Lista 2 - Arthur Gonçalves de Moraes

1)
1 - dir, esq, esq = versicolor
2 - esq = setosa
3 - dir, dir, esq = versicolor
4 - dir, dir, dir = virginiana
c)

2)
I) V. A árvore possui 5 folhas
II) V. Somente a Iris_Setosa tem cobertura de 100%
III) V. A classe Iris_Vigínica tem cobertura de 6,8%, mas não é a menor cobertura (Iris_Versicolor 2,7%)
c)

3)

	Precisão	Recall	F1 Score	TVP	TFN	TFP	TVN
A	10/17	10/17	10/17	10/17	7/17	7/105	98/105
B	15/23	15/18	30/41	15/18	3/18	8/104	96/104
C	20/26	20/30	5/7	20/30	10/30	6/92	86/92
D	50/56	50/57	100/113	50/57	7/57	6/65	59/65

*TVP = acertos da classe / total classe
*TFN = erros da classe / total classe
*TFP = classificados errado de outras classes / total - classe
*TVN = classificados corretamente de outras classes / total - classe
*Precisão = VP/(VP+FP)
*Recall = VP/(VP+FN)
*F1 Score = (2*recall*precisão)/(recall+precisão)

4)
4.1)
- Dados desbalanceados podem causar problemas de classificação no modelo, favorecendo a classificação da classe majoritária
- Soluções:
  - Alterar o tamanho do conjunto de dados, removendo ou adicionando instâncias às classes majoritária ou minoritária (oversampling e undersampling)
  - Utilizar diferentes custos de classificação
  - Aprendizado separado para cada classe

4.2)
- Dados ausentes podem ser causados por erros no equipamento de coleta, transmissão ou armazenamento, assim como pela inexistência de certo parâmetro em consultas anteriores ou por não terem sido informados

- Soluções:
  - Remover instâncias com dados ausentes
  - Preencher valores manualmente
  - Utilizar métodos para atribuir valores aos campos faltantes (moda, média, indução)
  - Utilizar algoritmos que lidam com dados ausentes

4.3)

- Dados inconsistentes são gerados no processo de integração de conjunto de dados e podem ser classificações diferentes para instâncias idênticas ou diferentes unidades de medida para o mesmo atributo
- Dados redundantes são causados por problemas na coleta, na entrada, no armazenamento, na integração ou na transmissão de dados e podem ser relacionados à instâncias ou atributos, sendo valores iguais, muito parecidos ou que podem ser induzidos por outros atributos
- Soluções:
  - Remoção

4.4)

- Para atributos com somente 2 opções utiliza-se 1 dígito binário
- Para atributos não oridinais utiliza-se um número C de bits, onde cada posição do valor 1 indica uma opção diferente, ou divisão em pseudoatributos caso a quantidade de opções seja alta
- Atributos ordinais são substituídos respeitando sua ordenação, normalmente usando valores inteiros ou reais. Caso seja necessário converter valores ordinais em valores binários, pode ser utilizado o código cinza ou o código termômetro.

4.5)

- Alguns algoritmos foram feitos para trabalhar com valores qualitativos
- Nesses casos, os valores quantitativos devem ser discretizados, utilizando técnicas supervisionadas (melhores resultados) ou não

4.6)

- Atributos podem ter limites inferiores e superiores muito diferentes ou utilizarem escalas diferentes
- Uma técnica muito utilizada é a normalização:
  - Por amplitude:
  Por reescala: Vnovo = min + (Vatual - menor)/(maor - menor)(max-min), sendo max e min os valores de máximo e mínimo desejados;
  *Por padronização: Vnovo = (Vatual - u) / o , onde u = média e o = desvio padrão

4.7)

- Número elevado de atributos causam problemas nos modelos (maldição da dimensionalidade)
- A combinação ou eliminação de atributos traz benefícios de desempenho, custo computacional e compreensão dos resultados

- Soluções:
  - Agregação: substituição de atributos por um único dado pela combinação desses. Ocasiona a perda dos valores originais
  - Seleção: elimina os atributos não selecionados. Traz muitos benefícios para o modelo por deixar o conjunto de dados mais específico
  - A seleção pode ser feita utilizando:
  *Embutido: a seleção é embutida ao próprio algoritmo
  *Baseado em filtro: um filtro é utilizado no conjunto de dados em uma etapa de pré-processamento
  *Baseada em wrapper: utiliza o próprio algoritmo como uma caixa preta, analisando a redução da taxa de erro em relação à redução de atributos