-- CONHECIMENTOS ESPECÍFICOS --

Considerando uma distribuição condicional expressa na forma de função de densidade de probabilidade $f(x|y) = ye^{-xy}$, em que e denota a constante de Euler, e x e y, valores reais positivos que representam, respectivamente, os pontos de suporte das variáveis aleatórias contínuas X e Y, julgue os itens a seguir.

51 Se Y seguir uma distribuição exponencial com média igual a 1, então, para $x > 0$, a função de densidade da variável aleatória X será $f(x) = (x+1)^{-2}$.

52 $P(X > 1 | Y = 2) = e^{-2}$.

53 A média da variável aleatória X, condicionada ao evento $Y = 5$, é igual a 5.

54 $P(X = 1) = P(Y = 10)$.

55 As variáveis aleatórias X e Y são independentes.

Os valores 2, 3, 1, 0, 2 constituem uma amostra aleatória simples de tamanho 5 retirada de uma distribuição discreta W, na qual $P(W = w) = p(1-p)^w$, com $w \in \{0, 1, 2, \dots\}$, sendo p um parâmetro que denota uma probabilidade.

Com base nas informações precedentes e no método de estimação por máxima verossimilhança, julgue os próximos itens.

56 O estimador de máxima verossimilhança para a variância de W é a variância amostral.

57 Não há estimador de máxima verossimilhança para a moda de W, já que o valor da moda não depende da probabilidade p.

58 Se $\hat{P}(W = 2)$ denota a estimativa de máxima verossimilhança da probabilidade $P(W = 2)$, então $\hat{P}(W = 2) = 0{,}4$.

59 A estimativa de máxima verossimilhança da probabilidade p é igual a 0,625.

60 Pelo método da máxima verossimilhança, a estimativa da média de W é igual a $\frac{8}{5}$.

Uma amostra aleatória simples de tamanho n será retirada, com reposição, de certa população para a estimação de um parâmetro populacional λ. O estimador, representado por T_n, possui as propriedades $E[T_n] = \frac{(n+2)\lambda}{n}$ e $\text{Var}[T_n] = \frac{\lambda^2}{n}$.

No que diz respeito ao estimador hipotético T_n do parâmetro λ, julgue os seguintes itens.

61 Se T_n seguir uma distribuição normal, então a razão $\frac{nT_n - (n+2)\lambda}{\sqrt{n}\lambda}$ será normal padrão.

62 T_n é estimador de λ assintoticamente não viciado.

63 O erro-padrão de T_n é igual a 1.

64 T_n é estimador consistente.

65 Se $n = 10$, então o erro quadrático médio de T_n será igual a $\frac{\lambda^2}{10}$.

Acerca das abordagens relacional e não relacional, entre outros conceitos relacionados a banco de dados, julgue os itens a seguir.

66 O exemplo a seguir exibe como os dados podem ser armazenados em um banco de dados relacional do tipo chave-valor.

```
{
  nome_orgao: "ANATEL",
  endereco: {estado: "Distrito Federal",
cidade: "Brasília"},
  numero_telefone: "55-61-1234-5678",
  atuacao: ["telecomunicação"],
}
```

67 Nos bancos de dados NoSQL orientados a colunas, é possível organizar os dados em famílias de colunas, o que facilita o agrupamento e o acesso a informações específicas, como, por exemplo, os dados de contato.

68 Em SQL, *triggers* são procedimentos automáticos executados em resposta a determinados eventos em uma tabela ou *view* e são disparados exclusivamente em resposta a ações DML, tais como `INSERT`, `UPDATE` ou `DELETE`.

69 Em um banco de dados, os índices são armazenados fisicamente na mesma ordem em que são definidos logicamente para permitir acesso eficiente aos dados.

70 Se executada, a consulta a seguir retorna a contagem de clientes com mais de 18 anos de idade que realizaram transações com valores superiores a 1.000, agrupados e ordenados de forma decrescente por idade.

```
SELECT COUNT(c.id), c.idade
FROM cliente c
WHERE c.idade > 18
  AND EXISTS (
    SELECT 1
    FROM transacao t
    WHERE t.cliente_id = c.id
  AND t.valor > 1000
  )
GROUP BY c.idade
ORDER BY c.idade DESC;
```

Julgue os itens que se seguem, relativos a conceitos de *data warehouse*, técnicas de modelagem dimensional e otimização de bases de dados para BI.

71 Nos *data warehouses*, os índices de junção associam as linhas da tabela de fatos com as respectivas colunas na tabela de dimensões, o que facilita as operações de consulta e análise de dados.

72 Uma tabela de fatos descreve as entidades de um negócio e apresenta uma ou mais colunas-chave que funcionam como um identificador único exclusivo, bem como colunas descritivas adicionais.

