Residência em Tecnologia da Informação Aplicada à Área Jurídica

Prova de Conhecimentos Específicos

Área de Concentração 2 - Analista de *Business Intelligence*

Candidato: _____

CPF:_____ Telefone: _____

Questões

1) Um armazém de dados (*data warehouse*) é um importante recurso de suporte à tomada de decisão em empreendimentos modernos. Sobre as características de um *data warehouse*, assinale a alternativa verdadeira:

a) Seus dados são limpos regularmente, evitando guardar dados de um longo período de tempo.
b) É orientado a temas de negócio.
c) É estruturado para otimizar principalmente as operações de carga de dados.
d) Sua modelagem segue a 3ª forma normal.

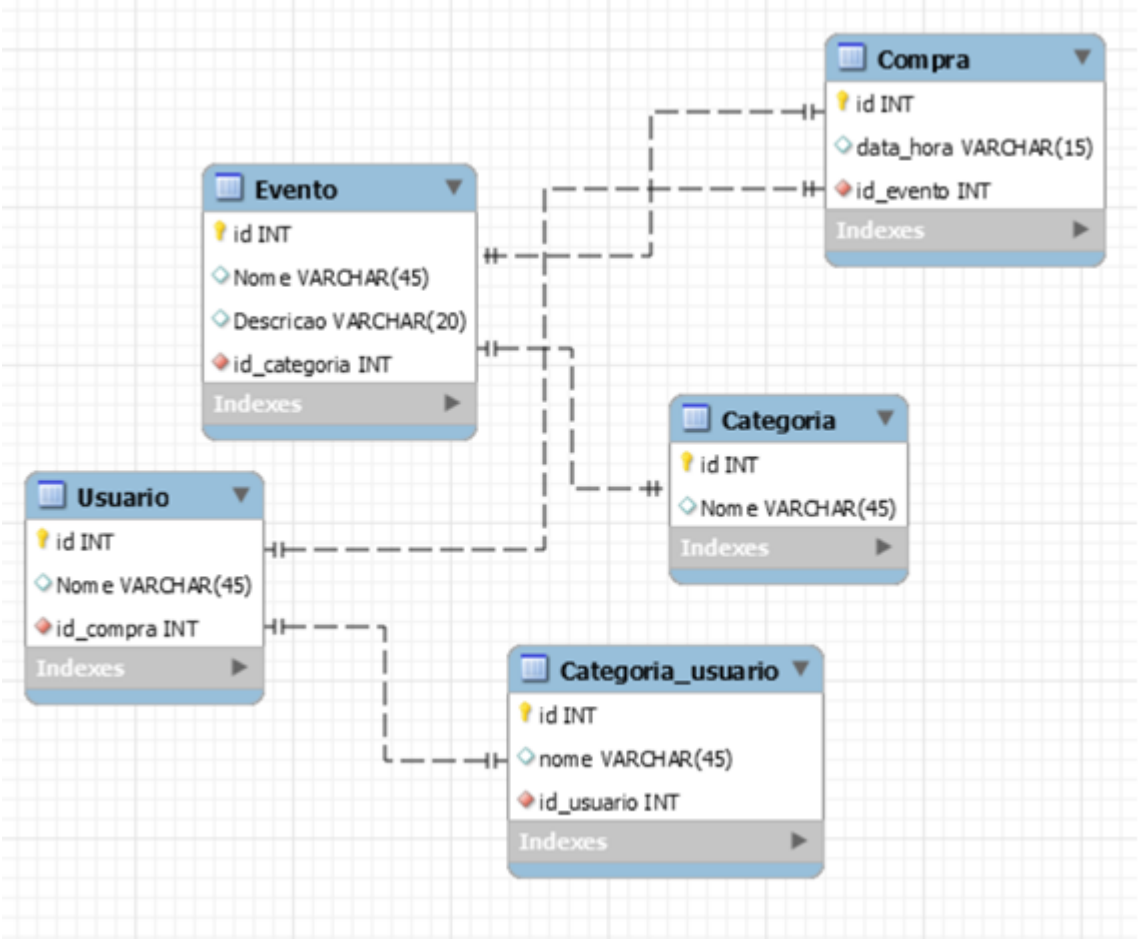
2) Considerando as boas práticas de modelagem de um data *warehouse* utilizando-se o esquema estrela, assinale a alternativa verdadeira sobre características usuais de uma tabela FATO:

a) Possui uma pequena quantidade de registros.
b) É diretamente vinculada a outras tabelas fato.
c) Não possui campos categóricos.
d) Está relacionada a uma única tabela dimensão.

