15

# Instituto de Pesquisa Econômica Aplicada

## PROVAS OBJETIVAS - TARDE
## Cargo: Técnico de Planejamento e Pesquisa

Concurso Público – IPEA-2023
EDITAL Nº 01/2023 - IPEA
de 29 de novembro de 2023.

## Perfil II - Planejamento e Gestão do Conhecimento e de Dados
### Especialidade: CIÊNCIA DE DADOS

## LEIA ATENTAMENTE AS INSTRUÇÕES ABAIXO.

**01 -** O candidato recebeu do fiscal o seguinte material:

**a)** este **Caderno de Questões**, com o enunciado das 70 (setenta) questões objetivas, sem repetição ou falha, com valor de 1,0 ponto cada.

**b) Cartão-Resposta** destinado às respostas das questões objetivas formuladas nas provas.

**02 -** O candidato deve verificar se este material está em ordem e se o seu nome e número de inscrição conferem com os que aparecem no **Cartão-Resposta**. Caso não esteja nessas condições, o fato deve ser **IMEDIATAMENTE** notificado ao fiscal.

**03 -** Após a conferência, o candidato deverá assinar, no espaço próprio do **Cartão-Resposta**, com **caneta esferográfica de tinta preta, fabricada em material transparente**.

**04 -** O candidato deve ter muito cuidado com o **Cartão-Resposta**, para não o **dobrar**, **amassar** ou **manchar**. O **Cartão-Resposta SOMENTE** poderá ser substituído se, no ato da entrega ao candidato, já estiver danificado.

**05 -** Logo após a autorização para o início das provas, o candidato deve conferir se este **Caderno de Questões** está em ordem e com todas as páginas. Caso não esteja nessas condições, o fato deve ser **IMEDIATAMENTE** notificado ao fiscal.

**06 -** Para cada uma das questões objetivas, são apresentadas 5 alternativas classificadas com as letras (A), (B), (C), (D) e (E); só uma responde adequadamente ao quesito proposto. O candidato só deve assinalar **UMA** letra no **Cartão-Resposta**, preenchendo todo o espaço compreendido pelos círculos, com **caneta esferográfica de tinta preta, fabricada em material transparente**, de forma contínua e densa. A leitura óptica do **Cartão-Resposta** é sensível a marcas escuras; portanto, os campos de marcação devem ser preenchidos completamente, sem deixar claros. A marcação em mais de uma alternativa anula a questão**, mesmo que uma das respostas esteja correta.**

Exemplo: Ⓐ ● Ⓒ Ⓓ Ⓔ

**07 -** As questões objetivas são identificadas pelo número que se situa acima do seu enunciado.

**08 - Será eliminado** deste Concurso Público o candidato que

**a)** for surpreendido, durante as provas, em qualquer tipo de comunicação com outro candidato;

**b)** portar ou usar, durante a realização das provas, aparelhos sonoros, fonográficos, de comunicação ou de registro, eletrônicos ou não, tais como agendas eletrônicas e(ou) similares, gravadores, *pen drive*, mp3 *player* e(ou) similar, fones de ouvido, chaves com alarme ou com qualquer outro componente eletrônico, relógios de qualquer natureza, telefones celulares, microcomputadores portáteis e(ou) similares;

**c)** se ausentar da sala em que se realizam as provas levando consigo o **Caderno de Questões** e(ou) o **Cartão-Resposta**;

**d)** se recusar a entregar o **Caderno de Questões** e/ou o **Cartão-Resposta**, quando terminar o tempo estabelecido;

**e)** não assinar a **lista de presença** e(ou) o **Cartão-Resposta**.

**Obs.** O candidato só poderá ausentar-se do recinto das provas após **duas horas** contadas a partir do efetivo início das mesmas. Por motivos de segurança, o candidato **NÃO poderá levar o Caderno de Questões**, a qualquer momento.

**09 -** O candidato deve reservar os 30 (trinta) minutos finais para marcar seu **Cartão-Resposta**. Os rascunhos e as marcações assinaladas no **Caderno de Questões NÃO serão levados em conta**.

**10 - O tempo disponível para estas provas de questões objetivas é de 4 (quatro) horas**, já incluído o tempo para marcação do seu **Cartão-Resposta**, findo o qual o candidato deverá, obrigatoriamente, entregar o **Cartão-Resposta**, o **Caderno de Questões** e **assinar a Lista de Presença**.

**11 -** As questões e os gabaritos das Provas Objetivas serão divulgados a partir do primeiro dia útil após sua realização, na página da **Fundação Cesgranrio** (**www.cesgranrio.org.br**).

FUNDAÇÃO cesgranrio

## CONHECIMENTOS ESPECÍFICOS

**1**

O processo de ingestão de dados é normalmente dividido em três etapas principais:

1 - Extração, ou coleta, de dados das fontes disponíveis;
2 - Transformação dos dados coletados para que atendam às necessidades específicas de processamento e análise; e
3 - Carga dos dados em algum repositório de destino, como um banco de dados relacional ou um data lake.

Essas três etapas podem variar dependendo de os dados serem estruturados ou não.

Nesse contexto, verifica-se que, na etapa de

(A) carga, os dados estruturados são sempre transferidos diretamente ao repositório de destino, sem necessidade de transformação.
(B) carga, os dados não estruturados são sempre convertidos em formatos estruturados antes de serem armazenados.
(C) extração, os dados estruturados são coletados exclusivamente através de APIs especializadas.
(D) transformação, os dados estruturados podem requerer conversão para um formato não estruturado para facilitar a análise avançada.
(E) transformação, os dados não estruturados podem necessitar de processamento de linguagem natural ou de técnicas de reconhecimento de imagens.

**2**

No contexto da ciência de dados em cloud computing, e considerando-se as soluções de Plataforma como Serviço (PaaS), as de Infraestrutura como Serviço (IaaS) e as de Software como Serviço (SaaS), constata-se que

(A) IaaS é mais adequado que PaaS para o desenvolvimento rápido de aplicações de ciência de dados, pois oferece ferramentas e bibliotecas específicas para ciência de dados pré-instaladas.
(B) PaaS oferece maior controle sobre o hardware e os sistemas operacionais, tornando-o mais adequado para tarefas de ciência de dados que requerem customizações profundas no nível do sistema.
(C) PaaS fornece um ambiente de desenvolvimento e execução pronto para uso, eliminando a necessidade de gerenciar a infraestrutura subjacente, o que é comum em IaaS.
(D) SaaS é geralmente mais flexível que PaaS no que diz respeito à instalação e à configuração de ambientes de ciência de dados, pois permite um maior grau de personalização e controle.
(E) SaaS oferece mais opções para escalabilidade e distribuição de dados do que PaaS, tornando-o a escolha preferencial para aplicações de ciência de dados que precisam de alta escalabilidade.

**3**

O Ecossistema Spark tem componentes que oferecem funcionalidades específicas que o tornam uma ferramenta versátil e eficiente para o processamento de grandes volumes de dados, a análise em tempo real, o aprendizado de máquina e muito mais. Essa integração e flexibilidade são algumas das razões pelas quais o Spark se tornou uma ferramenta amplamente utilizada em aplicações de Big Data. Os componentes da Plataforma Spark pertencem a dois grupos principais: os componentes básicos e os componentes especializados, que provêm funcionalidades mais avançadas. Dentre os componentes básicos, podemos destacar o Spark Core, também conhecido como “coração” do Ecossistema, e que é responsável pelas tarefas consideradas essenciais.

O componente Spark Core

(A) é um módulo que atua como sistema de gerenciamento de banco de dados relacional para processamento de dados estruturados e distribuídos em um contexto Big Data.
(B) é uma biblioteca para processamento de gráficos e análises de dados distribuídos, sem esquema e não estruturados, características principais em um ambiente Big Data.
(C) é uma ferramenta responsável pelo tratamento e análise de streams de dados em tempo real, específica de grandes volumes, com velocidade e variedade, características típicas em sistemas de Big Data.
(D) corresponde ao módulo que contempla funcionalidades de computação distribuída, incluindo agendamento de tarefas e gerenciamento de memória, para aplicações Big Data.
(E) diz respeito ao principal add-on da plataforma que, por ser um framework orientado a objetos, tem esse hot-spot para melhorar a eficiência do armazenamento de dados distribuídos de aplicações Big Data em clusters.

**4**

Existem várias abordagens para a ingestão de dados, sendo cada uma delas adequada para determinado tipo de necessidade e de cenário.

No caso da ingestão de dados em tempo real, streaming, os dados são

(A) coletados e processados em intervalos regulares, por exemplo, diariamente ou semanalmente.
(B) capturados e processados continuamente à medida que são gerados.
(C) processados em pequenos lotes, com o processamento ocorrendo em intervalos curtos, mas não instantâneos.
(D) processados apenas após um evento específico ser acionado, como, por exemplo, uma transação em banco de dados ou um clique de usuário.
(E) armazenados em um data lake ou data warehouse, antes de qualquer forma de processamento ou de análise.

