Lista #5
Curso: Ciência da Computação
Disciplina: Inteligência Artificial
Profª. Cristiane Neri Nobre
Data de entrega: 30/03
Valor: 1 ponto

Questão 01

Uma árvore de decisão pode se beneficiar da técnica de poda (pruning) para melhorar a generalização do modelo. Considere o comportamento esperado de uma árvore de decisão que **não utiliza poda** em um conjunto de dados com muito ruído.

Avalie as afirmativas a seguir:

I. A árvore de decisão tende a se ajustar excessivamente aos dados de treinamento, criando um modelo superajustado (overfitting) que se adapta ao ruído.
II. A profundidade da árvore pode aumentar desnecessariamente, o que pode levar a um desempenho fraco em novos dados de teste.
III. A ausência de poda permite que a árvore capture padrões reais nos dados, mesmo em um conjunto com muito ruído, resultando em uma melhor performance de generalização.

Quais das afirmações estão corretas?

A) Apenas I e II
B) Apenas II e III
C) Apenas I e III
D) Todas estão corretas
E) Apenas II está incorreta

Questão 02

Os algoritmos ID3 e C4.5 são utilizados na construção de árvores de decisão, mas apresentam diferenças importantes. Considere as seguintes afirmativas sobre as diferenças entre os dois algoritmos:

I. O algoritmo ID3 utiliza o ganho de informação como métrica para escolher os atributos, enquanto o C4.5 utiliza o ganho de informação normalizado (razão de ganho) para compensar o viés do ID3 em favor de atributos com muitos valores.
II. O algoritmo C4.5 pode lidar com atributos contínuos, enquanto o ID3 só trabalha com atributos discretos.
III. O C4.5 é capaz de lidar com dados ausentes, enquanto o ID3 exige que os dados estejam completos para a construção da árvore.

Quais das afirmações estão corretas?

A) Apenas I e II
B) Apenas II e III
C) Apenas I e III
D) Todas estão corretas
E) Apenas I está incorreta

Questão 03

Os algoritmos ID3, C4.5 e CART são amplamente utilizados na construção de árvores de decisão, cada um com características específicas.

Considere as seguintes afirmativas sobre esses algoritmos:

I. Tanto o ID3 quanto o C4.5 utilizam o ganho de informação ou o índice de ganho para a escolha dos atributos, enquanto o CART utiliza o critério de Gini ou entropia para essa tarefa.
II. O algoritmo C4.5 pode lidar com atributos contínuos e discretos, enquanto o ID3 trabalha apenas com atributos discretos. O CART também lida com ambos os tipos de atributos.
III. Diferentemente do ID3 e do C4.5, o algoritmo CART gera árvores de decisão binárias, ou seja, em cada nó, a divisão é feita sempre em dois ramos.

Quais das afirmações estão corretas?

A) Apenas I e II
B) Apenas II e III
C) Apenas I e III
D) Todas estão corretas
E) Apenas I está incorreta

Questão 04

Sobre a construção de árvores no Random Forest, analise as seguintes afirmações:

I. Cada árvore na floresta é construída a partir de uma amostra aleatória do conjunto de dados com reposição (bootstrap sampling).
II. Para cada divisão de nó, o algoritmo considera um subconjunto aleatório de features, o que aumenta a diversidade entre as árvores.
III. Todas as árvores no Random Forest são treinadas usando exatamente o mesmo subconjunto de features para cada nó.

Quais afirmações estão corretas?

A) Apenas I e II estão corretas
B) Apenas II está correta
C) Apenas I e III estão corretas
D) Todas estão corretas

Questão 05

Sobre o uso do F1-score, considere as afirmações:

I. O F1-score é uma métrica útil quando há um desbalanceamento entre as classes, pois combina precisão e recall em uma única métrica.
II. O F1-score será alto se tanto a precisão quanto o recall forem altos, e será baixo se uma dessas métricas for baixa, mesmo que a outra seja alta.
III. O F1-score é preferido em contextos onde os falsos negativos são mais custosos que os falsos positivos.

Quais estão corretas?

A) Apenas I está correta
B) Apenas II está correta
C) Apenas I e II estão corretas
D) Apenas II e III estão corretas
E) Todas estão corretas

Questão 06

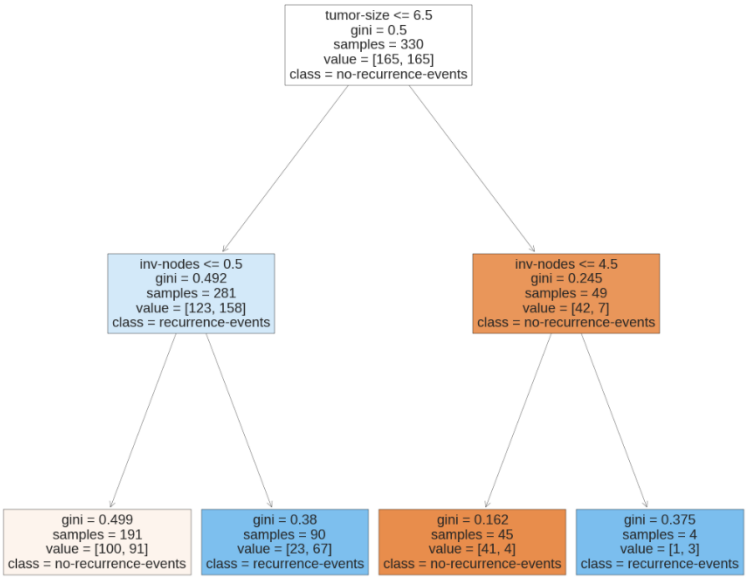
Considere a seguinte matriz de confusão para um classificador com 3 classes A, B e C:

