Trabalho Prático

Análise de Dados em Informática

Análise de Desempenho De Técnicas de Aprendizagem Automática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2024/2025

---

1. Objetivos

1.1. Objetivo Geral:

- Análise de Desempenho de técnicas de aprendizagem automática

1.2. Objetivos Específicos:

- Definir a metodologia de trabalho
- Análise e Discussão dos Resultados com recurso ao Python
- Escrita de artigo científico

2. Calendarização

Entrega do trabalho: até **15 de junho de 2025 pelas 23:59**
Defesa e discussão: em data a marcar pelo Professor de TP

3. Normas

- Deverá resolver as tarefas propostas usando o Python.
- A **data final de ENTREGA** do trabalho é dia **15 de junho de 2025 pelas 23:59,** no moodle. Independentemente destes prazos, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.
- A entrega do trabalho consta de um artigo científico (**máx. 8 páginas**) conforme *template* disponibilizado no moodle, apresentação *powerpoint* com resumo do trabalho realizado, entre outros. Deverá submeter todos os documentos num ficheiro compactado. O zip file deve conter:
    - artigo científico em pdf
    - dados utilizados em formato csv
    - Notebook completo (e comentado) do código criado em Python para resolver as tarefas propostas
    - apresentação PowerPoint com resumo do artigo para 10 minutos (ppt)
- O nome do ficheiro zip deverá seguir a seguinte notação:
  ANADI_YYY_XXX_Nºaluno1_Nºaluno2_Nºaluno3.zip, onde **YYY** representa a sigla do docente das TP, e **XXX** representa a turma TP.

  Exemplo: **ANADI_EFG_3AD3BD_7777777_8888888_9999999.zip.**
- Trabalhos cujo nome não respeite a notação indicada **serão penalizados em 10%.**
- **A entrega do trabalho deverá ser submetida no moodle até à data de entrega definida**. **Não serão aceites trabalhos fora do prazo**.
- A apresentação, **em formato de comunicação (10 minutos),** e discussão dos trabalhos decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes e apresentar uma das componentes do trabalho realizado e sistematizado na apresentação **ppt**. **Serão colocadas questões sobre o trabalho para as quais todos os elementos do grupo deverão responder.** Os elementos ausentes ou que não sigam as orientações definidas para a realização da apresentação/defesa não terão classificação.

- A avaliação do trabalho será realizada pelo docente das aulas teórico-práticas (TP).
- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas TP.
- Código de conduta: (cf. Regulamento Disciplinar dos Estudantes do IPP)
    - Nenhum estudante ou grupo pode assumir pertença de trabalho realizado por outrem ou desenvolvido em conluio.
    - É expressamente proibido o uso de materiais, artefactos ou código de outrem sem a devida, e explícita indicação de origem.
    - Código de outras fontes deve ser claramente identificado no próprio código, indicando a fonte.
    - Casos de apropriação ilícita de materiais, artefactos e/ou código, sujeito a avaliação, serão reportados à Presidência do ISEP.
    - A utilização de ferramentas com IA de assistência à codificação/desenho (e.g. chatGPT) deve ser mencionada
- É obrigatório o uso da ferramenta de controle de versões GitHub. Devem partilhar o repositório com os vossos professores de TP's.

3.1. Artigo Científico

No Artigo Científico (máx. 8 páginas) deverão ser documentadas todas as fases da metodologia de trabalho seguida, contextualização do tema, exploração, preparação dos dados, análise e discussão dos resultados e conclusões.

Deve ser seguido o *template* IEEE disponibilizado no moodle (Word ou Latex).

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos:

- Breve revisão do estado da arte (algoritmos de aprendizagem automática e análise de desempenho);
- Desenvolvimento de modelos de Aprendizagem Automática;
- A qualidade do processo de análise de dados seguido, a organização do código, a avaliação dos modelos criados, análise e discussão dos resultados e as conclusões alcançadas;
- Organização, qualidade da escrita, apresentação e clareza do artigo científico;
- A comunicação e discussão;
- Participação individual de cada um dos elementos em %.

Contextualização (Abstract, Introdução (motivação, objetivos e metodologia seguida))	**2 valores**
Análise de desempenho de técnicas de aprendizagem automática (código Python – 30%; artigo científico (análise exploratória de dados, definição e avaliação dos modelos, análise e discussão dos resultados) – 70%)	**12 valores**
Conclusões	**2 valores**
Apresentação e Discussão	**4 valores**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua % de participação. No momento da defesa do trabalho será validada a participação de cada um dos elementos do grupo, na concretização dos objetivos do trabalho e do grupo.

4. Descrição do Trabalho

O objetivo principal deste trabalho consiste na aplicação de algoritmos de aprendizagem automática na exploração de dados e respetiva comparação usando os testes estatísticos mais adequados. Deve ser produzido um artigo científico (português ou inglês), conforme *template* indicado, com o estado da arte sobre os diferentes algoritmos, os modelos desenvolvidos, os resultados obtidos, a análise e discussão dos resultados e as conclusões gerais do trabalho (síntese das conclusões).

