Universidade de São Paulo

vencerás pela educação

ASD

RH nº 002/2025

Analista de Sistemas (Ciência de Dados)

Instruções

1. **Só abra este caderno quando o fiscal autorizar**.
2. Verifique se o seu nome está correto na capa deste caderno e se a folha de respostas pertence ao **grupo ASD**. Informe ao fiscal de sala eventuais divergências.
3. Durante a prova, são **vedadas** a comunicação entre candidatos e a utilização de qualquer material de consulta e de aparelhos de telecomunicação.
4. Duração da prova: 4 horas. Cabe ao candidato controlar o tempo com base nas informações fornecidas pelo fiscal. O(A) candidato(a) poderá retirar-se da sala definitivamente apenas a partir das 15 h. Não haverá tempo adicional para preenchimento da folha de respostas.
5. O(A) candidato(a) deverá seguir as orientações estabelecidas pela FUVEST a respeito dos procedimentos adotados para a aplicação deste concurso.
6. Lembre-se de que a FUVEST se reserva ao direito de efetuar procedimentos adicionais de identificação e controle do processo, visando a garantir a plena integridade do exame. Assim, durante a realização da prova, será coletada por um fiscal uma **foto** do(a) candidato(a) para fins de reconhecimento facial, para uso exclusivo da USP e da FUVEST. A imagem não será divulgada nem utilizada para quaisquer outras finalidades, nos termos da lei.
7. Após a autorização do fiscal da sala, verifique se o caderno está completo. Ele deve conter **60** questões objetivas, com 5 alternativas cada, e **1** questão dissertativa. Informe ao fiscal de sala eventuais divergências.
8. Preencha a folha de respostas com cuidado, utilizando caneta esferográfica de **tinta azul ou preta**. Essa folha **não será substituída** em caso de rasura.
9. Ao final da prova, é **obrigatória** a devolução da folha de respostas acompanhada deste caderno de questões.

Declaração

Declaro que li e estou ciente das informações que constam na capa desta prova, na folha de respostas, bem como dos avisos que foram transmitidos pelo fiscal de sala.

_____

ASSINATURA

O(a) candidato(a) que não assinar a capa da prova será considerado(a) ausente da prova.

Texto para as questões 01 e 02

Em silêncio

Precisava de silêncio para pensar, ordenar sua vida e rumos. Juntou poucas coisas, navegou até uma ilha deserta. Mas a gritaria das aves marinhas fundia-se com o farfalhar do vento nas palmeiras, e quando ambos se calavam, batiam inevitáveis as ondas contra as pedras. Silêncio não havia. Tomou suas coisas, voltou ao continente, recolheu-se numa gruta em montanha distante. Embora isolado, logo se viu rodeado de ruídos, pequenos alguns, minúsculos outros, que o aparente silêncio circundante agigantava. Era o gotejar do excesso de umidade, o esvoejar dos morcegos ao anoitecer, o zumbir de um ou outro inseto, um gorjear lá fora, um escavar cá dentro, um rastejar, e o ronco majestoso dos trovões, o estalar dos relâmpagos. Novamente arrebanhou seus poucos pertences. E desceu a montanha, regressou à cidade. As chaves da sua casa tilintavam no bolso, não atendeu ao apelo. Tomou ônibus e metrô, caminhou até a praça mais central. Ali, onde tantos passavam e as buzinas dos carros e os apitos dos guardas e o gritar dos ambulantes e o chamado das sirenes se entrecruzavam, sentou-se. Assim como havia ignorado as chaves, ignorou os sons todos que lhe atingiam a cabeça, esqueceu os ouvidos. E, vagarosamente, começou a descida em seu silêncio interior.

Marina Colasanti. *Hora de alimentar serpentes*. Global, 2013

01

No conto, a busca do protagonista está relacionada à tentativa de

(A) demonstrar que o silêncio perfeito é uma construção imaginária e artificial.
(B) isolar-se do mundo externo para criar uma realidade paralela e inexorável.
(C) comprovar que a natureza é menos ruidosa do que o ambiente urbano.
(D) encontrar uma quietude absoluta que, paradoxalmente, revela-se inalcançável.
(E) estabelecer um contraponto entre a vida solitária e a vida social da cidade.

V

02

O sufixo “-ejar”, presente em “esvoejar”, desempenha papel semântico específico na construção do verbo, conferindo-lhe a ideia de:

(A) Estado contínuo e permanente.
(B) Movimento leve e intermitente.
(C) Intensificação de uma ação.
(D) Formação de substantivos abstratos.
(E) Relação de causa e consequência.

Texto para as questões de 03 a 05

Cuidar da nossa saúde às vezes lembra aquela olhadela que damos na cabine do avião a caminho de nosso assento. Por todo lado só vemos coisas complicadas: telas, indicadores, alavancas, luzes piscantes, manivelas, interruptores, mais alavancas... botões do lado esquerdo, botões do lado direito, botões no teto (não, fala sério, Por que eles põem botões no teto?). Desviamos o olhar, agradecidos pelo fato de os pilotos saberem o que estão fazendo. Como passageiros tudo que nos importa é se o avião vai ficar no céu. Quando a questão é nosso corpo, somos nós os passageiros ignorantes. Porém - reviravolta na história -, os pilotos também somos nós. E quando não sabemos como nosso corpo funciona, é como se estivéssemos em voo cego. Nós sabemos como queremos nos sentir. Queremos acordar com um sorriso, animados e empolgados para o novo dia. Queremos ter uma alegria no andar, livres de qualquer dor. Queremos passar momentos agradáveis com nossa família, com uma sensação de gratidão positividade. Mas pode ser complicado descobrir como chegar lá. São tantos botões que nos sentimos esmagados. O que fazer? Por onde começar? Temos que começar pela glicose. Por quê? Porque ela é a alavanca da cabine com o maior custo-benefício. É a mais fácil de compreender (graças aos monitores contínuos de glicose), afeta instantaneamente nossas sensações (porque influencia nossa fome e nosso humor), e muita coisa passa a se encaixar a partir do momento em que conseguimos controlá-la.

Adaptado de Inchauspé, Jessie. *A revolução da glicose: equilibre os níveis de açúcar no sangue e mude sua saúde e sua vida*. Trad. André Fontenelle. Objetiva, 2022.

03

No texto, a relação entre a complexidade da cabine de um avião e a administração da saúde humana evidencia

(A) a inutilidade de tentar entender processos fisiológicos, excessivamente complicados.
(B) a dificuldade de compreender o próprio corpo e o conhecimento para controlá-lo.
(C) o fato de que apenas profissionais especializados podem lidar com questões de saúde.
(D) a impossibilidade de pessoas comuns poderem interferir no próprio bem-estar.
(E) a necessidade de confiar em terceiros para a regulação da saúde, física e emocional.

V

04

No trecho “Como passageiros tudo que nos importa é se o avião vai ficar no céu”, a inclusão do termo “o” antes de “que” tem como efeito:

(A) Tornar “o que” equivalente a “aquilo que”, funcionando como pronome demonstrativo.
(B) Alterar a função sintática de “que”, transformando-o em conjunção integrante.
(C) Modificar o sentido original da frase, tornando-a ambígua.
(D) Indicar uma relação de posse, característica de pronomes relativos.
(E) Introduzir ideia de consequência, assumindo o sentido de vulnerabilidade.

05

O texto se organiza essencialmente por meio de:

(A) Descrições objetivas e técnicas, abrangendo fisiologia humana.
(B) Argumentação exclusivamente baseada em dados estatísticos.
(C) Estruturas formais, utilizadas em textos científicos acadêmicos.
(D) Afirmações e questionamentos sem embasamento lógico ou científico.
(E) Comparações e metáforas, simplificando conceitos complexos.

