



PROGRAMA DE RESIDÊNCIA EM TECNOLOGIA DA INFORMAÇÃO
TRIBUNAL DE CONTAS DO ESTADO DO RIO GRANDE DO NORTE

EDITAL 013/2019 – PROVA DE CONHECIMENTOS ESPECÍFICOS
ÁREA DE CONCENTRAÇÃO 3: BUSINESS INTELLIGENCE AND ANALYTICS
16 / 06 / 2019

Identificação do Candidato	
Nome completo:	
CPF:	Assinatura:

Leia com atenção as seguintes instruções:

1. Aguarde a autorização do(s) fiscal(is) para poder iniciar a Prova.
2. Não esqueça de colocar seu **nome completo** (preferencialmente em letras maiúsculas) e de assinar o campo acima.
3. Este Caderno de Prova, com páginas numeradas de 1 a 11, é constituído de **30 (trinta) questões** de múltipla escolha, cada uma com quatro alternativas. Verifique se o Caderno de Prova está completo, sem falhas de impressão ou problemas que comprometam sua leitura. Caso necessário, solicite imediatamente ao(s) fiscal(is) a substituição do Caderno de Prova completo.
4. Confira se este Caderno de Prova corresponde à área de concentração para a qual foi inscrito. Caso haja alguma divergência, notifique imediatamente o(s) fiscal(is).
5. Leia com atenção o enunciado das questões antes de responde-las.
6. Cada questão possui **apenas uma** alternativa correta. Você deverá marcar a resposta que julgar correta usando caneta esferográfica de tinta na cor azul ou preta no local correspondente à respectiva questão. A interpretação das questões faz parte da Prova, de modo que não será permitida qualquer tipo de pergunta ou explicação ao(s) fiscal(is).
7. Não serão computadas questões sem marcação de resposta ou que contenham mais de uma marcação, marcação rasurada ou emendada.
8. Tenha cuidado ao manusear este Caderno de Prova, evitando rasuras, pois ele **não será substituído** por esse motivo. Também não é permitido destacar quaisquer das folhas que compõem este Caderno de Prova.
9. O tempo máximo para resolução desta Prova é de **2 (duas) horas**, para o qual **não haverá prorrogação**. Transcorrido esse tempo, o Caderno de Prova será recolhido pelo(s) fiscal(is).
10. Terminada a realização da Prova, este Caderno de Prova deverá ser **obrigatoriamente** entregue ao(s) fiscal(is) antes de se retirar da sala de realização.



QUESTÕES

1. Considerando a definição da tabela a seguir, selecione a consulta SQL que retorna os estudantes cujo nome começam com João.

TABLE students

id INTEGER PRIMARY KEY,
name VARCHAR(30) NOT NULL

- a) SELECT sum(*) FROM students WHERE name = "João%"
 - b) SELECT count(*) FROM students WHERE name like "João%"
 - c) SELECT count(*) FROM students WHERE name = "João%"
 - d) SELECT sum(*) FROM students WHERE name like "João%"
2. Visões podem ser criadas em SQL através de consultas pré-definidas. Nesse contexto, qual é a maneira correta de se criar uma VIEW chamada all_students para retornar todos os nomes de estudantes armazenados na tabela students apresentada na questão anterior?
- a) CREATE VIEW all_students AS SELECT name FROM CLIENTES
 - b) CREATE all_students AS VIEW name FROM CLIENTES
 - c) CREATE all_studentscli AS SELECT name FROM CLIENTES
 - d) CREATE VIEW all_students, SELECT name FROM CLIENTES
3. Colunas com identificadores únicos na área de negócio podem ser modeladas em um banco de dados de forma que a coluna respectiva a esse campo não permita a inserção de valores repetidos. Qual das cláusulas abaixo deverá ser usada no comando SQL (Structured Query Language) para aplicar essa restrição no momento da criação da coluna?
- a) CHECK
 - b) DEFAULT
 - c) UNIQUE
 - d) DISTINCT
4. Constraints são restrições que podem ser utilizadas para garantir uma maior integridade dos dados de um banco de dados. Nesse contexto, assinale a alternativa verdadeira sobre como definir restrições de chave estrangeiras (*constraints*) em um banco de dados relacional através de comandos SQL.
- a) O uso do CASCADE pode ser usado tanto para deletar como para atualizar registros em uma tabela filha, que referencia uma tabela pai.
 - b) A cláusula SET NULL é utilizada para impedir uma operação na tabela pai.
 - c) Uma tabela só pode possuir uma *constraint*, já que só tem uma chave primária.
 - d) As *constraints* são criadas nas tabelas pai, as quais são referenciadas pelas tabelas filho.