**5**

Para melhorar o processamento de grandes volumes de dados através de computação paralela ou distribuída, pode-se utilizar sharding, técnica que divide os dados em partes menores, chamadas shards. Essas partes são normalmente armazenadas em diferentes nós, ou sítios, de processamento em um sistema distribuído.

O sharding é necessário para garantir a

(A) privacidade dos dados
(B) eficiência do processamento de dados
(C) disponibilidade dos dados em caso de falha de um nó
(D) escalabilidade do sistema de processamento de dados
(E) consistência dos dados em caso de alteração de um shard

**6**

Apesar de existirem diversas aplicações reais em que há necessidade de ingestão periódica de dados, em algumas a ingestão em lote pode não ser vantajosa, como, por exemplo, em

(A) análise e planejamento de consumo de energia.
(B) análises históricas e de tendências no mercado de ações.
(C) integrações de sistemas empresariais, como o CRM e o ERP.
(D) monitoramento e gerenciamento de tráfego urbano ou de fenômenos climáticos.
(E) relatórios corporativos para análise de tendências e de desempenho de vendas.

RASCUNHO

**7**

O Banco Nossa Caixa é o primeiro banco do hemisfério sul a implantar o GDPS (Processamento Paralelo em Locais Geograficamente Distantes, na sigla em inglês), sistema que processa, simultaneamente, todos os dados da instituição em dois locais fisicamente separados. A tecnologia garante o armazenamento e a continuidade do funcionamento de todos os canais de atendimento e de negócios do banco, ainda que a operação de um dos equipamentos responsáveis por essas funções seja interrompida por um blecaute ou incêndio, por exemplo. O banco investiu R$ 80 milhões no projeto, que levou quatro anos desde a concepção até a implantação.

Do Banco Nossa Caixa

Disponível em: https://www.saopaulo.sp.gov.br/ultimas-noticias/nossa-caixa-usa-sistema-pioneiro-de-processamento-de-dados/. Acesso em: 5 jan. 2024.

No modelo de arquitetura Massively Parallel Processor, MPP, extensibilidade e escalabilidade são características comuns e vantajosas no armazenamento e no processamento de dados paralelos.

Nesse contexto, qual modelo de processamento se enquadra à arquitetura MPP?

(A) disco compartilhado (shared-disk)
(B) memória compartilhada (shared-memory)
(C) memória distribuída (shared-nothing)
(D) três esquemas (ANSI/SPARC)
(E) visões materializadas (materialized view)

**8**

Sistemas de bancos de dados apresentam benefícios e desafios potencializados quando é possível adotar uma solução de gerência distribuída, coordenada por um sistema de banco de dados distribuído.

Nesse contexto, uma importante questão de projeto refere-se à

(A) autonomia, dimensão a ser considerada na arquitetura de um banco de dados distribuído, pois indica como a distribuição dos dados será processada, com a designação de fragmentos por nó.
(B) confiabilidade do sistema, que pode, em função da natureza dos dados e da natureza das aplicações clientes, ser relaxada, em prol do desempenho de execução das consultas.
(C) fragmentação vertical de dados, uma técnica que particiona um conjunto de dados em tuplas, ou registros, entre os nós do sistema distribuído.
(D) replicação de dados seja ela total ou parcial, pois demanda a implementação de protocolos que garantam a consistência das réplicas.
(E) execução de consultas, que se concentram na minimização dos custos de processador e de entrada e saída, sendo os custos de comunicação parcela de menor impacto no total.

**9**

Nas últimas décadas, a automatização e a inserção de máquinas agrícolas transformaram profundamente o panorama do trabalho nas áreas rurais [...] e, em menos de 50 anos, a produtividade do agronegócio brasileiro aumentou 400%. [...]

Esses resultados vieram com a adoção da tecnologia nos processos cotidianos, e também com o investimento em pesquisas [...]. Agora, o agronegócio pode estar diante de um novo salto de produtividade; big data e machine learning são ferramentas que estão ganhando espaço e que podem, novamente, transformar o cenário do campo.

Disponível em: https://summitagro.estadao.com.br/tendencias-e--tecnologia/como-big-data-e-machine-learning-sao-aplicados-no--agronegocio/. Acesso em: 5 jan. 2024. Adaptado.

A utilização da plataforma paralela de processamento MapReduce aplica-se adequadamente como um framework de processamento de Big Data, visando à escalabilidade para as aplicações.

Nesse contexto, uma característica inerente à MapReduce é a

(A) sua abordagem, que reflete uma solução simplificada de processamento de dados paralelos em um cluster.
(B) sua função map(), que gera, para cada registro de entrada, um ou mais pares (chave,valor).
(C) sua função reduce(), que gera um resultado combinado em valores que compartilham chaves distintas.
(D) sua demanda pela existência de um esquema de dados.
(E) natureza estática de suas funções, que devem respeitar a especificação original, caracterizando uma restrição inerente a essa abordagem.

**10**

Sistemas Gerenciadores de Bancos de Dados, SGBD, via de regra, permitem acesso às suas funcionalidades por meio de interfaces interativas e, também, por meio de programas de apoio ou comandos específicos. O SGBD PostgreSQL possui um conjunto de programas de apoio. O programa a ser utilizado para realizar o backup de um banco de dados ou de um esquema, no PostgreSQL, é o pg_dump.

Nesse contexto, considere o seguinte comando:

pg_dump -a -Fp Alfa > Beta

Esse comando gera um backup

(A) das definições (esquema) do banco de dados Alfa, apenas
(B) das definições (esquema) dos bancos de dados Alfa e Beta, em formato proprietário, apenas
(C) de todo o banco de dados Beta (esquema e dados) em formato proprietário
(D) dos dados do esquema Beta do banco de dados Alfa em formato textual (comandos SQL)
(E) dos dados do banco de dados Alfa em formato textual (comandos SQL)

**11**

Uma das principais características de sistemas gerenciadores de bancos de dados (SGBD) NoSQL, quando comparados aos sistemas gerenciadores bancos de dados relacionais (SGBDR), é que seu esquema é considerado flexível ou não existente (schemaless).

O esquema de um SGBD NoSQL ser flexível ou não existente tem como consequência o fato de que

(A) a qualidade de dados pode ser mais difícil de ser garantida, quando comparado a um SGBDR.
(B) a realização de adaptações para persistir conteúdos distintos é mais demorada do que em um SGBDR.
(C) o desempenho de uma consulta aos dados é maior do que em um SGBDR.
(D) os formatos de data são padronizados, da mesma forma que em SGBDR.
(E) dados binários de qualquer natureza podem ser persistidos, o que um SGBDR não consegue fazer.

RASCUNHO

**12**

Para um estudo do tema Educação, foram coletados dados de escolas e de professores em todos os municípios brasileiros. Esses dados foram armazenados em duas relações (tabelas), organizadas de acordo com o seguinte modelo relacional:

ESCOLA (cod-escola, nome-escola, cod-municipio, quantidade-alunos)

PROFESSOR (CPF, nome-prof, data-nascimento, cod-municipio-residencia, cod-escola-prof)

A chave primária de ESCOLA é cod-escola, e a de PROFESSOR é CPF. A coluna cod-escola-prof em PROFESSOR é uma chave estrangeira e indica em que escola o professor leciona. Considere a utilização dos operadores de Projeção ($\pi$ ou PROJETE), Seleção ($\sigma$ ou SELECIONE) e Junção ( $\bowtie$ ou JUNTE) da Álgebra Relacional.

Que sequência de operações, em Álgebra Relacional, produz como resultado uma relação R-X com CPF e nome dos professores que **NÃO** residem no mesmo município onde lecionam?

(A) R-X ← SELECIONE CPF, nome-prof de PROFESSOR onde cod-municipio-residencia é diferente de cod-municipio

(B) R-A ← JUNTE PROFESSOR a ESCOLA onde cod-municipio-residencia é diferente de cod-municipio
R-X ← PROJETE CPF, nome-prof de R-A

(C) R-A ← SELECIONE CPF, nome-prof de PROFESSOR
R-X ← JUNTE R-A a ESCOLA onde cod-escola-prof = cod-escola

(D) R-A ← JUNTE PROFESSOR a ESCOLA onde cod-escola-prof = cod-escola
R-B ← SELECIONE R-A onde cod-municipio-residencia é diferente de cod-municipio
R-X ← PROJETE CPF, nome-prof de R-B

(E) R-A ← PROJETE CPF, nome-prof de PROFESSOR
R-B ← JUNTE R-A a ESCOLA onde cod-escola-prof = cod-escola
R-X ← SELECIONE R-B onde cod-municipio-residencia é diferente de cod-municipio

**13**

Para a avaliação de políticas públicas na área de Segurança Alimentar e Nutricional, um município brasileiro utilizou dados persistidos em três relações (tabelas) organizadas de acordo com o seguinte modelo relacional:

PRODUTO (cod-produto, nome-produto, grupo-alimentar)
FORNECEDOR (CNPJ, nome-empresa, tipo)
COMPRADO (CNPJ, cod-produto, data, quantidade, valor)

Os atributos que formam as chaves primárias de cada tabela estão sublinhados.

