**PROGRAMA DE RESIDÊNCIA EM TECNOLOGIA DA INFORMAÇÃO**
**TRIBUNAL REGIONAL ELEITORAL DO RIO GRANDE DO NORTE**
**EDITAL 001/2022 – PROVA DE CONHECIMENTOS ESPECÍFICOS**

**ÁREA DE CONCENTRAÇÃO 2 – BUSINESS INTELLIGENCE E ANALYTICS**

Candidato: _____
CPF: _____ Telefone: _____

---

**QUESTÕES**

**01.** Ao utilizar uma base de dados, é comum que muitos dos dados não estejam presentes. A falta de alguns valores nas bases de dados se deve, muitas vezes, por falta de preenchimento nos cadastros, falha em sensores, entre outros motivos. Para que técnicas de aprendizado de máquina possam utilizar corretamente os dados, em geral, é necessária a correção desse tipo de problema. Dentre as abordagens utilizadas para tratar esse problema, analise as seguintes afirmativas:

I) É possível eliminar os exemplos que possuem valores faltando em seus atributos.
II) Pode-se substituir o campo faltoso por valores retirados a partir dos próprios dados, como a média ou a moda.
III) É possível utilizar um modelo de classificação para preencher automaticamente os dados.

Sobre as afirmativas anteriores, é correto dizer que:
A) Somente a afirmativa III está correta.
B) Todas as afirmativas estão corretas.
C) Nenhuma afirmativa está correta.
D) Somente a afirmativa I está correta.

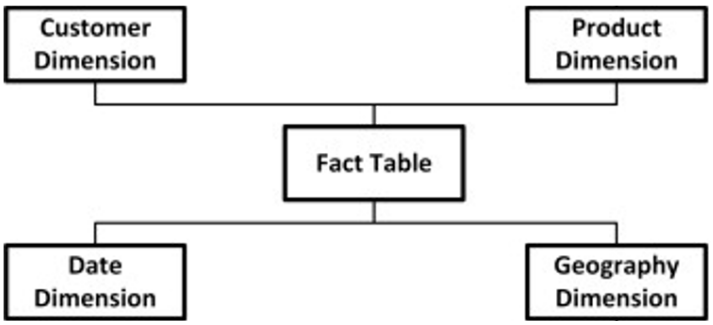
---

**02.** Ao se preparar para construir um *data warehouse* para uma grande empresa, é comum analisar diferentes abordagens para auxiliar esse processo, como as *bottom-up* e a *top-down*. Sobre os benefícios e limitações da abordagem *top-down* em relação à bottom-up, analise as seguintes afirmativas:

I. Provê um armazenamento central e único de dados.
II. Tende a levar mais tempo para ser construído.
III. Baixo risco de falhar.

Estão corretas as afirmativas:
A) I e II, apenas.
B) I e III, apenas.
C) II e III, apenas.
D) I, II e III.

**03.** Considere a seguinte imagem relativa a modelagem Esquema Estrela de um dado *data warehouse* de vendas.




Acerca dessa modelagem, assinale a alternativa que mantenha a modelagem conceitualmente correta:
A) Coloca-se na dimensão `Product` a informação de quantos produtos foram vendidos.
B) Coloca-se na dimensão `Date` informações agregadas como mês e ano, mas não o dia exato da venda.
C) A dimensão `Customer` não precisa de um identificador (chave primária), mas apenas os dados do cliente.
D) A dimensão `Geography` não possui chave estrangeira.

---

**04.** Dentre as abordagens para se realizar uma modelagem dimensional estão a modelagem Esquema Estrela e a Floco de Neves. Acerca das características da modelagem Floco de Neves quando comparada ao Esquema Estrela, analise as seguintes afirmativas:

I. Menor quantidade de dados duplicados.
II. Menor quantidade de registros na tabela fato.
III. Maior quantidade de junções de tabelas em certas consultas.

Estão corretas:
A) I e II, apenas.
B) I e III, apenas.
C) II e III, apenas.
D) I, II e III.

---

**05.** Acerca da modelagem dimensional, analise as seguintes afirmativas:

I. Fatos representam aspectos de um evento como quando e quanto.
II. Dimensões representam aspectos de um evento como quem, como e onde.

Assinale a alternativa correta acerca da veracidade das afirmativas I e II, respectivamente:
A) Falsa, Falsa.
B) Verdadeira, Falsa.
C) Falsa, Verdadeira.
D) Verdadeira, Verdadeira.

