Trabalho Prático

Iteração 2

Análise de Dados em Informática

Técnicas de Aprendizagem Automática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2020/2021

---

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 **20 de junho de 2021 pelas 23:55**
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- Deverá ser usado a ferramenta R.

- A **data final de ENTREGA** do trabalho é **20 de junho de 2021 pelas 23:55**, 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 presencialmente ou por videoconferência. Nesta última situação 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.

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 *Machine Learning*;
- A qualidade do processo de análise de dados seguido, a organização do código, a avaliação dos modelos criados, a 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) e Revisão da Literatura)	**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 % participação identificada. 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 Direção-Geral da Saúde e outras entidades a nível internacional, têm elaborado relatórios e estudos com o objetivo de informar os decisores e técnicos das entidades de saúde diretamente envolvidos na gestão da pandemia, e no planeamento das medidas de mitigação. Parte desta informação, é disponibilizada, semanalmente, num relatório de situação sobre a curva epidémica e os parâmetros de transmissibilidade da infeção por SARS-CoV-2. Uma das ferramentas usadas é a matriz de risco que relaciona a taxa de Transmissibilidade R(t) com a Incidência.

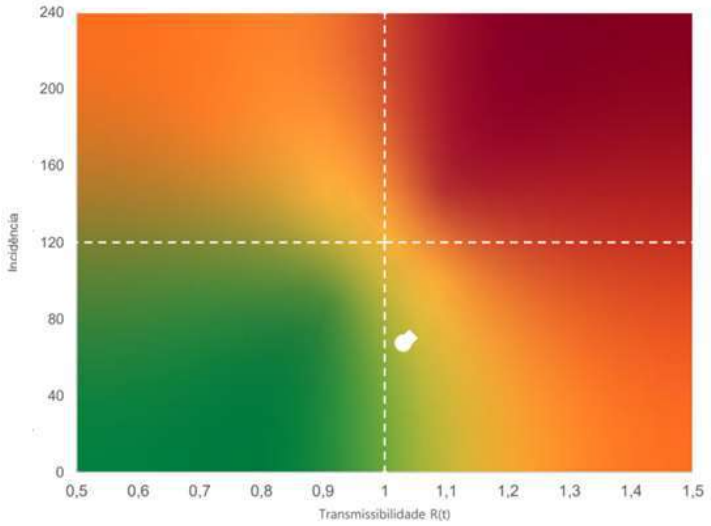

Figura 1 - Matriz de Risco

No âmbito da 2ªiteração do Trabalho Prático, pretende-se que realizem a análise de vários indicadores sobre os dados da pandemia COVID-19 a nível mundial, através de modelos de classificação/regressão usando os algoritmos de aprendizagem automática estudados: regressão linear, árvores de regressão/decisão, k-vizinhos-mais-próximos e redes neuronais.
É usado o ficheiro ”**countryagregatedata.xlsx**”, disponível no moodle, agrupados por país, num dado período. O ficheiro contém dados do *dataset* facultado com dados reais, retirado da base dados internacional "*Our World in Data*” [3], dinamizada pela Universidade *Johns Hopkins University (JHU)*.

O conjunto de dados a analisar neste trabalho diz respeito a vários indicadores (nº total de casos, nº médio de novos casos, nº total de mortos, taxa de Transmissibilidade R(t) e outros relativos a cada país).

4.1. *Regressão*

1. Comece por carregar o ficheiro ("**countryagregatedata.xlsx**") para o ambiente do R, verifique a sua dimensão e obtenha um sumário dos dados.
2. Crie um diagrama de correlação entre todos os atributos e comente o que se observa.
3. Obtenha um modelo de regressão linear simples para a variável objetivo para determinar o "*total_deaths*" usando o número de novos casos ("*new_cases*")
   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.
4. Tendo em conta o conjunto de dados apresentado, pretende-se prever a esperança de vida ('*life_expectancy*'), 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.

   Compare os resultados obtidos pelos modelos referidos na questão 4, usando o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE). 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

5. Derive um novo atributo **NiveldeRisco**, discretizando o atributo "stringency_index" - em 2 classes: "low" e "high" usando como valor de corte a média do atributo.

6. Estude a capacidade preditiva relativamente a este novo atributo **NiveldeRisco** usando os seguintes métodos:
   a) árvore de decisão;
   b) rede neuronal;
   c) K-vizinhos-mais-próximos.

   Usando o método ***k-fold cross validation*** obtenha a média e o desvio padrão da taxa de acerto da previsão do atributo **NiveldeRisco** com os dois melhores modelos. 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.

7. Derive um novo atributo **ClassedeRisco**, discretizando o atributo "reproduction_rate" - **Taxa de Transmissibilidade R(t)** e o atributo "incidence" - **Incidência** em 3 classes (conforme Figura 1): "Vermelho", "Amarelo" e "Verde". Por simplificação considera-se que a Incidência corresponde à razão entre o atributo "total_cases" e "population" multiplicado por 100.000 habitantes.

8. Avalie a capacidade preditiva relativamente a este novo atributo **ClassedeRisco** usando os métodos:
   a) árvore de decisão
   b) rede neuronal
   c) K-vizinhos-mais-próximos

   Compare os resultados dos modelos anteriores. Discuta em detalhe qual o modelo que apresentou o melhor e o 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.
Na secção de conclusões do artigo devem ser sintetizadas e sistematizadas as conclusões de cada secção: Regressão e Classificação.

5. Referências Bibliográficas

[1]. Christopher Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[2]. Tom Mitchell, Machine Learning. McGraw-Hill, 1997.
[3]. Coronavirus Source Data - Our World in Data, https://ourworldindata.org/coronavirus-source-data