V

Texto para as questões de 06 a 08

Mal o CEO da Meta, Mark Zuckerberg, anunciou as mudanças nas políticas de moderação de suas plataformas, muitos educadores, comunicadores e jornalistas presentes nos diferentes grupos dos quais faço parte começaram a questionar a eficácia da Educação Midiática. O que podemos diante de um Musk e um Zuckerberg? De que adianta educar para a checagem de notícias se agora "abriram-se as porteiras" e nenhum de nós vai dar conta de distinguir o que é verdadeiro ou falso, de remover conteúdos agressivos, preconceituosos, de construir referenciais seguros para obtermos informações íntegras e confiáveis? É enxugar gelo, nadar contra a corrente, melhor a gente se preparar para viver no caos, diziam alguns, já ameaçando sair de vez das redes, boicotar a Meta, banir o digital de vez do seu cotidiano. Entendo a Educação Midiática como um importante e potente elemento para que possamos lidar com todos os desafios presentes no mundo digital – e de resto, no mundo real, que o reflete. Ela é uma alternativa viável e segura que todos nós, que desejamos continuar a viver civilizadamente em sociedade, podemos tomar em nossas mãos. Se as chamadas big techs nos abandonam à própria sorte, cabe a cada um de nós entender qual é o nosso papel nesse ecossistema.

Adaptado de: Januária Cristina Alves. "Novas diretrizes da Meta: será o fim da Educação Midiática?" Nexo Jornal. 16 de janeiro de 2025.

06

No trecho "Mal o CEO da Meta, Mark Zuckerberg, anunciou as mudanças nas políticas de moderação de suas plataformas", o termo "Mal" estabelece uma relação de

(A) comparação, equiparando dois acontecimentos simultâneos.
(B) condição, introduzindo uma possibilidade de mudança.
(C) consequência, indicando um efeito direto da ação posterior.
(D) oposição, contrastando ideias consideradas divergentes.
(E) tempo, indicando uma ação ocorrida imediatamente após outra.

07

No trecho, e em relação ao contexto em que se insere, "o mundo real, que o reflete", a oração após a vírgula tem a função de

(A) servir como uma explicação para ser utilizada pelas *big techs*.
(B) apresentar um aposto que exemplifica o conceito anterior.
(C) introduzir uma nova informação independente da anterior.
(D) retomar "o mundo digital", explicando sua relação com o mundo real.
(E) estabelecer um juízo de valor sobre o funcionamento das redes sociais.

V

08

No período em que se encontra, a sequência textual "melhor a gente se preparar" apresenta-se gramaticalmente como

(A) uma inversão sintática, típica da linguagem formal.
(B) um erro gramatical, com a falta do verbo principal na oração.
(C) uma construção elíptica, simplificando o segmento sintático.
(D) um exemplo de hipérbole, intensificando a situação descrita.
(E) uma construção arcaica, comum em textos literários antigos.

Texto para as questões de 09 a 14

Rain Is Coming to Burning Los Angeles and Will Bring Its Own Risks

Rain is forecast to begin as soon as Saturday afternoon and to continue as late as Monday evening, says meteorologist Kristan Lund of the National Weather Service's Los Angeles office. The area desperately needs the precipitation, but experts are warily monitoring the situation because rain poses its own risks in recently burned areas—most notably the potential occurrence of mudslides and similar hazards. "Rain is good because we've been so dry," Lund says. "However, if we get heavier rain rates or we get the thunderstorms, it's actually a lot more dangerous because you can get debris flows."

Fires do a couple of different things to the landscape that can increase the risk of burned material, soil and detritus hurtling out of control. When fires burn hot or long enough, they leave an invisible layer of waxy material just under the surface of the ground. This develops from decomposing leaves and other organic material, which contain naturally hydrophobic or water-repellent compounds. Fire can vaporize this litter, and the resulting gas seeps into the upper soil—where it quickly cools and condenses, forming the slippery layer.

When rain falls on ground that has been affected by this phenomenon, it can't sink beyond the hydrophobic layer—so the water flows away, often hauling debris with it. "All of the trees, branches, everything that's been burned—unfortunately, if it rains, that stuff just floats," Lund says. "It's really concerning." Even a fire that isn't severe enough to create a hydrophobic layer can still cause debris flows, says Danielle Touma, a climate scientist at the University of Texas at Austin. Under normal conditions, trees and other plants usually trap some rain above the surface, slowing the water's downward journey. But on freshly burned land there's much less greenery to interfere; all the rain immediately hits the ground. [...]

Fortunately, the rain should also help firefighters tame the blazes that remain active. The largest, the Palisades Fire, is currently 77 percent contained. The second largest, the Eaton Fire, is 95 percent contained. The Hughes Fire is third largest and only 56 percent contained. A fire can be fully contained but still burning. The containment percentage refers to the amount of the perimeter that has barriers that firefighters expect will prevent further spread.

Scientific American. January 27th, 2025. Adaptado.

09

Com base no primeiro parágrafo e na opinião dos especialistas, qual das seguintes inferências pode ser feita?

(A) A área queimada apresenta sérios danos, e as chuvas fortes podem auxiliar na recuperação do meio ambiente.
(B) O meteorologista Kristan Lund acredita que a chuva será um problema, independentemente de sua intensidade.
(C) Embora a chuva prevista para o fim de semana seja bem-vinda, há também preocupações devido aos riscos de deslizamentos de terra.
(D) Como a possibilidade de deslizamentos de terra ocorre apenas nos locais não afetados por incêndios, os malefícios são imperceptíveis.
(E) Uma vez que a região está bem-preparada para lidar com a ocorrência de chuvas, a área não está suscetível a riscos.

10

O termo "litter", no parágrafo 2, refere-se

(A) ao gás que é liberado pela decomposição das folhas durante as queimadas.
(B) à substância que resulta na maior fertilidade do solo após o incêndio.
(C) ao material utilizado pelos bombeiros para controlar incêndios florestais.
(D) ao acumulado de cinzas que se forma após a queima do material decomposto.
(E) à camada de material orgânico que cobre a superfície do solo.

11

Considerando a oração "[...] it can't sink beyond the hydrophobic layer—so the water flows away [...]" (3º parágrafo), o termo "so" pode ser substituído, sem prejuízo de sentido, por

(A) moreover.
(B) therefore.
(C) nevertheless.
(D) conversely.
(E) furthermore.

12

Segundo Danielle Touma, uma especialista em ciências climáticas da Universidade do Texas em Austin,

(A) mesmo em terrenos queimados, a chuva não consegue causar deslizamentos de detritos.
(B) qualquer queimada pode criar uma camada que repele a água, dificultando sua absorção.
(C) até incêndios menos graves podem resultar em deslizamentos de detritos em períodos de chuva devido à falta de vegetação.
(D) a chuva é mais eficaz em terrenos queimados, pois a camada hidrofóbica impede a perda de água.
(E) sob condições normais, a mata impede as águas pluviais de penetrar no solo de maneira eficiente.

13

Na oração "[...] the rain should also help firefighters tame the blazes that remain active. [...]" (4º parágrafo) o uso do verbo modal should indica

(A) conselho.
(B) capacidade.
(C) condição.
(D) expectativa.
(E) obrigação.

14

Considerado o contexto, ao usar o termo “Fortunately” (4º parágrafo), o autor

(A) demonstra apreensão ao sugerir que os incêndios ainda estão fora de controle e a chuva não será suficiente para ajudar.
(B) almeja suavizar a situação descrita, visto que a chuva sozinha não resolve o problema e existem incêndios ainda não contidos.
(C) transmite a necessidade de ações rápidas, destacando que a contenção dos incêndios precisa ser acelerada para evitar mais danos.
(D) expressa preocupação e questiona a capacidade dos bombeiros em controlar os incêndios de maneira eficaz.
(E) mostra-se indiferente, sem se aprofundar nos detalhes ou nas implicações dos incêndios que ainda estão ocorrendo.

