UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
UNIDADE ACADÊMICA ESPECIALIZADA EM CIÊNCIAS AGRÁRIAS
CURSO DE ANÁLISE E DESENVOLVIMENTO DE SISTEMAS
COMPONENTE CURRICULAR:
FUNDAMENTOS E TÉCNICAS EM CIÊNCIAS DE DADOS
Prof. Josenalde Oliveira

Lista de Exercícios 2

1) Com relação aos tipos de bancos NoSQL, analise as afirmativas a seguir:

I. Bancos de dados de documentos armazenam dados como documentos (JSON, XML, etc.). Um exemplo de banco deste tipo é o MongoDB.
II. O banco de dados Neo4J é de um tipo de banco que possuem vértices e arestas representando as relações entre esses vértices.
III. Banco de dados colunares guardam colunas juntas, ao invés de linhas, sendo o tipo de banco do Neo4J.

Estão corretas as afirmativas:
a) I, III
b) I, II
c) II, III
d) I, II, III

2) Considerando os comandos disponíveis no shell do MongoDB e uma coleção de dados alunos não vazia, com atributos nome, mediaGeral e avaliacao, analise as seguintes afirmativas:

I. O seguinte comando retorna uma lista vazia, uma vez que os critérios de busca não foram definidos:
`db.alunos.find( {} )`
II. No comando abaixo, o programador colocou erroneamente o `$in` no lugar do `$or` para encontrar alunos baseado nos valores de sua nota média geral:
`db.alunos.find( { mediaGeral: { $in: [ 7, 10 ] } } )`
III. Para atualizar na coleção alunos o atributo avaliação de todos os alunos com média maior que 9, podemos executar o seguinte comando:
`db.alunos.updateMany( { mediaGeral: { $gt: 9 } }, { $set: { avaliacao: "Ótimo desempenho!" } } )`
IV. Para inserir um novo registro na coleção "alunos", podemos executar o seguinte comando:
`db.alunos.insertOne( { nome: "João", mediaGeral: 7, avaliacao: "Na média" } )`

Estão corretas as afirmativas:
a) I e II
b) II e III
c) III e IV
d) I e IV
Sugestão: crie um banco de dados no mongoDB chamado escola, com a coleção alunos, e insira alguns documentos com as chaves nome, mediaGeral e avaliação, para testar os comandos acima e validar suas respostas.

3) Classifique as bibliotecas abaixo em função de sua maior aplicabilidade/direcionamento, conforme classes abaixo:

a) visualização, plotagem

b) computação científica
c) aquisição, tratamento e análise/consulta de dados
d) aprendizagem de máquina
e) big data

( ) Pandas ( ) SciPy ( ) Spark ( ) pyTorch ( ) NumPy
( ) scikit-Learn ( ) Hive ( ) matplotlib ( ) seaborn ( ) pymongo
( ) tensorFlow ( ) Selenium

4) Entre as funcionalidades do NumPy, podemos destacar:
a) tratamento flexível para dados ausentes
b) melhor performance em seus arrays do que os tipos primitivos de Python
c) facilita agregação de dados
d) é um concorrente do Pandas

5) Qual dos itens a seguir NÃO é característica do ambiente Jupyter:
a) permite execução iterativa por meio do IPython
b) é baseado no modelo de desenvolvimento edição-compilação-execução
c) permite integrar equações em LaTex e tags HTML em células Markdown
d) O Google Colab é baseado em Jupyter

6) Sobre arquivos de imagem enquanto fontes de dados, podemos afirmar que:
a) são dados semi-estruturados, de esquema flexível
b) dado ser um dado binário, embora ineficiente, pode ser salvo como um tipo binData no mongodb
c) são dados não estruturados, com um esquema rígido
d) em nenhuma hipótese necessitam ser transformados em dados estruturados para análise, após o PDI

7) Quais são os 3Vs mais importantes do Big Data?
a) volume, velocidade, viabilidade
b) volume, velocidade, variedade
c) valor, velocidade, viabilidade
d) velocidade, variedade, valor

8) Em um banco de dados o que é *missing data*, usualmente preenchido com NaN ao importar dados para um dataframe em Pandas?
a) o mesmo que outlier
b) dados faltantes, toda e qualquer falha na obtenção de respostas sobre os elementos selecionados e designados para pertencerem à amostra
c) ocorre quando não há falta de informação no banco de dados
d) são dados antigos e ultrapassados

