

-- CONHECIMENTOS ESPECÍFICOS --

O esquema lógico para a criação de um banco de dados no modelo relacional de um sistema de beneficiamento de produção de café tem as seguintes entidades: Fazenda, Funcionario, Supervisor, Plantacao, Lote_de_Cafe, Analista_de_qualidade, Etapa_de_Beneficiamento. Nesse esquema, tem-se que:

- uma fazenda pode ter várias plantações;
- um funcionário pode ser um supervisor ou um analista de qualidade;
- uma plantação pode produzir vários lotes de café;
- um lote de café pode passar por várias etapas de beneficiamento;
- cada etapa de beneficiamento tem um supervisor responsável;
- um lote de café pode ser analisado quanto à qualidade por vários analistas.

Com base no cenário apresentado, julgue os itens a seguir.

- 76** Para garantir a integridade referencial e a restrição de que somente supervisores sejam responsáveis pelo processo de beneficiamento, a tabela `Etapa_de_Beneficiamento` deve ter uma chave estrangeira referenciando a tabela `Supervisor`, que deve conter uma chave estrangeira referenciando a tabela `Funcionario`.
- 77** Para a correta implementação do projeto físico do banco de dados, serão necessárias duas tabelas de associação.

Julgue os próximos itens, a respeito de bancos de dados relacionais e não relacionais.

- 78** Bancos de dados do tipo colunar oferecem alta escalabilidade devido à sua capacidade de particionar dados verticalmente, o que permite a distribuição eficiente das colunas em múltiplos nós em ambientes distribuídos.
- 79** Em algumas implementações de bancos de dados relacionais, como PostgreSQL, é possível criar atributos do tipo *array*, o que permite armazenar múltiplos valores em uma única coluna, contradizendo o princípio original do modelo relacional de não permitir atributos multivalorados, pois cada coluna de uma tabela deve armazenar apenas um valor por atributo.
- 80** Bancos de dados NO-SQL são mais adequados para situações em que os dados tenham de ser altamente consistentes e transacionais, uma vez que sua estrutura flexível permite validações de integridade referencial e controles complexos de relacionamento entre os dados.
- 81** O modelo relacional, embora seja eficiente para operações de leitura e escrita simples, pode se tornar ineficiente em situações que exijam processamento de grandes volumes de dados semiestruturados, como *logs* ou eventos. Já o modelo orientado a documentos é mais flexível para consultas complexas que envolvam múltiplas junções entre documentos, oferecendo melhor desempenho em comparação ao modelo relacional.

Em relação a BI (*Business Intelligence*) e visualização de dados, julgue os itens que se seguem.

- 82** O gráfico de dispersão pode ser uma ferramenta útil para explorar a relação entre as variáveis, possibilitando uma análise visual inicial; no entanto, técnicas adicionais, como suavização por regressão local, transformação de variáveis ou o uso de gráficos alternativos (como histogramas bivariados), podem ser necessárias para revelar padrões mais claramente.
- 83** Para um *data warehouse* com grande volume de dados históricos e consultas frequentes, o esquema estrela geralmente apresenta melhor desempenho que o esquema *snowflake*, pois minimiza a complexidade das junções entre tabelas.
- 84** Em uma arquitetura de BI típica, o *data warehouse* integra dados de várias fontes externas, realizando extração, transformação e carga (ETL) antes de eles serem disponibilizados para análise.

Acerca de técnicas utilizadas na ciência de dados, julgue os itens a seguir.

- 85** O algoritmo Naive Bayes é inadequado para, por exemplo, a análise de dados de produção agrícola na previsão da qualidade de grãos de café, pois exige uma independência completa entre variáveis como pH do solo, umidade e temperatura, e não apresenta bons resultados quando essas variáveis são correlacionadas.
- 86** A PCA (análise de componentes principais) é uma técnica que transforma variáveis correlacionadas em componentes principais ortogonais, o que permite a redução da dimensionalidade dos dados; a seleção dos componentes principais é realizada com base na variância explicada por cada componente.
- 87** O algoritmo Apriori gera primeiro todos os *itemsets* individuais (itens) e, em seguida, combina-os para gerar *itemsets* de tamanho maior. Já o algoritmo FP-growth constrói uma estrutura de dados compacta denominada FP-tree, permitindo a mineração de *itemsets* frequentes sem gerar candidatos. Quanto à eficiência, o Apriori é mais rápido que o FP-growth devido à simplicidade de sua abordagem.
- 88** O K-means exige a definição do número de *clusters* como parâmetro de entrada e tem um desempenho eficiente em grandes conjuntos de dados, mas é sensível a *outliers* e só funciona bem para *clusters* esféricos e de densidade semelhante.
- 89** As árvores de decisão são uma técnica de aprendizado supervisionado que é adequada para analisar, por exemplo, a influência de fatores na qualidade da carne bovina em um sistema de produção, permitindo avaliar variáveis tais como alimentação, genética, manejo e peso ao abate e gerando regras interpretáveis, que auxiliam na tomada de decisão com base nessas características.
- 90** A métrica *lift* é usada para medir a dependência entre dois itens em uma regra de associação: se o *lift* de uma regra de associação entre dois itens A e B for maior que 1, então A e B obrigatoriamente aparecerão em conjunto mais frequentemente do que seria esperado se A e B fossem independentes.

Julgue os próximos itens, a respeito de gestão e governança de dados.

