Inteligência Artificial

812839 - Vinícius Miranda de Araújo

Lista de Exercícios 05

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

R: Letra A

Explicação para a questão

I - Correto. Sem poda, a árvore pode crescer excessivamente e ajustar-se não apenas aos padrões reais dos dados, mas também ao ruído. Isso resulta em overfitting, o que prejudica a generalização.

II - Correto. Árvores sem poda podem continuar dividindo os dados até que todos os exemplos sejam perfeitamente classificados, resultando em um modelo excessivamente complexo que não generaliza bem para novos dados.

III - Incorreto. Se o conjunto de dados contém muito ruído, uma árvore sem poda pode aprender padrões espúrios, prejudicando a generalização.

A poda é usada para evitar que a árvore cresça excessivamente e se ajuste ao ruído dos dados. Sem poda, a árvore pode ficar muito profunda e se tornar altamente especializada nos dados de treino (overfitting). Isso prejudica o desempenho em novos dados, pois o modelo memoriza detalhes irrelevantes ao invés de identificar padrões generalizáveis.

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

R: Letra C

Explicação para a questão

I - O ID3 escolhe atributos com base no ganho de informação, que pode favorecer atributos com muitos valores distintos. O C4.5 corrige esse problema usando a razão de ganho (gain ratio), que normaliza o ganho de informação para evitar esse viés.

II - Correto. O ID3 só lida com atributos discretos, enquanto o C4.5 pode dividir atributos contínuos em intervalos, permitindo que sejam usados na construção da árvore.

III - Correto. O ID3 não lida com valores ausentes, exigindo dados completos. O C4.5, por outro lado, pode tratar valores ausentes ao calcular a probabilidade de ocorrência dos valores e distribuir os exemplos proporcionalmente.

A principal diferença entre ID3 e C4.5 está nas melhorias feitas no C4.5 para resolver limitações do ID3. O ID3 usa o ganho de informação puro, favorecendo atributos com muitos valores distintos, enquanto o C4.5 usa a razão de ganho para corrigir esse viés. Além disso, o C4.5 suporta atributos contínuos e dados ausentes, tornando-o mais robusto e aplicável a diferentes cenários.

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

R: Letra C

Explicação para a questão

I - Correto. Tanto o ID3 quanto o C4.5 utilizam o ganho de informação para selecionar os atributos, mas o C4.5 usa a razão de ganho (gain ratio) para melhorar a escolha e evitar a preferência por atributos com muitos valores distintos. O CART, por outro lado, utiliza o índice de Gini ou entropia para dividir os dados em cada nó.

II - Correto. O ID3 só trabalha com atributos discretos, enquanto o C4.5 pode lidar com atributos contínuos, dividindo-os em intervalos. O CART também pode lidar com atributos contínuos e discretos.

III - Correto. O CART sempre gera árvores binárias, ou seja, cada divisão em um nó resulta em dois ramos. Diferentemente do ID3 e C4.5, que podem gerar árvores com mais de dois ramos em um nó.

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

R: Letra A

Explicação para a questão

I - Correto. No Random Forest, cada árvore é construída usando uma amostra aleatória com reposição do conjunto de dados, conhecida como bootstrap sampling. Isso significa que algumas instâncias podem ser repetidas, e outras podem ser deixadas de fora de cada árvore.

II - Correto. Para a divisão de cada nó, o algoritmo considera um subconjunto aleatório de features. Isso aumenta a diversidade entre as árvores, tornando o modelo mais robusto e menos suscetível ao overfitting, pois cada árvore não é influenciada por todas as features disponíveis.

III - Incorreto. As árvores no Random Forest não utilizam exatamente o mesmo subconjunto de features para cada nó. Pelo contrário, um subconjunto aleatório de features é escolhido para cada divisão de nó, o que promove a diversidade entre as árvores, evitando a correlação entre elas.

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

R: Letra C

Explicação para a questão

I - Correto. O F1-score é uma métrica que combina precisão e recall em uma única medida, o que a torna útil em situações de desbalanceamento entre as classes. Quando há uma classe predominante, o F1-score ajuda a equilibrar a avaliação de ambos os aspectos, em vez de depender apenas de uma métrica isolada.

II - Correto. O F1-score será alto quando precisão e recall forem altos. Caso um deles seja baixo, o F1-score também será baixo, independentemente de o outro ser alto. Isso ocorre porque o F1-score é a média harmônica entre essas duas métricas, e uma métrica baixa diminui significativamente o valor do F1-score.

