**Lista 5 – Inteligência Artificial – Joao Paulo Dias Estevao**

**Questão 1:**

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

1. A árvore de decisão tende a se ajustar excessivamente aos dados de treinamento, criando um modelo superajustado (overfitting) que se adapta ao ruído. **V**
2. A profundidade da árvore pode aumentar desnecessariamente, o que pode levar a um desempenho fraco em novos dados de teste. **V**
3. 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. **F**

Quais das afirmações estão corretas?

Resposta:

1. **Apenas I e II**

**Questão 2:**

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:

1. 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. **V**
2. O algoritmo C4.5 pode lidar com atributos contínuos, enquanto o ID3 só trabalha com atributos discretos. **V**
3. 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? **V**

Quais das afirmações estão corretas?

Resposta:

**D) Todas estão corretas**

**Questão 3:**

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:

1. 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. **V**
2. 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. **V**
3. 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. **V**

Quais das afirmações estão corretas?

Resposta:

**D) Todas estão corretas**

**Questão 4:**

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

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

Quais das afirmações estão corretas?

Resposta:

1. **Apenas I e II estão corretas**

**Questão 5:**

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

1. O F1-score é uma métrica útil quando há um desbalanceamento entre as classes, pois combina precisão e recall em uma única métrica. **V**
2. 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. **V**
3. O F1-score é preferido em contextos onde os falsos negativos são mais custosos que os falsos positivos. **F**

Quais das afirmações estão corretas?

Resposta:

1. **Apenas I e II estão corretas**

**Questão 6:**

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

**Tabela

O conteúdo gerado por IA pode estar incorreto.**

Calcule a Taxa de Verdadeiro Negativo (TVN) da classe A.

Não é A = Total (B) + Total (C) = 65 + 70 = 135

TVN significa que não era e não foi classificado sendo, portanto:

TVN = (40 + 10 + 10 + 55)/135 = 115/135 = 0.85

Resposta:

**c) 0,85**

**Questão 7:**

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.

**Diagrama, Linha do tempo

O conteúdo gerado por IA pode estar incorreto.**

Considerando-se as seguintes afirmações:

1. Quanto maior o valor do gini, mais puro é o nó. **F**
2. A maior cobertura por classe gerada a partir das regras é de aproximadamente 61%. **V**
3. A atributo “tumor-size é o que tem maior entropia nesta base de dados. **F**

Cobertura das classes (Afirmação II):

C1 = 100/165 = 0.606; C2 = 67/165 = 0.406; C3 = 41/165 = 0.248; C4 = 3/165 = 0.018

Resposta:

1. **II, apenas**

**Questão 8:**

Em um problema de classificação com classes desbalanceadas, é comum aplicar técnicas de oversampling ou undersampling.

Considere as afirmativas a seguir:

1. O oversampling aumenta a quantidade de instâncias da classe minoritária, o que pode levar ao overfitting, especialmente quando se duplicam instâncias existentes. **V**
2. O undersampling reduz a quantidade de instâncias da classe majoritária, o que pode levar à perda de informação relevante. **V**
3. 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. **V**

Quais afirmativas estão corretas?

**d) Todas estão corretas**

**Questão 9:**

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

1. 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). **V**
2. O undersampling pode ser útil em conjuntos de dados muito grandes, pois reduz o tempo de treinamento ao diminuir o número de exemplos. **V**
3. Um modelo treinado em dados balanceados geralmente apresenta melhor recall para a classe minoritária em comparação com um modelo treinado em dados desbalanceados. **V**

Resposta:

**d) Todas estão corretas**

**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:

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

Resposta:

1. **Apenas I e II**

**Questão 11:**

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

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

Resposta:

1. **Apenas I e III**

**Questão 12:**

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

1. 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. **V**
2. O One-Hot Encoding pode aumentar significativamente a dimensionalidade do conjunto de dados, especialmente em variáveis com muitos níveis (cardinalidade alta). **V**
3. O Frequency Encoding substitui cada categoria pela frequência relativa de sua ocorrência, preservando informações de ordem entre as categorias. **F**

Resposta:

1. **Apenas I e II**

**Questão 13:**

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

1. 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. **V**
2. Após o treinamento, é possível prever os valores de teste com o método .predict(X\_test) em ambos os modelos. **V**
3. A escolha de criterion='entropy' ou criterion='gini' está disponível apenas para o DecisionTreeClassifier. **F**

Resposta:

1. **Apenas I e II**

**Questão 14:**

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

Analise:

1. O hiperparâmetro max\_depth controla a profundidade máxima de cada árvore tanto na DecisionTreeClassifier quanto na RandomForestClassifier. **V**
2. O parâmetro n\_estimators define o número de árvores utilizadas no RandomForestClassifier. **V**
3. O parâmetro random\_state garante reprodutibilidade dos resultados em ambos os modelos ao fixar a semente dos geradores aleatórios. **V**

Resposta:

1. **Todas estão corretas**

**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?

**Tabela

O conteúdo gerado por IA pode estar incorreto.**

**Tabela

O conteúdo gerado por IA pode estar incorreto.**

Resposta:

1. **A raiz da árvore é o atributo Interesse com ganho de 0,235**

**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

**Tabela

O conteúdo gerado por IA pode estar incorreto.**

Gosta: 9/17

Exp Alta: 4/9  
Interesse Alto: 6/9  
Horas Baixa: 4/9  
Soma = 9/17 \* 4/9 \* 6/9 \* 4/9 = 0.0697

Não Gosta: 8/17

Exp Alta: 1/8  
Interesse Alto: 1/8  
Horas Baixa: 5/8  
Soma = 8/17 \* 1/8 \* 1/8 \* 5/8 = 0.0045

Total = 0.0697 + 0.0045 = 0.0742

**P(Gosta)** = 0.0697/0.0742 \* 100 = 93,9%  
**P(Não Gosta)** = 0.0045/0.0742 \* 100 = 6,06%

Resposta:

**a) 93,82% e 6,18%**