15

A soma dos 5 elementos de uma progressão geométrica (PG) de razão igual a 2 é 651. O último termo dessa PG é

(A) 312.
(B) 320.
(C) 324.
(D) 332.
(E) 336.

16

Um triângulo isósceles possui lados iguais a x (dois dos lados) e y (um lado). Sabendo-se que $x + y = 10$, $x \cdot y = 24$ e $x > y$, a área desse triângulo é

(A) $6\sqrt{2}$.

(B) $8\sqrt{2}$.

(C) $9\sqrt{2}$.

(D) $10\sqrt{2}$.

(E) $14\sqrt{2}$.

17

Seja θ um ângulo entre 90 e 180 graus. Se o seno de θ for $\frac{3}{5}$, o seu cosseno será:

(A) $-\frac{2}{5}$

(B) $\frac{4}{5}$

(C) $-\frac{4}{5}$

(D) $-\frac{3}{5}$

(E) $\frac{3}{5}$

18

Dentre as alternativas a seguir, aquela que apresenta o maior valor é:

(A) 25^{40}

(B) 10^{55}

(C) 6^{75}

(D) 12^{55}

(E) 15^{50}

Note e adote:
$\log_{10} 2 = 0{,}301$
$\log_{10} 3 = 0{,}477$
$\log_{10} 5 = 0{,}699$

19

Um fazendeiro possui nove vacas leiteiras que produzem, ao longo de 25 dias, 5.800 litros de leite. Suponha que ele compre mais seis vacas que tenham a mesma produção média diária de leite que as anteriores. A produção de leite dessas quinze vacas, ao longo de 45 dias, será

(A) 15.600 litros.
(B) 16.000 litros.
(C) 16.800 litros.
(D) 17.400 litros.
(E) 18.200 litros.

20

Um reservatório de água tem o formato de uma pirâmide de altura 6 metros e base quadrada de lado 4 metros.
Quando esse reservatório estiver cheio até $\frac{3}{4}$ de sua altura, o volume de água será, em metros cúbicos:

(A) $\frac{63}{2}$

(B) $\frac{59}{2}$

(C) 27

(D) $\frac{53}{2}$

(E) 24

V

21

Um banco de dados possui 15 questões de matemática e 12 questões de português para serem sorteadas para uma prova contendo três questões de cada uma das disciplinas. Com esses dados, o número de provas distintas possíveis é

(A) 80.200.
(B) 86.000.
(C) 92.500.
(D) 96.000.
(E) 100.100.

V

22

Um banco cobra, em seu cheque especial, a taxa de 10% ao mês, e a dívida é atualizada no primeiro dia de cada mês subsequente à utilização. Se um cliente ficou negativado em 1.000 reais no dia primeiro de fevereiro de 2025 e, desde então, não conseguiu fazer nenhum pagamento, a sua dívida no dia primeiro de julho de 2025 será

(A) 1.100,00 reais.
(B) 1.500,00 reais.
(C) 1.610,51 reais.
(D) 1.712,35 reais.
(E) 1.800,00 reais.

23

"ChatGPT, DeepSeek e similares pertencem à classe de LLMs, avançados modelos de linguagem treinados a partir de grandes bancos de dados – majoritariamente em inglês. Os mais populares pertencem a empresas norte-americanas. E assim como os algoritmos de pesquisa e redes sociais, não são neutros. Ou seja, podem reproduzir vieses, preconceitos e estereótipos de seus programadores, que por sua vez podem receber ordens dos donos das empresas e de outros atores – na China, por exemplo, empresas devem passar por análise de segurança e obter aprovações do governo antes de lançar produtos (...).
Nesse cenário, como ficam os países que não têm plataformas nacionais de alcance global, como é o caso do Brasil? E mais: o que acontecerá com a História e a memória desses países diante de uma população cada vez mais conectada a LLMs estrangeiras globais que acredita mais no que encontra nos apps e sites de busca do que nos livros de História? "

Luciana Garbin, IAs estão apagando e reescrevendo pedaços da História. E o Brasil com isso?, *O Estado de S. Paulo* (on-line), 29/01/2025 (Adaptado)

O texto apresentado traz uma crítica

(A) ao trabalho dos programadores, que estabelecem mecanismos frágeis de intelecção das máquinas dotadas de inteligência.
(B) às chamadas *big techs* que visam apenas o lucro a partir dos serviços que disponibilizam, gerando dificuldades para o usuário.
(C) às empresas de tecnologia brasileiras, que não desenvolveram LLMs de alcance global, o que seria possível a despeito do alcance da língua portuguesa no mundo.
(D) aos governos das nações desenvolvidas, por estabelecerem padrões desiguais de controle de segurança para os produtos como o ChatGPT e o DeepSeek.
(E) às pessoas que, em grande parte, passaram a confiar em fontes e a acreditar em informações não contextualizadas historicamente.

V

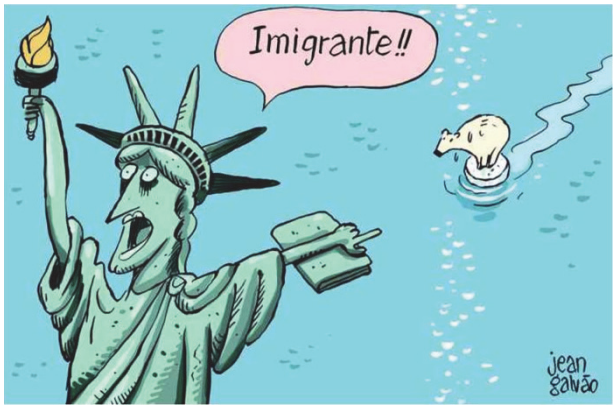
24

Em *Ideias para adiar o fim do Mundo*, Ailton Krenak coloca ênfase no papel que o rio Watu desempenha para a unidade do povo *krenak*, o povo "cabeça da terra". O rio foi palco de um evento que marcou nosso país. De posse dessas informações e com base na leitura do livro, assinale a alternativa que indica o nome do rio, em português, e o evento mencionado.

(A) Solimões – seca de 2023.
(B) Doce – rompimento da barragem do Fundão.
(C) Tietê – enchente de São Paulo em 1929.
(D) São Francisco – transposição de suas águas.
(E) Paraopeba – rompimento da barragem de Brumadinho.

25

Observe a charge a seguir:



Folha de São Paulo, 26.01.2025

Assinale a alternativa que melhor descreve as situações às quais a charge se refere.

(A) Denuncismo, confusão relativa às categorias sociais e mudanças climáticas.
(B) Estigmas sociais, alarmismo social e mobilidade social.
(C) Alarmismo social, identificação dos imigrantes como animais e mudanças climáticas.
(D) Racismo, degelo da Antártida e luta de classes.
(E) Denuncismo, extinção de espécimes da fauna e mudanças sociais.

V

26

Em *O Perigo de uma História Única*, Chimamanda Ngozi Adichie afirma que "Há pouco tempo dei uma palestra numa universidade e um aluno me disse que era uma grande pena que os homens nigerianos fossem agressivos como o personagem do pai no meu romance. Eu disse a ele que tinha acabado de ler um livro chamado *O psicopata americano* e que achava uma grande pena que os jovens americanos fossem assassinos em série. Bem, obviamente eu disse isso num leve ataque de irritação. Mas jamais teria me ocorrido pensar que, só porque li um romance no qual o personagem era um assassino em série, ele de alguma maneira representava todos os americanos. Não digo isso porque me considero uma pessoa melhor do que esse aluno (...). Já tinha lido Tyler, Updike, Steinbeck e Gaitskill. Não tinha uma história única dos Estados Unidos".
Nesse livro, como no trecho de O Estado de S. Paulo citado na questão 23, pode-se afirmar que a versão de um fato será tanto mais disseminada quanto

(A) maior for o poder econômico e cultural de quem a comunica.
(B) mais críveis forem os seus contornos narrativos.
(C) menos verossímeis forem os valores ínsitos à mensagem.
(D) maior for o poder militar e científico de quem a comunica.
(E) menores forem os riscos de conter inverdades.

