Trabalho Prático

Análise de Dados em Informática

Análise de Desempenho

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2022/2023

---

1. Objetivos

Objetivo Geral:

- Análise de Desempenho de técnicas de aprendizagem automática

Objetivos Específicos:

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

2. Calendarização

Entrega do trabalho: até **18 de junho de 2023 pelas 23:59**
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- Deverá ser usada a ferramenta R.
- A **data final de ENTREGA** do trabalho é **18 de junho de 2023 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
  - script completo (e comentado) do código criado em R para resolver o problema
  - 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_AMD_3AD_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**. A defesa e apresentação da comunicação **poderá** ser realizada numa plataforma de vídeo conferência (Zoom ou MSTeams) e todos os elementos do grupo devem ter a câmara e microfones ligados. 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 Bitbucket. 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 (código R – 40%, artigo científico (definição e avaliação dos modelos, análise e discussão dos resultados**)** – 60%)	**14 valores**
Conclusões	**2 valores**
Apresentação e Discussão	**2 valores**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua % participação. No momento da defesa do trabalho (que poderá ser via videoconferência), 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).

Foi realizada uma recolha de informação em ciclistas profissionais. Este conjunto de dados inclui elementos relativos aos treinos no período de pré-temporada. Pretende-se validar possíveis relações entre os dados recolhidos e os medições obtidas para diferentes métricas indicadoras do desempenho do atleta.

São disponibilizados dados estruturados referentes à preparação de ciclistas para uma nova temporada de competições, após o seu treino de inverno. O d*ataSet* consiste em *1000* observações de *11* variáveis.

ID	**Gender**	**Team**
Background	**Pro_level**	**Winter Training Camp**
Altitude_results	**vo2_results**	**hr_results**
dob	**Continent**	

As variáveis são, genericamente, autoexplicativas, havendo, no entanto, a necessidade de clarificar alguns atributos mais específicos deste domínio:

- **Background** – Tipo de perfil do ciclista
- **Pro_Leve**l – Nível de competição do ciclista
- **Altitude_results** – resultado do treino de altitude
- **vo2_results** – resultado do teste de volume de oxigênio máximo
- **hr_results** – resultado do teste de frequência cardíaca e **hr** results
- **dob** – data de nascimento

No âmbito da 2ª iteração do Trabalho Prático, pretende-se que realizem a análise dos dados da preparação de ciclistas para a nova temporada de competições através de modelos de classificação/regressão usando os algoritmos de aprendizagem automática estudados: regressão linear, árvores de decisão, k-vizinhos-mais-próximos e redes neuronais.
Deve ser usado o ficheiro "**ciclismo.csv**", disponível no moodle.

4.1. *Regressão*

1. Comece por carregar o ficheiro ("**ciclismo.csv**") para o ambiente do R, verifique a sua dimensão e obtenha um sumário dos dados.
2. Derive um novo atributo **Age** usando como valor do atributo **dob**

3. Analise os atributos do conjunto de dados mais significativos, usando gráficos, análises estatísticas e/ou outros métodos apropriados.

4. Realize o pré-processamento dos dados:
   a) Faça a identificação de NA e limpe o *dataSet*, se aplicável
   b) Identifique dados inconsistentes e *outliers*, se aplicável
   c) Implemente a seleção de atributos, se aplicável
   d) Implemente a normalização dos dados, se necessário

5. Crie um diagrama de correlação entre todos os atributos. Comente o que observa.

6. Obtenha um modelo de regressão linear simples para determinar a variável "**Altitude_results** " usando o valor relativo à componente dos resultados de frequência cardíaca ("**hr_results**"):
   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 sobre os 30% casos de teste.
   d) Teste se é possível obter resultados melhores utilizando um modelo mais complexo.

7. Tendo em conta o conjunto de dados apresentado, pretende-se prever o atributo "**vo2_results**" relativo ao resultado do teste de volume de oxigênio máximo, aplicando:
   a) Regressão linear múltipla.
   b) Árvore de regressão, usando a função *rpart*. Apresente a árvore de regressão obtida.
   c) Rede neuronal usando a função *neuralnet*, fazendo variar os parâmetros. Apresente a rede obtida.

8. Compare os resultados obtidos pelos modelos referidos na questão 7, usando o erro médio absoluto (MAE) e a raiz quadrada do erro médio (RMSE).

9. Justifique se os resultados obtidos para os dois melhores modelos são estatisticamente significativos (para um nível de significância de 5%). Identifique o modelo que apresenta o melhor desempenho.

4.2. *Classificação*

1. Estude a capacidade preditiva relativamente ao atributo "**Pro_level**" usando os seguintes métodos:
   - árvore de decisão;
   - rede neuronal;
   - K-vizinhos-mais-próximos.

a) Usando o método ***k-fold cross validation*** obtenha a média e o desvio padrão da taxa de acerto da previsão do atributo "**Pro_level**" com os dois melhores modelos obtidos na alínea anterior.

b) Dos três modelos, um é conhecido por ter uma forma de aprendizagem conhecida como "**Lazy Learning**", identifique o modelo e as implicações deste tipo de modelos.

c) 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.

d) 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.

2. Estude a capacidade preditiva relativamente ao atributo "**Winter_training_camp**" usando os seguintes métodos:

- árvore de decisão;
- rede neuronal;

a) Usando o método ***k-fold cross validation*** obtenha a média e o desvio padrão da taxa de acerto da previsão do atributo "**Winter_training_camp**" com os dois melhores modelos obtidos na alínea anterior.

b) Verifique se existe diferença significativa no desempenho dos dois melhores modelos obtidos anteriormente (use um nível de significância de 5%).

c) Compare os resultados dos modelos. Identifique o modelo que apresenta o melhor desempenho, de acordo com os critérios: Accuracy; Sensitivity; Specificity e F1.

3. Estude a capacidade preditiva relativamente ao atributo "**Gender**" usando os seguintes métodos:

- rede neuronal;
- K-vizinhos-mais-próximos.

a) Usando o método ***k-fold cross validation*** obtenha a média e o desvio padrão da taxa de acerto da previsão do atributo "**Gender**" com os dois melhores modelos obtidos na alínea anterior.

b) Verifique se existe diferença significativa no desempenho dos dois melhores modelos obtidos anteriormente (use um nível de significância de 5%).

c) Compare os resultados dos modelos. Identifique o modelo que apresenta o melhor 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.

5. Referências Bibliográficas

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