5. Acerca da terminologia utilizada no sistema MongoDB, assinale a alternativa correta.
- a) Um *database* agrupa registros de itens similares.
 - b) Uma *collection* é um conjunto de databases utilizados por uma aplicação.
 - c) O controle de acesso dos usuários é por *database* e não por *collections*.
 - d) Um *document* é uma estrutura de dados compostas por pares de atributo-valor.
6. Consultas no MongoDB podem realizar operações de filtragem e agrupamento assim como acontece nos bancos de dados relacionais. No caso do MongoDB, avalie qual dos seguintes códigos apresenta uma consulta sintaticamente válida.
- a)

```
db.orders.aggregate([  
  { $match: { status: "A" } },  
  { $group: { _id: "$cust_id", total: { $sum: "$amount" } } }  
])
```
 - b)

```
db.orders.search([  
  { $match: { status: "A" } },  
  { $group: { _id: "$cust_id", total: { $sum: "$amount" } } }  
])
```
 - c)

```
db.orders.aggregate([  
  { $matchRecord: { status: "A" } },  
  { $groupById: { _id: "$cust_id", total: { $sum: "$amount" } } }  
])
```
 - d)

```
db.orders.mapReduce([  
  function match() { status: "A" },  
  function group() { _id: "$cust_id", total: { $sum: "$amount" } }  
])
```



-
7. Bancos de dados baseados em grafos são grandes ferramentas para se realizar consultas sobre relacionamentos entre os dados. Neo4J é um banco que suporta tais consultas. Acerca das consultas suportadas pelo Neo4J através do Cypher, assinale a alternativa cuja consulta está sintaticamente correta visando encontrar uma pessoa b que tem um amigo em comum c com a pessoa a de nome Sidney.
- a) `SELECT (a:Person {name:'Sidney'})-[:KNOWS]->(b)-[:KNOWS]->(c), (c)-[:KNOWS]->(b) RETURN b, c`
 - b) `FIND {a:Person {name:'Sidney'}}-[:KNOWS]->(b)-[:KNOWS]->(c), (a)-[:KNOWS]->(c) RETURN b, c`
 - c) `SELECT (a:Person {name:'Sidney'})-[:KNOWS]->(b)-[:KNOWS]->(c), (a)-[:KNOWS]->(c) RETURN b, c`
 - d) `MATCH (a:Person {name:'Sidney'})-[:KNOWS]->(b)-[:KNOWS]->(c), (a)-[:KNOWS]->(c) RETURN b, c`
8. 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.
9. 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.



-
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 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).
 - d) 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).
11. Sobre o processo de ETL, aplicado a data warehouse, é correto afirmar que:
- a) a fase de extração de dados consiste em obter os dados do servidor do data warehouse.
 - b) a fase de transformação consiste em realizar modificações nos dados carregados, adequando seus valores ao modelo definido para o data warehouse.
 - c) as fases de extração e carga de dados são realizadas de forma simultânea.
 - d) a fase de carga de dados visa eliminar valores nulos contidos nos bancos de dados transacionais da empresa.
12. No processo ETL de um data warehouse: (I) a tradução de valores codificados conhecida como limpeza de dados (por exemplo, supondo que o sistema de origem armazena 1 para sexo masculino e 2 para feminino, mas o data warehouse armazena M para masculino e F para feminino) e (II) a geração de valores de chaves substitutas (surrogate keys) são:
- a) integrantes da etapa de transformação.
 - b) integrantes das etapas de transformação e carga, respectivamente.
 - c) manipulações de dados, aplicáveis no processo pivoting.
 - d) integrantes da etapa de carga.
13. Um Data Warehouse pode ser subdividido em pequenas porções para atender um departamento da empresa, por exemplo. Essa subdivisão do Data Warehouse é conhecida como
- a) Database.
 - b) Data Mart.
 - c) Data Center.
 - d) Data Store.
-