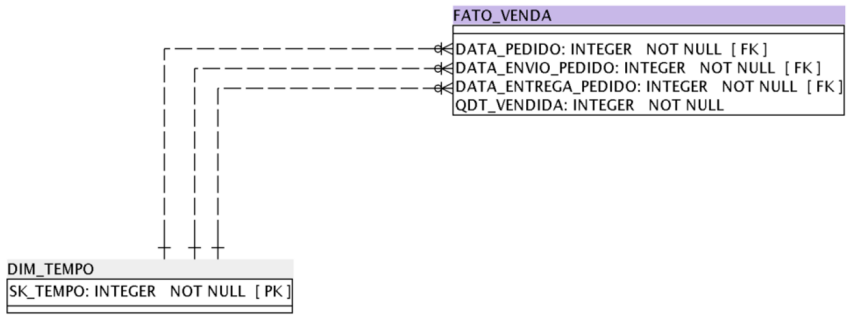
**06.** Esquema estrela, ETL e *drill down* podem, respectivamente, ser definidos como:
A) Técnica de modelagem multidimensional, sigla em inglês para ambiente de teste e carga, técnica para criação de cubos.
B) Técnica de projeto de banco de dados, linguagem para manipulação de dados de bancos multidimensionais, técnica de mineração de dados.
C) Técnica de otimização de banco multidimensionais, sigla em inglês para processo de extração, transformação e carga de dados, operação OLAP para mostrar o detalhe dos dados (ir num nível abaixo da hierarquia da dimensão que se está observando).
D) Técnica de modelagem multidimensional, processo de especificação, transferência e limpeza de dados, operação OLAP para agregar os dados (subindo num nível da hierarquia da dimensão que se está observando).

---

**07.** *On-Line Analytical Processing* (OLAP) é uma categoria de tecnologia de *software* que habilita uma análise de dados rápida, consistente e interativa. Acerca dos conceitos relacionados a essa categoria de tecnologia, selecione a alternativa correta:
A) ROLAP é uma das principais alternativas ao MOLAP.
B) O modelo MOLAP é utilizado para se mapear dados permitindo operações OLAP em um banco de dados relacional.
C) A operação de *drill-down* abstrai detalhes e agrega dados a partir de categorias.
D) A operação de *roll-up* apresenta níveis mais detalhados de informação.

---

**08.** Considerando uma modelagem dimensional e a figura abaixo, podemos afirmar que `FATO_VENDA` possui quantas dimensões?




A) Uma única dimensão.
B) Duas dimensões.
C) Três dimensões.
D) Quatro dimensões.

---

**09.** Na literatura, é possível encontrar os termos *data warehouse* e *data mart*. Acerca desses termos, selecione a alternativa correta:
A) *Data mart* é um sinônimo de *data warehouse*.
B) Um *data mart* não tem nenhuma relação com um *data warehouse*.
C) Um *data mart* pode ser entendido como uma combinação de *data warehouses*.
D) Um *data warehouse* pode ser visto como uma combinação de *data marts*.

**10.** Os *data warehouses* se distinguem dos sistemas tradicionais de diferentes formas. Sobre esse assunto, analise as seguintes afirmativas sobre as características de um *data warehouse*:

I. Forma de estruturação é focada nas operações de carga.
II. Geralmente possui múltiplas fontes de origem de dados.
III. Dados usualmente não são mantidos no nível mais detalhado possível.

Estão corretas as seguintes afirmativas:
A) I e II, apenas.
B) I e III, apenas.
C) II e III, apenas.
D) I, II e III.

---

**11.** Ao se montar um *data warehouse*, é necessário realizar um processo de extração, transformação e carga (ETL) de dados. Sobre esse tema, selecione a alternativa correta:
A) *Staging areas* são áreas de armazenamento temporário de dados que visam simplificar o processo de ETL.
B) A extração de dados é realizada a partir de fontes confiáveis e de integridade garantida, como bancos de dados, e nunca de planilhas ou arquivos de texto.
C) O processo de carga deve sempre se dar de maneira sequencial, registro a registro, garantindo assim a consistência dos dados finais.
D) A estrutura dos dados definida na *staging area* é determinante para o desempenho das consultas de um *data warehouse*.

---

**12.** Sobre *dashboards* em soluções de *Business Intelligence*, temos as seguintes afirmações:

I. O objetivo principal de um *dashboard* é mostrar todos os dados possíveis para que o usuário possa tomar a melhor decisão.
II. o público alvo não possui relevância na hora de escolher os dados a serem mostrados.
III. existem tipos de gráficos que são mais adequados para visualizar tipos específicos de dados.
IV. a construção de um *dashboard* deve começar a partir das perguntas que precisam ser respondidas.

