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

Nº _____ CURSO _____

NOME eduardovski

_____

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 validacao 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 supervisionada que pode ser utilizado tanto para problemas de regressao como de classificação.

QUESTÃO 3 - O *ensemble* learning, random forest 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 random forest 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 voting ou a media.

GRUPO 2 (4 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO EM FOLHA DE TESTE SEPARADA.

Atendendo ao *fine tuning* afinação de hiperparâmetros:

no word

a) explique o conceito de hiperparâmetros em modelos de aprendizagem automática;
b) discuta a importância do *fine tuning* de hiperparâmetros;
c) enumere dois métodos de *fine tuning* de hiperparâmetros e forneça uma breve explicação de cada um.

GRUPO 3 (4 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO EM FOLHA DE TESTE SEPARADA.

no word

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.

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

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.

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.

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.

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')
```

Código Corrigido:

1. **[3]** `df.drop_duplicates(inplace=False)`
   - **Correção**: Altere `inplace=False` para `inplace=True` para garantir que as duplicatas sejam removidas no próprio `DataFrame`.
   - **Final**: `df.drop_duplicates(inplace=True)`
2. **[9]** `df_group.groupby(by=['Class', 'Proline']).mean()`
   - **Correção**: `df_group` não foi definido no código. Substitua `df_group` por `df` para usar o `DataFrame` correto.
   - **Final**: `df.groupby(by=['Class', 'Proline']).mean()`
3. **[10]** `print(df_group.groupby(by=['Alcohol']).agg(pandas.Series.mode))`
   - **Correção**: Assim como em [9], substitua `df_group` por `df`. Adicionalmente, corrija `pandas.Series.mode` para `pandas.Series.mode()`.
   - **Final**: `print(df.groupby(by=['Alcohol']).agg(pandas.Series.mode))`
4. **[11]** `print(estimator.bin_Edges_[0])`
   - **Correção**: O nome correto do atributo é `bin_edges_` e não `bin_Edges_`.
   - **Final**: `print(estimator.bin_edges_[0])`
5. **[12]** `df['alcohol_binned'] = estimator.fit_transform(df[['Alcohol']])`
   - **Correção**: O método `fit_transform` retorna uma matriz, então é necessário converter para um `DataFrame` ou ajustar o formato.
   - **Final**: `df['alcohol_binned'] = estimator.fit_transform(df[['Alcohol']]).flatten()`

[12] Linha Atualizada Sem `flatten()`

O método `fit_transform` retorna uma matriz `numpy`. Para evitar `.flatten()`, podemos usar o método `ravel()` ou converter diretamente o array em uma coluna:

python — Copiar código

```
df['alcohol_binned'] = estimator.fit_transform(df[['Alcohol']])[:, 0]
```

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:

[ ] 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.

Respostas:

1. **Verdadeiro (V)**: Técnicas de regressão, como regressão linear ou logística, são amplamente utilizadas para prever valores contínuos.
2. **Falso (F)**: Prever resultados discretos é mais comum em problemas de classificação, não de regressão.
3. **Falso (F)**: A regressão pode ser usada mesmo quando os dados de treino incluem variáveis discretas (ex.: variáveis categóricas, após codificação).
4. **Falso (F)**: Não é obrigatório que todos os dados de treino sejam discretos ou contínuos para aplicar regressão; é possível lidar com misturas.

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;
[ ] 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.

Respostas:

1. **Falso (F)**: O modelo com a precisão máxima não determina o resultado final em Max Voting.
2. **Falso (F)**: Cálculo da média é usado em métodos como bagging, não em Max Voting.
3. **Verdadeiro (V)**: Max Voting ocorre quando cada modelo no conjunto faz uma predição e a classe com mais votos é selecionada.
4. **Falso (F)**: Max Voting não envolve selecionar um único modelo com base no desempenho.