**Universidade do Minho**
Departamento de Informática
Mestrado [integrado] em Engenharia Informática
Mestrado em Matemática e Computação

N° _____ CURSO _____

NOME _____

_____

Dados e Aprendizagem Automática
1° Ano, 1° Semestre
Edição 2021/2022

Prova escrita, 17 de dezembro, 2021

**GRUPO 1**
(4 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO EM FOLHA DE TESTE SEPARADA.

QUESTÃO 1

No desenvolvimento de sistemas de aprendizagem automática (*machine learning*) podem ser utilizados diferentes paradigmas de aprendizagem.

**Supervisionada tem informação sobre os resultados, Nao supervisonada não se sabe o resultado Aprendizagem por reforço programamos as caracteristicas do problema para aprender**

Neste contexto pretende-se que:

a) caracterize os paradigmas de aprendizagem supervisionada, não supervisionada e por reforço;

b) apresente dois exemplos de técnicas de cada paradigma, ilustrando-os com casos de aplicação.

**Regrassão linear e logística**
**K-means e K-medois**
**Q-learning e sarsa**

QUESTÃO 2

O processo de desenvolvimento de uma solução de aprendizagem automática envolve diversas etapas, que podem diferir de acordo com a metodologia escolhida.

Tendo em consideração a metodologia CRISP-DM, pretende-se que enumere e descreva as suas etapas.

**Business understanding.....**

**GRUPO 2**
(4 valores)

Responda às questões deste grupo no espaço reservado PREENCHENDO OS ESPAÇOS VAZIOS com as expressões devidas de modo que a afirmação seja correta.

QUESTÃO 1

No contexto da utilização de técnicas de aprendizagem automática (*machine learning*), a adoção de uma metodologia para a extração de conhecimento descreve e cria _____**Etapas**_____ pelos quais deverá passar o desenvolvimento de um projeto de extração de conhecimento para _____**tomadas de decisão**_____.

QUESTÃO 2

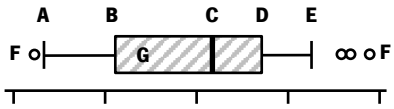
A metodologia de extração de conhecimento que se desenvolve em 5 etapas, a saber, _____**SAMPLE**_____, _____**EXPLORE**_____, _____**MODIFY**_____, _____**MODEL**_____ e _____**ASSESS**_____, denomina-se SEMMA.

QUESTÃO 3

Máquina de Vetores de Suporte (*Support Vector Machine*) é uma técnica _____**Supervisionada**_____ de aprendizagem automática que pode ser utilizada para resolver problemas de _____**classificação**_____ e de _____**regressão**_____.

QUESTÃO 4

Num diagrama de caixa (*boxplot*), como no exemplo à direita, o ponto **C** corresponde à _____**Mediana**_____, a caixa **G** representa _____**intervalo interquartil**_____ dos dados do estudo, e os círculos **F** identificam os valores _____**outliers**_____ do *dataset*.



N° _____

**GRUPO 3**
(6 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO NO ESPAÇO RESERVADO.

Considere o *dataset Titanic*, utilizado por diversas vezes ao longo do semestre. Considere também o excerto de código apresentado na Figura 1, onde é apresentada a construção e avaliação de um modelo de aprendizagem automática.

QUESTÃO 1

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

```
[1]  df = pd.read_csv('titanic_dataset.csv')
[2]  X = df.drop(['Survived', 'Age', 'PassengerId', 'Name',
           'Ticket', 'Cabin', 'Embarked', 'Sex'], axis=1)
[3]  y = df['Survived']
[4]  sex_ohe = pd.merge(df['Sex'], drop_first=True)
[5]  embarked_ohe = pd.merge(df['Embarked'], drop_first=True)
[6]  X = pd.concat([X, sex_ohe, embarked_ohe], axis=1)
[7]  X_train, X_test, y_train, y_test =
                train_test_split(y, X, test_size=0.3)
[8]  model = Sequential()
[9]  model.add(Dense(16, input_dim=y.shape[1],
                                activation='relu'))
[10] model.add(Dense(8, activation='relu'))
[11] model.add(Dense(1, activation='sigmoid'))
[12] model.compile( loss = 'binary_crossentropy',
                    optimizer = 'adam',
                    metrics = ['mse'])
[13] model.transform(X_train, y_train, epochs=50,
                batch_size=32)
[14] loss, acc = model.evaluate(X_train, y_train)
```

Se as variaveis forem irrelevantes ok, está bem

Não é merge é Dumies

O y está trocado com o x

Deveria ser X.shape

Figura 1. Excerto de um modelo de aprendizagem.

QUESTÃO 2

Identifique a técnica de aprendizagem utilizada no excerto de código apresentado na Figura 1, e indique quatro hiperparâmetros passíveis de serem modificados para afinar o modelo.

Redes neuronais

optimizer, epochs, numero de camadas e numero de neuronios por camada

QUESTÃO 3

Admita que o *dataset Titanic* não está balanceado. Descreva de que forma este desbalanceamento influencia o modelo.

Overfitting

N° _____

**GRUPO 4**
(6 valores)

Comente as afirmações seguintes, assinalando a sua veracidade (**V**) ou falsidade (**F**), justificando a resposta EXCLUSIVAMENTE no espaço disponibilizado.
NÃO SÃO CONSIDERADAS respostas para as quais não exista justificação expressa.

QUESTÃO 1

**V** No desenvolvimento de sistemas de aprendizagem automática, a fase de preparação de dados tem particular importância porque os dados obtidos do «mundo físico» são incompletos, contêm lixo e são falsos.

**Sim, os dados contem ruido e é preciso tira lo**

QUESTÃO 2

**F** Técnicas de aprendizagem automática baseadas no desenvolvimento de árvores de decisão são utilizadas exclusivamente para a resolução de problemas de classificação.

**Tambem dá para regressão**

QUESTÃO 3

**V** Paradigmas de aprendizagem com supervisão exigem maior intervenção humana do que qualquer outro paradigma uma vez que necessitam de quem desempenhe o papel de supervisor.

**Precisa de humanos**

QUESTÃO 4

**V** O tratamento de valores nulos (*missing values*) existentes num *dataset* pode envolver a remoção de observações/registos ou de atributos/características.

**Se o registo tiver muitos missing values pode se remover o registo**

QUESTÃO 5

**F** A matriz de confusão à direita apresenta um valor de *accuracy* de $^{165}/_{150}$.

**165/150 é só estupido, deveria ser 150/165**

n=165

| | | PREVISÃO | | |
|---|---|---|---|---|
| | | NÃO | SIM | |
| ATUAL | NÃO | 50 | 10 | 60 |
| | SIM | 5 | 100 | 105 |
| | | 55 | 110 | |

QUESTÃO 6

**F** Num processo de aprendizagem automática, a qualidade dos dados não afeta os resultados do processo uma vez que na fase de preparação de dados serão resolvidos todos os problemas como, por exemplo, ruído, *outliers*, dados falsos ou dados duplicados.

**Podem nao ser resolvidos devido a elevada complexidade**