III - Incorreto. O F1-score é uma métrica equilibrada entre precisão e recall, mas não é necessariamente preferido em contextos onde falsos negativos são mais custosos. Nessas situações, pode-se preferir o recall, pois ele foca em minimizar os falsos negativos, enquanto o F1-score considera tanto falsos positivos quanto falsos negativos.

Questão 06

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

Considere as seguintes afirmativas sobre esses algoritmos:

	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

R: Letra C

Explicação para a questão

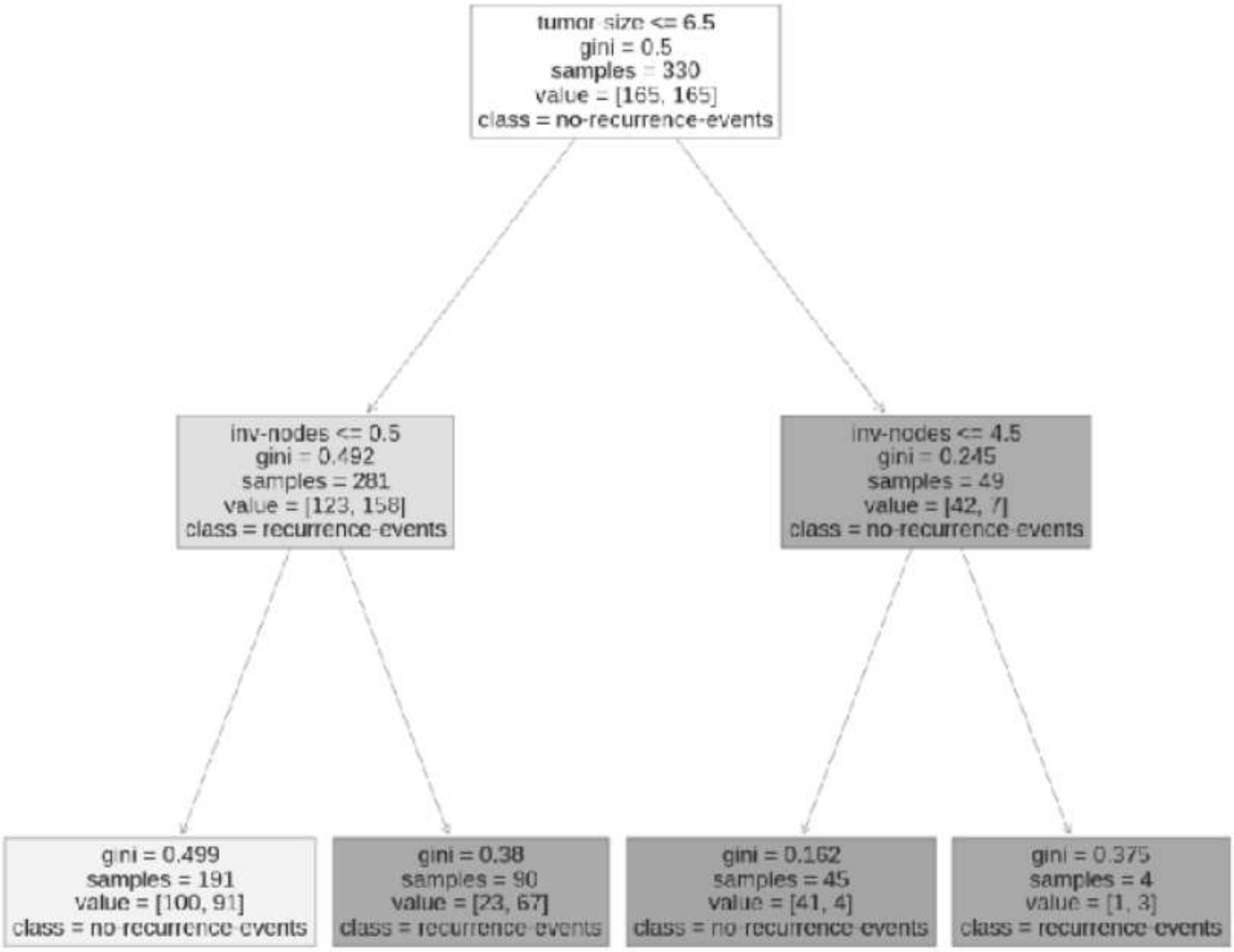
A TVN significa quem é verdadeiro e foi classificado como não sendo, nesse caso a soma da linha B é 65 e da linha C é 70, logo $65 + 70 = 135$.

Somando da linha B e C, as colunas de B e C, temos: $40 + 10 + 10 + 50 = 115$

Portanto, $\frac{115}{135} \sim 0,85$

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.

