Lista de Exercícios

Data Science

UCAM – 2021 – 1

1. Quais são os 5 vs principais do *Big Data*?

   Volume, Velocidade, Variedade , Veracidade , Valor e Variabilidade

2. Em projetos de big data geralmente existe um dos 5 V's que se destacam. Nos projetos abaixo quais são aqueles que a velocidade é a principal característica?

   A. Dados da disseminação da covid-19; dados da temperatura do núcleo de uma usina nuclear;

   B. Dados de mercado financeiro; dados de provas online;

   C. Dados de vendas em um erp; dados de umidade de solo;

   D. Dados de batimento cardíaco; dados de peso de uma pessoa;

   E. Dados de vacinação da covid-19; dados de vendas mensais;

3. Quais são as fases do ciclo de vida dos dados?

   Coleta, Armazenamento, Recuperação e Descarte

4. A governança de dados aplicada pelo framework *DAMA-DMBOK* introduz vários pilares que regimentam o uso de dados.
Em um projeto de big data um cientista de dados verifica que existem duas bases de dados, a primeira possui informações sobre vendas de produtos e a segunda sobre clientes.
As bases de dados possuem dados completos, sem erros ou falta de dados parciais e com o campo id_cliente semelhantes entre elas.
Quais pilares estão sendo avaliados no contexto descrito no enunciado?

   A. Qualidade e integração

   B. Qualidade e privacidade

   C. Qualidade e segurança

   D. Integração e privacidade

   E. Integração e direitos autorais

5. Dado um conjunto de dados de vendas de imóveis contendo:
Variável dependente: valores de vendas
Variáveis independentes: área; número de quartos; número de banheiros;
Qual tipo de aprendizado de máquina ideal para elaboração de uma análise de predição de valores?

A. Aprendizado supervisionado

B. Aprendizado não supervisionado

C. Aprendizado por reforço

D. Aprendizado semi-supervisionado

E. Nenhuma das outras alternativas

6. Quais são as etapas do ciclo de vida de um projeto de Big Data indicado no estudo de FAYYAD et al., 1996?

Seleção, Processamento , Transformação, Mineração de dados e Interpretação e Avaliação

7. Um hospital está conduzindo um estudo para diagnosticar câncer a partir de sintomas de seus pacientes.
Um médico recém-formado e que não possui especialização na doença identificou os sintomas e os resultados dos diagnósticos.
Nesse contexto, pode-se afirmar:

A. Nenhuma das outras alternativas

B. Os dados coletados são ideais para realização de um estudo utilizando um algoritmo de classificação

C. Os dados coletados são ideais para realização de um estudo utilizando um algoritmo de regressão

D. Os dados coletados são ideais para realização de um estudo utilizando um algoritmo de clusterização

E. Os dados coletados são ideais para realização de um estudo utilizando um algoritmo de associação

8. Considerando o seguinte dataset:

Idades = [4, 10, 20, 21, 26, 35, 48, 45, 49, 15, 11, 18, 17, 5, 2, 14, 88, 79, 55, 64, 47, 49]

Qual o valor da média, mediana e moda?
Media: 32,8 / Mediana: 23,5 / Moda: 49
Qual o valor da amplitude, variância e desvio padrão?
Amplitude: 86 / Variância: 609,20 / Desvio Padrão: 24,68

A resolução deve ser elaborada utilizando Python;
Construa um dataframe em Pandas a partir da lista;

9. Em um projeto de big data construiu-se um algoritmo capaz de coletar comentários de redes sociais. Com quais tipos de dados esse projeto necessita lidar? Dados Não Estruturados

10. Uma empresa de e-mails está sofrendo reclamações constante de seus clientes devido a quantidade de spans recebidos. Essa portanto, toma a decisão de criar um sistema de detecção de *SPAN*.
A empresa já possui um histórico de e-mails identificados como *SPAN*.
Baseado no contexto acima, quais tipos de treinamentos e algoritmos recomendados para criação de um sistema de identificação de *SPAN*? Justifique sua resposta. Aprendizado Supervisionado. Algoritmos de classificação:KNN, Árvores de Decisão, SVM, Regressão Logística.

11. Utilizando o dataset abaixo, faça:

- Ler o arquivo CSV contendo os dados;
- Criar funções para conversão dos dados de valores para float;
- Criar funções para correção das datas para um formato aceitável pelo Python;
- Informar o somatório de vendas total;
- Informar a média de venda mensal de vendas;
- Criar uma coluna com a representatividade (porcentagem) de venda de cada mês em relação ao total das vendas;
- Criar uma coluna com a representatividade (porcentagem) acumulada de venda de cada mês em relação ao total das vendas;
- Criar gráfico de barras contendo a venda de cada mês;
- Utilizando o gráfico acima, crie um novo gráfico onde além das barras haverá uma linha indicando a média das vendas;

MÊS_ANO	VALOR_VENDA_MENSAL
jan/20	R$ 11.007.043,23
fev/20	R$ 10.931.711,78
mar/20	R$ 14.825.160,14
abr/20	R$ 13.858.179,77
mai/20	R$ 14.097.146,98
jun/20	R$ 15.114.495,86
jul/20	R$ 14.359.248,57
ago/20	R$ 15.094.411,39
set/20	R$ 16.055.703,02
out/20	R$ 15.990.863,10
nov/20	R$ 16.137.095,80
dez/20	R$ 18.409.089,37

OBSERVAÇÕES DO PROFESSOR:

A LISTA NÃO É INDIVIDUAL, PORTANTO, VOCÊS PODEM TRABALHAR EM EQUIPE PARA SUA RESOLUÇÃO. LEMBREM-SE: DIVIDIR PARA CONQUISTAR É UMA ÓTIMA TÁTICA E, NO DESENVOLVIMENTO DE SOFTWARE, NÓS NUNCA TRABALHAMOS SOZINHOS

PARA AQUELES QUE ESTÃO COM DIFICULDADES E NÃO ENTREGARAM NENHUM TRABALHO ATÉ O MOMENTO, SUGIRO MUITO DEDICAR-SE A RESOLUÇÃO DA LISTA.

Qual deles você quer ser?!