Sobre as afirmativas anteriores, é correto dizer que:
A) As afirmativas III e IV estão corretas.
B) As afirmativas I e III estão corretas.
C) As afirmativas II, III e IV estão corretas.
D) As afirmativas I e II estão corretas.

---

**13.** Sobre o aprendizado de máquina supervisionado, considere as seguintes afirmações:

I. As árvores de decisão são um exemplo de algoritmo utilizado para se construir um modelo e realizar a tarefa de classificação.
II. Não são necessários rótulos de dados (classes) para que os algoritmos consigam realizar a construção do modelo de classificação.

III. O principal objetivo do aprendizado supervisionado é realizar a separação dos dados em grupos distintos baseado na similaridade entre os objetos.
IV. A construção de um modelo eficiente depende da qualidade e representatividade dos dados utilizados.

São corretas as afirmações:
A) Somente I.
B) I e IV.
C) II e IV.
D) I, II e III.

---

**14.** O Aprendizado de Máquina é um dos campos de estudo mais promissores dentre aqueles existentes no momento. Dentro desse contexto, analise as seguintes afirmações:

I. Para construir bons modelos computacionais, o volume de dados representa uma questão importante na qualidade do resultado.
II. Os algoritmos utilizados para gerar os modelos inteligentes não são afetados pela dimensão dos dados de entrada.
III. A intervenção do ser humano é dispensável para construção de todo modelo inteligente.
IV. Todo algoritmo de aprendizado de máquina precisa de dados para construir os modelos inteligentes.

Estão corretas as afirmações:
A) I e II.
B) I e IV.
C) II e III.
D) III e IV.

---

**15.** Em grandes corporações ou órgãos públicos, é comum que existam diferentes sistemas em operação, cada um com sua própria base de dados. Para ajudar na tomada de decisões, o *Business Intelligence* busca integrar os dados vindos de diferentes fontes para dar uma visão geral através de informações extraídas a partir dos próprios dados. Analise as afirmações a seguir referentes a etapa de integração dos dados:

I. Os dados obtidos de diferentes fontes podem provocar inconsistências na base final.
II. O formato dos dados não representa um problema na integração das bases.
III. A integração dos dados não é uma etapa que pode ser automatizada.
IV. A normalização é fundamental para garantir a consistência das informações extraídas dos dados.

Sobre as afirmativas anteriores, é correto dizer que:
A) As afirmações I e III estão corretas.
B) As afirmações I e IV estão corretas.
C) As afirmações II e III estão corretas.
D) Todas as afirmações estão corretas.

**16.** Dentro do contexto de Mineração de Dados, podemos dizer afirmar:

I. A limpeza dos dados é uma das etapas do processo.
II. O treinamento é uma parte fundamental para efetuar a limpeza dos dados.
III. A busca por padrões consiste em um dos possíveis objetivos da Mineração de Dados.
IV. O escalonamento dos valores é uma etapa possível na preparação dos dados.

Sobre as afirmativas anteriores, é correto dizer que:
A) Somente a alternativa I está correta.
B) Estão corretas somente as afirmativas III e IV.
C) Estão incorretas as afirmativas II e IV.
D) Somente a alternativa II está incorreta.

---

**17.** Durante o processo de construção de um modelo de classificação, faz-se necessário dividir o conjunto de dados em partes para treinar, validar e testar o modelo. A partir dessa informação, analise as seguintes afirmações:

I. O conjunto de treinamento é utilizado para ajustar os parâmetros do modelo.
II. A acurácia do modelo medida com o conjunto de treinamento é uma medida confiável.
III. O conjunto de testes é utilizado para medir a acurácia do modelo considerando amostras que nunca foram vistas antes.
IV. O conjunto de validação deve ser utilizado em conjunto com a fase de treinamento do modelo para verificar sua evolução e a existência de *overfitting*.

São válidas as afirmativas:
A) Somente I, III e IV
B) Somente I, II e III
C) Somente I e III
D) Somente II e IV

---

**18.** Quando queremos extrair informações básicas sobre os dados, recorremos a medidas estatísticas. Dentre as medidas mais comuns, temos a média, mediana e a moda. Sobre elas, temos as seguintes afirmativas:

I. Todas são medidas de tendência central.
II. A média é afetada pela presença de outliers nos dados.
III. Os resultados obtidos no cálculo da moda e mediana são sempre os mesmos.
IV. A mediana é igual ao segundo quartil dos dados.

Analisando as afirmativas, podemos dizer que:
A) Todas as afirmativas estão corretas.
B) Somente a afirmativa IV está correta.
C) Somente as afirmativas II e III estão corretas.
D) Somente as afirmativas I, II e IV estão corretas.