Nesse contexto, considere o comando SQL apresentado a seguir.

```
SELECT P.cod-produto, SUM (quantidade)
FROM PRODUTO P, FORNECEDOR F, COMPRADO C
WHERE P.cod-produto = C.cod-produto
AND C.CNPJ = F.CNPJ
AND F.tipo = 'agricultura familiar'
GROUP BY P.cod-produto
HAVING SUM (quantidade) > 10000
```

Os resultados produzidos pela execução desse comando apresentam o código do produto e a soma das quantidades compradas dos produtos de

(A) fornecedores com mais de 10.000 produtos distintos.
(B) fornecedores do tipo 'agricultura familiar' que tiveram mais de 10.000 unidades compradas.
(C) fornecedores do tipo 'agricultura familiar' que fornecem mais de 10.000 produtos distintos.
(D) todos os fornecedores do tipo 'agricultura familiar'.
(E) produtos que tiveram mais de 10.000 unidades compradas.

**14**

Um cientista de dados precisa discretizar uma variável, representando distâncias entre cidades em quilômetros em 10 intervalos com, aproximadamente, o mesmo número de observações.

Nesse contexto, a técnica mais adequada é a discretização

(A) Binária
(B) K-Means
(C) de Entropia
(D) de Largura Igual
(E) de Frequência Igual

**15**

A limpeza de dados, *data cleansing*, é uma tarefa importante que pode ser complexa e demorada, no entanto é um investimento fundamental que pode melhorar a qualidade e a utilidade dos dados para futuras análises.
Seja um conjunto de dados com informações de saúde referentes a uma população. Pode-se limpar esses dados para identificar e tratar valores extremos, discrepantes, contraditórios ou inválidos. Com isso, há maior confiabilidade para estimar a prevalência, a incidência, a mortalidade e os fatores de risco de uma doença naquela população representada por aqueles dados.
Por exemplo, seja o conjunto de dados abaixo referente a uma amostra de 5 indivíduos em uma mesma cidade, na qual um analista percebeu a necessidade de limpeza de dados por conta de potenciais inconsistências.

Indivíduo 1: Sexo: Feminino; Idade: 8 anos; Altura: 1,15m; Peso: 40kg; Batimento Cardíaco em Repouso: 85 bpm
Indivíduo 2: Sexo: Masculino; Idade: 22 anos; Altura: 1,60m; Peso: 60kg; Batimento Cardíaco em Repouso: 72 bpm
Indivíduo 3: Sexo: Feminino; Idade: 40 anos; Altura: 1,60m; Peso: 55kg; Batimento Cardíaco em Repouso: 10 bpm
Indivíduo 4: Sexo: Masculino; Idade: 55 anos; Altura: 1,90m; Peso: 100kg; Batimento Cardíaco em Repouso: 70 bpm
Indivíduo 5: Sexo: Feminino; Idade: 70 anos; Altura: 1,50m; Peso: 60kg; Batimento Cardíaco em Repouso: 70 bpm

Qual ação é a única claramente necessária para realizar data cleansing neste conjunto de dados específico?

(A) Corrigir o valor da altura no indivíduo 1, pois parece anormalmente alto.
(B) Verificar o peso no indivíduo 2, pois nessa idade e altura deveria ser um valor maior.
(C) Confirmar o valor do batimento cardíaco no indivíduo 3, que parece anormalmente baixo.
(D) Alterar o peso no indivíduo 4, pois não condiz com os valores de sexo, idade e altura indicados.
(E) Modificar o valor da altura do indivíduo 5, pois é anormalmente baixo para a idade e sexo.

**16**

Algoritmos fuzzy matching em processamento de linguagem natural são métodos que permitem encontrar correspondências aproximadas entre strings, ou seja, sequências de caracteres, como palavras ou frases. Esses algoritmos são úteis para lidar com situações, tais como erros de digitação, variações ortográficas, sinônimos, abreviações. Eles também podem ser aplicados para comparar textos, extrair informações, classificar sentimentos, entre outras finalidades. Existem diferentes tipos de algoritmos fuzzy matching, como a Similaridade de Jaccard, que mede a proporção de elementos comuns entre dois conjuntos de strings.

Qual das palavras a seguir apresenta o maior valor da similaridade de Jaccard, quando comparada com a palavra "computador"?

(A) amputar
(B) amplificador
(C) calcular
(D) deputado
(E) senador

**17**

A partir de dados da pesquisa Perfil do Processado e Produção de Provas nas Ações Criminais por Tráfico de Drogas, realizada em dezembro de 2023 pelo Instituto de Pesquisa Econômica Aplicada (Ipea), é possível levantar informações sociodemográficas sobre os bairros em que o direito à inviolabilidade domiciliar é relativizado. Os resultados revelam que os bairros mais ricos e aqueles de população predominantemente branca são praticamente imunes às entradas em domicílio, as quais se concentram substancialmente nos bairros mais pobres e naqueles com população predominantemente negra ou minoritariamente branca.
Qual técnica de desidentificação de dados sensíveis é a mais adequada para preservar a privacidade dos indivíduos processados, permitindo, ainda, a análise sociodemográfica dos bairros?

(A) Anonimização: remover todos os identificadores diretos dos indivíduos, incluindo nome, CPF, RG e endereço, e também remover qualquer referência a bairro ou características sociodemográficas.
(B) Pseudonimização: substituir os identificadores diretos dos indivíduos por códigos ou símbolos, mas publicar a tabela de correspondência junto com os dados.
(C) Agregação: agrupar os dados dos indivíduos em categorias amplas, como faixa etária e classe social, mas excluir informações sobre cor ou raça e bairro.
(D) Perturbação: adicionar ruído ou alterar aleatoriamente todos os valores dos dados, incluindo informações sociodemográficas e de bairro, de forma a tornar os dados inúteis para análise específica de bairro.
(E) Generalização: substituir informações detalhadas de identificação pessoal por categorias mais amplas, mantendo dados sobre bairro e características sociodemográficas, como faixa etária, classe social, cor ou raça.

**18**

A deduplicação de dados é uma técnica importante no gerenciamento de informações, especialmente em ambientes onde grandes volumes de dados são gerados e armazenados. Essa técnica é necessária em ambientes onde grandes volumes de dados são gerados porque pode ajudar a reduzir o consumo de armazenamento e a aumentar a eficiência dos processos de análise de dados.

A deduplicação de dados é útil, por exemplo, no domínio da medicina, em que há grandes conjuntos de dados genômicos que são analisados para identificar padrões e mutações associadas a doenças específicas. Nesse cenário, a deduplicação é vital para assegurar a precisão das análises, pois, se amostras de DNA de um mesmo paciente são coletadas e sequenciadas em diferentes momentos e locais, pode haver uma repetição inadvertida dessas amostras no banco de dados. Nesse contexto, a deduplicação de dados é crucial para a integridade da pesquisa, pois dados duplicados podem levar a interpretações errôneas, como a superestimação da prevalência de uma mutação genética rara.

A técnica de deduplicação de dados consiste em um processo de

(A) agregação de dados, que combina duplicatas em um único conjunto de dados.
(B) compactação de dados, que reduz o tamanho de um conjunto de dados, eliminando duplicatas.
(C) restauração de dados, que recupera dados duplicados e otimiza o processo de backup.
(D) replicação de dados, que permite gerar cópias de um conjunto de dados, aumentando a disponibilidade dos mesmos.
(E) redução de dados, que remove duplicatas de um conjunto de dados, mantendo apenas uma única instância de cada dado.

**19**

Considere um conjunto de dados que inclui as variáveis idade, altura e peso. Os dados de idade estão entre 0 e 100 anos, os dados de altura estão entre 1,50 e 2,00 metros e os dados de peso estão entre 50 e 100 kg.

Qual das seguintes técnicas de normalização numérica é mais adequada para esse conjunto de dados?

(A) Normalização Min-Max, para transformar os dados de modo que os valores fiquem entre 0 e 1.
(B) Normalização Z-Score, para transformar os dados de modo que a média seja 0 e o desvio padrão seja 1.
(C) Normalização de Binário, para transformar os dados de modo que os valores fiquem entre 0 e 1, sendo 0 para valores menores ou iguais a um determinado limite e 1 para valores maiores que esse limite.
(D) Normalização de Decil, para transformar os dados de modo que os valores fiquem entre 0 e 100, sendo 0 para o menor valor e 100 para o maior valor.
(E) Normalização de Quantil, para transformar os dados de modo que os valores fiquem entre 0 e 100, sendo 0 para o primeiro percentil e 100 para o último percentil.

**20**

Em um estudo de saúde pública, um pesquisador está analisando um conjunto de dados que inclui informações sobre idade, peso, altura e nível de atividade física dos participantes, no entanto ele percebe que alguns dados referentes ao peso estão ausentes.

