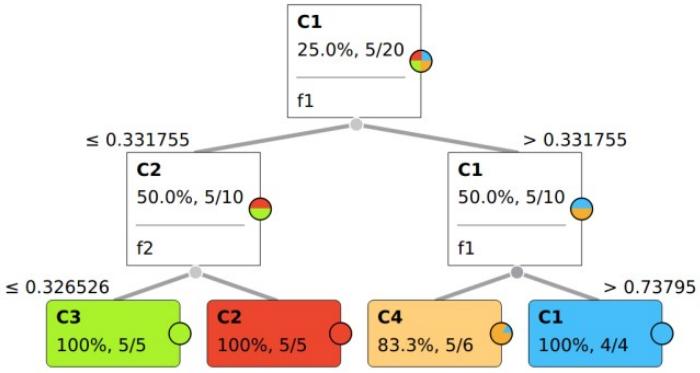
Lista de exercícios – Ciência dos Dados

1. Em Aprendizado de Máquina, árvore de decisão é uma das abordagens de modelagem preditiva mais utilizadas. Esse tipo de abordagem indutiva utiliza uma árvore de decisão (como modelo preditivo ou hipótese) para inferir, a partir de valores de atributos de exemplos (representados nos ramos), valores de uma variável alvo (representados nas folhas), podendo resolver problemas de classificação e regressão.

Considerando a hipótese abaixo evidenciada, marque a alternativa que melhor descreve as características do problema.



a) O problema em questão é de classificação, tem quatro *features* e o modelo foi treinado com 20 exemplos. Apesar de ter quatro *features*, apenas duas foram utilizadas para construir a fronteira de decisão.

b) O problema tem pelo menos 4 classes e a hipótese foi construída a partir dos valores de duas *features (f1 e f2)*, não sendo possível determinar se mais *features* estavam presentes nos exemplos.

c) O problema em questão é de regressão e tem quatro *features*, no entanto, apenas duas foram utilizadas para construir a fronteira de decisão.

d) O problema em questão é de regressão e a hipótese foi construída a partir dos valores de duas *features (f1 e f2)*, não sendo possível determinar se mais *features* estavam presentes nos exemplos.

e) O problema tem pelo menos 4 classes e a hipótese foi construída a partir dos valores de duas *features (f1 e f2)*. Pode-se garantir que não haviam outras *features* nos exemplos.

2) A validação cruzada é uma das várias técnicas de validação de modelos para verificar como os resultados de uma análise estatística serão generalizados para um conjunto de dados independente. Nesse processo exemplos do conjunto de treinamento são avaliados com a garantia que os mesmos são desconhecidos para o modelo.

A respeito da quantidade de exemplos de um conjunto de treinamento que são avaliados em uma validação cruzada, assinale a afirmativa correta:

a) No método de validação cruzada todos os exemplos do conjunto de treinamento são avaliados.
b) No método de validação cruzada apenas os exemplos de uma das *folds* do conjunto de treinamento são avaliados.
c) No método de validação cruzada os exemplos de uma das *folds* do conjunto de treinamento não são avaliados.
d) No método de validação cruzada nenhum dos exemplos do conjunto de treinamento é avaliado.

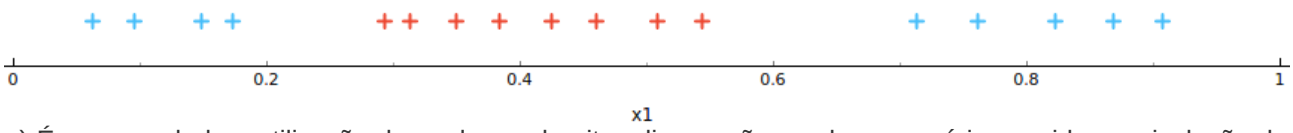
3. Qual dos parâmetros abaixo não é do algoritmo de Árvore de Decisão:

a) Profundidade máxima.

b) Quantidade mínima de exemplos nos folhas.

c) Número de árvores.

4. Considerando os dados ilustrados na figura abaixo, que possuem apenas uma *feature* (x1) e duas classes (cruz azul e cruz vermelha), marque a alterativa que considera mais correta a respeito da definição de uma fronteira de decisão para os mesmos.



a) É recomendada a utilização de qualquer algoritmo linear, não sendo necessário considerar a inclusão de outras *features*.

b) Classificadores de Árvores de Decisão conseguiriam estabelecer essa fronteira, mas seria necessário incluir outras *features*.

