dataset pequeno, isso quer dizer que o nosso modelo está muito preso aos dados do dataset pequeno, e isso pode (ou não) significar que o nosso modelo pode não ser mt compativel com o dataset maior, e que a classificação não vai ser tão boa

TAREFA DATASET COMPETIÇÃO:

- Para além do *dataset* selecionado na tarefa anterior, os grupos deverão trabalhar o *dataset* disponível em **https://www.kaggle.com/c/daasbstp2023**: 1 elemento cria o grupo e depois envia o convite para nós nos registarmos no grupo
    - O *link* anterior redireciona para a plataforma *Kaggle* onde foi criada uma competição. O *dataset* a utilizar na competição, assim como todos os detalhes e funcionamento da mesma, estão disponíveis no referido *link*;

- O primeiro passo consiste em aceder à plataforma *Kaggle*, utilizando o seguinte *link* para se inscreverem na competição:
  https://www.kaggle.com/t/f0810933a3bc4182a66aae2ad32f6872

  Devem, de seguida, formar equipas com os restantes elementos do grupo de trabalho. O nome da equipa deverá seguir o formato ***GRUPO_<CURSO>_<X>*** onde ***<CURSO>*** corresponde ao curso de mestrado (MMC, MEI ou MIEI) e ***<X>*** ao número do grupo. Não poderão efetuar submissões na plataforma *Kaggle* enquanto o grupo se apresentar incompleto.

- Explorar, analisar e preparar o *dataset* da competição, procurando extrair conhecimento relevante no contexto do problema em questão;
- Conceção e otimização de modelos de *Machine Learning* para o *dataset* da competição:
  - Deverão submeter os resultados obtidos na plataforma *Kaggle* de forma a obter a *accuracy* do modelo;
  - Existe um **limite diário de 3 submissões válidas** pelo que deverão procurar começar as submissões assim que possível. A competição encerra no final do dia **08 de janeiro de 2024**.
- Obtenção e análise crítica de resultados;
- Interpretação dos resultados adquiridos e definição da sua utilidade no contexto do problema subjacente ao *dataset* trabalhado. Determinar e explicitar os resultados mais relevantes.

Entrega e Avaliação

1 relatório onde se descrevem as duas tarefas

Os resultados obtidos deverão ser objeto de 1 relatório, limitado a 20 páginas, que apresente, entre outros:

- Quais os domínios a tratar, quais os objetivos e como se propõe a atingi-los;
- Qual a metodologia seguida e como foi aplicada;
- Descrição e exploração detalhada de ambos os *datasets* e de todo e qualquer tratamento efetuado;
- Descrição dos modelos desenvolvidos, quais as suas características, como e sobre que parâmetros foi realizado o *tuning* do modelo, características do treino, entre outros detalhes que seja oportuno fornecer;
- Sumário dos resultados obtidos e respetiva análise crítica;
- Apresentação de sugestões e recomendações após análise dos resultados obtidos e dos modelos desenvolvidos.

Todo o processo deverá ser acompanhado de exemplos e indicações que permitam reproduzir todos os passos realizados assim como os resultados obtidos.

Durante o período de aulas do dia **23 de novembro de 2023** decorrerá a avaliação da TAREFA DATASET GRUPO da componente prática de avaliação em grupo. No referido dia será feito um checkpoint ao trabalho desenvolvido pelos grupos de trabalho, devendo cada grupo utilizar os meios que considerar mais adequados para demonstrar os resultados obtidos.

Nos dias **11 e 12 de janeiro de 2024** decorrerão as sessões de apresentação do trabalho desenvolvido em ambas as TAREFAS. Os grupos de trabalho deverão escolher o *slot* desejado para realização da apresentação, sendo que esses *slots* serão disponibilizados nas próximas semanas. Cada grupo disporá de 10 minutos para realizar a apresentação, utilizando os meios que considerar mais adequados.

O relatório, assim como os restantes elementos produzidos, deverão ser compactados num único ficheiro zip que deverá ser submetido, por um elemento do grupo, até ao dia **10 de janeiro de 2024** na plataforma de e-learning da Universidade do Minho (em "*Conteúdo/Instrumentos de Avaliação em Grupo/Submissão TPG*").

Avaliação por pares

Cada grupo deverá realizar uma análise coletiva sobre o contributo e esforço que cada elemento deu para o avanço do trabalho. Dessa análise devem conseguir identificar os membros que trabalharam acima, na e abaixo da média. Para esta componente de avaliação está previsto 1 valor para cada aluno que reflete a sua contribuição individual no desenvolvimento deste instrumento de avaliação.

Assim, um elemento do grupo deverá enviar um email, colocando em CC os restantes elementos do grupo, para valves@di.uminho.pt, analide@di.uminho.pt, filipa.ferraz@di.uminho.pt, dad@di.uminho.pt e bruno.fernandes@algoritmi.uminho.pt. O assunto deverá ser "**AP DAA - Avaliação por pares**".

No texto do email deverão indicar, para cada elemento do grupo, o respetivo delta (parcela a somar à nota desta componente). Lembra-se que os deltas podem ser negativos, nulos ou positivos e que, em cada grupo, o somatório dos deltas deve ser sempre igual a 0.00 e, individualmente, nunca podem ultrapassar a unidade.

Exemplo 1 (corresponde a um esforço igual entre todos):

PG1234 João DELTA = 0
PG5678 António DELTA = 0
PG9123 Maria DELTA = 0
PG4567 Rita DELTA = 0

Exemplo 2 (o António recebe 1 valor adicional, a Rita mantém a classificação, ao João e à Maria são descontados 0.5 valores a cada):

PG1234 João DELTA = -0.5
PG5678 António DELTA = 1
PG9123 Maria DELTA = -0.5
PG4567 Rita DELTA = 0

Código de Conduta

Os participantes do presente trabalho académico declaram ter atuado com integridade e confirmam que não recorreram à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração. Mais declaram que conhecem e respeitaram o Código de Conduta Ética da Universidade do Minho.

Referências Bibliográficas

Além do material disponibilizado nas aulas, aconselha-se a consulta de fontes como:

- Machine Learning. T. Michell, McGraw Hill, ISBN ISBN: 978-1259096952, 2017.
- Introduction to Machine Learning. Alpaydin, E. ISBN: 978-0-262-02818-9. Published by The MIT Press, 2014.
- Computational Intelligence: An Introduction. Engelbrecht A., Wiley & Sons. 2nd Edition, ISBN: 978-0470035610, 2007.
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Hastie, T., R. Tibshirani, J. Friedman, 12nd Edition, Springer, ISBN: 978-0387848570, 2016.
- Machine Learning: A Probabilistic Perspective. K.P. Murphy, 4th Edition, The MIT Press, ISBN: 978-0262018029, 2012.