27

Considere o art. 2º do Estatuto da USP:

> Artigo 2º – São fins da USP:
>
> I – promover e desenvolver todas as formas de conhecimento, por meio do ensino e da pesquisa;
> II – ministrar o ensino superior visando à formação de pessoas capacitadas ao exercício da investigação e do magistério em todas as áreas do conhecimento, bem como à qualificação para as atividades profissionais;
> III – estender à sociedade serviços indissociáveis das atividades de ensino e de pesquisa.

As alternativas a seguir indicam as cinco Pró-Reitorias existentes na USP. Assinale aquela cujas atividades NÃO se relacionam diretamente com os fins da Universidade.

(A) Pró-Reitoria de Cultura e Extensão Universitária.
(B) Pró-Reitoria de Graduação.
(C) Pró-Reitoria de Inclusão e Pertencimento.
(D) Pró-Reitoria de Pesquisa e Inovação.
(E) Pró-Reitoria de Pós-Graduação.

V

28

Considere a seguinte situação: Ênio Oliveira, Vice-Reitor da USP, falece em um acidente. Nesse caso, Edna Cruz, a Reitora, deverá

(A) indicar um novo Vice-Reitor, que será homologado pelo Conselho Universitário em até 15 dias.
(B) dar início ao processo eleitoral, para a escolha de um novo Vice-Reitor, que cumprirá um mandato novo, de 4 anos.
(C) ser substituída, em suas ausências ou impedimentos, pelo decano do Conselho Universitário.
(D) dar início ao processo eleitoral, para a escolha de um novo Vice-Reitor, que exercerá tal função pelo tempo que restava de mandato para Ênio.
(E) indicar um novo Vice-Reitor, que deverá ser nomeado pelo Governador de SP em até 15 dias.

V

29

Uma Unidade tem 8 Professores Titulares, todos membros natos da Congregação. O número de representantes dos Professores Associados e dos Professores Doutores é, respectivamente,

(A) 2 e 1.
(B) 4 e 3.
(C) 5 e 3.
(D) 4 e 2.
(E) 6 e 3.

30

Aproximando-se as inscrições para Diretor de um Instituto, a comunidade local sabe que Lucas, Ana e Maria pretendem ser candidatos, tendo como candidatos a Vice-Diretor, respectivamente, Sara, Lia e Marcos. Sabendo que Marcos é Professor Associado 2 e todos os demais são Professores Titulares, é possível afirmar que a Chapa Maria e Marcos

(A) não pode se candidatar, em nenhuma hipótese.
(B) pode se candidatar, sem qualquer restrição.
(C) pode se candidatar numa eventual segunda fase de inscrições, mas apenas se Lucas e Sara ou Ana e Lia deixarem de se inscrever.
(D) pode se candidatar numa eventual segunda fase de inscrições, mesmo que as chapas Lucas e Sara e Ana e Lia se inscrevam.
(E) pode se candidatar desde logo, mas só concorrerão se Lucas e Sara ou Ana e Lia deixarem de se inscrever.

V

31

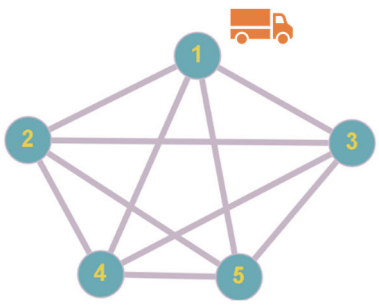
Na biblioteca *Scipy*, existem várias funções utilizadas para avaliar o comportamento estatístico de amostras de dados. Assinale a alternativa que descreve uma função que pode ser usada para verificar se um grande volume de dados segue uma distribuição comum, como a normal, sem exigir que os parâmetros da distribuição sejam previamente conhecidos.

(A) `scipy.stats.anderson()`
(B) `scipy.stats.ks_2samp()`
(C) `scipy.stats.shapiro()`
(D) `scipy.stats.pearsonr()`
(E) `scipy.stats.ttest_1samp()`

V

32

A análise de dados espaciais auxilia na solução de problemas de roteamento e na escolha de trajetos ótimos realizados por veículos. Uma empresa de logística precisa definir a melhor rota para um entregador que deve visitar clientes diferentes em uma cidade, retornando ao ponto de origem após a última entrega. O objetivo é minimizar a distância percorrida, garantindo que cada cliente seja visitado exatamente uma vez, conforme pode ser observado no grafoa seguir, onde o caminhão deve partir do ponto 1 e retornar para esse mesmo ponto, após todas as entregas.



Assinale a alternativa que apresenta a categoria de análise espacial na qual esse problema se encaixa.

(A) Algoritmo de Dijkstra, que encontra o caminho mais curto entre dois pontos em um grafo.
(B) Método de interpolação espacial, que estima valores desconhecidos com base em pontos de referência próximos.
(C) Interpolação Krigagem, usada para estimar valores em áreas não amostradas.
(D) Análise de Sensibilidade Espacial, usada para medir impactos de mudanças geográficas.
(E) Problema do Caixeiro Viajante, um problema de otimização combinatória que busca minimizar a distância total percorrida ao visitar um conjunto de pontos.

V

33

No Power BI, a modelagem de dados é essencial para garantir desempenho e a correta interpretação das informações. Um modelo, no Power BI, consiste em uma ou mais tabelas e diversas relações entre elas (quando existir mais de uma tabela). Para garantir granularidade e eficiência nas visualizações e relatórios, a escolha do esquema de dados é fundamental.
Uma empresa está implementando um *dashboard* no Power BI para monitorar as vendas de seus produtos em diversas regiões do país. O banco de dados contém informações sobre:

- Vendas realizadas (data, valor, quantidade, produto vendido, vendedor e região).
- Detalhes dos produtos (código, categoria, marca e preço unitário).
- Informações dos clientes (nome, CPF, idade, estado civil e cidade).
- Registros de vendedores (nome, código do vendedor e equipe de vendas).

Considerando as melhores práticas de modelagem de dados no Power BI, qual esquema de dados é mais adequado para estruturar esse modelo e garantir performance e facilidade de análise?

(A) Modelo plano, pois consolidar todas as informações em uma única tabela elimina a necessidade de relacionamentos, simplificando as consultas.
(B) Modelo estrela, pois permite organizar os dados com uma tabela fato de vendas conectada a tabelas dimensão, otimizando a performance e a flexibilidade das análises.
(C) Modelo floco de neve, pois ao normalizar as tabelas dimensão, reduz a redundância e melhora a velocidade das consultas no Power BI.
(D) Modelo plano, pois a ausência de *joins* melhora a escalabilidade do modelo ao lidar com grandes volumes de dados.
(E) Modelo floco de neve, pois evita qualquer redundância ao dividir as dimensões em tabelas menores, garantindo um modelo mais eficiente.

V

34

Considere um Analista de Sistemas especializado em Ciência de Dados, designado para analisar grandes volumes de textos livres oriundos de interações dos clientes com a empresa, incluindo mensagens enviadas por *chat* e redes sociais, com o objetivo de extrair *insights* sobre a satisfação dos consumidores. Esses textos são classificados como dados

(A) estruturados.
(B) semiestruturados.
(C) quantitativos.
(D) não estruturados.
(E) ordinais.

35

Uma grande empresa do setor financeiro decidiu modernizar sua infraestrutura de dados para suportar análises preditivas e relatórios gerenciais avançados, além de manter a eficiência nas transações diárias de seus clientes. Atualmente, a empresa possui um banco de dados relacional tradicional que armazena transações bancárias em tempo real, mas enfrenta dificuldades ao executar consultas analíticas complexas, como identificação de padrões de fraude e segmentação de clientes com base no histórico de gastos.
Diante desse cenário, a empresa considera a separação da sua arquitetura de dados em dois ambientes distintos: um banco de dados transacional (OLTP) e um ambiente analítico (OLAP).