R: Letra A

Explicação para a questão

I - Incorreta. Quanto maior o valor do Gini, mais impuro é o nó, e não o contrário. O valor de Gini indica a diversidade das classes dentro do nó:

- Gini próximo de 0 → Nó mais puro (predominância de uma classe).
- Gini próximo de 0.5 → Nó mais impuro (classes equilibradas).

II - Correta. A maior cobertura por classe gerada pelas regras pode ser estimada a partir dos valores de distribuição das classes nos nós terminais.

- $\frac{100}{165} = 0,601$
- $0,61 * 100 = 60.1\% \sim 61\%$

III - Incorreto. O atributo com maior ganho de informação (ou seja, menor entropia após a divisão) é escolhido como raiz, nesse caso, como "tumor-size" é a raíz ele não pode ter menor entropia.

Questão 8

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

R: Letra D

Explicação para a questão

I - Correto. O oversampling aumenta a quantidade de instâncias da classe minoritária, o que pode resultar em overfitting, principalmente quando as instâncias são simplesmente duplicadas. Isso ocorre porque o modelo pode memorizar essas instâncias repetidas, em vez de aprender padrões generalizáveis.

II - Correto. O undersampling reduz a quantidade de instâncias da classe majoritária, o que pode levar à perda de informação relevante, pois algumas instâncias da classe majoritária, que poderiam ser importantes para o modelo, são descartadas.

III - Correto. O SMOTE (Synthetic Minority Over-sampling Technique) é uma técnica de oversampling que cria instâncias sintéticas da classe minoritária, o que pode ajudar a melhorar a generalização do modelo, pois aumenta a diversidade das instâncias sem causar overfitting.

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

R: Letra D

Explicação para a questão

I - Correto. Técnicas de oversampling, como SMOTE, devem ser aplicadas após o split dos dados em treino e teste para evitar data leakage. Isso porque, se o balanceamento for feito antes do split, pode haver uma contaminação dos dados de teste com instâncias sintéticas, o que prejudica a avaliação real do modelo.

II - Correto. O undersampling pode ser útil em conjuntos de dados muito grandes, pois ao reduzir o número de exemplos, diminui o tempo de treinamento, tornando o processo mais eficiente, especialmente quando a classe majoritária é muito predominante e o tempo de treinamento se torna um problema.

III - Correto. Um modelo treinado em dados balanceados geralmente apresenta melhor recall para a classe minoritária, pois o balanceamento ajuda o modelo a aprender melhor as características da classe minoritária, diminuindo o viés em favor da classe majoritária.

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

R: Letra E

Explicação para a questão

I - Incorreto. A imputação pela média ou mediana pode, de fato, distorcer a distribuição original dos dados, especialmente quando há outliers, mas isso não significa que essa técnica de

imputação seja sempre inadequada. Ela é uma escolha simples e prática em muitos casos, embora tenha limitações, especialmente quando os dados são assimétricos ou possuem valores extremos.

II - Correto. A técnica de imputação mais apropriada depende do tipo de variável (numérica, categórica, ordinal) e da natureza do problema. Por exemplo, variáveis numéricas podem ser imputadas com a média ou mediana, enquanto variáveis categóricas podem ser imputadas com o valor mais frequente ou por outras técnicas específicas.

III - Incorreto. Ignorar os valores ausentes e remover diretamente as linhas com dados faltantes pode ser uma prática válida em algumas situações, especialmente se a quantidade de dados ausentes for pequena e sua remoção não afetar substancialmente o desempenho do modelo. A remoção deve ser avaliada com base no contexto e na quantidade de dados faltantes.

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

R: Letra C

Explicação para a questão

I - Correto. Técnicas avançadas como KNN imputation e regressão multivariada consideram a correlação entre variáveis para estimar os valores ausentes. Essas técnicas tentam prever os

valores ausentes com base nas relações entre as variáveis, aproveitando a similaridade entre as instâncias ou as dependências entre as variáveis.

II - Incorreto. A imputação não deve ser aplicada antes do split dos dados em treino e teste. Na verdade, a imputação deve ser feita após o split para evitar vazamento de dados (data leakage). Se a imputação for feita antes do split, os dados de teste podem influenciar a imputação, comprometendo a avaliação real do modelo.

III - Correto. Em pipelines profissionais, a imputação é frequentemente combinada com validação cruzada para garantir que o processo de imputação não cause vazamento de dados. Ao realizar a imputação durante a validação cruzada, cada fold realiza a imputação de maneira independente, evitando que os dados de treino influenciem os de teste.