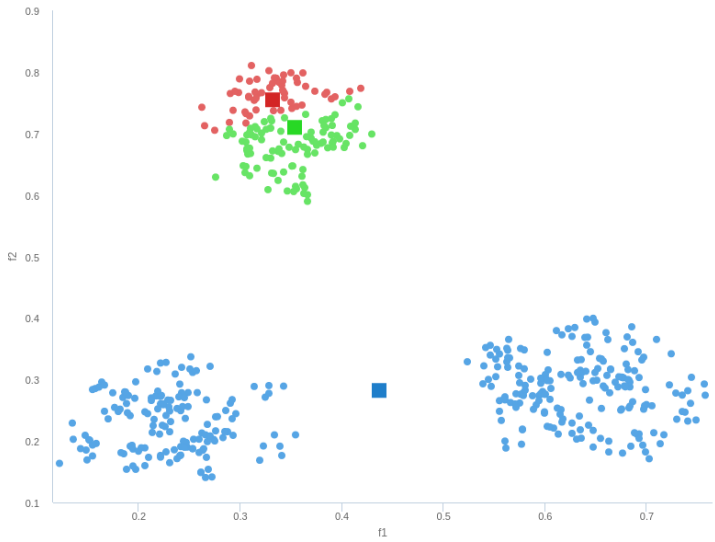
c) Pode-se estabelecer essa fronteira utilizando algoritmos lineares, mas seria necessário incluir pelo menos mais uma *feature (x2)*, aplicando a mesma uma função quadrática a partir de x1. Exemplo: x2 = f(x1) = x1^2.

d) Classificadores de Árvores de Decisão conseguiriam estabelecer essa fronteira, mas seria necessário incluir pelo menos mais uma *feature (x2)*, aplicando a mesma uma função quadrática a partir de x1. Exemplo: x2 = f(x1) = x1^2.

e) Não é possível estabelecer essa fronteira de decisão utilizando os algoritmos de Aprendizado de Máquina estudados na disciplina até o momento.

5) Na figura abaixo é ilustrado um espaço de observações com duas dimensões (f1 e f2) em que claramente parecem existir 3 *clusters*. Foi inicializada a execução do algoritmo K-Means para k=3 e de forma aleatória os quadrados, que representam os 3 *centroids*, foram posicionados no espaço bidimensional.

A respeito do cenário descrito acima, assinale a afirmativa que melhor define o resultado que o algoritmo vai obter para essa execução.



a) Serão encontrados três *clusters*, mas o resultado não vai parecer satisfatório, pois o algoritmo vai dividir observações que aparentam estar em apenas um *cluster* em dois *clusters*.

b) Serão encontrados três *clusters* e o resultado vai parecer satisfatório.

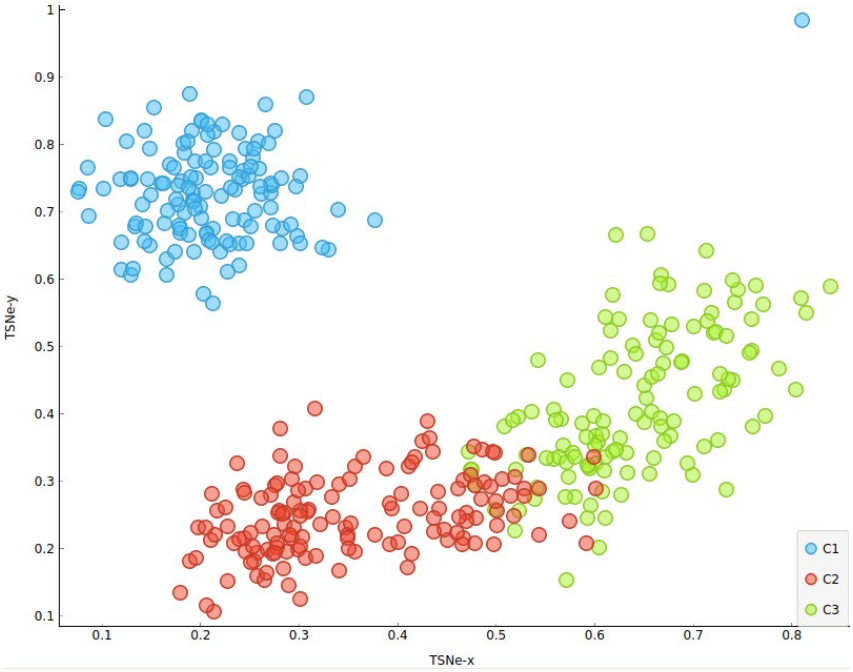
c) Serão encontrados dois *clusters, mas o resultado não vai parecer satisfatório, pois o algoritmo vai dividir observações que aparentam estar em apenas um cluster em dois clusters.*

d) Serão encontrados dois *clusters* e o resultado vai parecer satisfatório.

6. A respeito de modelos *Ensemble*, descreva suas principais vantagens e cenários mais indicados de aplicação.

7. O T-distributed Stochastic Neighbor Embedding (t-SNE) é um algoritmo de aprendizado de máquina para visualização de dados desenvolvido por Laurens van der Maaten e Geoffrey Hinton. É uma técnica de redução de dimensionalidade não linear, bem adequada para incorporar dados de alta dimensão para visualização em um espaço de baixa dimensão de duas ou três dimensões. Especificamente, ele modela cada objeto de alta dimensão por um ponto bidimensional ou tridimensional, de modo que objetos semelhantes sejam modelados por pontos próximos e objetos diferentes sejam modelados por pontos distantes com alta probabilidade.