3) Considerando as boas práticas de modelagem de um *data warehouse* utilizando-se o esquema estrela, assinale a alternativa verdadeira sobre as características esperadas de uma tabela DIMENSÃO:

a) Possuir uma grande quantidade de registros comparado a tabelas fato.
b) Não possuir campos categóricos.
c) Ser diretamente vinculada a outras tabelas dimensão.
d) Ser vinculada a mais de uma tabela fato.

Analise a seguir a modelagem de banco de dados realizada para um determinado sistema transacional, responsável por registrar as compras de ingressos a eventos de uma casa de festas. Com base nessa figura, responda as questões que se seguem.



4) Sobre a modelagem de tabelas fatos do sistema ilustrado, é uma tabela fato aquela que representa:

a) Os usuários do sistema.
b) As compras realizadas.
c) As categorias existentes.
d) O nome dos eventos realizados.

5) Sobre a modelagem estrela do sistema ilustrado, uma relação deverá ser criada entre uma tabela dimensão e uma tabela fato para representar a relação entre:

a) O nome do evento e sua categoria.
b) O usuário e a descrição do evento que participou.
c) O usuário e a compra realizada.
d) O usuário e sua categoria.

6) Consultas SQL são realizadas para acessar os dados de *data warehouses* modelados através de um esquema estrela. Sobre essas consultas, assinale o item que não é comumente utilizado após a palavra-chave **SELECT**:

a) Atributos da tabela fato.
b) Atributos da tabela dimensão.
c) Chaves primárias.
d) Funções para calcular somas ou médias.

7) Sobre as Consultas SQL realizadas para se acessar os dados de *data warehouses* modelados por um esquema estrela, assinale o item que é comumente utilizado após as palavras-chaves **GROUP BY**:

a) Atributo de dimensão.
b) Atributo de fato.
c) Chave primária.
d) Funções para calcular somas ou médias.

8) Sobre as diferenças entre as modelagens esquema estrela e floco de neves, assinale a alternativa verdadeira:

a) Não existem diferenças entre esses modelos, são diferentes nomes para a mesma abordagem.
b) O esquema floco de neves é um modelo mais normalizado (formas normais).
c) O esquema estrela é um modelo mais normalizado (formas normais).
d) Consultas ao esquema estrela são mais ineficientes.

9) *Dashboards* são ferramentas que auxiliam os usuários do negócio a visualizarem de forma rápida os dados. Nesse contexto, assinale a afirmativa 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.

10) 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).

11) Os processos de ETL são muito comuns para coletar e organizar dados de várias fontes. Um importante componente desses processos, denominado *control flow*, permite que:

a) As fontes de dados (*sources*) sejam monitoradas no que tange o desempenho computacional na entrega dos dados.
b) As operações que compõem as tarefas de transformação de dados sejam estabelecidas.
c) As tarefas de *data flow* sejam coordenadas por meio de regras e restrições.
d) Os diferentes graus de acesso que as tarefas de *data flow* devem adquirir durante a fase de extração, sejam gerenciados.

12) Por meio da ferramenta *Pentaho Data Integration* (PDI), é possível realizar processos de ETL (*Extraction, Transformation and Load*). A ferramenta gráfica que faz parte do PDI, usada para modelar as *transformations* e os *jobs*, é denominada:

a) *Spoon*.
b) *Kettle*.
c) *Weka*.
d) *Mondrian*.

13) *Pentaho* é um software de código aberto para inteligência empresarial, desenvolvido em Java. Assinale a alternativa que apresenta com quais bancos de dados o software trabalha nativamente:

a) *NoSQL, Hadoop e Oracle*.
b) *PostgreSQL e Oracle.*
c) *NoSQL e Hadoop*.
d) *PostgreSQL, Hadoop e Oracle*.

14) Considere as seguintes características de um projeto de banco de dados.

I. O modelo de dados é conhecido a priori e é estável;
II. A integridade dos dados deve ser rigorosamente mantida;
III. Velocidade e escalabilidade são preponderantes.

Dessas características, o emprego de bancos de dados *NoSQL* é favorecido somente por:

a) II e III.
b) III.
c) II.
d) I e II.

