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

Nº _____ CURSO _____

NOME _____

_____

Dados e Aprendizagem Automática
1º Ano, 1º Semestre
Edição 2023/2024
Prova Escrita, 14 de dezembro, 2023

OBS: OS TERMOS EM INGLÊS CUJA TRADUÇÃO PODERIA GERAR CONFUSÃO FORAM MANTIDOS EM *ITÁLICO*.

GRUPO 1 (4 valores)

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

QUESTÃO 1 - O método de validação de modelos denominado *hold out validation* é um método de _____partição_____ de dados que divide o *dataset* em duas partes: uma parte _____treino_____ e outra parte _____teste_____.

QUESTÃO 2 - O *Support Vetor Machine* é um algoritmo de aprendizagem automática _____supervisionado_____ que pode ser utilizado tanto para problemas de _____classificaçãop_____ como de _____regressão_____.

QUESTÃO 3 - O *ensemble* _____learning_____ envolve o aproveitamento das previsões de vários modelos fracos, normalmente árvores de decisão, para criar um conjunto robusto e preciso de previsões. Cada árvore de decisão na _____florest_____ é treinada num subconjunto diferente dos dados de treino e a previsão final é efetuada agregando as previsões individuais de todas as árvores através de técnicas como o cálculo _____da média_____ ou a _____votação_____.

GRUPO 2 (4 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO EM FOLHA DE TESTE SEPARADA.

Atendendo ao *fine tuning* afinação de hiperparâmetros:

a) explique o conceito de hiperparâmetros em modelos de aprendizagem automática; controlam o comportamento de um algoritmo de aprendizado de máquina
b) discuta a importância do *fine tuning* de hiperparâmetros; fundamental para obter um modelo eficiente e bem ajustado
c) enumere dois métodos de *fine tuning* de hiperparâmetros e forneça uma breve explicação de cada um.
Grid search : método sistemático que testa todas as combinações possíveis de um conjunto pré-definido de valores para os hiperparâmetros
Random search: seleciona combinações aleatórias de hiperparâmetros dentro de intervalos definidos, em vez de testar todas as combinações possíveis

GRUPO 3 (4 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO EM FOLHA DE TESTE SEPARADA.

Comente as afirmações seguintes, assinalando a veracidade (V) ou a falsidade (F), justificando a resposta. NÃO SÃO CONSIDERADAS respostas para as quais não exista justificação expressa.

V QUESTÃO 1 - É possível utilizar técnicas de aprendizagem não supervisionada mesmo quando os casos de treino contêm informação sobre os resultados pretendidos.

F QUESTÃO 2 - A precisão é uma métrica que mede a capacidade de um modelo de classificação para capturar todas as instâncias relevantes, incluindo os falsos positivos. **Falsos negativos**

F QUESTÃO 3 - Modelos baseados em árvores, como Árvores de Decisão, tendem a ser menos propensos a *overfitting* quando comparado com modelos mais complexos, como Redes Neuronais Artificiais **Muito propensas a overfitting as arvores**

F QUESTÃO 4 - Em *ensemble learning*, o *bagging* e o *boosting* são técnicas utilizadas para combinar as previsões de vários modelos, que seguem o mesmo princípio subjacente. **O bagging nao utiliza varios modelos**

N° _____

GRUPO 4 (6 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO NO ESPAÇO RESERVADO.

Considere o *dataset* "wine.csv", usado diversas vezes no decurso do semestre, com o intuito de treinar um modelo de aprendizagem com capacidade de classificar o vinho em 1 das 3 classes, de acordo com algumas características.

Considere, ainda, o excerto de código abaixo, onde se apresenta a preparação dos dados para a construção de um modelo de aprendizagem automática.

O excerto de código apresentado contém imprecisões. Identifique 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('wine.csv')
[2] print(df.duplicated().sum())
[3] df.drop_duplicates(inplace=False)
[4] df.rename(columns={"OD280/OD315 of diluted wines":
    "Protein Concentration"}, inplace=True)
[5] df_clean = df.drop(df.loc[(df['Ash']<2) &
    (df['Alcalinity of ash']>15)].index)
[6] print(f"Histogram: {df['Magnesium'].hist()}")
[7] print(f"Skewness: {df['Magnesium'].skew()}")
[8] print(f"Kurtosis: {df['Magnesium'].kurt()}")
[9] df_group.groupby(by=['Class', 'Proline']).mean()
[10] print(df_group.groupby(by=
     ['Alcohol']).agg(pandas.Series.mode))
[11] print(estimator.bin_Edges_[0])
[12] df['alcohol_binned'] =
     estimator.fit_transform(df[['Alcohol']])
[13] estimator =
     sklearn.preprocessing.KBinsDiscretizer(n_bins=3,
     encode='ordinal', strategy='quantile')
```

[1] _____

[2] _____

[3] Tinha que ser true

[4] _____

[5] _____

[6] _____

[7] _____

[8] _____

[9] df_group é df_clean

[10] edges

[11] Falta definir o estimator

[12] _____

[13] _____

GRUPO 5 (2 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO NO ESPAÇO RESERVADO.

Assinale a veracidade (V) ou a falsidade (F) de cada uma das afirmações que se apresentam. Para cada questão, uma afirmação INCORRETAMENTE assinalada ANULA uma resposta assinalada corretamente.

QUESTÃO 1 - Em *machine learning*, técnicas de regressão:

- [X] São usadas para prever resultados contínuos;
- [ ] São usadas para prever resultados discretos;
- [ ] São usadas quando todos os dados de treino são contínuos;
- [ ] São usadas quando todos os dados de treino são discretos.

QUESTÃO 2 - Qual das seguintes opções descreve a técnica de *Max Voting* na aprendizagem de conjuntos?

- [ ] É um método em que o modelo com a precisão máxima determina o resultado final;
- [ ] Envolve o cálculo da média das previsões de cada modelo no conjunto;
- [X] É uma técnica em que cada modelo do conjunto vota numa classe e a classe com mais votos é escolhida como previsão final;
- [ ] Refere-se à seleção do melhor modelo do conjunto com base no seu desempenho num conjunto de validação.