Considerando-se a necessidade de manter a precisão e a confiabilidade do estudo, qual das seguintes abordagens seria a mais apropriada para tratar esses dados ausentes sobre o peso dos participantes?

(A) Interpolação de dados, para estimar os valores de peso ausentes, com base em um modelo matemático que leva em conta as tendências e os padrões dos dados existentes.
(B) Remoção de registros, para excluir todos os registros de participantes que não incluíram informações de peso, mesmo que as outras informações estejam completas.
(C) Substituição por zero, para substituir os dados de peso ausentes por zero, assumindo que a ausência de informação indica a menor medida possível.
(D) Substituição por valores extremos, para substituir os dados de peso ausentes pelos valores máximos ou mínimos observados no conjunto de dados.
(E) Uso de dados de outra pesquisa, para substituir os dados de peso ausentes por dados de peso de um estudo similar.

**21**

Outliers são pontos ou observações em um conjunto de dados que diferem significativamente da maioria dos demais outros pontos. Eles podem ser resultados de variações na medição, erros de entrada de dados ou, ainda, podem indicar uma variação genuína da fonte de coleta.

Em preparação para análise de um conjunto de dados, o tratamento de outliers

(A) é sempre necessário, independentemente do tamanho do conjunto de dados ou do objetivo da análise.
(B) é sempre uma tarefa simples que pode ser realizada por qualquer analista de dados, sem risco de perder informações valiosas.
(C) é sempre uma tarefa complexa que requer um conhecimento profundo de estatística e que independe do conjunto de dados e do objetivo da análise.
(D) deve ser realizado antes de realizar agregações, pois os outliers podem afetar os resultados da análise que inclua uma agregação.
(E) deve ser realizado após realizar agregações, pois os outliers podem obscurecer os resultados da agregação.

RASCUNHO

Continua

**22**

Considere o seguinte texto sobre integração de dados.

> Como viabilizar o compartilhamento efetivo de dados e informações das cadeias agropecuárias entre instituições de governo e dessas com a sociedade? Esta foi a principal questão que os participantes do 1º Painel de Cadeias Agropecuárias e Dados Abertos buscaram responder na tarde de quinta-feira (2/12), durante webinar realizado pelo Instituto de Pesquisa Econômica Aplicada (Ipea).
>
> Disponível em: https://www.ipea.gov.br/portal/categorias/45-todas-as-noticias/noticias/11394-especialistas-debatem-abertura-e-integracao-de-dados-de-cadeias-agropecuarias?highlight=WyJhYmFzdGVjaW1lbnRvIiwiYWd1YSIsIidcdTAwZTFndWEiLCJhZ3VhJywiXQ==. Acesso em: 5 jan. 2024.

Considerando-se o questionamento apresentado no texto e sabendo-se que, quando da integração de conjuntos de dados de múltiplas fontes, matching é uma questão relevante, o problema de identificação de entidades em múltiplas fontes de dados remete ao desafio de

(A) agregar os objetos do conjunto de dados, oferecendo versões consolidadas de suas ocorrências.
(B) detectar a ocorrência de uma mesma entidade do mundo real.
(C) determinar a correlação entre os identificadores dos objetos.
(D) reduzir o conjunto de dados a ser submetido ao processo de análise.
(E) selecionar as variáveis mais significativas do conjunto de dados, eliminando as irrelevantes ou redundantes.

**23**

Considere um conjunto de dados estruturados composto por colunas, que refletem as características desses dados, e por linhas, que combinam essas características.

No tratamento desses dados, o processo de enriquecimento consiste em

(A) adicionar dados externos ao conjunto de dados.
(B) expandir ou adernar os dados de um conjunto, sem interferência externa.
(C) verificar a integridade dos relacionamentos entre as colunas.
(D) normalizar os dados, mitigando a diferença de escala entre as colunas.
(E) imputar dados em linhas de dados que apresentem valores ausentes.

RASCUNHO

**24**

Fazendo seu primeiro programa na linguagem Scala, um programador experiente em Python criou o seguinte código:

```
object SumProgram {
  def main(args: Array[String]): Unit = {
    val a = 5
    val b = 10
    val result = sum(a, b)

   println(s"A soma de $a e $b é: $result")
  }

  def sum(a: Int, b: Int): Int = {
    return a + b
  }
}
```

De acordo com a especificação de Scala, a palavra-chave `return`, nesse caso, é

(A) exigida, e sua ausência causa erro de compilação.
(B) exigida, e sua ausência causa erro de execução.
(C) errada, e sua presença causa erro de compilação.
(D) errada, e sua presença causa erro de execução.
(E) possível, mas desnecessária.

**25**

Em Python 3, existem várias formas de se imprimir um número float com um formato específico. Por exemplo, para imprimir o número 12.45667 aproximado para 1 casa decimal, isto é, como 12.5, é possível usar a formatação no estilo da linguagem C, como em

```
print("%.1f" % 12.45667)
```

Outra forma correta de, em Python, imprimir o número 12.45667 exatamente como 12.5 é

(A) `print("{:.1f}".format(12.45667))`
(B) `print(f"{12.45667:.1g}")`
(C) `print(f"{12.45667:.1e}")`
(D) `print(round(12.45667,0))`
(E) `print("%1f" % 12.45667)`

**26**

A paralelização em rotinas de ciência de dados traz benefícios importantes, especialmente quando é necessário tratar uma grande quantidade de dados.

O principal motivador para paralelizar uma rotina é

(A) aumentar a segurança sobre o valor correto do resultado.
(B) reduzir a memória total utilizada.
(C) reduzir o tempo para que as rotinas sejam completadas.
(D) simplificar o código das rotinas.
(E) diminuir o custo total do processamento.

**27**

Considere a função em Python a seguir.

```
def sopares(dados):
  for i in range(len(dados)):
    if dados[i] % 2 != 0:
      return False
  return True
```

Qual função retorna os mesmos valores quando recebe uma lista de inteiros positivos?

(A)
```
def sopares(dados):
  return all(x % 2 == 0 for x in dados)
```

(B)
```
def sopares(dados):
  return all(x % 2 != 0 for x in dados)
```

(C)
```
def sopares(dados):
  return all(map(lambda x: x % 2 != 0, dados))
```

(D)
```
def sopares(dados):
  return next((True for x in dados if x % 2 == 0), False)
```

(E)
```
def sopares(dados):
  for i in dados:
    if i % 2 == 0:
      return True
  return False
```

**28**

Analise o seguinte programa em Python:

```
class ClasseX:
    def mostrar(self):
        print("X",end="")

class ClasseY:
    def mostrar(self):
        print("Y",end="")

class ClasseZ(ClasseY, ClasseX):
    def mostrar(self):
      super().mostrar()
      print("Z",end="")

obj_z = ClasseZ()
obj_z.mostrar()
```

Quando esse programa for executado, sua saída será

(A) XY
(B) YX
(C) YZ
(D) ZY
(E) Z

**29**

Em Python, a classe `list` implementa o conceito genérico de lista, que tem como uma de suas características permitir acessar cada item unicamente por um índice.

Outra propriedade importante das listas em Python é

(A) exigir que todos os objetos da lista sejam da mesma classe.
(B) garantir que cada elemento só esteja presente uma vez na lista.
(C) guardar pares chave-valor, que permitem achar um valor por meio da chave.
(D) manter os elementos sempre ordenados, do menor para o maior.
(E) permitir a alteração do tamanho da lista pela inserção e pela remoção de elementos.

RASCUNHO

Continua

**30**

Um analista de dados está utilizando R e o pacote ggplot2 para criar um gráfico XY apenas com pontos. Ele possui dois vetores, `vetor_x` e `vetor_y`, ambos com o mesmo tamanho, representando os valores dos eixos X e Y, respectivamente.

Uma forma correta de criar um gráfico XY básico, apenas com pontos, passando esses vetores como parâmetros é

(A) `ggplot(data = vetor_y, aes(x = vetor_x, y = vetor_y))`
(B) `ggplot(vetor_x, vetor_y, geom = "point")`
(C) `ggplot(aes(x = vetor_x, y = vetor_y))`
(D) `ggplot()+ geom_point(aes(x = vetor_x, y = vetor_y))`
(E) `ggplot()+ geom_point (x = vetor_x, y = vetor_y)`

**31**

Um cientista de dados está trabalhando com um conjunto de dados e usando a linguagem de programação R. O conjunto de dados, chamado `dados`, consiste em três colunas: `ID, Idade e Salario`, em que `Idade` indica um grupo de idade. O cientista de dados precisa calcular a média de `Salario` para cada grupo de idade. No entanto, ele deve realizar essa tarefa utilizando apenas as funções base do R, sem pacotes adicionais.

Qual função ou conjunto de funções do R ele deve usar para realizar essa tarefa de forma eficiente?