Em relação ao contexto apresentado, assinale a alternativa correta.

(A) A empresa pode continuar utilizando um único banco OLTP, desde que otimize seus índices e crie visões materializadas para melhorar a performance de consultas analíticas, eliminando a necessidade de um ambiente OLAP.
(B) A empresa deve optar por um banco de dados NoSQL em substituição ao OLTP tradicional, pois bancos relacionais não são capazes de lidar com transações financeiras de forma eficiente.
(C) O principal benefício de utilizar um OLAP nesse cenário é garantir alta disponibilidade e escalabilidade horizontal, melhorando a velocidade das transações diárias dos clientes.
(D) O ambiente OLTP deve ser utilizado para armazenar transações bancárias e consultas operacionais em tempo real, enquanto o OLAP será responsável por consultas analíticas complexas, como detecção de fraudes e previsão de tendências.
(E) Em uma arquitetura de Big Data moderna, não há mais distinção entre OLTP e OLAP, pois soluções como Apache Spark e Hadoop substituem qualquer necessidade de bancos de dados transacionais e analíticos separados.

V

36

Uma ONG, especializada na busca por gatos desaparecidos, contratou uma empresa de tecnologia para desenvolver um sistema de classificação de imagens baseado em Aprendizado Profundo. O objetivo é que o sistema identifique gatos em fotos enviadas por usuários. Para isso, a empresa optou por utilizar Redes Neurais Convolucionais (CNNs), dada sua capacidade de extrair automaticamente padrões visuais hierárquicos.
Durante o treinamento, os desenvolvedores perceberam que a rede estava obtendo alta acurácia no conjunto de treino, mas baixo desempenho no conjunto de teste. Além disso, ao inspecionar os mapas de ativação, notaram que a rede estava focando em características irrelevantes do fundo da imagem em vez de identificar os gatos corretamente.

E relação ao problema descrito, assinale a alternativa que apresenta a abordagem mais eficaz para aprimorar a capacidade de generalização de um modelo de aprendizado de máquina.

(A) Aumentar a complexidade do modelo, utilizando o máximo de parâmetros possíveis, sem restrições.
(B) Treinar o modelo exclusivamente com os dados de treinamento disponíveis, sem validação externa ou ajuste fino.
(C) Implementar técnicas de regularização, como L1 ou L2, e utilizar validação cruzada para avaliar o desempenho e ajustar hiperparâmetros.
(D) Reduzir drasticamente o conjunto de dados de treinamento para evitar sobreajuste, mesmo que isso comprometa a representatividade dos dados.
(E) Ignorar a fase de pré-processamento de dados e utilizar os dados brutos diretamente no treinamento do modelo.

V

37

Uma equipe de cientistas de dados desenvolve um modelo preditivo para estimar o preço de carros usados com base em variáveis como ano de fabricação, quilometragem, marca e número de proprietários anteriores.

Assinale a alternativa que apresenta a abordagem mais adequada para construir este modelo preditivo.

(A) Utilizar um modelo de regressão linear simples que leva em conta apenas a quilometragem do carro.
(B) Utilizar um modelo de regressão logística para classificar os carros como baratos ou caros.
(C) Utilizar apenas variáveis numéricas e excluir variáveis categóricas como a marca do carro.
(D) Utilizar um modelo de regressão polinomial que sempre considera a relação entre quilometragem e preço como uma curva quadrática.
(E) Utilizar um modelo de regressão linear múltipla que considera todas as variáveis fornecidas (ano de fabricação, quilometragem, marca e número de proprietários anteriores).

V

38

Analise o trecho de código Python a seguir:

```
1       a = [1,2,3,4,5,6,7,8,9,10]
2       for i in range (0, 10):
3           a[i] = a[i] + a[i-2]
4       print(a[i]*a[i-2] - a[i-1]*a[i-1])
```

Em relação ao trecho apresentado, assinale a alternativa que indica o conteúdo que será exibido na tela a partir da execução da linha 4 (*print*).

(A) o código não funciona, pois apresentará um erro em tempo de execução.
(B) 0
(C) -4
(D) 104
(E) 44

39

Os algoritmos de clusterização são utilizados na ciência de dados para agrupar elementos semelhantes com base em suas características. Um dos métodos mais comuns para medir a similaridade entre pontos é a distância Euclidiana, que calcula o quão próximos ou distantes os elementos estão em um espaço multidimensional. Essa métrica é a base para a determinação da formação dos clusters em algoritmos como K-Means e DBSCAN.
Uma empresa deseja agrupar clientes com base em seu comportamento de compra. Para isso, foram coletados dois atributos: a quantidade de produtos diferentes comprados no último mês (X) e o valor total gasto (em centenas de reais) (Y). A tabela, a seguir, apresenta os dados coletados de quatro clientes, que serão usados para gerar a matriz de distâncias com base na distância Euclidiana:

Cliente	Qtde de produtos (X)	Valor Gasto (Y)
A	2	3
B	5	7
C	1	4
D	6	2

Em relação à matriz de distância gerada, assinale a alternativa correta.

(A) O cliente B está mais próximo do cliente D do que do cliente C.
(B) O cliente A está mais próximo do cliente C do que do cliente D.
(C) A matriz de distâncias não pode ser simétrica.
(D) A maior distância entre dois clientes na matriz é 6.00.
(E) O cliente C tem a menor soma total de distâncias para os outros clientes.

V

40

Em aprendizado de máquina, a calibração de hiperparâmetros é um processo importante para otimizar o desempenho de um modelo. Considere o seguinte cenário:
Você está treinando um modelo de *Random Forest* para prever o preço de imóveis e percebe que o desempenho do modelo não está satisfatório. Após uma análise, você decide calibrar os hiperparâmetros para tentar melhorar o modelo. Para isso, você seleciona os seguintes hiperparâmetros para calibração:

- n_estimators (número de árvores na floresta);
- max_depth (profundidade máxima de cada árvore);
- min_samples_split (número mínimo de amostras necessárias para dividir um nó).

Assinale a alternativa que apresenta a melhor abordagem para encontrar a combinação ideal desses hiperparâmetros.

(A) Ajustar os hiperparâmetros manualmente, testando diferentes combinações de uma única vez, sem validação cruzada, até encontrar uma configuração que melhore o desempenho.
(B) Utilizar a técnica de pesquisa aleatória (*Random Search*), testando uma combinação aleatória de valores para os hiperparâmetros, sem avaliar o desempenho em diferentes subdivisões do conjunto de dados.
(C) Focar apenas no parâmetro n_estimators e testar os valores 50, 100 e 150, já que esse é o parâmetro mais importante para a *Random Forest*.
(D) Usar a técnica de pesquisa em grade (*Grid Search*), testando todas as combinações possíveis de valores para n_estimators, max_depth e min_samples_split, e avaliar a performance a partir da validação cruzada.
(E) Manter os hiperparâmetros padrões da biblioteca e esperar que o modelo se ajuste automaticamente, pois os hiperparâmetros padrões funcionam bem na maioria dos casos.

V

41

Um modelo de linguagem baseado em unigramas foi treinado em um grande volume de textos em português. Esse modelo atribui probabilidades a palavras individuais, sem levar em consideração a ordem em que aparecem na sentença. Sabendo-se que a perplexidade é uma métrica que mede quão bem um modelo de linguagem prediz um texto, assinale a alternativa que melhor representa a perplexidade do modelo nas frases "qual sanduíche Maria comeu" e "Maria comeu o sanduíche".