73 Em regra, as chaves substitutas são incorporadas nas tabelas de dimensões de um *data warehouse* relacional para atribuir um identificador único para cada registro nas tabelas de dimensões.

74 Nos *data warehouses*, as colunas de chave de dimensão definem a dimensão de uma tabela de fatos, enquanto os valores das chaves de dimensão estabelecem sua granularidade.

75 A relação entre uma tabela de fatos e suas respectivas tabelas de dimensões em um *data warehouse* é estabelecida pela chave primária da tabela de fatos e a chave estrangeira da tabela de dimensões.

76 A consolidação de dimensões de floco de neve em um modelo de tabela única pode implicar o armazenamento de dados não normalizados e redundantes, o que resulta em maior tamanho de armazenamento.

77 O processo de otimização de desempenho, em bases de dados, consiste em redimensionar o tamanho do modelo semântico, o que resulta em uma maior rapidez na atualização dos dados, cálculos e renderização dos elementos visuais em relatórios.

Em relação ao tratamento e à qualidade dos dados no sistema de gerenciamento de informações, julgue os itens subsequentes.

78 De acordo com a Lei Geral de Proteção de Dados Pessoais (LGPD), os dados pessoais devem ser retidos por um período de cinco anos, mesmo após a conclusão do seu processamento, desde que sejam cumpridos os limites técnicos das atividades em questão.

79 O agente de tratamento deve demonstrar a adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais, bem como demonstrar a eficácia dessas medidas.

A respeito de visualização, análise exploratória de dados e geoprocessamento, julgue os seguintes itens.

80 A definição de seções implícitas pode ser realizada pela utilização de formas coloridas e pela sobreposição de elementos visuais alinhados, o que confere uma clara distinção entre diferentes áreas do *layout* dos *dashboards*.

81 Os mapas temáticos incluem áreas geográficas delineadas por um ou mais polígonos, a exemplo dos mapas de uso do solo e mapas que indicam a capacidade agrícola de determinada região.

82 A aplicação de hierarquias de atributo em modelos analíticos é uma prática eficiente, que permite a organização de valores pré-agregados em cada nível e otimiza a análise de dados.

83 No contexto de um *layout* de relatório, o equilíbrio assimétrico consiste na distribuição do peso uniforme dos objetos em ambas as metades da página, independentemente do tamanho dos objetos.

No que se refere à governança de dados, julgue os próximos itens.

84 De acordo com a LGPD, dados pessoais relacionados a convicção religiosa são considerados sensíveis e só podem ser tratados com o consentimento específico e destacado do titular ou de seu representante legal.

85 Um dado de referência, como, por exemplo, os dados dos clientes, é uma informação crucial para a operação do negócio.

Julgue os próximos itens, relativos a Naive Bayes e *random forest*.

86 Tamanho do nó, número de árvores e número de recursos amostrados, ou número de preditores amostrados, são parâmetros de algoritmos *random forest*.

87 Naive Bayes é um algoritmo de classificação baseado na aprendizagem por reforço, em que um agente realiza uma ação e recebe uma recompensa de acordo com o resultado dessa ação por meio da implementação do teorema de Bayes, com o objetivo de encontrar a probabilidade a *posteriori*.

88 *Random forest* é um algoritmo de classificação que permite a realização de mineração dos dados por meio da criação de estruturas de aprendizagem a partir de uma base de dados na qual se utiliza uma única árvore de decisão para a classificação dos dados.

89 O algoritmo de classificação Naive Bayes pode ser utilizado para o cálculo da probabilidade de ocorrência de um evento, com base em probabilidades obtidas em eventos numéricos passados, e, por isso, não pode ser empregado em atividades de classificação textual.

90 Nas árvores de decisão e em *random forest*, são utilizadas técnicas estatísticas com o objetivo de se produzir, a partir de um conjunto de observações, uma predição de valores em função de uma ou mais variáveis independentes contínuas e(ou) binárias.

A respeito de KNN (*k-nearest neighbours*), SVM (*support vector machines*), *deep learning* e técnicas de agrupamento, julgue os itens a seguir.

91 *Deep learning* é um tipo de aprendizado de máquina que usa redes neurais artificiais para permitir que sistemas digitais aprendam e tomem decisões com base em dados não estruturados e não rotulados.

92 KNN é um algoritmo de aprendizado supervisionado não paramétrico que não pode ser utilizado em problemas de classificação, uma vez que seu objetivo é prever valores numéricos e não valores categóricos.

93 SVM é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para desafios de classificação ou regressão.

94 Uma das formas de se realizar um agrupamento é por meio de técnicas de agrupamento baseadas em hierarquia, em que se pode criar estrutura hierárquica de acordo com a proximidade entre os indivíduos, o que resulta em uma árvore binária.

95 O SVM classifica os dados encontrando uma linha ou hiperplano ideal; essa linha de separação é encontrada entre duas classes distintas pela análise dos dois pontos, um de cada grupo, mais próximos da outra classe.