(A) `mean(dados$Salario)`
(B) `split(dados$Salario, dados$Idade)`
(C) `tapply(dados$Salario, dados$Idade, sum)`
(D) `aggregate(Salario ~ Idade, dados, mean)`
(E) `sum(dados$Salario) / length(unique(dados$Idade))`

**32**

Um projeto de saúde utiliza a biblioteca Pandas, em Python, para analisar dados biométricos. O conjunto de dados é trabalhado em um DataFrame chamado `saude_df`, com várias colunas representando medidas típicas de saúde, como o peso e a pressão arterial. Cada linha representa o valor dessas medidas para uma pessoa.

Para selecionar aleatoriamente 20% das linhas do DataFrame e apenas as colunas `Peso` e `Pressão Arterial`, pode ser usada a expressão

(A) `saude_df['Peso', 'Pressão Arterial'].sample(0.2)`
(B) `saude_df[['Peso', 'Pressão Arterial']].sample(0.2)`
(C) `saude_df.sample(0.2)['Peso', 'Pressão Arterial']`
(D) `saude_df.sample(frac=0.2)['Peso', 'Pressão Arterial']`
(E) `saude_df.sample(frac=0.2)[['Peso', 'Pressão Arterial']]`

**33**

Um cientista de dados recebeu um vetor chamado `dados` contendo valores da renda mensal da população de uma cidade e resolveu obter algumas estatísticas que descrevessem os dados recebidos.

A linha de código em R que calcula corretamente a média do vetor dados é

(A) `media <- sum(dados)/len(dados)`
(B) `media <- mean(dados)`
(C) `media <- average(dados)`
(D) `media <- sapply(dados, function(x) length(x))`
(E) `media <- sapply(dados, function(x) x/len(dados))`

**34**

Na programação funcional, que é um paradigma suportado pela Linguagem Scala, uma das práticas fundamentais é o uso de funções puras.

A principal propriedade que caracteriza uma função pura é a(o)

(A) baixa coesão
(B) ausência de efeitos colaterais
(C) incapacidade de chamar outra função
(D) alto acoplamento
(E) encapsulamento aberto

**35**

Um programa em Python contém um vetor unidimensional, de tamanho 20, chamado `vetor`, criado com o Numpy. Deseja-se obter, em apenas uma expressão, uma matriz de 4 linhas e 5 colunas criada linha a linha a partir dos elementos desse vetor.

Para tal fim, a expressão a ser utilizada é

(A) `np.split(vetor, 4)`
(B) `vetor.reshape(5, 4)`
(C) `np.reshape(vetor, (4, 5))`
(D) `vetor.transpose(4, 5)`
(E) `np.array(vetor, shape=(4, 5))`

**36**

Python é dinamicamente tipado, o que significa que os tipos de variáveis são determinados em tempo de execução com base nos valores atribuídos, proporcionando flexibilidade durante o desenvolvimento. Essa dinamicidade permite a criação de código versátil e adaptável.

Nesse contexto, considere o seguinte trecho de código:

```
a = 5
b = 5
c = '5'
d = (a + b) + c
```

Qual é o resultado da variável d?

(A) '555'
(B) 15
(C) '105'
(D) 555
(E) erro

RASCUNHO

**37**

Uma lista é uma estrutura de dados versátil e fundamental usada para organizar e armazenar coleções de itens. As listas em Python oferecem uma maneira dinâmica e eficiente de lidar com sequências de itens. Por terem a capacidade de adicionar, remover e modificar elementos, essas listas desempenham um papel crucial em muitos programas em Python, tornando-se uma ferramenta essencial para tarefas que vão desde o armazenamento simples de dados até algoritmos mais complexos e manipulação de dados.

Nesse contexto, considere duas listas, `L1` e `L2`, que foram implementadas em Python. As configurações iniciais dessas listas são as seguintes:

```
L1 = [1, 2, 3]
L2 = [3, 4, 5]
```

Em seguida, foram feitas as seguintes operações:

```
L1.append(3)
L3 = L1
L3.append(7)
L1.append(8)
L4 = L1 + L2
```

Qual é o valor de `L4`?

(A) [1, 2, 3, 4, 5]
(B) [1, 2, 3, 3, 4, 5]
(C) [1, 2, 3, 3, 8, 3, 4, 5]
(D) [1, 2, 3, 3, 3, 4, 5, 7, 8]
(E) [1, 2, 3, 3, 7, 8, 3, 4, 5]

**38**

Em um estudo sobre tráfico de pessoas realizado pelo Ipea, os autores trazem as seguintes afirmações:

I – "Na medida em que haja desacordo sobre a definição de tráfico de pessoas entre diferentes países [...], os números globais do tráfico de pessoas são provavelmente enviesados" (p.23).

II – "Em tese, identificar as vítimas do tráfico de pessoas poderia ser difícil, pois estas fazem parte de populações com baixa visibilidade (por exemplo, profissionais do sexo, imigrantes indocumentados)" (p.24).

MORAES, R. F. de *et al*. **Uma solução em busca de um problema**: repensando o enfrentamento ao tráfico de pessoas no Brasil. Rio de Janeiro: Ipea, 2022.

Ambas as afirmações se referem a vieses de seleção da amostra, e esses vieses

(A) são de tipos diferentes, sendo o primeiro um exemplo de *reporting bias* e o segundo, de *measurement bias*.
(B) são de tipos diferentes, sendo o primeiro um exemplo de *measurement bias* e o segundo, de *reporting bias*.
(C) são ambos exemplos de *reporting bias*.
(D) são ambos exemplos de *measurement bias*.
(E) são ambos exemplos de *attrition bias*.

**39**
Sejam Y e Z variáveis aleatórias independentes e identicamente distribuídas, assumindo valores inteiros.

Considere as seguintes informações:

* $Prob(Y + Z > 4) = 0$;
* $Prob(Y + Z > -3) = 1$;
* $Prob(YZ = 0) = 0$;
* $E(Z) = 0$;
* $E(Y^2 + Z^2) = 16/5$.

Quanto vale $Prob(Y^2 + Z^2 > 16/5)$?

(A) 1/2
(B) 5/16
(C) 3/5
(D) 9/25
(E) $\frac{1}{\sqrt{\pi}}$

**40**
Considere um exame para detectar o uso de uma droga. Suponha que, para um usuário da droga, o exame tenha probabilidade 0,95 de dar, corretamente, positivo. Suponha que, para um não usuário, o exame tenha probabilidade 0,9 de dar, corretamente, negativo.
Para um certo indivíduo, a probabilidade de que ele use a droga é estimada, antes do exame, como sendo de 0,8. Após a aplicação do exame, o resultado foi positivo. Seja P a probabilidade que se deve estimar, com os dados acima, de que o resultado do exame esteja errado, ou seja, de que o resultado seja um falso positivo.

Quanto vale, aproximadamente, P?

(A) $P < 0{,}01$
(B) $0{,}01 < P < 0{,}02$
(C) $0{,}02 < P < 0{,}05$
(D) $0{,}05 < P < 0{,}1$
(E) $P > 0{,}1$

**41**
Um dado comum tem seis faces equiprováveis numeradas de 1 a 6. Um jogador lança três dados comuns e independentes — um vermelho, um verde e um azul — e anota a soma dos três números obtidos.

Sabendo-se que o total é maior ou igual a 16, qual é a probabilidade condicional de que os dados verde e azul marquem o mesmo número?

(A) 1/6
(B) 1/5
(C) 1/3
(D) 2/5
(E) 1/2

**42**
Um baralho tem 26 cartas vermelhas e 26 cartas pretas. Um jogador embaralha as cartas e pousa 5 sobre a mesa, de tal forma que não se veja a cor. Ele então revela as 4 primeiras cartas: são todas vermelhas.

Qual é a probabilidade P de que a quinta carta também seja vermelha?

(A) $P = 1/2$, exatamente
(B) $9/20 < P < 1/2$
(C) $1/2 < P < 11/20$
(D) $P < 9/20$
(E) $P > 11/20$

**43**
Seja $I = [0,1]$ o intervalo unitário na reta real.

Sejam $(X_i)$ pontos aleatórios de I, independentes e com distribuição uniforme.

Qual é a probabilidade de que valha $X_1^2 + X_2^2 + X_3^2 < 1$?

(A) 5/12
(B) 1/2
(C) $\pi/6$
(D) $\frac{\sqrt{2}}{2}$
(E) $\pi\frac{\sqrt{3}}{8}$

**44**
Em uma maternidade, 400 bebês nasceram em uma semana. Sejam H e M os números de meninos e de meninas, respectivamente. Sabe-se, portanto, que $H + M = 400$. Suponha para esse problema que, para cada bebê, a probabilidade de que seja menino seja exatamente igual a 1/2; suponha também que os sexos dos bebês sejam perfeitamente independentes uns dos outros. Seja P a probabilidade condicional de que $H < 90$, dado que $H < 100$.

Aproximadamente, quanto vale P?

(A) $P < 1/400$
(B) $1/400 < P < 1/10$
(C) $1/10 < P < 1/4$
(D) $1/4 < P < 1/2$
(E) $P > 1/2$

**45**
Considere uma comunidade na qual o coeficiente de variação populacional (CV) da renda de seus indivíduos é de 20%.