(A) As perplexidades das duas frases serão iguais, pois ambas contêm as mesmas palavras.
(B) A perplexidade de "Maria comeu o sanduíche" será maior, pois a presença de "o" torna a sequência mais previsível.
(C) A perplexidade de "qual sanduíche Maria comeu" será maior, pois a palavra "qual" tem uma probabilidade menor de ocorrência e é menos comum.
(D) A perplexidade de "Maria comeu o sanduíche" será menor, pois o modelo tende a atribuir maior probabilidade à sequência de palavras com maior frequência.
(E) A perplexidade de "qual sanduíche Maria comeu" será menor, pois a palavra "qual" ajuda a contextualizar melhor a sequência de palavras.

V

42

O pré-processamento de textos é uma etapa importante no processo de análise e classificação de dados textuais. Ele visa transformar textos brutos em um formato adequado para ser utilizado em algoritmos de aprendizado de máquina. Entre as técnicas mais comuns no pré-processamento de textos, estão a remoção de *stop words*, a tokenização, a lematização e o estemização. Considere o texto original a seguir:
"O carro estava muito sujo, então ele decidiu limpar o carro depois de um longo dia de trabalho. O carro ficou brilhante após a limpeza."

Com base nas técnicas de pré-processamento citadas, como ficará o texto original após a aplicação de tokenização e remoção de *stop words*?

(A) ["carro", "sujo", "decidir", "limpeza", "brilhante"]
(B) ["carro", "sujo", "decidiu", "limpar", "carro", "longo", "dia", "trabalho", "carro", "brilhante", "limpeza"]
(C) ["carro", "sujo", "decidiu", "limpar", "carro", "longo", "dia", "trabalho", "carro", "brilhante", "após", "limpeza"]
(D) ["carro", "limpar", "brilhante", "carro"]
(E) ["carro", "limpeza", "brilhante", "decidiu", "trabalho"]

43

Considere a classe No, implementada em Python, que será a base de formação de uma Lista Simplesmente Encadeada:

```
class No:
        def __init__(self, dado):
        self.dado = dado
        self.proximo = None
```

Considere ainda o trecho de código em Python que manipula a Lista Simplesmente Encadeada e que está declarado dentro da classe ListaEncadeada:

```
class ListaEncadeada:
        def __init__(self):
             self.cabeca = None

        def metodoX(self):
            if not self.cabeca:
                return None
            Y = self.cabeca
            atual = self.cabeca
            anterior = None
            anterior_Y = None
            while atual:
                if atual.dado > Y.dado:
                        Y = atual
                        anterior_Y = anterior
                anterior = atual
                atual = atual.proximo
            if anterior_Y is None:
                self.cabeca = Y.proximo
            else:
                anterior_Y.proximo = Y.proximo
            return Y.dado

        def metodoZ(L):
            K = ListaEncadeada()
            while L.cabeca:
               W = L.metodoX()
               K.inserir_fim(W)
            return K
```

A classe ListaEncadeada contém outros métodos que permitem a sua completa manipulação, como inserir elemento no início, inserir elemento no final, exibir conteúdo da lista e remover elementos. Assinale a alternativa que apresenta o conteúdo retornado pelo metodoZ, quando for enviado como parâmetro a seguinte Lista Ligada: `[15, 28, 2, 10, 50, 14, 77]`

(A) `[2, 10, 14, 15, 28, 50, 77]`
(B) `[77, 50, 28, 15, 14, 10, 2]`
(C) `[77, 28, 15, 50, 14, 10, 2]`
(D) `[50, 28, 15, 77, 14, 10, 2]`
(E) `[77, 28, 14, 15, 50, 10, 2]`

44

Considere o seguinte código Python utilizando a biblioteca Pandas:

```
import pandas as pd
dados = {'Nome': ['Ivo','Iza','Ney','Ana'],
 'Idade': [28, 34, 23, 21],
 'Salario': [3000, 4000, 1500, 2000]}

df = pd.DataFrame(dados)

t_idade = df['Idade'].dtype
print(f'O tipo da coluna Idade é: {t_idade}')
```

No código apresentado, foram utilizados dois conceitos importantes sobre o Pandas: DataFrame e dtype. Assinale a alternativa que contém a saída correta do comando print quando o código for executado.

(A) `O tipo da coluna "Idade" é: int64`
(B) `O tipo da coluna "Idade" é: object`
(C) `O tipo da coluna "Idade" é: int`
(D) `O tipo da coluna "Idade" é: objectInt`
(E) `O tipo da coluna "Idade" é: float64`

45

Suponha o desenvolvimento de uma plataforma que integra informações de diversas fontes governamentais para promover a transparência e facilitar a análise de políticas públicas. Para garantir que os dados utilizados atendam aos princípios de dados abertos, é fundamental compreender suas características essenciais. Em relação ao contexto descrito, assinale a alternativa que apresenta uma característica fundamental para que um conjunto de dados seja considerado "dado aberto".

(A) Disponibilidade restrita a usuários autorizados.
(B) Licenciamento que permite livre uso, modificação e compartilhamento.
(C) Formato proprietário que requer software específico para acesso.
(D) Proteção por direitos autorais que impede redistribuição.
(E) Acesso mediante pagamento de taxa.

46

Considere o seguinte código na linguagem R que utiliza estruturas de repetição para processar um vetor numérico:

```
numeros <- c(2, 4, 6, 8, 10)
resultado <- 0

for (i in seq_along(numeros)) {
  if (numeros[i] %% 4 == 0) {
    resultado <- resultado + numeros[i]
  }
}
print(resultado)
```

Com base na execução desse código, assinale a alternativa que apresenta a saída impressa pelo comando `print (resultado)`.

(A) 0
(B) 2
(C) 4
(D) 12
(E) 18

47

Uma empresa multinacional lida com grandes volumes de dados provenientes de diversas fontes, incluindo bancos de dados transacionais, sensores IoT, logs de servidores e redes sociais, envolvendo dados estruturados e não estruturados. Durante o processo de armazenamento e recuperação de dados, a organização enfrenta desafios de desempenho e consistência.

Considerando o cenário descrito, assinale a alternativa que apresenta a abordagem mais adequada para otimizar a recuperação eficiente e garantir a integridade dos dados.

(A) Utilizar exclusivamente bancos de dados relacionais tradicionais, pois garantem integridade referencial e eliminam problemas de latência, independentemente do volume de dados.

(B) Eliminar a necessidade de armazenamento distribuído, consolidando todos os dados em um único servidor de alta performance, garantindo acesso mais rápido sem necessidade de replicação.

(C) Priorizar a coleta e o armazenamento dos dados, pois a recuperação pode ser ajustada posteriormente sem impacto significativo no desempenho dos sistemas.

(D) Armazenar todos os dados no formato bruto (*raw data*) sem processamento prévio, pois isso permite maior flexibilidade na recuperação sem a necessidade de esquemas predefinidos ou otimizações de consulta.

(E) Implementar um sistema de armazenamento híbrido, combinando bancos de dados relacionais e não relacionais, além de técnicas como particionamento de dados e indexação para otimizar a recuperação.

V

48

Considere um cenário onde uma empresa precisa implementar uma solução de *Business Intelligence* (BI) para análise avançada de dados e relatórios interativos. O time de dados avalia Tableau e Power BI como principais opções. Assinale a alternativa que apresenta a afirmação correta sobre as diferenças e/ou semelhanças técnicas entre essas ferramentas.

(A) O Tableau permite a criação de modelos de dados relacionais internamente, enquanto o Power BI depende exclusivamente de bancos de dados externos para modelagem.

(B) No Tableau, os *dashboards* são atualizados automaticamente em tempo real sem depender da configuração do usuário, enquanto no Power BI é necessário definir atualizações manuais para manter os dados sincronizados.

(C) O Power BI usa a linguagem DAX (*Data Analysis Expressions*) para criar cálculos avançados e métricas, enquanto o Tableau utiliza o conceito de *Calculated Fields* para expressões analíticas.

