Lista 2 IA

1.

Caminho de cada instância:

Instância 1 : Direita -> Esquerda -> Esquerda -> iris versicolor

Instância 2 : Esquerda -> iris setosa

Instância 3 : Direita -> Direita -> Esquerda -> iris versicolor

Instância 4 : Direita -> Direita -> Direita -> iris virginica

Resposta: Letra C

2.

I. Esta árvore possui 5 regras de classificação - Verdade, existem 5 resultados possíveis nessa árvore de decisão

II. Das regras geradas, há apenas uma com cobertura por classe de 100% - Verdade, apenas a classe da Iris_Setosa possui cobertura de classe de 100%

III. A menor cobertura por classe é de 6.8% e corresponde à classe Iris_Virgínica - Falso, apesar da menor cobertura de classe da Iris_Virginica realmente ser 6.8%, esta não é a menor cobertura de classe. A menor cobertura de classe é de 2.7% da Iris_Versicolor

Resposta: Letra C

3.

	Precisão	Recall	F1Score	TVP	TFN	TFP	TVN
A	0.58	0.58	0.58	0.58	0.41	0.06	0.93
B	0.65	0.83	0.73	0.83	0.16	0.07	0.92
C	0.77	0.66	0.71	0.66	0.33	0.06	0.93
D	0.89	0.87	0.88	0.87	0.12	0.09	0.90

4.

A métrica Gini é utilizada pelo algoritmo CART (Classification and Regression Trees), um algoritmo popular para construção de árvores de decisão. O Índice de Gini mede a impureza de um nó da árvore de decisão, e seu objetivo é ajudar a determinar qual atributo deve ser utilizado para dividir os dados em nós sucessivos.

O algoritmo CART usa o índice de Gini como critério de seleção para dividir os nós em uma árvore de decisão. O índice de Gini avalia quão bem um nó de decisão divide o conjunto de

dados entre as diferentes classes. A ideia é minimizar a impureza em cada divisão, resultando em nós o mais puros possível.

O índice de Gini varia de 0 a 0.5:

- 0: Significa que o nó é puro, ou seja, contém exemplos de apenas uma classe.
- 0.5: Significa que o nó está completamente impuro, ou seja, há uma distribuição balanceada entre as classes.

O algoritmo CART tenta dividir os dados em dois ou mais nós (dependendo do tipo de árvore) e mede a redução da impureza após a divisão. Ele calcula a impureza ponderada dos nós filhos, com base no número de amostras em cada nó:

O CART testa várias divisões possíveis e escolhe aquela que minimiza o índice de Gini da divisão, ou seja, que maximiza a pureza dos nós filhos. Isso é feito ao tentar diferentes variáveis e diferentes pontos de corte (thresholds) para cada variável, buscando a melhor divisão.

O algoritmo CART prefere o índice de Gini porque ele permite tomar decisões rápidas e eficientes sobre a divisão dos dados. A cada nó, a métrica é usada para avaliar várias divisões possíveis, escolhendo a divisão que minimiza a impureza e maximiza a discriminação entre as classes.

5.

Parte 1 - Processsamento – Balanceamento

O problema de dados desbalanceados acontece quando a distribuição das classes em um conjunto de dados é desigual, como em situações onde uma classe, como a de pacientes saudáveis, é muito mais frequente do que a classe de pacientes doentes. Esse desbalanceamento pode prejudicar o desempenho de algoritmos de aprendizado de máquina, que tendem a favorecer a classe majoritária, ignorando a classe minoritária.

Parte 2 - Processamento - Dados ausentes

O tratamento de dados ausentes é essencial para garantir a qualidade dos dados em um conjunto. Esses dados podem estar ausentes por diversos motivos, como problemas de coleta, transmissão, armazenamento ou falhas humanas no preenchimento. Algumas técnicas de aprendizado de máquina conseguem lidar com dados ausentes, mas outras exigem tratamento prévio.

Parte 3 - Processamento - Dados inconsistentes e redundantes

Dados inconsistentes são aqueles que apresentam valores conflitantes nos atributos. Isso pode ocorrer em processos de integração de dados, como o uso de diferentes unidades de medida (metros e centímetros) ou classificações divergentes para instâncias semelhantes.

Parte 4 - Processamento - Conversão simbólica-numérica

Atributos binários: Quando um atributo simbólico tem apenas dois valores (como "sim" ou "não"), ele pode ser facilmente convertido em 0 e 1. Atributos nominais (com mais de dois valores): A conversão utiliza a codificação 1-de-c (ou codificação canônica), onde cada categoria é representada por uma sequência binária exclusiva. Exemplo: as cores "vermelho", "azul" e "verde" são codificadas como três sequências de bits diferentes, com apenas um bit igual a 1 por vez. Atributos ordinais: Para atributos que têm uma ordem natural (como "pequeno", "médio", "grande"), a conversão deve manter essa ordem, podendo usar números inteiros ou reais que refletem a hierarquia.

Parte 5 - Processamento - Conversão numérico-simbólica

A solução é a discretização dos atributos numéricos, onde os valores contínuos são divididos em intervalos que se tornam categorias simbólicas. Existem métodos paramétricos, onde o usuário pode definir parâmetros, como o número de intervalos, e não paramétricos, que utilizam apenas os dados dos atributos. Os métodos de discretização podem ser supervisionados (utilizando a informação sobre a classe dos exemplos) ou não supervisionados. Os supervisionados tendem a produzir melhores resultados, já que consideram as classes ao definir os intervalos.

Parte 6 - Processamento - transformação de atributos numéricos

O texto aborda a transformação de atributos numéricos, que ocorre quando os valores de um atributo variam significativamente ou quando diferentes atributos estão em escalas distintas. Essa transformação é essencial para evitar que um atributo tenha mais influência sobre os outros durante o processamento de dados. Uma técnica comum é a normalização, recomendada quando há grande disparidade nos valores dos atributos

Parte 7 - Processamento - Redução de dimensionalidade

A redução de dimensionalidade ajuda a lidar com conjuntos de dados com muitos atributos, reduzindo a complexidade e melhorando o desempenho dos algoritmos de aprendizado de máquina. Ela pode ser feita de duas formas principais:

-Agregação: Combina atributos em novos, reduzindo o número total, mas podendo perder informações.

-Seleção de atributos: Mantém apenas os atributos mais relevantes e descarta os demais.