15) Sobre banco de dados *NoSQL* é correto afirmar que:

I. Como forma de permitir as buscas em documentos semiestruturados, um banco de dados *NoSQL* do tipo orientado a documentos armazena objetos indexados por chaves utilizando tabelas de *hash* distribuídas.
II. Suas estruturas não permitem o uso de linguagens do tipo do *SQL* para recuperação de dados.
III. Privilegiam a rapidez de acesso e a disponibilidade dos dados em detrimento das regras de consistência das transações.

a) I e II estão corretas.
b) I, II e III estão corretas.
c) II e III estão corretas.
d) III está correta.

16) Bancos de dados conhecidos como *NoSQL* podem ser particionados em diferentes servidores, o que introduz o problema de processar consultas que envolvem múltiplos nós de processamento. Um modelo usualmente empregado nessas circunstâncias é conhecido como:

a) *CAP Theorem*.
b) *Map/Reduce*.
c) *Hash tables*.
d) *Clustered columns*.

17) De acordo com o esquema relacional abaixo, assinale a consulta em *SQL* para obter uma lista de grupos que tenham mais de dez produtos relacionados:

GRUPO (codGrupo, grupo)
PRODUTO (codProduto, produto, codGrupo)

a) SELECT Grupo.Grupo, (select count(*) FROM Produto
where Grupo.CodGrupo = Produto.CodGrupo) FROM Grupo
WHERE count(*) > 10

b) SELECT Grupo.grupo, count(*) FROM Grupo
INNER JOIN Produto ON Grupo.CodGrupo = Produto.CodGrupo
GROUP BY Grupo.grupo WHERE count(*) > 10

c) SELECT Grupo.grupo, count(*) FROM Grupo
INNER JOIN Produto ON Grupo.CodGrupo = Produto.CodGrupo
GROUP BY Grupo.grupo HAVING count(*) > 10

d) SELECT Grupo.grupo, (SELECT count(*) FROM Produto
where Grupo.CodGrupo = Produto.CodGrupo) FROM Grupo
WHERE EXISTS SELECT count(Produto.CodGrupo) > 10 FROM Produto

18) A solução correta para que uma consulta *SQL* retorne as agências que possuem média dos saldos aplicados em conta maior que 1200 é:

a) select nome_agencia, avg(saldo) from conta
group by nome_agencia
having avg(saldo) > 1200

b) select nome_agencia, avg(saldo) from conta
where (having avg(saldo) > 1200)

c) select nome_agencia, avg(saldo) from conta
where avg(saldo) > 1200
group by nome_agencia

d) select nome_agencia, avg(saldo) from conta
group by nome_agencia
having saldo > 1200

19) Uma das características inerentes ao modelo chave-valor de bancos de dados *NoSQL* é a(o):

a) suporte à compreensão da semântica do valor associado à chave.
b) favorecimento à evolução de esquemas conceituais.
c) dependência de linguagem de consulta específica.
d) estrutura de armazenamento interna complexa.

20) ETL é o método mais comum para transferir dados de uma fonte de dados OLTP para um *data warehouse*. Contudo, pode-se também empregar o processo de extração, carga e transformação no formato ELT. Para tanto, faz-se necessário o uso de:

a) *Staging tables* ou tabelas intermediárias no banco de dados de destino (*data warehouse*).
b) Tecnologias distintas de banco de dados de origem (OLTP) e de destino (*data warehouse*).
c) Metodologia *Kimbell* e esquemas estrela normalizados para a modelagem do *data warehouse*.
d) Extração *Full* de todos os dados tanto das tabelas de dimensões quanto de fatos.

21) Quais das técnicas abaixo não é utilizada para realizar redução de dimensionalidade nos dados:

a) *Principal Component Analysis*.
b) Análise de variância de atributos.
c) *Bootstrap*.
d) Análise de correlação entre atributos.

22) Os histogramas são representações gráficas de:

a) Distribuição dos dados.
b) Correlação entre variáveis.
c) Acurácia de classificação.
d) Predição de classes.

23) Antes da construção de um modelo de classificação, é comum realizar uma etapa de pré-processamento. Quais das etapas a seguir não pode ser considerada como um pré-processamento:

a) Balanceamento dos dados.
b) Amostragem.
c) Transformação de tipos.
d) Treinamento.