(D) O Power BI permite a conexão direta a qualquer banco de dados sem necessidade de configuração adicional, enquanto o Tableau exige sempre a importação dos dados para sua plataforma antes da análise.

(E) O Tableau e o Power BI não permitem integração com serviços de nuvem para armazenamento e processamento de dados, sendo obrigatória a utilização de servidores locais para o processamento das informações.

49

Uma empresa de transporte urbano deseja analisar a distribuição de pontos de ônibus e a demanda de passageiros em uma cidade para otimizar suas rotas. O objetivo é identificar quais áreas possuem alta densidade de demanda e onde há necessidade de expansão da infraestrutura de transporte. Assinale a alternativa que apresenta a abordagem mais eficaz para realizar essa análise geoespacial e para identificar os padrões de distribuição de demanda.

(A) Utilizar análise de pontos quentes (*Hot Spot Analysis*) com a técnica de *Kernel Density Estimation* (KDE) para identificar áreas de alta concentração de passageiros, levando em consideração a distribuição espacial da demanda e a densidade de pontos de ônibus.

(B) Aplicar o algoritmo K-Means para a clusterização dos pontos de ônibus e a demanda de passageiros, considerando as distâncias euclidianas entre os pontos de interesse e utilizando um número fixo de clusters previamente definido.

(C) Implementar uma regressão logística para prever as áreas de alta demanda com base em variáveis socioeconômicas, sem considerar a localização geográfica dos pontos de ônibus e a distribuição de passageiros

(D) Utilizar o algoritmo DBSCAN para identificar a densidade de passageiros por área, mas não considerar as distâncias entre pontos de ônibus e áreas de maior demanda, ignorando a conexão espacial entre os pontos.

(E) Realizar uma análise espacial utilizando apenas métodos de interpolação espacial, sem aplicar nenhuma técnica de clusterização ou análise de densidade, para determinar onde a demanda é maior.

V

50

Uma empresa de *e-commerce* deseja identificar atividades fraudulentas em transações financeiras analisando padrões incomuns nos dados. Para isso, a equipe de ciência de dados decide aplicar técnicas de detecção de anomalias. Qual alternativa apresenta a abordagem mais eficaz para detectar essas anomalias em um grande volume de dados transacionais?

(A) Utilizar um modelo de regressão linear para prever valores esperados de transações e classificar como anômalo qualquer valor acima da média.

(B) Aplicar técnicas baseadas em aprendizado não supervisionado, como o algoritmo Isolation Forest, que separa anomalias baseando-se no número de divisões necessárias para isolá-las.

(C) Usar exclusivamente um banco de dados SQL tradicional e criar regras manuais para definir limites fixos de valores anômalos.

(D) Implementar um modelo de aprendizado supervisionado que detecta anomalias apenas considerando transações já rotuladas como fraudulentas, sem necessidade de análise de novos padrões.

(E) Aplicar um modelo de classificação tradicional, como Naïve Bayes, pois esse tipo de algoritmo é ideal para detectar fraudes em grandes bases de dados, mesmo sem considerar padrões temporais ou comportamentais.

51

No processo de modelagem de um banco de dados relacional, é importante seguir boas práticas para garantir integridade, eficiência e escalabilidade. Qual das alternativas, a seguir, representa uma prática correta ao projetar um banco de dados relacional?

(A) Criar colunas duplicadas em tabelas diferentes para facilitar a busca dos dados e reduzir o tempo de execução das consultas.
(B) Evitar a criação de relações entre tabelas, pois junções (JOINs) podem prejudicar a performance das consultas em bancos relacionais.
(C) Armazenar todos os dados em um único arquivo CSV, pois facilita a manipulação sem necessidade de um sistema gerenciador de banco de dados (SGBD).
(D) Utilizar chaves primárias (PK) para identificar unicamente cada registro em uma tabela, garantindo integridade e evitando registros duplicados.
(E) Utilizar sempre nomes genéricos para tabelas e colunas, como "Tabela1" e "DadoX", pois simplifica a manutenção do banco de dados.

v

52

Uma empresa do setor varejista deseja melhorar sua tomada de decisão utilizando técnicas de mineração de dados. A equipe de ciência de dados está avaliando análises descritivas e preditivas para diferentes necessidades. Assinale a alternativa que caracteriza, corretamente, esses dois tipos de análise.

(A) A análise descritiva busca prever eventos futuros com base em padrões históricos, enquanto a análise preditiva apenas resume os dados sem fazer inferências.
(B) A análise preditiva é usada para compreender padrões históricos e tendências passadas, enquanto a análise descritiva emprega algoritmos de aprendizado de máquina para prever eventos futuros.
(C) A análise descritiva fornece um resumo dos dados históricos para identificar padrões e tendências, enquanto a análise preditiva utiliza esses padrões para prever comportamentos e eventos futuros.
(D) A análise preditiva é baseada apenas em estatísticas descritivas e não emprega técnicas como aprendizado de máquina ou modelagem estatística.
(E) A análise descritiva e a análise preditiva são termos intercambiáveis, pois ambas se limitam a descrever dados sem gerar *insights* para o futuro.

v

53

Uma empresa de *e-commerce* processa 10 milhões de transações diárias e deseja identificar compras anômalas que possam indicar fraude. Para lidar com esse grande volume de dados, a equipe de ciência de dados decide utilizar o Apache Spark para processar os dados de forma distribuída.
A equipe analisou um subconjunto de transações (em dólares):

```
[35, 42, 38, 40, 1500, 37, 39, 41, 36, 2500, 43, 5000, 38, 44, 3700]
```

Utilizando o Spark SQL, calcularam a média e o desvio padrão amostral das compras. Um valor é considerado anomalia se estiver acima de 2 desvios padrão da média.

Em relação à situação proposta e à análise, assinale a alternativa que apresenta as transações que podem ser classificadas como anômalas.

(A) `[1500, 2500, 5000]`
(B) `[5000]`
(C) `[2500, 5000, 3700]`
(D) `[1500, 2500, 3700, 5000]`
(E) `[1500, 2500, 3700]`

v

54

Em aprendizado de máquina, *underfitting* (subajuste) e *overfitting* (sobreajuste) são problemas que afetam o desempenho dos modelos. Considerando as definições apresentadas, assinale a alternativa que descreve a diferença entre esses dois problemas.

(A) *Underfitting* ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, enquanto *overfitting* ocorre quando o modelo não aprende o suficiente e generaliza bem para novos dados.
(B) *Overfitting* acontece quando o modelo é muito simples e não consegue capturar padrões nos dados, enquanto *underfitting* ocorre quando o modelo é muito complexo e memoriza os dados de treinamento.
(C) *Underfitting* ocorre quando o modelo é muito simples e não consegue capturar padrões nos dados, enquanto *overfitting* ocorre quando o modelo memoriza os dados de treinamento e tem baixo desempenho em novos dados.
(D) *Underfitting* e *overfitting* são problemas opostos, mas ambos ocorrem apenas quando os dados de treinamento contêm ruídos ou inconsistências.
(E) *Underfitting* e *overfitting* são sinônimos e indicam que um modelo está generalizando mal os dados de teste, independentemente da complexidade do modelo.

v

55

No ecossistema Python, diversas bibliotecas são amplamente utilizadas para diferentes tarefas em ciência de dados, aprendizado de máquina e processamento de linguagem natural (PLN). Considerando as características e aplicações dessas bibliotecas, assinale a alternativa que descreve a funcionalidade principal de uma delas?

