Trabalho Prático

Análise de Dados em Informática

Análise de Desempenho De Técnicas de Aprendizagem Automática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2023/2024

---

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 **9 de junho de 2023 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 **9 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
    - 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**. 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 automática (código Python – 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 % 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).

Os dados consistem na estimativa dos níveis de obesidade em pessoas com idades entre os 14 e os 61 anos e diversos hábitos alimentares e condições físicas. Foram recolhidos dados de 2111 pessoas, nos quais foram obtidos 17 atributos relacionados com hábitos alimentares e condição física dos participantes.

Os atributos relacionados com os hábitos alimentares são:

- FCCAC - Frequência de Consumo de Comida Altamente Calórica
- FCV - Frequência de Consumo de Vegetais
- NRP - Número de Refeições Principais
- CCER - Consumo de Comida Entre Refeições
- CA - Consumo de Água
- CBA - Consumo de Bebidas Alcoólicas
- MCC - Monitorização do Consumo Calorias
- Histórico de Obesidade Familiar

Os atributos relacionados com a condição física são:

- Género
- Idade
- Peso
- Altura
- FAF - Frequência de Atividade Física
- TUDE - Tempo de Utilização de Dispositivos Eletrónicos
- Fumador
- TRANS - Transporte utilizado

O atributo *Label* refere-se à categoria de risco de obesidade de cada indivíduo.

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

4.1. Regressão

1. Comece por carregar o ficheiro "`Dados_Trabalho_TP2.csv`" para o ambiente do Python, verifique a sua dimensão e obtenha um sumário dos dados.
2. Derive um novo atributo, "IMC" usando a informação dos atributos do Peso e Altura.
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 e comente o que observa.
6. Obtenha um modelo de regressão linear simples para a variável "IMC" usando o atributo relativo à "Idade" de cada registo:
   - **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 20% casos de teste.

d) Teste se é possível obter um modelo de regressão linear simples com melhor resultado, utilizando outra variável dos preditores disponíveis no *dataset*.

7. Tendo em conta o conjunto de dados apresentado, pretende-se prever o atributo "IMC", aplicando os seguintes modelos (para os modelos sugeridos, o conjunto de atributos a utilizar pode variar):

a) Regressão linear múltipla.

b) Árvore de regressão, usando a função ***DecisionTreeRegressor***. Apresente a árvore de regressão obtida.

c) Rede neuronal usando a função ***MLPRegressor***, fazendo variar os parâmetros e arquitetura do modelo. 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. Utilizando o *dataset* fornecido, obtenha modelos de classificação para prever o risco de obesidade de cada indivíduo. Os modelos desenvolvidos devem ser avaliados utilizando o método de ***k-fold cross validation***. Devem apresentar como resultados a média e desvio padrão da métrica que considerem mais adequada ao problema em análise. Os métodos a utilizar são os seguintes:

- **Árvores de Decisão:** Utilizando a função ***DecisionTreeClassifier*** desenvolva um modelo de classificação que responda ao problema proposto. Deve fazer o ajuste dos parâmetros do modelo, de forma a garantir que não está a ocorrer ***overfitting*** nos dados de treino;
- **SVM:** Utilizando a função ***SVC*** desenvolva um modelo de classificação que responda ao problema proposto. Deve testar todos os *Kernels* possíveis, e fazer ajuste de parâmetros, de forma a garantir que não está a ocorrer ***overfitting*** nos dados de treino. No final, deve identificar o modelo com o melhor desempenho, justificando;

- **Rede Neuronal:** Utilizando o ***package*** do ***Keras***, deve desenvolver um modelo de classificação que responda ao problema proposto. Deve fazer uma otimização da arquitetura e dos parâmetros do modelo. No final deve apresentar a arquitetura encontrada, e os parâmetros escolhidos, tentando justificar as escolhas efetuadas durante o processo de otimização;
- **K-vizinhos-mais-próximos:** Utilizando a função ***KNeighborsClassifier*** desenvolva um modelo de classificação que responda ao problema proposto. Deve fazer o ajuste dos parâmetros do modelo, de forma a otimizar o seu desempenho.

a) 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.

b) Compare os resultados dos modelos anteriores. Discuta em detalhe qual o modelo que apresentou melhor e pior desempenho de acordo com as seguintes métricas: ***Accuracy***; ***Sensitivity***; ***Specificity*** e ***F1***.

c) Seleção de atributos: Os modelos de classificação desenvolvidos podem obter um melhor desempenho se utilizarem apenas alguns dos atributos disponíveis? Avalie esta hipótese realizando uma seleção de atributos, justificando as escolhas realizadas. Pode basear-se no diagrama de correlação obtido anteriormente, para justificar as escolhas efetuadas.

2. Novos preditores: A partir dos preditores que compõem o *dataset* disponibilizado, derive novos preditores, que considere que possam ser úteis para utilizar em modelos de classificação. Avalie se com a utilização dos novos preditores, nos dois melhores modelos obtidos anteriormente, existe diferença significativa no desempenho (use um nível de significância de 5%).

3. Estude a capacidade preditiva relativamente ao atributo "Genero" usando os métodos:

- Rede Neuronal;
- SVM

a) Usando o método ***k-fold cross validation*** obtenha a média e o desvio padrão da taxa de acerto da previsão do atributo "Genero" 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.

Referências Bibliográficas

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