-
14. Tabelas fato e tabelas dimensão são partes integrantes:
- a) do roll-up.
 - b) da mineração de dados.
 - c) do esquema estrela.
 - d) do drill-down.
15. No âmbito das agregações sql em data warehouse (DW), cube e rollup são extensões da cláusula:
- a) having.
 - b) grouping sets.
 - c) avg.
 - d) group by.
16. OLAP é uma tecnologia que habilita analistas de negócios a extraírem e analisarem dados a partir de diferentes pontos de vista. Sobre as operações suportadas, assinale a alternativa correta.
- a) Roll-up adiciona mais dimensões na análise sendo realizada.
 - b) Drill-down é uma operação que mover a análise para um nível menos detalhado de informação, reduzindo dimensões.
 - c) Slice é uma operação que pode filtrar os dados por alguma dimensão específica.
 - d) Dice rotaciona os eixos de análise, não modificando os dados analisados.
17. O QlikView trabalha com diferentes tipos de arquivos. Sobre as características desses tipos de arquivo, assinale a alternativa correta.
- a) Um arquivo QVD é um arquivo que contém uma tabela de dados exportada.
 - b) QVW é a extensão do arquivo que contém a especificação de relatórios, sem informações de scripts e da modelagem de dados.
 - c) Arquivos QVD não são otimizados para leitura de dados.
 - d) Arquivos QVW são conectores para intercâmbio de dados com outras plataformas.
18. Qual dos algoritmos abaixo não pode ser considerado um comitê (ensemble)?
- a) AdaBoost
 - b) Multi-layer perceptron (MLP)
 - c) Random forests
 - d) Perceptron
-



-
19. No contexto de processamento de linguagem natural, qual das técnicas abaixo não pode ser considerada pré-processamento?
- a) Análise de sentimentos
 - b) Remoção de stopwords
 - c) Stemização
 - d) Lematização
20. Qual das técnicas abaixo não deve ser aplicada a dados não-linearmente separáveis?
- a) Support vector machines (SVM)
 - b) Regressão linear
 - c) Multi-layer perceptron (MLP)
 - d) Random forests
21. São opções válidas para reduzir o overfitting de um modelo, EXCETO:
- a) Configurar os hiperparâmetros do estimador
 - b) Revisar o balanceamento de classes
 - c) Validar o treinamento diretamente no conjunto de teste
 - d) Aplicar técnicas de redução de dimensionalidade
22. A barragem móvel de Maeslant, nos Países Baixos, foi projetada para proteger o Porto e a cidade de Roterdã em caso de enchentes, tempestades ou tsunamis. O fechamento da barragem é controlado por um sistema de previsão de desastres, que foi cuidadosamente planejado considerando que o fechamento da barragem bloqueia o acesso de navios ao porto. Considerando as características do sistema em questão, pode-se afirmar que:
- a) Falsos negativos e falsos positivos são considerados um problema neste caso, mas falsos negativos são considerados mais importantes que falsos positivos.
 - b) Falsos negativos e falsos positivos são considerados um problema neste caso, mas falsos negativos são considerados menos importantes que falsos positivos.
 - c) Falsos negativos e falsos positivos são considerados um problema neste caso e são considerados igualmente importantes.
 - d) Falsos negativos e falsos positivos não são considerados um problema neste caso.



23. Sobre a ferramenta Weka, podemos afirmar que:

- a) é um software livre implementado em Python e que pode ser utilizado somente em dispositivos Linux.
- b) possui um conjunto de algoritmos de aprendizado de máquina implementados para serem utilizados livremente para resolver problemas de mineração de dados.
- c) não é possível integrar os algoritmos implementados em Java com um código desenvolvido pelo usuário.
- d) não possui interface gráfica para utilização pelo usuário.

24. A mineração de dados é o processo de extração de informações de maneira inteligente a partir de um grande volume de dados, de maneira que essas informações possam ser utilizadas posteriormente. Diante disso, analise as seguintes afirmações:

- I. a mineração de dados é uma área interdisciplinar que envolve, entre outras, as áreas de Ciência da Computação e Estatística.
- II. o agrupamento, a classificação e a associação de dados estão dentre as tarefas de Mineração de Dados possíveis.
- III. Aprendizado de máquina e mineração de dados são a mesma coisa.

Estão corretas as seguintes afirmações:

- a) I e III
- b) II e III
- c) I e II
- d) somente II

25. O agrupamento de dados é uma das tarefas de mineração de dados. Sobre essa tarefa, não podemos afirmar que:

- a) a medida de similaridade ou proximidade é um fator crítico e que pode alterar diretamente o resultado do agrupamento.
- b) distância e correlação são exemplos de medida de similaridade utilizadas para realizar agrupamentos.
- c) quando usamos medidas de distância para medir similaridade, a magnitude dos dados é um fator que impacta no resultado final.
- d) os algoritmos de agrupamento buscam separar os dados em conjuntos de forma que sempre os dados em um grupo estejam mais próximos entre si do que de dados em grupos diferentes.