24) Sobre o processo de amostragem dos dados, analise as seguintes afirmativas:

I) Para ser considerada representativa, uma amostra deve possuir aproximadamente as mesmas propriedades dos dados originais.
II) O uso de uma amostra para construir um modelo de classificação tende a produzir um menor custo computacional.
III) Uma amostragem com reposição tem sempre o mesmo efeito que uma amostragem sem reposição.

IV) Para conseguir uma amostra representativa, seu tamanho deve ser sempre superior a 20% do total dos dados.

Estão corretas as afirmativas:

a) I e II
b) III e IV
c) II e III
d) II e IV

25) Sobre a análise de agrupamentos, é possível afirmar que:

a) O objetivo principal é encontrar um conjunto de associações entre os itens de uma base de dados.
b) Tem como meta a sumarização dos conjunto de dados em valores mais simples.
c) Identificar elementos semelhantes e separá-los para uma análise exploratória mais profunda.
d) Seu uso requer que os dados estejam devidamente rotulados com as respetivas classes.

26) O sobre ajustamento dos modelos (*overfitting*) é um fenômeno comum de ocorrer em tarefas de classificação. Dentre as alternativas abaixo, qual não pode ser apontada como uma maneira de evitar o fenômeno:

a) Manter o modelo simples, reduzindo o número de variáveis e de possíveis ruídos nos dados.
b) Manter o nível acurácia de acurácia maior possível durante a fase de treinamento.
c) Usar técnicas de validação cruzada.
d) Usar um conjunto de validação para indicar o momento de interromper o treinamento.

27) O *Weka* é uma ferramenta computacional que possui um conjunto de algoritmos de aprendizado de máquina para realizar análise de dados. A partir de uma base de dados de uma rede varejista de supermercados, contendo informações sobre todas as compras feitas nos caixas das lojas, considere as seguintes afirmações do que é possível fazer utilizando o *Weka*:

I) Preparar os dados para utilização por algoritmos de classificação.
II) Utilizar a API do Weka para incorporar um algoritmo de agrupamento a um software de análise de dados, cuja linguagem de programação seja compatível com o *Weka*.
III) Obter automaticamente o significado de dados não conhecidos presentes na base.

IV) Identificar possíveis *outliers* nos dados.

Estão corretas as afirmativas:

a) I e II
b) II, III e IV
c) Somente a alternativa I
d) I, II e IV

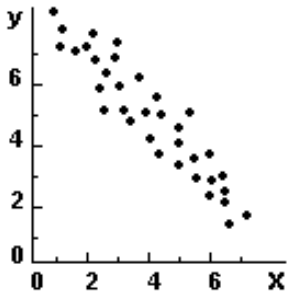
28) Ao utilizar uma base de dados, é comum que muitos dos dados não estejam presentes. A falta de alguns valores nas bases de dados se deve, muitas vezes, por falta de preenchimento nos cadastros, falha em sensores, entre outros motivos. Para que técnicas de aprendizado de máquina possam utilizar corretamente 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.

29) Analise a figura mostrada a seguir, que apresenta uma visão da ocorrência conjunta dos valores de duas variáveis, X e Y.



Segundo a estatística, assinale a alternativa correta sobre a correlação observada entre as variáveis X e Y:

a) Não existe correlação.
b) Existe uma correlação negativa.
b) Existe uma correlação perfeita.
d) Existe uma correlação linear positiva.

30) Ao analisar um conjunto de dados, um analista consegue obter as seguintes informações:

I) As escalas dos dados que descrevem as variáveis são muito diferentes.
II) Existe um nível de correlação alto entre mais de duas variáveis do conjunto de dados.
III) Dentre as variáveis, existe, pelo menos, uma que é categórica (não numérica).
IV) O conjunto possui dados faltosos para algumas instâncias.

A partir dessas afirmações, marque a alternativa que faz mais sentido no contexto de um problema de classificação utilizando redes neurais:

a) Não é preciso fazer qualquer alteração sobre os dados, dado que as redes neurais estão preparadas para tratar todas as situações descritas.
b) As redes neurais precisam de valores de entrada em todos os casos e, por isso, necessariamente deve-se eliminar ou preencher os dados faltosos antes de passar os dados para a rede.
c) Só é necessário se preocupar com as informações contidas nos itens I), II) e IV), dado que o tipo de dado não faz diferença para as redes neurais.
d) A escala dos dados e a correlação entre eles não impactam no resultado.

Boa prova!