Universidade do Minho
Departamento de Informática
Mestrado [integrado] em Engenharia Informática

Nº _____ CURSO _____

NOME _____

_____

Dados e Aprendizagem Automática
1º Ano, 1º Semestre
Edição 2022/2023

Prova escrita, 5 de janeiro, 2023

GRUPO 1
(4 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO NO ESPAÇO RESERVADO PREENCHENDO OS ESPAÇOS VAZIOS COM AS EXPRESSÕES CORRETAS.

QUESTÃO 1

Numa metodologia de análise de dados como o **CRISP-DM**, a preparação de dados é uma tarefa anterior à **Modelação** e é preponderante visto que os dados recolhidos do mundo real podem apresentar-se **Com ruido**.

QUESTÃO 2

Algoritmos de *Clustering*, tais como **K-Means** e **K-Medoids**, implementam uma técnica de aprendizagem **Nao supervisionada** com o objetivo de agrupar um conjunto de casos de estudo, de tal forma que os objetos no mesmo grupo apresentam mais **semelhanças** entre si do que com outros grupos.

QUESTÃO 3

Feature Engineering permite a criação de **campos** a partir da informação disponível, como forma de auxiliar o **modelo** a realizar previsões mais **precisas**.

GRUPO 2
(4 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO EM FOLHA DE TESTE SEPARADA.

QUESTÃO 1

Em alguns algoritmos de *Machine Learning* é usada a técnica de descida por gradiente (*gradient descent*) no processo de otimização dos parâmetros do algoritmo.

a) Quais poderão ser os motivos para esta convergir lentamente ou não convergir? **Se for muito alto não converge. Se for baixo converge lentamente**

b) Indique 2 exemplos de algoritmos de *Machine Learning* que façam uso desta técnica. **Regressão linear e logisitca**

N° _____

GRUPO 3
(4 valores)

PARA CADA AFIRMAÇÃO, RESPONDA ASSINALANDO A SUA VERACIDADE (**V**) OU FALSIDADE (**F**).
JUSTIFIQUE A RESPOSTA EXCLUSIVAMENTE NO ESPAÇO DISPONIBILIZADO.
NÃO SÃO CONSIDERADAS RESPOSTAS PARA AS QUAIS NÃO EXISTA JUSTIFICAÇÃO.

QUESTÃO 1

F O algoritmo de aprendizagem *Decision Tree* apresenta normalmente um melhor desempenho quando comparado com o algoritmo *Random Forest*, apresentando características que possibilitem mitigar o problema de *overfit* de dados.

Decision tree tem problemas de overffitting

QUESTÃO 2

F A *Off-Policy Learning* verificada nos algoritmos de *Reinforcement Learning* considera a avaliação e a otimização da respetiva *policy* aplicada para a seleção das ações do algoritmo inteligente.

Estaria certo se fosse para On-policy

QUESTÃO 3

F Uma matriz de confusão é uma métrica de avaliação de desempenho de modelos de *Reinforcement Learning*.

Apenas para os supervisonados

QUESTÃO 4

F Em todos os algoritmos de *clustering* é necessário justificar a quantidade de *clusters* a procurar nos dados.

Só é o K-means e o medois

N° _____

GRUPO 4
(6 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO NO ESPAÇO RESERVADO.

Considere o *dataset* "breast_cancer", usado diversas vezes no decorrer das aulas, com o intuito de treinar um modelo de classificação com capacidade de prever a existência de um tumor mamário, de acordo com alguns dados clínicos do paciente.

Considere, ainda, o excerto de código abaixo, onde se apresenta a construção e avaliação de um modelo de aprendizagem automática.

QUESTÃO 1

O excerto de código apresentado contém imprecisões. Identifique-as e corrija-as utilizando o espaço disponível ao lado do excerto (não deve copiar todo o excerto, mas apenas aquilo que corrigiu).

```
[1]  df = pandas.read_csv('breast_cancer_dataset.csv')
[2]  df['diagnosis'] =
            df['diagnosis'].substitute(['B', 'M'], [0, 1])
[3]  X = df.drop(['diagnosis', 'id'], axis=1)
[4]  y = df['diagnosis']
[5]  X_train, X_test, y_train, y_test =
         train_test_split(X, y, test_size=1, random_state=2023)
[6]  model = RandomForestClassifier(random_state=2023)
[7]  model.predict(X_train, y_train)
[8]  inferences = model.fit(X_test)
[9]  accuracy = accuracy_score(y_train, inferences)
[10] mse = MSE(y_test, inferences)
[11] print(classification_report(y_test, inferences))
[12] print(confusion_matrix(y_test, inferences))
```

Tem que ser menor que 1 para haver split dos dados

trocar para predict por fit

deveria ser y_test

Figura 1. Excerto de um modelo de aprendizagem.

GRUPO 5
(2 valores)

ASSINALE A VERACIDADE (**V**) OU FALSIDADE (**F**) DE CADA UMA DAS AFIRMAÇÕES QUE SE APRESENTAM.
UMA AFIRMAÇÃO INCORRETAMENTE ASSINALADA ANULA UMA RESPOSTA ASSINALADA CORRETAMENTE.

QUESTÃO 1

Qual o significado de '*boosting*' no contexto de modelos de previsão?

- [ ] Fazer diferentes modelos "votar" para obter uma solução final;
- [ ] Validar um modelo utilizando conjuntos de dados maiores;
- [V] Treinar modelos iterativamente de acordo com os erros de classificação;
- [ ] Dividir aleatoriamente um conjunto de dados para produzir modelos alternativos.

QUESTÃO 2

Qual o significado de 'categórico' quando nos referimos a uma variável num conjunto de dados?

- [ ] Uma variável categórica não pode ser transformada;
- [ ] Não se usam valores numéricos para codificar uma variável categórica;
- [ ] Uma variável categórica não pode ser utilizada como variável dependente/*target*;
- [V] Uma variável categórica não pode ser utilizada como um número/quantidade.