26. Quais dos algoritmos a seguir não pode ser utilizados para realizar agrupamentos:

- a) k-means
- b) mistura finita de gaussianas
- c) árvores de decisão
- d) DBSCAN

27. Para avaliar técnicas de aprendizado de máquina supervisionado, podemos utilizar as seguintes medidas, exceto:

- a) precisão e revocação (precision e recall)
- b) acurácia
- c) curvas Receiver Operating Characteristic (ROC)
- d) coeficiente silhueta

28. A legislação brasileira busca incluir as pessoas com deficiência em todas as áreas de atuação da sociedade. Na educação superior, por exemplo, pessoas com deficiência já se tornaram parte do corpo discente de muitas universidades. O gráfico abaixo apresenta o número pessoas com deficiência que ingressaram em cursos de graduação na UFRN nos últimos anos. Um acontecimento que impactou a presença de pessoas com deficiência na UFRN foi a criação do curso de Licenciatura em Letras Português-LIBRAS, pois LIBRAS é a primeira língua, a língua nativa, de pessoas surdas.

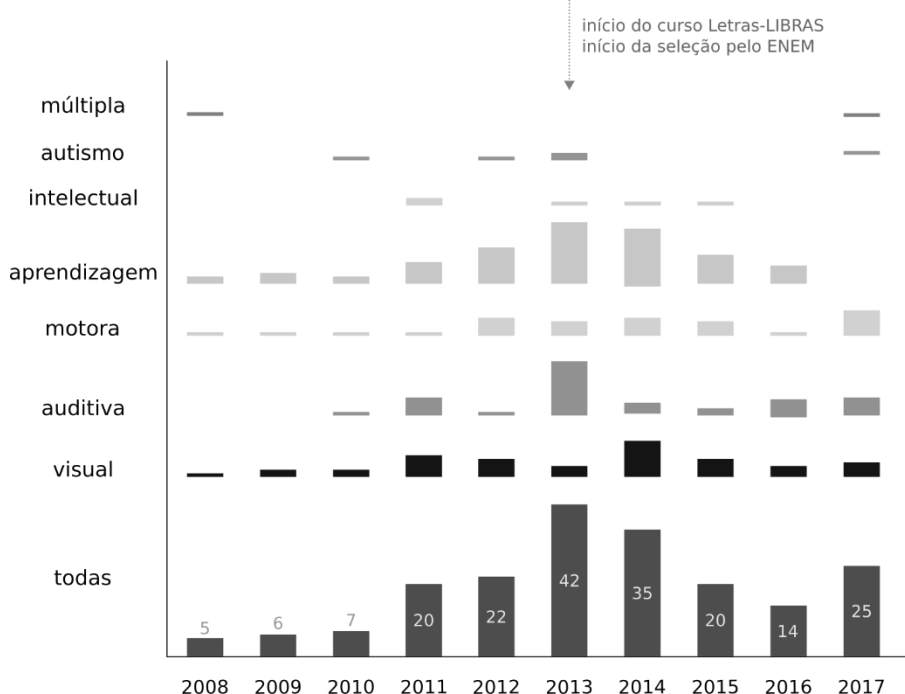


Figura 1. Quantidade de alunos com deficiência ingressantes na UFRN. Fonte: CAENE.

Que conclusão correta você consegue chegar analisando o gráfico acima?

- a) A criação do curso de Letras Português-LIBRAS fez aumentar significativamente o número de alunos com deficiência na UFRN nos últimos anos.
- b) Os deficientes auditivos formam o maior grupo de alunos com deficiência na universidade nos últimos anos.
- c) Com o passar dos anos, a tendência de termos alunos com deficiência na UFRN só aumenta.
- d) A criação do curso de Letras-Libras atendeu uma grande demanda reprimida de pessoas com deficiência auditiva, mas logo o ingresso desse público voltou a se equilibrar na UFRN.

29. Quando se elabora um gráfico, deve existir uma relação entre o valor representado e a forma de representação. Dentre os vários recursos visuais utilizados no gráfico abaixo, o ano foi associado à cor e a área da circunferência foi associada à quantidade de projetos.

