Lista #3

Curso: Ciência de Dados

Disciplina: Aprendizado de Máquina I

Profª. Cristiane Neri Nobre

Data de entrega: 22/09

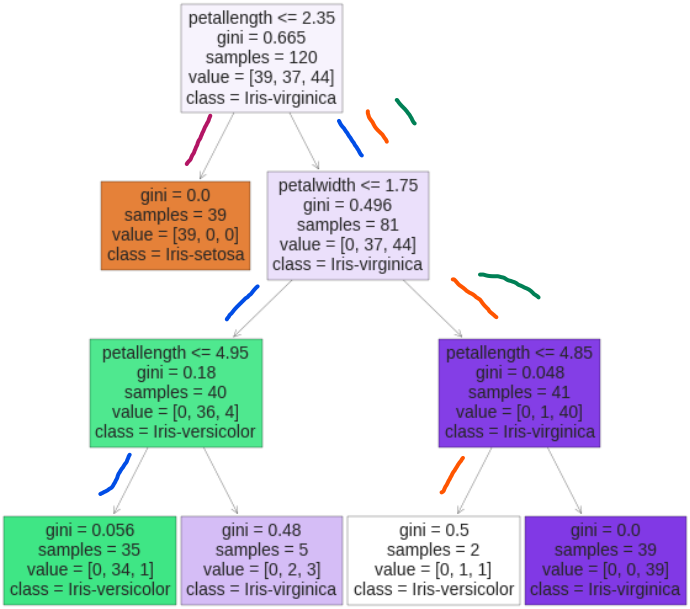
Valor: 2,0 pontos

Nome: Gustavo Costa

Questão 01

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 descritas por quatro atributos: **Tamanho** da Pétala, **largura** da Pétala, **Tamanho** da Sépala e **Largura** da Sépala.

Dado um objeto de classe desconhecida, essa árvore classifica o objeto nas classes: **Iris_Setosa, Iris_Virgínica e Iris_Versicolor**. Esta árvore foi gerada com os hiperparâmetros (DecisionTreeClassifier(criterion='gini', max_depth=3)), usando a linguagem Python.



Com base nestas informações, qual as saídas da árvore para os seguintes **registros de teste**, respectivamente?

Registros de teste	Tamanho da Pétala	Largura da Pétala	Tamanho da Sépala	Largura da Sépala	
Instância 1	3.46	0.87	2.45	1.78	Iris-Versicolor

Instância 2	1.67	1.89	0.78	1.32 Iris Setosa
Instância 3	2.56	2.34	2.45	1.78 Iris Versicolor
Instância 4	6.67	2.34	2.45	1.78 Iris Virginica

a) Iris_Virgínica, íris_Setosa, Iris_Versicolor, Iris_Virgínica
b) Iris_ Setosa, íris_Setosa, Iris_Virgínica , Iris_Versicolor
c) Iris_Versicolor, íris_Setosa, Iris_Versicolor, Iris_Virgínica
d) Íris_Setosa, Iris_Virgínica, Iris_Virgínica , Iris_Versicolor
e) Iris_Versicolor, Íris_Setosa, Iris_Versicolor, Íris_Setosa

Questão 02

Considerando a árvore da questão anterior, e as seguintes afirmações:

I. Esta árvore possui 5 regras de classificação
II. Das regras geradas, há apenas uma com cobertura por classe de 100%
III. A menor cobertura por classe é de 6.8% e corresponde à classe Iris_Virgínica

É **correto** o que se afirma em:

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

Questão 03

Considere a seguinte matriz de confusão obtida por meio do classificador, **Árvore de decisão**, para um problema de quatro classes:

		Foi classificado como A	B	C	D
Era da classe	**A**	10	4	2	1
	B	1	15	2	0
	C	2	3	20	5
	D	4	1	2	50

Quais os valores para as métricas abaixo para cada uma das classes A, B, C e D?

	Precisão	Recall	F1Score	TVP	TFN	TFP	TVN
A	10/17	10/17	0.588	10/17	7/17	7/105	98/105
B	15/ 23	15/18	0.732	15/18	3/18	8/104	96/104
C	20/26	20/30	0.714	20/30	10/30	6/92	86/92
D	50/56	50/57	0.885	50/57	7/57	6/65	59/65

Questão 04

Investigue o funcionamento da métrica GINI utilizada pelo algoritmo CART.