A respeito do espaço de baixa dimensão (duas dimensões) ilustrado abaixo para um problema de classificação (com três classes C1, C2 e C3), obtido a partir da aplicação de t-SNE em um espaço com 2048 dimensões, marque o item que melhor descreve as conclusões que podem ser tiradas do resultado dessa redução de dimensionalidade.



a) Modelos de classificação terão dificuldade de prever exemplos da classe C1, pois ela possui um outlier.

b) Modelos de classificação não são capazes de definir as fronteiras de decisão para esse problema, pois C2 e C3 não são separáveis.

c) É aceitável que modelos de classificação tenham dificuldade de separar exemplos de C2 e C3, mas espera-se que no caso de C1, todas as previsões sejam acertadas em uma validação cruzada.

d) É aceitável que modelos de classificação tenham dificuldade de separar exemplos de C2 e C3, mas espera-se que no caso de C1, a maior parte das previsões sejam acertadas em uma validação cruzada.

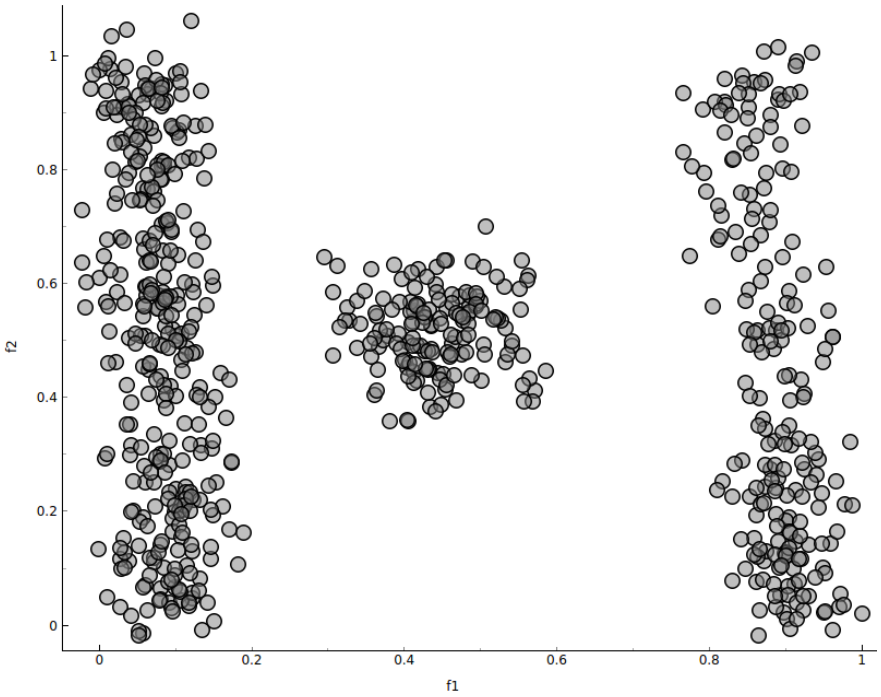
e) É aceitável que modelos de classificação tenham dificuldade de separar exemplos de C2 e C3, e espera-se que no caso de C1, nenhuma das previsões sejam acertadas em uma validação cruzada.

8) Assinale os algoritmos abaixo que podem ser considerados determinísticos?

( ) K-Means ( ) Regressão Linear ( ) Regressão logística ( ) DBSCAN

9) O algoritmo k-means é um método de quantização vetorial, originalmente do processamento de sinais, que visa particionar *n* observações em *k clusters* nos quais cada observação pertence ao *cluster* com a média mais próxima (centros de *cluster* ou *centróide* de *cluster*), servindo como um protótipo do *cluster*. É muito popular para análise de *cluster* em mineração de dados.

A respeito do espaço de observações ilustrado abaixo (com duas *features* f1 e f2), um especialista de domínio havia mencionado anteriormente que existem três grandes grupos para esses dados, mas que dois deles eram bem distintos entre si. Marque a alternativa que melhor descreve as conclusões que podem ser tiradas a partir da observação desses dados.



a) O algoritmo K-Means (k=3) é recomendável para esse tipo de problema.

b) O algoritmo K-Means (k=2) é recomendável para esse tipo de problema.

c) O algoritmo K-Means não é adequado para esse tipo de problema, dessa maneira eu devo utilizar outro algoritmo.

10. Bancos e operadoras de cartões de crédito estiveram entre os primeiros a usar a aprendizagem de máquina. Eles costumam usar a tecnologia para identificar transações que podem ser fraudulentas. Se a sua operadora de cartão de crédito o telefonar para validar uma compra específica que você tenha feito recentemente, a empresa provavelmente usou o aprendizado de máquina para sinalizar uma transação suspeita em sua conta.

A partir do contexto apresentado acima, descreva como o mesmo poderia ser modelado como um problema de Aprendizado de Máquina. Explique quais seriam as Classes, Exemplos e possíveis atributos.