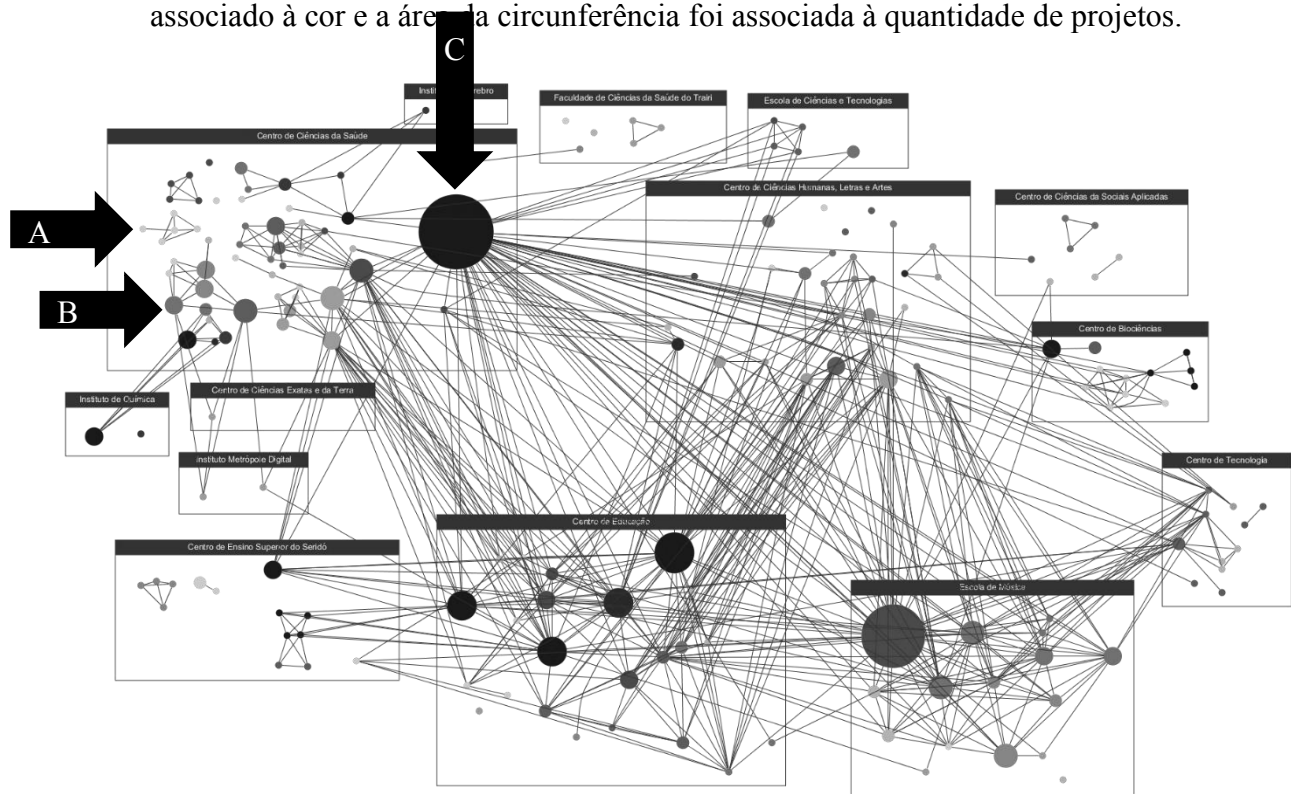


Figura 2. Colaboração entre docentes em projetos de extensão na UFRN entre 2008 e 2017 relacionados com deficiência. Fonte: Dados abertos da UFRN e SIGAA



Escolha uma opção de resposta abaixo cujos valores sejam mais coerentes com as representações das circunferências indicadas pelas setas A, B e C no gráfico acima.

A

	ano	quantidade de projetos
A	2014	34
B	2012	10
C	2007	56

c)

	ano	quantidade de projetos
A	2014	9
B	2012	26
C	2007	72

b)

	ano	quantidade de projetos
A	2013	34
B	2015	10
C	2007	56

d)

	ano	quantidade de projetos
A	2008	23
B	2012	40
C	2015	12

30. Estatísticas descritivas são medidas que sumarizam dados amostrais, sendo algumas delas mais ou menos sensíveis a valores extremos. Em termos de medidas de tendência central, qual das seguintes medidas tem seu valor mais sensível a valores extremos?

- a) Média
- b) Mediana
- c) Amplitude
- d) Desvio padrão