A métrica GINI mede a pureza dos nós quando a árvore vai ser construida. O objetivo do CART é criar divisões nos dados que maximizem a separação das classes nos nós, segue o mesmo sentido do ID3, C4.5. O índice de GINI quantifica o quão "puro" ou "impuro" um conjunto de dados entre as classes. Ele varia de 0 a 0.5, 0.5 representa a máxima impureza.

Questão 05

Faça um resumo dos arquivos que estão no CANVAS sobre etapas de pré-processamento:

1. Parte 1 - Processsamento – Balanceamento
2. Parte 2 - Processamento - Dados ausentes
3. Parte 3 - Processamento - Dados inconsistentes e redundantes
4. Parte 4 - Processamento - Conversão simbólica-numérica
5. Parte 5 - Processamento - Conversão numérico-simbólica
6. Parte 6 - Processamento - transformação de atributos numéricos
7. Parte 7 - Processamento - Redução de dimensionalidade

Importante:

1. Você deve apresentar todas as discussões necessárias para uma completa compreensão do que foi feito, em todas as questões.
2. Adicionar também o link para o código desenvolvido

Parte 1 - Processamento - Balanceamento

Quando a base é desbalanceada, o modelo aprenderá a classe majoritária com mais facilidade. Dito isso, é preciso realizar algumas "manobras" para o modelo não possuir um Viés. Há duas formas de se fazer isso:

1° Forma: Retirar instâncias da classe majoritária (UNDERSAMPLING) -> Retirar de forma aleatória ou utilizar métodos heurísticos para selecionar as melhores, as mais representativas.
2° Forma: Adicionar instâncias da classe minoritária (OVERSAMPLING) -> Adicionar aleatoriamente ou também utilizar métodos heurísticos como o uso do Smote.

Parte 2 - Processamento - Dados Ausentes

Nem sempre é viável apenas descartar os dados ausentes. Dependendo, isso impacta na quantidade de dados totais. Há 4 formas principais de lidar com essa situação:

1° Forma: Tentar preencher as ausências, nem sempre é possível também porque nem sempre é possível ir atrás da fonte dos dados.
2° Forma: Imputar a média para dados numéricos e a moda para dados categóricos. Não é recomendado porque gera muitos dados inconsistentes.
3° Forma: Imputar valores fixos, não é interessante porque você está atribuindo uma informação que não representa aquela instância.
4° Forma: Imputar valores utilizando algoritmos de Machine Learning como KNN Imputer, Miss Forest, etc. É o mais recomendado porque ele encontra as melhores informações por similaridade.

Parte 3 - Processamento - Dados Inconsistentes e Redundantes

Dados inconsistentes são JOGADOS FORA. Não há como escolher uma instância certa.
A redundância são instâncias idênticas. Isso deixa o modelo VICIADO, portanto, precisam ser eliminadas as repetições.

Parte 4 - Processamento - Conversão Simbólica-Numérica

Dados Nominais:
-> Dicotômico (binário): 0 e 1, LabelEncoder
-> Não ordinal com poucas opções: OneHotEncoder
-> Ordinal: LabelEncoder
-> Não ordinal com muitas opções: Criar uma nova codificação, novos atributos que caracterizem um objeto.

Parte 5 - Processamento - Conversão Numérica-Simbólica

O objetivo é reverter a conversão anterior, transformando dados numéricos de volta para representações categóricas. Isso é ser útil para a interpretação de resultados, facilitando a compreensão dos dados e suas implicações.

Parte 6 - Processamento - Transformação de atributos numéricos.

São aplicadas técnicas para modificar atributos numéricos, por exemplo: a normalização, padronização e transformação logarítmica. Essas transformações ajudam a melhorar o desempenho dos modelos, garantindo que as variáveis estejam em escalas comparáveis e reduzindo a influência de outliers.

Parte 7 - Processamento - Redução de Dimensionalidade

A redução de dimensionalidade envolve técnicas como PCA (Análise de Componentes Principais) que visam simplificar conjuntos de dados complexos, mantendo a maior parte da informação relevante