Acerca da avaliação de modelos de classificação, julgue os itens que se seguem.

96 A acurácia é uma métrica adequada para a avaliação de modelos quando não há desbalanceamento de classes, pois reflete com precisão a capacidade geral do modelo de fazer previsões corretas em todas as classes.

97 A matriz de confusão, em problemas de classificação multiclasses, é uma tabela com duas linhas e duas colunas; na diagonal principal dessa matriz quadrada, estão os valores corretos e, na matriz secundária, os erros cometidos pelo modelo.

98 Um modelo de classificação que apresenta alta revocação é útil em contextos em que seja crucial identificar a maior quantidade possível de casos positivos, mesmo que isso resulte em um número maior de falsos positivos.

99 A área sob a curva ROC (*receiver operating characteristic*) é uma métrica de qualidade útil para avaliar um modelo: quanto mais próxima a curva estiver do canto superior direito do gráfico, melhor será a predição do modelo.

A respeito de técnicas de redução de dimensionalidade, julgue os itens subsecutivos.

100 Quando da configuração dos parâmetros do *autoencoder*, o tamanho do espaço latente é uma informação crucial, pois determina o tamanho do espaço onde os dados de entrada serão comprimidos.

101 Para utilizar de forma adequada a análise de componentes principais (PCA, na sigla em inglês), é essencial normalizar os dados; se as variáveis não estiverem na mesma escala, aquelas com maior variância terão maior impacto, distorcendo o resultado da PCA.

Julgue os próximos itens, referentes ao processamento de linguagem natural.

102 A saída do Word2Vec consiste em vetores densos de baixa dimensão que representam palavras em um espaço contínuo, onde cada palavra é mapeada para um vetor numérico no qual cada dimensão captura uma característica da palavra.

103 A similaridade de cosseno é uma métrica pela qual se avalia a similaridade entre dois vetores com base no ângulo entre eles em um espaço vetorial, de forma que, à medida que os vetores se aproximarem, aumentará a similaridade de cosseno.

104 A lematização prescinde do POS *tagging* para que as palavras sejam reduzidas corretamente, pois todas as palavras são reduzidas ao mesmo *lemma*, independentemente de sua classe gramatical.

105 Na redução de palavras ao radical, ocorre *under-stemming* quando duas palavras separadas são reduzidas erroneamente à mesma raiz e, com isso, ocorre a perda de distinção semântica entre palavras com significados diferentes.

A respeito de *Big Data*, julgue os próximos itens.

106 *Pipelines* de dados apresentam uma única estrutura para o recebimento dos dados originados de uma fonte não confiável.

107 Em *Big Data*, ruídos consistem em informações extras que acabam deturpando as análises, enquanto *overfitting* designa a interpretação equivocada dos ruídos como dados legítimos.

108 A coleta de dados por meio de aplicativos é considerada explícita, porque o usuário a autoriza.

109 O YARN (*Yet Another Resource Negotiator*) é um sistema de arquivos distribuídos que faz parte do *framework* Hadoop.

110 No modelo SaaS (*software as a service*) da computação em nuvem utilizado para *Big Data*, a aplicação e os dados são gerenciados pelo provedor da nuvem.

111 Projetos de *Big Data*, quando necessário, crescem horizontalmente, com a inclusão de novos nodos, e verticalmente, com o acréscimo de mais memória.

112 Em um *data lake*, os dados são depositados em estado bruto, sem terem passado por qualquer análise e mesmo sem terem uma governança.

113 Subconjunto de um *data warehouse*, o *data mart* é especializado em uma área específica de uma organização.

114 O processo de ELT, devido às suas etapas, exige maior definição de regras, estruturas e relações do que a abordagem ETL.

115 No processamento ROLAP, bancos de dados relacionais são utilizados como local de armazenamento para agregação, enquanto, nos processamentos MOLAP e HOLAP, utilizam-se bancos de dados multidimensionais.

Acerca do fluxo de *Big Data*, julgue os itens que se seguem.

116 Na apresentação de dados, a extração de subcoleções e a consulta de parâmetros permitem a navegação em diversos cenários da visualização.

117 *Streaming processing* é uma tecnologia de *Big Data* exclusiva para atender processamentos de serviços de *streaming* de áudio e vídeo.

118 O serviço ElasticSearch utiliza índices divididos em fragmentos, de maneira que cada nó armazena diversos fragmentos e atua na coordenação das operações nos vários fragmentos.

119 As funções do MapReduce transformam um volume grande de dados em grupamentos segmentados, mantendo na saída a mesma quantidade de dados da entrada.

120 Na etapa de captura de *Big Data*, grandes volumes de dados são armazenados em bancos de dados NoSQL, devido à sua escalabilidade e à sua flexibilidade.