- 91** A Infraestrutura Nacional de Dados Espaciais (INDE) é um componente estratégico para a governança de dados espaciais no Brasil, pois define padrões e promove a integração, a interoperabilidade e o compartilhamento de dados geoespaciais entre órgãos públicos, setores privados e a sociedade.
- 92** Os metadados são fundamentais para a governança de dados, pois fornecem informações sobre a origem, a estrutura, o significado e a qualidade dos dados, contribuindo para a rastreabilidade, segurança, interoperabilidade e conformidade regulatória.
- 93** O DMBok (*Data Management Body of Knowledge*) é um *framework* que descreve as melhores práticas de gestão de dados, focando nas tecnologias de armazenamento e processamento de dados, sem considerar aspectos relacionados a governança, qualidade ou segurança dos dados.

Considerando os métodos HTTP utilizados em APIs REST, julgue os próximos itens, a respeito de integração de dados e mecanismos de interoperabilidade.

- 94** Os métodos GET e HEAD são considerados seguros, pois sua execução não deve modificar os dados armazenados no servidor, embora possa gerar efeitos colaterais indiretos, como registros de *logs*.
- 95** O método DELETE não é idempotente, pois, a partir da segunda execução, pode gerar respostas inconsistentes do servidor.
- 96** O método POST é seguro e idempotente, pois a execução de múltiplas requisições resulta no mesmo estado final dos dados.

Acerca de inteligência artificial e *machine learning*, julgue os itens que se seguem.

- 97** Diferentemente do aprendizado supervisionado e não supervisionado, o aprendizado por reforço baseia-se em um agente que interage com um ambiente e recebe recompensas ou penalidades conforme suas ações, buscando maximizar um retorno cumulativo ao longo do tempo.
- 98** Para avaliar o desempenho de um classificador em problemas de classificação com classes significativamente desbalanceadas, a métrica acurácia é a mais adequada, pois considera a proporção de previsões corretas em relação ao total de amostras.
- 99** O aprendizado não supervisionado se caracteriza pela utilização de dados previamente rotulados para treinar um modelo de *machine learning*, permitindo que ele aprenda padrões automaticamente a partir dos rótulos desses dados.

Com referência à matemática computacional e à ciência da computação aplicadas, julgue os itens a seguir.

- 100** A convolução é uma operação matemática que combina duas funções para produzir uma terceira, modificando a forma de uma função com base em outra.
- 101** A notação Big O é utilizada para descrever o comportamento assintótico de um algoritmo, fornecendo um limite superior para o tempo de execução ou uso de memória em função do tamanho da entrada.

Com relação a processos de banco de dados ETL (extração, transformação e carga), julgue os itens subsequentes.

- 102** Na modelagem de *data warehouses*, a abordagem *top-down* torna o processo ETL mais flexível e adaptável a mudanças nos requisitos de negócio, enquanto a abordagem *bottom-up* exige um ETL rígido e pouco adaptável a novas necessidades.
- 103** Na extração incremental, apenas os dados alterados desde a última carga são extraídos, tornando o processo mais eficiente que a extração *full*, que recupera todos os dados da fonte.

Um pesquisador está analisando a relação entre duas variáveis: X (quantidade de certo fertilizante aplicado em uma produção) e Y (produtividade agrícola). Sabendo que o par (X, Y) segue, conjuntamente, uma distribuição normal bivariada, ele obteve a média condicional na forma $E[Y|X = x] = 500 + 30x$ para descrever a produtividade esperada com base na quantidade x desse tipo de fertilizante, além das informações apresentadas a seguir.

| | |
|-----------------------|--------------------|
| valor esperado de X | 50 kg por hectare |
| desvio padrão de X | 5 kg por hectare |
| desvio padrão de Y | 200 kg por hectare |

Com base nessas informações, julgue os itens a seguir.

- 104** O desvio padrão de $E[Y|X]$ é igual a 150 kg por hectare.
- 105** A variância de $Y + X$ é igual a 2.525 (kg/hectare)².
- 106** A correlação linear entre X e Y é superior a 0,80.
- 107** O valor esperado de Y é igual a 2.000 kg por hectare.

Um pesquisador está avaliando a diferença na produtividade agrícola entre duas técnicas A e B de manejo do solo, testando a hipótese nula $H_0: \mu_A - \mu_B = 0$ contra a hipótese alternativa $H_1: \mu_A - \mu_B > 0$, em que μ_A e μ_B denotam as produtividades médias populacionais proporcionadas por essas técnicas.

A partir dessa situação hipotética, julgue os próximos itens.

- 108** Se \bar{X}_A e \bar{X}_B denotarem as produtividades médias amostrais correspondentes às técnicas de manejo A e B , a hipótese nula deverá ser rejeitada se $\bar{X}_A - \bar{X}_B > c$, em que $c > 0$ representa um valor crítico que depende do nível de significância do teste e da forma da distribuição amostral da diferença $\bar{X}_A - \bar{X}_B$.
- 109** O poder de um teste de hipóteses é definido como a probabilidade de se aceitar a hipótese H_0 quando a hipótese H_0 , de fato, for verdadeira.
- 110** O nível descritivo do teste (p -valor) representa a probabilidade de a hipótese nula ser verdadeira; se, por exemplo, o p -valor fosse igual a 0,001, haveria evidências fortes para se rejeitar H_0 .