Um conjunto de políticas públicas que elevasse a renda de todos esses indivíduos em 10% provocaria, no CV,

(A) uma redução, passando a ser de 10%
(B) uma redução, passando a ser de 18%
(C) um aumento, passando a ser de 22%
(D) um aumento, passando a ser de 30%
(E) nem redução, nem aumento

**46**

Considere a matéria a seguir, sobre a dura realidade do trabalho infantil no Brasil.

**Quase 5% das crianças e adolescentes do país estão em situação de trabalho infantil, aponta IBGE**

O Brasil tem um total de 38,365 milhões de pessoas com idade de 5 a 17 anos. Destas, 2,103 milhões realizam atividades econômicas ou de autoconsumo, estando 1,881 milhão em situação de trabalho infantil, o equivalente a 4,9% do total de pessoas entre 5 e 17 anos no país, segundo dados da Pesquisa Nacional por Amostra de Domicílios Contínua 2022, divulgados pelo Instituto Brasileiro de Geografia e Estatística (IBGE). A pesquisa também apontou que, em 2022, 756 mil crianças e adolescentes exerciam atividades da **Lista TIP**, do governo federal, que elenca as **piores formas de trabalho infantil no país**. No geral, são serviços que envolvem risco de acidentes ou são prejudiciais à saúde, como trabalho na construção civil, em matadouros, oficinas mecânicas, comércio ambulante em locais públicos, coleta de lixo, venda de bebidas alcoólicas, entre outras atividades.

A pesquisa do IBGE considera duas categorias de atividades:

- **econômica**, que é a de **quem trabalhou pelo menos 1 hora completa e foi remunerado** em dinheiro, produtos, benefícios, etc., **ou que não teve remuneração direta, mas atuou para ajudar a atividade econômica de algum parente**.
- **de autoconsumo**, que incluem pesca, criação de animais, fabricação de roupas, construção de imóveis e outras que sejam para **uso exclusivo da pessoa ou de parentes**.

E nem todas as pessoas de 5 a 17 anos que exercem essas atividades se enquadram na situação de trabalho infantil.

Disponível em: https://g1.globo.com/trabalho-e-carreira/noticia/2023/12/20/quase-5percent-das-criancas-e-adolescentes-do-pais-estao-em-situacao-de-trabalho-infantil-aponta-ibge.ghtml. Acesso em: 29 dez 2023. Adaptado.

Nesse contexto, considere que uma pessoa de 5 a 17 anos é escolhida ao acaso e que se deseja estimar a probabilidade de que essa pessoa exerça trabalho infantil, mas não em uma de suas piores formas, que são elencadas na lista TIP, dado que ela realiza atividades econômicas ou de autoconsumo.

Essa probabilidade é, aproximadamente, de

(A) 2,9%
(B) 35,9%
(C) 46,5%
(D) 53,5%
(E) 59,8%

**47**

Ao fim do ano de 2022, 21 instituições financeiras fizeram estimativas para o Ibovespa, principal índice de ações da bolsa brasileira (B3), mostradas na notícia do Jornal Valor Econômico de 22/12/2022, que também apresenta a média aritmética e a mediana dos dados.

**Previsão para a bolsa é de alta em 2023**

Casa mais otimista projeta Ibovespa em 150 mil pontos; rumo dos juros será decisivo

Disponível em: https://valor.globo.com/financas/noticia/2022/12/22/previsao-para-a-bolsa-e-de-alta-em-20233.ghtml. Acesso em: 29 dez. 2023. Adaptado.

No útimo pregão do ano de 2023, o Ibovespa fechou em 134.185,23 pontos, conforme dados da B3.

Considerando-se os dados apresentados, a medida de posição das estimativas das instituições financeiras que mais se aproximou do resultado real, apresentado pelo Ibovespa no último pregão do ano, foi a:

(A) mediana
(B) média geométrica
(C) média harmônica
(D) menor moda
(E) maior moda

**48**

Um técnico, a fim de monitorar a mobilidade social de renda em uma região geográfica, criou um modelo agrupando as tradicionais classes sociais A, B, C, D e E em três extratos:

Classes A e B: Alta renda
Classe C: Média renda
Classes D e E: Baixa renda

A partir de uma extensa pesquisa, verificou que as probabilidades de um indivíduo transitar de um extrato para outro ou permanecer no mesmo extrato dependiam apenas de seu extrato atual e permaneciam constantes ao longo do tempo.

Levantou, assim, para cada extrato, as seguintes probabilidades de permanência ou mudança de extrato, após o período de tempo de um ano:

Alta renda
Permanência: 80%
Mudança para média renda: 10%
Mudança para baixa renda: 10%

Média renda
Permanência: 60%
Mudança para alta renda: 20%
Mudança para baixa renda: 20%

Baixa renda
Permanência: 70%
Mudança para alta renda: 10%
Mudança para média renda: 20%

A probabilidade de um indivíduo, inicialmente de média renda, ter mobilidade social de renda (ascenso ou descenso), terminando num extrato diferente do inicial, após um período de 2 anos é de

(A) 12% (B) 42% (C) 58% (D) 67% (E) 88%

**49**

Até o Censo de 2010, a região Norte ainda possuía mais homens do que mulheres, mas, segundo os dados do Censo 2022, divulgados em outubro de 2023 pelo IBGE, as mulheres já são a maioria em todas as regiões do Brasil.
A matéria a seguir mostra os municípios do Brasil com maior proporção de mulheres em relação a homens.

**Santos (SP) tem o maior percentual de mulheres no Brasil; veja lista**

Santos, no litoral paulista, é a cidade brasileira com a maior proporção de mulheres em relação a homens, segundo dados do Censo 2022 divulgados hoje pelo IBGE.
A cidade paulista lidera o *ranking* de municípios com mais mulheres, com uma proporção de 54,68%.

1. Santos (SP) - 54,68%
2. Salvador (BA) - 54,49%
3. São Caetano do Sul (SP) - 54,32%
4. Niterói (RJ) - 54,19%
5. Aracaju (SE) - 54,11%

Disponível em: https://noticias.uol.com.br/cotidiano/ultimas-noticias/2023/10/27/santos-ibge-censo-2022-cidades-com-mais-mulheres.htm. Acesso em: 11 jan. 2024. Adaptado.

Suponha que a Prefeitura de Santos, cidade brasileira com o maior percentual de mulheres, de aproximadamente 55%, realize um sorteio de 4 habitantes que cadastraram suas notas fiscais num aplicativo do município. Considere, ainda, que cada munícipe efetue tais cadastramentos de forma independente dos demais habitantes.

A probabilidade de que o número de mulheres sorteadas seja maior que o de homens sorteados é de aproximadamente:

(A) 30%
(B) 32%
(C) 39%
(D) 55%
(E) 76%

Dado
$0,45^2 = 0,2025$
$0,55^2 = 0,3025$

**50**

Considere que o salário médio de empregados de um determinado setor de serviços do Brasil é R$ 2.300,00, com um desvio padrão conhecido de R$ 400,00. Uma empresa desse setor é selecionada ao acaso, e dela é selecionada uma amostra de 36 funcionários, resultando em um salário médio de R$ 2.500,00. Um pesquisador decide realizar um teste de hipótese unilateral com um nível de confiança de 95% para verificar se a empresa selecionada paga salários médios maiores do que a média do setor.

Considere o extrato da Tabela a seguir.

| Tabela da Distribuição Normal Padrão | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| z | 0,0 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
| 0,0 | 0,5000 | 0,5040 | 0,5080 | 0,5120 | 0,5160 | 0,5199 | 0,5239 | 0,5279 | 0,5319 | 0,5359 |
| 0,1 | 0,5398 | 0,5438 | 0,5478 | 0,5517 | 0,5557 | 0,5596 | 0,5636 | 0,5675 | 0,5714 | 0,5753 |
| 0,2 | 0,5793 | 0,5832 | 0,5871 | 0,5910 | 0,5948 | 0,5987 | 0,6026 | 0,6064 | 0,6103 | 0,6141 |
| ⋮ | | | | | | | | | | |
| 2,7 | 0,9965 | 0,9966 | 0,9967 | 0,9968 | 0,9969 | 0,9970 | 0,9971 | 0,9972 | 0,9973 | 0,9974 |
| 2,8 | 0,9974 | 0,9975 | 0,9976 | 0,9977 | 0,9977 | 0,9978 | 0,9979 | 0,9979 | 0,9980 | 0,9981 |
| 2,9 | 0,9981 | 0,9982 | 0,9982 | 0,9983 | 0,9984 | 0,9984 | 0,9985 | 0,9985 | 0,9986 | 0,9986 |
| 3,0 | 0,9987 | 0,9987 | 0,9987 | 0,9988 | 0,9988 | 0,9989 | 0,9989 | 0,9989 | 0,9990 | 0,9990 |
| 3,1 | 0,9990 | 0,9991 | 0,9991 | 0,9991 | 0,9992 | 0,9992 | 0,9992 | 0,9992 | 0,9993 | 0,9993 |
| 3,2 | 0,9993 | 0,9993 | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 0,9995 | 0,9995 | 0,9995 |

Qual o valor-p do teste aplicado pelo pesquisador?