**19.** *Dashboards* são soluções que auxiliam os usuários do negócio a visualizarem de forma rápida os dados. Nesse contexto, marque o item que NÃO possua uma característica desejável dos *dashboards*:
A) Informação em tempo real.
B) Apresentar os chamados indicadores chave de desempenho.
C) Informar em detalhes todos os dados do negócio.
D) Transparecer metas, prioridades e níveis de desempenho.

---

**20.** Sobre a análise de agrupamentos, é possível afirmar que:
A) O *k-Nearest Neighbor* (k-NN) é um popular algoritmo utilizado para realizar a tarefa de agrupamento.
B) Tem como principal objetivo realizar a classificação de novos dados utilizando um modelo construído a partir de uma base com dados rotulados.
C) Seu uso requer que os dados estejam devidamente rotulados com as respectivas classes.
D) Seu objetivo é identificar elementos semelhantes e separá-los para uma análise exploratória mais profunda.

---

**21.** Dentro da área de Mineração de Dados, algumas etapas são utilizadas para preparar os dados para análise. Em geral, um problema comum se refere ao volume dos dados existentes nas bases, antes ou depois da integração de diferentes fontes. Dentre as alternativas que existem para tratar o problema, marque aquela que NÃO é adequada:
A) Amostrar os dados para obter e utilizar uma porção representativa menor.
B) Aplicar técnicas de redução de dimensionalidade para reduzir o número de características que descrevem cada entidade representada.
C) Remover dados redundantes.
D) Descartar uma ou mais fontes de dados.

---

**22.** Sobre o Aprendizado de Máquina Profundo, analise as seguintes afirmações:

I. As *Random Forests* são os algoritmos mais utilizados nesse tipo de aprendizado.
II. Em geral, são necessários grandes volumes de dados para conseguir bons resultados.
III. Técnicas de aprendizado profundo não podem ser utilizadas para classificar dados.
IV. A construção de modelos sempre leva poucos minutos para acontecer.

Analisando as afirmativas, podemos dizer que:
A) Somente a afirmativa II está correta.
B) Somente as afirmativas I e III estão corretas.
C) Somente as afirmativas I, II e III estão corretas.
D) Somente as afirmativas II e III e IV estão corretas.

---

**23.** São exemplos de técnicas de mineração de dados mais apropriados para se criar um modelo de classificação:
A) Algoritmo genético e Árvores B.
B) Regressão linear simples e Regressão multivariada.
C) Árvore de decisão e Florestas Aleatórias (*Random Forest*).
D) k-médias e Árvores rubro-negras.

**24.** No momento de construir um *dashboard*, a escolha dos elementos visuais é uma tarefa muito importante para o sucesso da interpretação do que se quer visualizar. Dessa forma, podemos dizer que:

A) Os gráficos de pizza e de barras devem ser os únicos utilizados dentro de um *dashboard*.

B) Os gráficos de linhas são úteis quando queremos visualizar a evolução de uma determinada entidade (produto, evento, etc) ao longo do tempo.

C) Os gráficos de pizza ou rosca são indicados quando temos um número muito grande de categorias.

D) Os gráficos de dispersão são mais indicados quando cada entidade a ser representada possui muitas características ou dimensões.

---

**25.** Dentre as opções a seguir, marque a alternativa que NÃO indica ferramentas que podem ser utilizadas para construir dashboards no contexto de *Business Intelligence*:

A) Qlkview e Pentaho

B) Pentaho e Metabase

C) Power BI e Tableau

D) Power BI e Hadoop

---

Considere as seguintes informações para responder as questões de 26 a 30.

Um determinado sistema de controle de pedidos de compra foi implementado utilizando-se um banco de dados relacional. Esse banco possui a tabela de nome `Products`, representando o produto a ser comprado, a tabela `Orders` com os pedidos, e a tabela `OrderDetails` com detalhes do pedido, além de outras tabelas. As respectivas estruturas das tabelas citadas podem ser vistas a seguir.

| ProductID | ProductName | SupplierID | CategoryID | Unit | Price |
|---|---|---|---|---|---|
| 1 | Chais | 1 | 1 | 10 boxes x 20 bags | 18 |
| 2 | Chang | 1 | 1 | 24 - 12 oz bottles | 19 |
| 3 | Aniseed Syrup | 1 | 2 | 12 - 550 ml bottles | 10 |

| OrderID | CustomerID | EmployeeID | OrderDate | ShipperID |
|---|---|---|---|---|
| 10248 | 90 | 5 | 1996-07-04 | 3 |
| 10249 | 81 | 6 | 1996-07-05 | 1 |
| 10250 | 34 | 4 | 1996-07-08 | 2 |

| OrderDetailID | OrderID | ProductID | Quantity |
|---|---|---|---|
| 1 | 10248 | 11 | 12 |
| 2 | 10248 | 42 | 10 |
| 3 | 10248 | 72 | 5 |