A poluição ambiental é um dos maiores problemas deste século, estando na origem de enormes danos para o ecossistema biológico, condições meteorológicas, saúde humana e organismos vivos. De acordo com o relatório da Organização Mundial de Saúde, nove em cada dez pessoas respiram ar poluído, causando aproximadamente sete milhões de mortes em todo o mundo todos os anos (segundo a OMS). Na Europa estão identificados alguns dos principais poluentes atmosféricos: amoníaco (**NH_3**), dióxido de azoto (**NO_2**), partículas em suspensão com uma componente aerodinâmica de diâmetro inferior a 2.5µm (**PM2.5**) e ozono troposférico (**O_3**). Nas cidades, onde vive 74% da população da União Europeia, os níveis de **PM2.5** e O_3 têm efeitos nocivos na saúde humana associados a doenças respiratórias e doenças cardiovasculares. A poluição atmosférica também danifica ecossistemas, sendo **O_3** considerado o poluente atmosférico mais prejudicial á vegetação e biodiversidade.
Este trabalho tem como objetivo analisar os níveis de poluentes em diferentes zonas geográficas de países europeus. Pretende-se validar uma possível interação entre os níveis de poluentes e a ocorrência de doenças ou o nº de mortes prematuras.

Para a realização desta análise deve utilizar os dados disponíveis no ficheiro ***AIRPOL_data.csv*** contém dados de níveis dos poluentes atmosféricos (**PM2.5**, **NO_2**, **O_3**)

de várias regiões geográficas europeias, relativos ao ano de 2022**.** Informação sobre os rótulos das colunas encontram-se descritas na tabela seguinte.

Country	**Pais**
NUTS_Code	Código da região do país
Air_Pollutant	Poluente atmosférico (PM2.5, O3 e NO2)
Disease	Doença associada ao poluente
Affected_Population	População afetada pelo poluente
Populated_Area[km2]	Área populacional em km2
Air_Pollution_Average[ug/m3]	Nível médio de poluição do poluente em ug/m3
Premature_Deaths	Número de mortes prematuras associadas ao poluente

No âmbito da 2ª iteração do Trabalho Prático, pretende-se a realização da análise exploratória dos dados disponibilizados, desenvolvendo modelos de regressão e classificação que foram estudados na unidade curricular, ao longo do semestre: regressão linear árvores de decisão, k-vizinhos-mais-próximos, redes neuronais e SVM.

4.1. Análise Exploratória de Dados

1. Comece por carregar o ficheiro "**`AIRPOL_data.csv`**" verifique a sua dimensão e obtenha um sumário dos dados.
2. Faça a exploração dos dados através dos gráficos mais apropriados.
3. Realize o pré-processamento dos dados.
4. Agrupe os países em quatro regiões:
    - Western Europe: Austria, Belgium, France, Germany, Netherlands, Switzerland
    - Eastern Europe: Poland, Czechia, Hungary
    - Southern Europe: Greece, Spain, Italy, Portugal
    - Northern Europe: 'Sweden, Denmark, Northern Europe, Finland

4.2. Regressão

Considere apenas os países `Southern Europe`:

1. Crie um diagrama de correlação entre a variável `Premature_Deaths` e os restantes atributos e interprete.
2. Usando o método ***k-fold cross validation*** obtenha um modelo regressão linear simples para a variável *`Premature_Deaths`* usando a variável `Affected_Population`
   a) Apresente a função linear resultante
   b) Visualize a reta correspondente ao modelo de regressão linear simples e o respetivo diagrama de dispersão.
   c) Calcule o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) do modelo
3. Usando o método ***k-fold cross validation*** desenvolva modelos para prever `Premature_Deaths` usando:
   a) Regressão linear múltipla.
   b) Árvore de regressão. Otimize os parâmetros do modelo. Apresente a árvore de regressão obtida.
   c) SVM. Otimize o kernel.
   d) Rede neuronal. Otimize a configuração da rede.
4. Compare os resultados obtidos pelos modelos referidos na questão 5, usando o erro *médio absoluto (MAE) e a raiz quadrada do erro médio (RMSE).*
5. Justifique se os resultados obtidos para os dois melhores modelos são estatisticamente significativos (para um nível de significância de 5%) e identifique o(s) modelo(s) com melhor desempenho.

4.3. Classificação

Considere os países das quatro regiões anteriormente definidas.

1. Derive um novo atributo `RespDisease` que separa as doenças em respiratórias ('Asthma' 'Chronic obstructive pulmonary disease') e não respiratórias.
2. Usando o método ***k-fold cross validation*** desenvolva modelos de previsão de `RespDisease` usando os seguintes métodos:
   a) Árvore de decisão. Otimize os parâmetros do modelo.
   b) Rede neuronal. Otimize a configuração da rede.
   c) SVM. Otimize o kernel.
   d) K-vizinhos-mais-próximos. Otimize o parâmetro K.

3. Obtenha a média e o desvio padrão da *Accuracy*; Sensitivity; *Specificity* e *F1* do atributo `RespDisease` com os modelos obtidos na alínea anterior.
4. Verifique se existe diferença significativa no desempenho dos dois melhores modelos obtidos anteriormente (use um nível de significância de 5%). Identifique o modelo que apresenta o melhor desempenho.
5. Compare os resultados dos modelos. Discuta em detalhe qual o modelo que apresentou melhor e pior desempenho de acordo com os critérios: *Accuracy*; *Sensitivity*; *Specificity* e *F1*.

Ter em consideração que em todas as questões devem ser justificados os pressupostos assumidos, e os resultados devem ser interpretados e analisados. O artigo científico deve incluir a descrição de todos os modelos desenvolvidos, decisões assumidas na parametrização e a análise e interpretação dos resultados.

Referências Bibliográficas

- Christopher Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
- Tom Mitchell, Machine Learning. McGraw-Hill, 1997.