(A) 0,0010
(B) 0,0013
(C) 0,0017
(D) 0,0025
(E) 0,0031

**51**

O Quadro abaixo fornece informações sobre o índice de vendas no varejo por estado em agosto de 2023.

| Estado | AC | AL | AM | AP | BA | CE | DF | ES | GO |
|---|---|---|---|---|---|---|---|---|---|
| 2023.08 | 106,13 | 101,87 | 105,07 | 101,42 | 104,39 | 108,94 | 100,46 | 105,30 | 102,22 |
| Estado | MA | MG | MT | MS | PA | PB | PE | PI | PR |
| 2023.08 | 118,16 | 106,14 | 104,23 | 103,77 | 105,21 | 88,22 | 100,77 | 103,18 | 102,57 |
| Estado | RJ | RN | RO | RR | RS | SC | SE | SP | TO |
| 2023.08 | 100,04 | 100,37 | 98,45 | 103,15 | 103,34 | 101,80 | 102,25 | 99,03 | 113,00 |

Um analista de dados resolveu verificar se há presença de outliers nesse conjunto de índices e decidiu fazer isso por meio de um Box Plot dos dados fornecidos.

Com base na técnica escolhida pelo analista, quantos índices podem ser enquadrados como outliers?

(A) 0
(B) 1
(C) 2
(D) 3
(E) 4

**52**

Uma pesquisa recente estudou a distribuição de renda familiar *per capita*, em salários mínimos (s.m.), de duas comunidades, Alfa e Beta, com, aproximadamente, o mesmo número de habitantes. Considerando-se **e** a base dos logaritmos naturais ou neperianos, na comunidade Alfa, verificou-se que tal renda pode ser bem aproximada por uma variável aleatória contínua (v.a.c.) X com função densidade de probabilidade (f.d.p.) da forma $f(X = x) = r\ \mathbf{e}^{x}$, $0 \leq x \leq 1$ s.m.; já para a comunidade Beta, constatou-se que a renda em estudo seguia aproximadamente a distribuição de uma variável aleatória contínua (v.a.c.) Y com função densidade de probabilidade (f.d.p.) da forma $g(Y = y) = s\ y\ \mathbf{e}^{y}$, $0 \leq y \leq 1$ s.m.

Um técnico deve decidir por apenas uma das duas comunidades para receber um programa assistencial, qual seja, aquela que possua o maior número de habitantes com renda familiar *per capita* até meio salário mínimo.

Nesses termos, o técnico deverá optar pela comunidade:

Dado
$\sqrt{\mathbf{e}} \cong 1{,}645$

(A) Alfa, que tem aproximadamente o dobro de pessoas elegíveis em relação às da comunidade Beta.
(B) Alfa, que tem aproximadamente o triplo de pessoas elegíveis em relação às da comunidade Beta.
(C) Beta, que tem aproximadamente o dobro de pessoas elegíveis em relação às da comunidade Alfa.
(D) Beta, que tem aproximadamente o triplo de pessoas elegíveis em relação às da comunidade Alfa.
(E) Alfa ou Beta, indistintamente, já que ambas têm o mesmo número de pessoas elegíveis.

**53**

Ao criar um índice para um *corpus* de notícias políticas, um pesquisador decidiu usar o modelo vetorial com a medida TF-IDF.

Descobriu, porém, que essa medida pode ser calculada de várias formas, com a característica comum de calcular o peso de cada termo a partir da ideia de que esse termo

(A) esteja ou não esteja em um documento.
(B) tenha um peso diretamente proporcional à sua frequência no documento e inversamente proporcional à sua frequência nos documentos do *corpus*.
(C) tenha um peso inversamente proporcional à sua frequência no documento e diretamente proporcional à sua frequência nos documentos do *corpus*.
(D) seja representado por um vetor que indica um conjunto de outros termos que possuem o mesmo conceito semântico.
(E) seja representado por um vetor que é a soma das representações dos embeddings de todos os termos do documento.

**54**

Na avaliação de um modelo criado por aprendizado de máquina em um experimento que buscava identificar textos de opinião sobre o desempenho da economia, separando-os dos que não forneciam opinião alguma, só fatos e dados, foi encontrada a seguinte matriz de confusão:

| | Opinião | Não Opinião |
|---|---|---|
| Opinião | 440 | 60 |
| Não Opinião | 20 | 480 |

Considerando-se que, nessa matriz, as linhas indicam a resposta correta e as colunas indicam a previsão, a acurácia é de

(A) 8%
(B) 44%
(C) 48%
(D) 88%
(E) 92%

**55**

Um pesquisador iniciante em aprendizado de máquina trabalhava com um modelo de classificação binário com as duas classes equilibradas. Inicialmente, ele fez a avaliação de seu modelo, separando 20% dos dados disponíveis para a avaliação, e o treinou com 80% dos dados, fazendo o processo apenas uma vez. Depois, a pedido de seu chefe, ele trocou a forma de avaliação, separando o conjunto de dados em 10 partes e escolhendo, em 10 rodadas, uma parte diferente para avaliação e as outras para treinamento.

Essas duas formas de avaliar um modelo são conhecidas, respectivamente, como

(A) estratificação e hold-out
(B) hold-out e k-fold
(C) leave-one-out e estratificação
(D) leave-one-out e k-fold
(E) Monte Carlo e leave-p-out

**56**

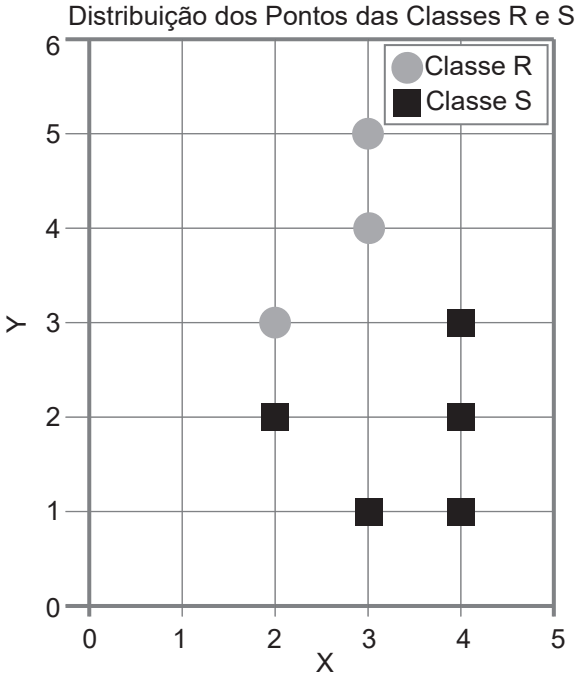
Em processamento de linguagem natural, o modelo Skip-Gram é uma técnica popular para treinar word embeddings.

O treinamento do modelo Skip-Gram destaca-se de outras técnicas, como o Continuous Bag of Words (CBOW), por ter a seguinte característica:

(A) utilizar contextos adjacentes para prever uma palavra-alvo.
(B) focar em prever palavras de contexto a partir de uma palavra-alvo.
(C) treinar cada palavra em uma janela de contexto separadamente.
(D) gerar embeddings, baseados em dependências sintáticas.
(E) empregar uma abordagem de bag-of-words para a geração de embeddings.

**57**

No gráfico XY, são apresentados pontos que representam duas propriedades de elementos de duas classes, R e S. Os pontos da classe R, representados como círculos, são [(3,5),(3,4),(2,3)], enquanto os pontos da classe S, representados como quadrados, são [(4,3),(4,2),(4,1),(3,1),(2,2)]. É necessário classificar pontos novos, de acordo com o algoritmo K-NN, com K=3, considerando a distância euclidiana.




Nesse contexto, as classes dos pontos [(3,2),(3,3) e (4,4)] são, respectivamente:

(A) [R,R,S] (B) [R,S,R] (C) [S,R,R] (D) [S,R,S] (E) [R,R,R]

**58**

Na arquitetura de redes neurais, diferentes funções de ativação são utilizadas nas camadas de neurônios para aplicar transformações não lineares aos dados. Uma dessas funções é a ReLU, conhecida por sua eficácia em diversos modelos de aprendizado profundo.

Ao implementar a função ReLU, um pesquisador deve seguir a fórmula:

(A) $f(x) = \frac{1}{1+e^{-x}}$

(B) $f(x) = \begin{cases} x, \text{ se } x > 0 \\ 0, \text{ se } x \leq 0 \end{cases}$

(C) $f(x) = \tanh(x)$

(D) $f(x) = \max(0,x) - \max(0,-x)$

(E) $f(x) = \sqrt{x^2+1} - 1$

**59**

Em redes convolucionais, o tamanho do passo normalmente é menor que o tamanho do filtro.

Se o tamanho do passo for maior que o tamanho do filtro, é possível que

(A) a camada de pooling fique sensível demais a pequenas variações.
(B) o modelo apresente características de overfitting.
(C) informações importantes sejam perdidas.
(D) um aumento significativo na dimensão espacial da entrada aconteça.
(E) um aumento significativo na dimensão espacial da saída aconteça.