**26.** Para se visualizar a quantidade de pedidos (`Orders`) realizados por dia, fazemos uso da seguinte consulta SQL:

A) `SELECT OrderDate, count(*) FROM Orders GROUP BY (OrderDate)`

B) `SELECT DISTINCT OrderDate, count(*) FROM Orders`

C) `SELECT OrderDate, sum(*) FROM Orders GROUP BY (OrderDate)`

D) `SELECT DISTINCT OrderDate, sum(*) FROM Orders`

**27.** Para se visualizar a lista de produtos e suas respectivas quantidades médias de unidades solicitadas por pedido, a consulta SQL que podemos utilizar é a:

A) `SELECT p.ProductName, avg(d.Quantity) FROM OrderDetails d, Products p WHERE d.ProductID=p.ProductID GROUP BY p.ProductName`

B) `SELECT p.ProductName, mean(d.Quantity) FROM OrderDetails d, Products p WHERE d.ProductID=p.ProductID GROUP BY p.ProductName`

C) `SELECT p.ProductName, mean(d.Quantity) FROM OrderDetails d, Products p WHERE d.ProductID=p.ProductID ORDER BY p.ProductName`

D) `SELECT p.ProductName, avg(d.Quantity) FROM Orders d, Products p WHERE d.ProductID=p.ProductID GROUP BY p.ProductName`

**28.** Para se buscar o nome do produto mais comprado pelo cliente de ID 71, é mais apropriado o uso da seguinte consulta SQL:

A) `SELECT p.ProductName FROM Orders o, OrderDetails d, Products p WHERE o.OrderID=d.OrderID AND o.CustomerID=71 ORDER BY d.Quantity DESC LIMIT 1`

B) `SELECT p.ProductName FROM Orders o, OrderDetails d, Products p WHERE o.OrderID=d.OrderID AND p.ProductID = d.ProductID AND o.CustomerID=71 ORDER BY d.Quantity DESC LIMIT 1`

C) `SELECT p.ProductName, d.Quantity FROM Orders o, OrderDetails d, Products p WHERE o.OrderID=d.OrderID AND p.ProductID = d.ProductID and o.CustomerID=71 ORDER BY d.Quantity ASC`

D) `SELECT p.ProductName, d.Quantity FROM Orders o, OrderDetails d, Products p WHERE o.OrderID=d.OrderID AND p.ProductID = d.ProductID and o.CustomerID=71 ORDER BY d.Quantity DEC`

**29.** Para se recuperar a lista completa de produtos com suas respectivas quantidades vendidas, sendo o valor nulo considerado igual a 0, selecione a alternativa com o comando mais apropriado para ser utilizado:

A) `SELECT p.ProductName, sum(d.Quantity) FROM Products p JOIN OrderDetails d ON p.ProductID=d.ProductID GROUP BY p.ProductName ORDER BY p.ProductName`

B) `SELECT p.ProductName, sum(d.Quantity) FROM Products p RIGHT JOIN OrderDetails d ON p.ProductID=d.ProductID GROUP BY p.ProductName ORDER BY p.ProductName`

C) `SELECT p.ProductName, sum(d.Quantity) FROM Products p LEFT JOIN OrderDetails d ON p.ProductID=d.ProductID GROUP BY p.ProductName ORDER BY p.ProductName`

D) `SELECT p.ProductName, sum(d.Quantity) FROM Products p INNER JOIN OrderDetails d ON p.ProductID=d.ProductID GROUP BY p.ProductName ORDER BY p.ProductName`

**30.** Considere que as tabelas relacionais do sistema de pedidos apresentado anteriormente foram migradas para um banco de dados não relacional como o MongoDB, por exemplo. Nesse contexto, analise as seguintes afirmativas:

I. Uma vez migrado os dados para o banco de dados não relacional, não será mais possível realizar operações de agregação através de consultas a esse novo banco.
II. Por questão de desempenho, para algumas consultas, é possível criar uma única coleção `MyOrders` para armazenar o conteúdo das seguintes tabelas do banco relacional em questão: `Orders`, `OrderDetails`, `Products`.
III. Um índice composto pelo nome do produto e pela data de compra pode ser criado, agilizando as consultas que envolvam esses atributos.

Estão corretas as seguintes afirmativas:
A) I e II, apenas.
B) I e III, apenas.
C) II e III, apenas.
D) I, II e III.