	Predito: A	Predito: B	Predito: C
Real: A	50	10	5
Real: B	15	40	10
Real: C	5	10	55

Calcule a **Taxa de Verdadeiro Negativo** (TVN) da classe A.
A. 0,71
B. 0,77
C. 0,85
D. 0,80
E. 0,83

Questão 07

A figura abaixo mostra uma árvore de decisão construída por um algoritmo de aprendizado indutivo a partir de um conjunto de dados em que as instâncias são classificadas em **Câncer recorrente** ou **Câncer não Recorrente.**



Considerando-se as seguintes afirmações:

I. Quanto maior o valor do gini, mais puro é o nó
II. A maior cobertura por classe gerada a partir das regras é de aproximadamente 61%
III. A atributo "tumor-size é o que tem maior entropia nesta base de dados

É **correto** o que se afirma em:

a) II, apenas.
b) III, apenas.
c) I e II, apenas.
d) I e III, apenas.
e) I, II e III.

Questão 08

Em um problema de classificação com classes desbalanceadas, é comum aplicar técnicas de oversampling ou undersampling. Considere as afirmativas a seguir:

I. O oversampling aumenta a quantidade de instâncias da classe minoritária, o que pode levar ao overfitting, especialmente quando se duplicam instâncias existentes.
II. O undersampling reduz a quantidade de instâncias da classe majoritária, o que pode levar à perda de informação relevante.
III. O uso de técnicas de balanceamento como SMOTE (Synthetic Minority Over-sampling Technique) pode ajudar a criar instâncias sintéticas da classe minoritária e melhorar a generalização do modelo.

Quais afirmativas estão corretas?

a) Apenas I e II
b) Apenas II e III
c) Apenas I e III
d) Todas estão corretas
e) Apenas III está correta

Questão 09

Em relação à aplicação prática das técnicas de oversampling e undersampling no aprendizado supervisionado, analise as afirmativas:

I. Técnicas de oversampling como SMOTE devem ser aplicadas apenas após o split dos dados em treino e teste para evitar vazamento de dados (data leakage).
II. O undersampling pode ser útil em conjuntos de dados muito grandes, pois reduz o tempo de treinamento ao diminuir o número de exemplos.
III. Um modelo treinado em dados balanceados geralmente apresenta melhor recall para a classe minoritária em comparação com um modelo treinado em dados desbalanceados.

Quais afirmativas estão corretas?

a) Apenas I e II
b) Apenas II e III
c) Apenas I e III
d) Todas estão corretas
e) Apenas I está correta

Questão 10

No pré-processamento de dados para modelos de aprendizado de máquina, a imputação de valores ausentes é uma etapa fundamental. Analise as afirmativas a seguir:

- **I.** A imputação pela média ou mediana pode distorcer a distribuição original dos dados, especialmente se houver outliers.
- **II.** A técnica de imputação mais apropriada depende do tipo de variável (numérica, categórica, ordinal) e da natureza do problema
- **III.** Ignorar os valores ausentes e remover diretamente as linhas com dados faltantes nunca é uma boa prática e deve ser evitado em qualquer circunstância.

Quais afirmativas estão corretas?

a) Apenas I e II
b) Apenas II e III
c) Apenas I e III

d) Todas estão corretas
e) Apenas II está correta

Questão 11

Considere agora algumas técnicas e boas práticas em imputação de dados ausentes:

I. Técnicas avançadas como KNN imputation e regressão multivariada consideram a correlação entre variáveis para estimar os valores ausentes.
II. A imputação deve ser aplicada antes do split dos dados em treino e teste para garantir consistência estatística no processo.
III. Em pipelines profissionais, a imputação é frequentemente combinada com validação cruzada para evitar o vazamento de dados (data leakage).

Quais afirmativas estão corretas?
a) Apenas I e II
b) Apenas II e III
c) Apenas I e III
d) Todas estão corretas
e) Apenas I está correta

Questão 12

Em relação às técnicas de codificação de variáveis categóricas em aprendizado de máquina, analise as afirmativas:

I. O Label Encoding pode induzir um modelo a assumir uma ordem inexistente entre categorias, o que pode ser problemático em algoritmos baseados em distância ou regressão.

II. O One-Hot Encoding pode aumentar significativamente a dimensionalidade do conjunto de dados, especialmente em variáveis com muitos níveis (cardinalidade alta).