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

R: Letra A

Explicação para a questão

I - Correto. O Label Encoding pode induzir um modelo a assumir uma ordem inexistente entre categorias, o que pode ser problemático em algoritmos baseados em distância (como KNN) ou regressão. Isso acontece porque o Label Encoding atribui números inteiros para as categorias, o

que pode fazer com que o modelo interprete essas variáveis como contínuas ou ordinais, mesmo que elas sejam nominalmente distintas.

II - Correto. O One-Hot Encoding pode aumentar significativamente a dimensionalidade do conjunto de dados, especialmente quando a variável categórica tem muitos níveis (alta cardinalidade). Cada categoria é representada por uma coluna binária, o que pode resultar em uma grande quantidade de novas colunas e, consequentemente, em um aumento considerável da dimensionalidade.

III - Incorreto. O Frequency Encoding substitui cada categoria pela frequência relativa de sua ocorrência, mas não preserva informações de ordem entre as categorias. Como as categorias são substituídas por números baseados na frequência, o modelo pode interpretar esses valores como ordinais, o que pode não refletir a natureza das categorias, especialmente se a frequência não tiver uma relação de ordem significativa.

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

R: Letra D

Explicação para a questão

I - Correto. O método .fit(X, y) é utilizado tanto no DecisionTreeClassifier quanto no RandomForestClassifier para treinar os modelos com os dados de entrada X e os rótulos y.

II - Correto. Após o treinamento, o método .predict(X_test) é usado em ambos os modelos para prever os valores de teste X_test.

III - Correto. A escolha de criterion='entropy' ou criterion='gini' está disponível apenas para o DecisionTreeClassifier.

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

R: Letra D

Explicação para a questão

I - Correto. O hiperparâmetro max_depth controla a profundidade máxima de cada árvore, tanto no DecisionTreeClassifier quanto no RandomForestClassifier. Ele limita a quantidade de divisões que uma árvore pode fazer.

II - Correto. O parâmetro n_estimators define o número de árvores a serem usadas no RandomForestClassifier. Esse parâmetro controla o tamanho da floresta.

III - Correto. O parâmetro random_state garante a reprodutibilidade dos resultados ao fixar a semente do gerador de números aleatórios, o que assegura que o modelo produza os mesmos resultados em execuções subsequentes.

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.

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

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

R: Letra B

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
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

Letra B

$$Exp \begin{cases} B - \frac{6}{17} & NG - \frac{5}{6} & G - \frac{1}{6} \\ M - \frac{6}{17} & NG - \frac{2}{6} & G - \frac{4}{6} \\ A - \frac{5}{17} & NG - \frac{1}{5} & G - \frac{4}{5} \end{cases}$$

$$Int \begin{cases} B - \frac{10}{17} & NG - \frac{7}{10} & G - \frac{3}{10} \\ A - \frac{7}{17} & NG - \frac{1}{7} & G - \frac{6}{7} \end{cases}$$

$$H \begin{cases} B - \frac{9}{17} & NG - \frac{5}{9} & G - \frac{4}{9} \\ A - \frac{8}{17} & NG - \frac{3}{8} & G - \frac{5}{8} \end{cases}$$

$$E(Classe) = \overset{NG}{\frac{8}{17}} \cdot \log_2\left(\frac{8}{17}\right) - \overset{G}{\frac{9}{17}} \cdot \log_2\left(\frac{9}{17}\right)$$

$$= 0,5117 + 0,4857$$

$$= 0,9975$$

$$E(Exp) = \overset{B}{\frac{6}{17}} \cdot E\left(\frac{5}{6}, \frac{1}{6}\right) + \overset{M}{\frac{6}{17}} \cdot E\left(\frac{2}{6}, \frac{4}{6}\right) + \overset{A}{\frac{5}{17}} \cdot E\left(\frac{1}{5}, \frac{4}{5}\right)$$

$$\underbrace{\frac{6}{17} \cdot \left(\underbrace{-\frac{5}{6} \cdot \log_2\left(\frac{5}{6}\right)}_{0,2192} \underbrace{-\frac{1}{6} \cdot \log_2\left(\frac{1}{6}\right)}_{0,4308} \right)}_{0,2294} + \underbrace{\frac{6}{17} \cdot \left(\underbrace{-\frac{2}{6} \cdot \log_2\left(\frac{2}{6}\right)}_{0,5283} \underbrace{-\frac{4}{6} \cdot \log_2\left(\frac{4}{6}\right)}_{0,3899} \right)}_{0,324} + \underbrace{\frac{6}{17} \cdot \left(\underbrace{-\frac{1}{5} \cdot \log_2\left(\frac{1}{5}\right)}_{0,4643} \underbrace{-\frac{4}{5} \cdot \log_2\left(\frac{4}{5}\right)}_{0,2575} \right)}_{0,2547}$$

