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 2022/2023

Trabalho Prático de Grupo
Outubro, 2022

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 modelação de incidentes rodoviários é um conhecido problema de características estocásticas, não-lineares. Tem, contudo, aparecido na literatura um conjunto de modelos que demonstram um potencial assinalável neste tipo de previsões. Com isso em consideração, foi construído um *dataset* que contém dados referentes à quantidade e características dos incidentes rodoviários que ocorreram numa cidade portuguesa em 2021. O objetivo deste trabalho passa por, entre outros, desenvolver modelos de *Machine Learning* capazes de prever o nível de incidentes rodoviários, numa determinada hora, na referida cidade.

Este enunciado prático engloba 2 TAREFAS.

TAREFA DATASET GRUPO:

- Consultar, analisar e selecionar um *dataset* de entre os que estão acessíveis a partir de fontes externas como, por exemplo, o *Google Dataset Search* ou *Kaggle*;
- Explorar, analisar e preparar o *dataset* selecionado, procurando extrair conhecimento relevante no contexto do problema em questão;
- Conceção e otimização de múltiplos modelos de *Machine Learning*;
- Obtenção e análise crítica de resultados.

TAREFA DATASET COMPETIÇÃO:

- Para além do *dataset* selecionado na tarefa anterior, os grupos deverão trabalhar o *dataset* disponível em **https://www.kaggle.com/c/sbstpdaa2223**:
    - 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/b8750c6af7ff42b9bec09ada72aa30b7**
    - 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. **Os grupos de trabalho deverão ser constituídos por, no máximo, 4 elementos**.

- 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 dia 07 de janeiro de 2023.
- 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

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 propõem 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 24 de novembro de 2022 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. Este checkpoint decorrerá de forma remota, sendo utilizada a ferramenta *Blackboard Collaborate Ultra* da plataforma de e-learning da Universidade do Minho.

Na semana de 16-21 de janeiro de 2023 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, 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, d7266@di.uminho.pt, d7646@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.

Exemplo 1 (todos recebem 1 valor, correspondendo 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 2 valores, a Rita recebe 1 valor, e o João e a Maria recebem 0.5 valores):

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.