III. O Frequency Encoding substitui cada categoria pela frequência relativa de sua ocorrência, preservando informações de ordem entre as categorias.

Quais afirmativas estão corretas?

a) Apenas I e II
b) Apenas II e III
c) Apenas I e III
d) Todas estão corretas
e) Apenas III está correta

Questão 13

Considere o uso das classes DecisionTreeClassifier e RandomForestClassifier do módulo sklearn.ensemble e sklearn.tree. Analise as afirmativas abaixo:

I. O método .fit(X, y) é usado tanto em DecisionTreeClassifier quanto em RandomForestClassifier para treinar o modelo com os dados de entrada X e os rótulos y.
II. Após o treinamento, é possível prever os valores de teste com o método .predict(X_test) em ambos os modelos.
III. A escolha de criterion='entropy' ou criterion='gini' está disponível apenas para o DecisionTreeClassifier.

Quais afirmativas estão corretas?

a) Apenas I e II

b) Apenas II e III

c) Apenas I e III

d) Todas estão corretas

e) Apenas III está correta

Questão 14

Considere a configuração de hiperparâmetros nos algoritmos de árvore de decisão e floresta aleatória (RandomForestClassifier).

Analise:

I. O hiperparâmetro max_depth controla a profundidade máxima de cada árvore tanto na DecisionTreeClassifier quanto na RandomForestClassifier.

II. O parâmetro n_estimators define o número de árvores utilizadas no RandomForestClassifier.

III. O parâmetro random_state garante reprodutibilidade dos resultados em ambos os modelos ao fixar a semente dos geradores aleatórios.

Quais afirmativas estão corretas?

a) Apenas I e II

b) Apenas II e III

c) Apenas I e III

d) Todas estão corretas

e) Apenas I está correta

Questão 15

Considerando a base de dados abaixo e o algoritmo de **Árvore de decisão ID3**, qual a raiz da árvore e qual o ganho de informação do atributo, respectivamente?

Obs:
É necessário apresentar todos os cálculos. Ou seja, não será considerada questão sem apresentação dos cálculos.

Entropia da classe:
E(9/17, 8/17) = 0,9975025464

Ganho de Experiencia:
0,9975 -(6/17 x E(5/6, 1/6) + 6/17 x E(2/6, 4/6) + 5/17 x E(1/5, 4/5)) = 0,252795391

Ganho de Interesse:
0,9975-(10/17 x E(7/10, 3/10) + 7/10 x E(1/7, 6/7)) = 0,235

Atributos da Base de dados:

1. **Experiência** com Programação (experiência): [Baixa, Média, Alta]
2. **Interesse** em Tecnologia (interesse): [Baixo, Alto]
3. **Horas** de Estudo por Semana (horas): [Baixas, Altas]
4. **Classe: Gosta ou não de IA**

Experiência	Interesse	Horas	Gosta de IA (Classe)
Baixa	Baixo	Baixas	Não Gosta
Baixa	Baixo	Baixas	Não Gosta
Média	Baixo	Baixas	Não Gosta
Baixa	Baixo	Altas	Não Gosta
Baixa	Alto	Baixas	Não Gosta
Alta	Baixo	Baixas	Não Gosta
Média	Baixo	Altas	Não Gosta
Baixa	Baixo	Altas	Não Gosta
Média	Baixo	Baixas	Gosta

Ganho de Horas:
0,9975-(9/17 x E(5/9, 4/9) + 8/17 x E(3/8, 5/8)) = 0,0236698075

Alta	Baixo	Baixas	Gosta
Média	Alto	Altas	Gosta
Baixa	Alto	Altas	Gosta
Alta	Alto	Baixas	Gosta
Média	Alto	Altas	Gosta
Alta	Alto	Altas	Gosta
Média	Alto	Baixas	Gosta
Alta	Baixo	Altas	Gosta

a) A raiz da árvore é o atributo **Experiência** com ganho de 0,232
b) A raiz da árvore é o atributo **Interesse** com ganho de 0,235
c) A raiz da árvore é o atributo **Horas** com ganho de 0,421
d) A raiz da árvore é o atributo **Interesse** com ganho de 0,194
e) A raiz da árvore é o atributo **Experiência** com ganho de 0,15

Questão 16

Utilizando-se a mesma base de dados anterior, e o algoritmo **Naive Bayes,** qual a probabilidade de a pessoa **GOSTAR ou não JOGAR de IA,** respectivamente, para o seguinte registro:

Experiência	Alta
Interesse	Alto
Horas	Baixas

Obs: Apresentar os cálculos necessários para a solução da questão.

a) 93,82% e 6,18%
b) 69,32% e 30,68%
c) 74,01% e 25,99%
d) 5,56% e 94,44%
e) 95,32% e 4,68%

P(Gostar):
(9/17)x(4/9)x(6/9)x(4/9) = 0,0697167756
0,0697167756 / 0,0743123638 x 100
=>93,8158%

P(Não Gostar):
(8/17)x(1/8)x(1/8)x(5/8) =0,0045955882
0,0045955882 / 0,0743123638 x 100
=> 6,1841%

S = 0,0743123638