$$E(Exp) = 0,8081 \qquad G(Exp) = 0,9975 - 0,8081 = 0,1894$$

$$E(Int) = \frac{10}{17} \cdot E\left(\frac{7}{10}, \frac{3}{10}\right) + \frac{7}{17} \cdot E\left(\frac{1}{7}, \frac{6}{7}\right)$$

$$= \underbrace{\frac{10}{17} \cdot \left(\underbrace{-\frac{7}{10} \cdot \log_2\left(\frac{7}{10}\right)}_{0,3602} \underbrace{-\frac{3}{10} \cdot \log_2\left(\frac{3}{10}\right)}_{0,5208} \right)}_{0,5184} + \underbrace{\frac{7}{17} \cdot \left(\underbrace{-\frac{1}{7} \cdot \log_2\left(\frac{1}{7}\right)}_{0,4011} \underbrace{-\frac{6}{7} \cdot \log_2\left(\frac{6}{7}\right)}_{0,1906} \right)}_{0,2436}$$

$$E(Int) = 0,762 \qquad G(Int) = 0,9975 - 0,762 = 0,2355$$

$$E(H) = \frac{9}{17} \cdot E\left(\frac{5}{9}, \frac{4}{9}\right) + \frac{8}{17} \cdot E\left(\frac{3}{8}, \frac{5}{8}\right)$$

$$= \underbrace{\frac{9}{17} \cdot \left(\underbrace{-\frac{5}{9} \cdot \log_2\left(\frac{5}{9}\right)}_{0,4711} \underbrace{-\frac{4}{9} \cdot \log_2\left(\frac{4}{9}\right)}_{0,599} \right)}_{0,5246} + \underbrace{\frac{8}{17} \cdot \left(\underbrace{-\frac{3}{8} \cdot \log_2\left(\frac{3}{8}\right)}_{0,5306} \underbrace{-\frac{5}{8} \cdot \log_2\left(\frac{5}{8}\right)}_{0,4237} \right)}_{0,4491}$$

$$E(H) = 0,9737 \qquad G(H) = 0,9975 - 0,9737 = 0,0237$$

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%

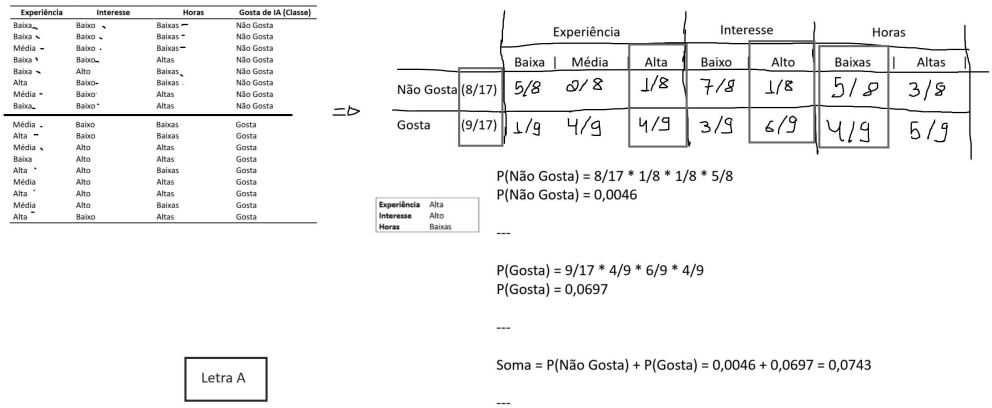
B) 69,32% e 30,68%

C) 74,01% e 25,99%

D) 5,56% e 94,44%

E) 95,32% e 4,68%

R: Letra A



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
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

		Experiência			Interesse		Horas	
		Baixa	Média	Alta	Baixo	Alto	Baixas	Altas
Não Gosta	(8/17)	5/8	2/8	1/8	7/8	1/8	5/8	3/8
Gosta	(9/17)	1/9	4/9	4/9	3/9	6/9	4/9	5/9

Experiência	Alta
Interesse	Alto
Horas	Baixas

P(Não Gosta) = 8/17 * 1/8 * 1/8 * 5/8
P(Não Gosta) = 0,0046

P(Gosta) = 9/17 * 4/9 * 6/9 * 4/9
P(Gosta) = 0,0697

Soma = P(Não Gosta) + P(Gosta) = 0,0046 + 0,0697 = 0,0743

Letra A