9) Qual seria uma boa definição para cientista de dados:
a) alguém com competências em programação e estatística, sem necessariamente conhecimento do negócio
b) alguém que embora conheça o negócio muito bem e programe tão bem quanto, não conhece técnicas estatísticas nem interpreta minimamente os resultados
c) alguém que equilibra a seleção de técnicas de programação e técnicas estatísticas, aplicadas a um negócio com o qual procura interagir com especialistas para melhor proposta de solução
d) alguém que domina os bancos de dados NoSQL

10) Sobre conceitos gerais da área de ciências de dados, marque V ou F. Quando for F, justifique:

( ) aprendizagem de máquina (ML – Machine Learning) e mineração de dados (DM – Data Mining) podem ser consideradas a mesma coisa no âmbito de ciência de dados
( ) técnicas como clusterização, detecção de anomalias e classificação são normalmente associadas ao resultado de técnicas de ML
( ) uma nuvem de palavras é uma técnica gráfica simples de realizar análise exploratória de dados
( ) a ciência de dados não é apenas ML e DM, sendo estas parte da etapa de transformação no ciclo de vida do dado
( ) O termo ETL (Extração, Transformação e Carga) está presente na etapa de produção do ciclo do dado, sendo processo comum em data warehouses
( ) Dashboard é uma técnica de visualização de resultados
( ) Estão entre os motores para o desenvolvimento da ciência de dados: IaaS, PaaS e SaaS
( ) O MongoDB Atlas é um exemplo de DBaaS no contexto cloud computing
( ) O sistema de arquivos distribuído HDFS é a base para a plafatorma Hadoop e para o Google File System

11) Para indexação de páginas web, o Google usa o MapReduce. Sobre a ação de Map, pode-se afirmar que:
a) agrega os resultados, gerando um resultado final
b) é executada em nós distribuídos, contabilizando e separando itens comuns
c) é executado em memória, tal como o Spark
d) não pode ser implementado em mongoDB

12) Tenho um problema com variável resposta numérica e variável independente numérica. Qual técnica não se aplica para a construção de um modelo de predição?
a) regressão linear
b) rede neural
c) random forest
d) regressão logística

13) Uma análise de variância seria aplicável à variáveis de entrada do tipo:
a) categóricas
b) numéricas
c) mistas
d) n.d.a

14) São variáveis quantitativas discretas (pode haver mais de uma opção):
a) números de filho de um casal
b) altura de uma pessoa
c) número de clientes num banco
d) número de poltronas num cinema

15) Na equação Y = ax + b, Y pode ser chamada de :
a) variável objetivo, variável instrumental, variável resposta, variável dependente
b) variável discordiana, variável pergunta, variável dependente
c) variável objetivo, variável target, variável resposta e variável dependente

16) são variáveis qualitativas ordinais:
a) classe social
b) voltagem elétrica

c) cor dos olhos
d) time de futebol

17) Sobre a etapa de modelagem, um modelo pode ser construído vários objetivos, menos o de:
a) ordenação
b) estimativa
c) previsão
d) decisão

18) Sobre modelos de armazenamento, julgue os itens a seguir em V e F. Se for F, justifique:
( ) o modelo baseado em grafos atual remete ao modelo hierárquico dos anos 60
( ) o modelo 'não apenas SQL' (No SQL) restringe o modelo relacional por não permitir relacionamentos
( ) o modelo chave-valor é baseado em tabelas hash e um de seus bancos é o Redis
( ) o modelo colunar Tall-narrow (TN) codifica o Timestamp binário no ID e possui poucas linhas e muitas colunas
( ) o DynamoDB é um exemplo famoso de banco baseado em grafos
( ) os modelos baseados em grafos são interessantes e bem aplicáveis em semântica web
( ) uma família de colunas no HBase pode ser extraída a partir de um RowKey
( ) a linguagem de consulta cypher é a base para extrações no mongoDB
( ) Um TimeStamp é um valor de 64 bits no mongoDB e permite registrar os milissegundos desde 01.01.1970 até o instante da persistência do dado
( ) Um dicionário Python é um tipo admissível para persistência direta no mongoDB
( ) Uma dicionário em Python pode ser transmitido sem transformação através do MQTT
( ) o MongoDB possui o limite de 100 níveis de documentos aninhados, com 16MB por documento
( ) Para documentos > 16 MB, uma solução é usar o GridFS do mongoDB, que divide o documento em coleções