RASCUNHO

**60**
O método de POS-tagging, ou Part of Speech tagging, é uma tarefa do processamento de linguagem natural em que

(A) cada palavra do texto recebe um rótulo indicando sua classe gramatical.
(B) cada palavra do texto recebe um rótulo indicando sua função sintática na sentença.
(C) cada palavra do texto recebe um rótulo indicando sua função semântica na sentença.
(D) cada sentença do texto recebe um rótulo indicando a quem pertence a fala.
(E) cada sentença do texto recebe um rótulo indicando o tipo de ato de fala que representa.

**61**
Uma cientista de dados percebeu que, ao processar alguns documentos, seria melhor remover palavras que aparecem em quase todo texto, as stop-words.

Para começar sua lista de stop-words, ela pode escolher listar todos os

(A) adjetivos e preposições
(B) advérbios e pronomes
(C) artigos e preposições
(D) artigos e substantivos
(E) substantivos e verbos

**62**
Um pesquisador possui um conjunto de dados consistindo em características diversas, features, e suas respectivas classificações, labels. Ele deseja dividir esse conjunto de dados em conjuntos distintos, para treinamento e para teste, com o objetivo de validar a eficácia de um modelo de aprendizado de máquina.

Nesse contexto, qual função do SciKit-learn ele deve utilizar para realizar essa divisão de maneira eficiente e adequada?

(A) sample_test_train
(B) slice_train_test
(C) split_learning
(D) split_test_train
(E) train_test_split

**63**
Em um projeto de classificação de textos, um modelo de machine learning foi aplicado em um conjunto de teste e apresentou os seguintes resultados: uma precisão de 80% e uma revocação de 70%.

Com base nessas informações e considerando-se apenas a parte inteira da porcentagem, qual é o F1 Score desse modelo?

(A) 2%
(B) 18%
(C) 37%
(D) 74%
(E) 98%

**64**
Um cientista de dados está utilizando máquinas de vetor de suporte (SVM) em um projeto de classificação, pois deseja evitar o overfitting do modelo aos dados de treinamento.

Qual das seguintes técnicas auxilia a prevenir o overfitting em SVM?

(A) Apagar algumas ligações da rede.
(B) Garantir que o hiperplano divida perfeitamente os pontos.
(C) Misturar o conjunto de teste com o de treinamento.
(D) Usar todos os pontos para a tomada de decisão.
(E) Utilizar um kernel linear.

**65**
Alguns trabalhos publicados como notas técnicas pelo Ipea se utilizam do método de classificação denominado de Bayes Ingênuo.

No contexto do classificador Bayesiano Ingênuo, Naive Bayes, a ingenuidade do modelo é caracterizada pela(o)

(A) dependência direta entre todas as variáveis preditoras
(B) ênfase na complexidade do modelo
(C) suposição de independência condicional entre as variáveis preditoras
(D) utilização de distribuição de probabilidade uniforme
(E) uso exclusivo de variáveis contínuas

**66**
Em uma nota técnica do Ipea sobre emprego público nos governos subnacionais brasileiros, no ano de 2016, aparece menção sobre o fato de as bases utilizadas possuirem outliers, ou valores atípicos.

A construção de um modelo preditivo a partir dos dados dessas bases, usando árvores aleatórias, Random Forests,

(A) apresentaria uma redução de variância.
(B) seria inadequado, devido à sensibilidade a outliers de Random Forests.
(C) teria uma tendência a overfitting.
(D) teria propensão a underfitting.
(E) teria dependência de linearidade nos dados.

**67**
Em uma nota técnica publicada em 2022 pelo Ipea, sobre população em situação de rua, foi utilizada a técnica de análise de componente principal (PCA).

Na análise por PCA, a primeira componente principal de um conjunto de dados representa a

(A) média dos dados
(B) mediana dos dados
(C) soma total dos dados
(D) variância máxima dos dados
(E) correlação mínima entre os dados

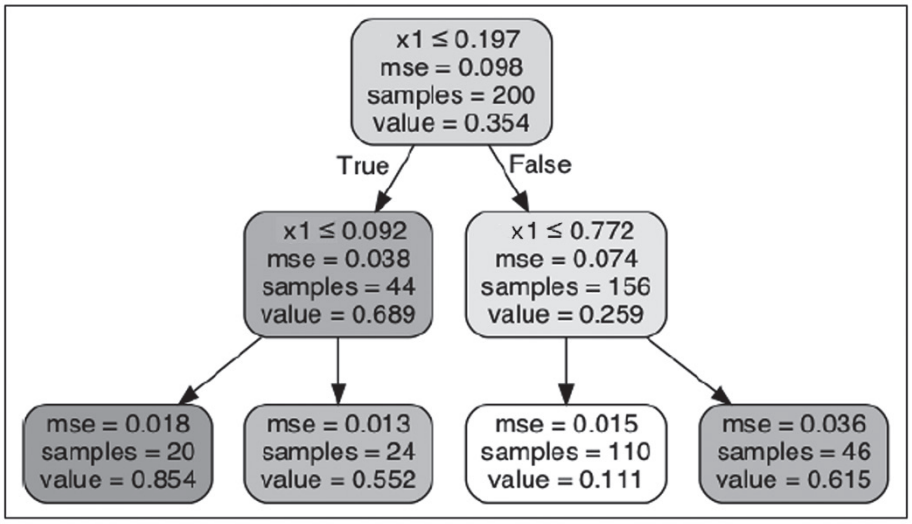
**68**

As árvores de decisão são um modelo de aprendizado de máquina que opera por meio da construção de uma estrutura em forma de árvore para tomar decisões e que oferece uma compreensão clara da lógica de decisão e da hierarquia de características que contribuem para as predições finais. Elas são versáteis e podem ser usadas tanto para tarefas de classificação quanto para as de regressão.

Nesse contexto, considere a construção de uma árvore de regressão usando a classe DecisionTreeRegressor do Scikit-Learn e seu treinamento em um conjunto de dados quadrático com max_depth=2, conforme mostrado a seguir:

```
from sklearn.tree import DecisionTreeRegressor
tree_reg = DecisionTreeRegressor(max_depth=2)
tree_reg.fit(X, y)
```

A árvore resultante é representada na Figura a seguir.




GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**: Concepts, Tools, and Techniques to Build Intelligent Systems. 2 ed. Sebastopol, CA: O'Reilly Media, Inc.: 2019, p. 183.

Considerando-se o cenário apresentado e que se deseja fazer uma predição para uma nova instância, com x1 = 0.6, qual será o valor predito?

(A) 0.015 (B) 0.074 (C) 0.111 (D) 0.259 (E) 110

**69**

A biblioteca Scikit-Learn emprega o algoritmo Classification And Regression Tree (CART) para treinar Árvores de Decisão. O algoritmo CART baseia-se na recursividade e na estratégia de divisão binária para construir uma árvore de decisão. Inicialmente, a árvore é representada por um único nó, que contém todos os dados de treinamento. A cada passo, o algoritmo busca a melhor maneira de dividir o conjunto de dados. A recursividade continua até que uma condição de parada seja atendida, como atingir uma profundidade máxima da árvore. Uma vez construída a árvore, a fase de predição ocorre ao percorrer a estrutura da árvore de acordo com as condições estabelecidas nos nós, levando a uma predição (inferência) para uma determinada entrada.

Considerando-se que n corresponde ao número de features e m ao número de instâncias, qual é a complexidade computacional assintótica de predição para árvores de decisão treinadas com o algoritmo CART?

(A) $O(m)$
(B) $O(m^2)$
(C) $O(n \times m \log(m))$
(D) $O(n^2 \times m \log(m))$
(E) $O(\log_2(m))$

**70**

No aprendizado não supervisionado, os dados de treinamento não têm rótulos. O objetivo é agrupar instâncias semelhantes em clusters. Nesse contexto, suponha que se deseja executar um algoritmo de agrupamento para tentar detectar grupos de visitantes semelhantes em um blog. Em nenhum momento é informado ao algoritmo a que grupo um visitante pertence, mas ele encontra essas conexões sem ajuda. Por exemplo, o algoritmo pode notar que 40% dos visitantes são homens que adoram histórias em quadrinhos e, geralmente, leem o blog à noite, enquanto 20% são jovens amantes de ficção científica que visitam o blog durante os fins de semana, e assim por diante. Deseja-se, nesse caso, usar um algoritmo de agrupamento hierárquico para subdividir cada grupo em grupos menores, o que pode ajudar a direcionar as postagens do blog para cada grupo específico.

Nesse cenário, qual é o algoritmo mais apropriado para fazer o agrupamento desejado?

(A) Agglomerative Hierarchical Clustering
(B) Principal Component Analysis (PCA)
(C) Linear Regression Clustering
(D) k-DBSCAN Clustering
(E) L-Means Clustering

RASCUNHO

RASCUNHO