(A) O spaCy é uma biblioteca especializada em Processamento de Linguagem Natural (PLN), oferecendo suporte para tokenização, lematização, reconhecimento de entidades nomeadas e modelos de linguagem pré-treinados.
(B) O TensorFlow é uma biblioteca voltada para manipulação e análise de dados tabulares, oferecendo suporte nativo para consultas SQL e operações eficientes com DataFrames.
(C) O Apache Arrow é um *framework* voltado para a criação e otimização de redes neurais profundas, fornecendo camadas e funções de ativação para treinamento de modelos de aprendizado profundo.
(D) A Scikit-learn (Sklearn) é uma biblioteca desenvolvida para visualização avançada de dados, com suporte a gráficos interativos e construção de *dashboards* dinâmicos.
(E) O PyTorch é uma biblioteca especializada exclusivamente na manipulação de arquivos e na aceleração de operações de leitura e escrita, sem aplicação para aprendizado de máquina ou redes neurais.

56

Uma plataforma de *e-commerce* deseja analisar automaticamente as avaliações deixadas pelos clientes nos produtos para determinar se são positivas ou negativas. Para isso, a equipe de ciência de dados está treinando um modelo de aprendizado de máquina para análise de sentimentos.
Dado que as avaliações são textos não estruturados, a equipe experimentou diferentes métodos de representação vetorial para transformar os textos em formatos que o modelo pode processar. Após testar diferentes abordagens, eles obtiveram os seguintes resultados em um modelo de classificação de sentimentos:

Representação Vetorial	Acurácia nos Dados de Treinamento	Acurácia nos Dados de Teste
Bag of Words (BoW)	95%	70%
TF-IDF	94%	73%
Word2Vec (CBOW)	90%	80%
BERT (Transformers)	89%	88%

Com base nos resultados apresentados, assinale a alternativa que descreve a melhor escolha de representação vetorial para este problema e sua justificativa.

(A) *Bag of Words* é a melhor escolha, pois obteve a maior acurácia nos dados de treinamento, garantindo que o modelo tenha aprendido melhor os padrões do conjunto de dados.
(B) TF-IDF é superior às outras técnicas, pois atribui pesos mais altos às palavras raras e, por isso, obteve um pequeno ganho de acurácia nos dados de teste em comparação ao BoW.
(C) Word2Vec (CBOW) é inferior ao BoW e ao TF-IDF, pois não captura bem as características estatísticas das palavras, o que resulta em modelos menos precisos para tarefas de classificação de sentimentos.
(D) *Bag of Words* e Word2Vec devem ser combinados para obter um modelo híbrido, pois BoW traz alta acurácia e Word2Vec melhora a generalização, compensando as fraquezas de cada abordagem.
(E) BERT (*Transformers*) é a melhor escolha, pois teve desempenho mais equilibrado entre os dados de treinamento e teste, indicando que o modelo não está sofrendo de sobreajuste e captura melhor o contexto do texto.

V

57

Em Python, a manipulação de arquivos é essencial para lidar com grandes volumes de dados de forma eficiente. Um Analista de Ciência de Dados precisa abrir, ler e processar um arquivo de texto contendo dados tabulares.
Com base nos conceitos corretos de manipulação de arquivos em Python, assinale a alternativa que apresenta a abordagem correta para manipular arquivos.

(A) Para abrir um arquivo para leitura, utiliza-se `open('arquivo.txt', 'r')`. Caso o arquivo não exista, ele será criado automaticamente para evitar erros.
(B) Para gravar dados em um arquivo sem sobrescrever seu conteúdo, deve-se abrir o arquivo no modo 'w', como em `open('arquivo.txt', 'w')`, pois esse modo adiciona novos dados ao final do arquivo sem apagar os já existentes.
(C) A função `readlines()` sempre lê e processa um arquivo de forma otimizada, independentemente do seu tamanho, pois armazena apenas uma linha por vez na memória, garantindo eficiência mesmo para arquivos extremamente grandes.
(D) O comando `with open('arquivo.txt', 'r') as f:` permite que o arquivo seja manipulado e garante seu fechamento automático após o uso, evitando vazamento de recursos. Além disso, essa abordagem permite a leitura de arquivos grandes sem carregar todo o conteúdo para a memória de uma só vez.
(E) Em Python, arquivos binários, como imagens e vídeos, não podem ser manipulados, pois a função `open()` suporta apenas a leitura de arquivos de texto.

V

58

Uma empresa do setor financeiro deseja prever a cotação diária de uma ação com base nos preços históricos. Para isso, a equipe de ciência de dados decide utilizar técnicas de modelagem de séries temporais.
A equipe analisou diferentes abordagens e encontraram os seguintes padrões nos dados:

- Os preços seguem uma tendência crescente ao longo do tempo;
- Há um padrão sazonal, com aumentos e quedas recorrentes em períodos específicos;
- Os valores atuais são fortemente influenciados pelos valores anteriores.

Com base nas características descritas, assinale a alternativa que apresenta a técnica de modelagem de séries temporais mais adequada para capturar esses padrões e gerar previsões precisas.

(A) Utilizar Regressão Linear Simples, pois ela assume que os preços das ações sempre seguem uma relação linear com o tempo, independentemente de tendências ou sazonalidades.
(B) Aplicar um modelo ARIMA (*AutoRegressive Integrated Moving Average*), pois ele é eficaz para qualquer tipo de série temporal, inclusive aquelas com sazonalidade complexa e não estacionárias, sem necessidade de ajustes adicionais.
(C) Utilizar K-Means Clustering, pois o agrupamento de dias com preços semelhantes permite prever diretamente os valores futuros da ação.
(D) Aplicar um Perceptron de Camada Única, pois redes neurais simples são suficientes para prever séries temporais sem a necessidade de considerar padrões sazonais ou de tendência.
(E) Utilizar Redes Neurais Recorrentes (RNN) ou LSTMs (*Long Short-Term Memory*), pois esses modelos são especializados em capturar dependências temporais longas e padrões sazonais complexos em séries temporais.

59

A tabela a seguir contém informações sobre pedidos de clientes em uma loja. No entanto, essa tabela apresenta redundâncias e dependências parciais, indicando que não está normalizada.

Tabela Pedidos (forma não normalizada):

Pedido_ID	Cliente_Nome	Cliente_Endereço	Produto_Nome	Quantidade	Preço_Unitário
1	João Silva	Rua A, 123	Produto A	2	50,00
1	João Silva	Rua A, 123	Produto B	1	30,00
2	Maria Souza	Rua B, 456	Produto A	3	50,00
3	Carlos Lima	Rua C, 789	Produto C	1	20,00
3	Carlos Lima	Rua C, 789	Produto B	2	30,00

Considerando apenas as regras da 1ª e 2ª formas normais (1FN e 2FN), assinale a alternativa que apresenta a correta normalização da tabela apresentada.

(A)

Tabela Clientes

Cliente_ID	Cliente_Nome	Cliente_Endereço
1	João Silva	Rua A, 123
2	Maria Souza	Rua B, 456
3	Carlos Lima	Rua C, 789

Tabela Produtos

Produto_ID	Produto_Nome	Preço_Unitário
1	Produto A	50,00
2	Produto B	30,00
3	Produto C	20,00

Tabela Pedidos

Pedido_ID	Cliente_ID
1	1
2	2
3	3

Tabela Itens_Pedido

Pedido_ID	Produto_ID	Quantidade
1	1	2
1	2	1
2	1	3
3	2	2
3	3	1

(B)

Tabela Clientes

Cliente_ID	Cliente_Nome	Cliente_Endereço
1	João Silva	Rua A, 123
2	Maria Souza	Rua B, 456
3	Carlos Lima	Rua C, 789

Tabela Pedidos

Pedido_ID	Cliente_Nome	Cliente_Endereço	Produto_Nome	Quantidade	Preço_Unitário
1	João Silva	Rua A, 123	Produto A	2	50,00
1	João Silva	Rua A, 123	Produto B	1	30,00
2	Maria Souza	Rua B, 456	Produto A	3	50,00
3	Carlos Lima	Rua C, 789	Produto C	1	20,00
3	Carlos Lima	Rua C, 789	Produto B	2	30,00

(C)

Tabela Pedidos

Pedido_ID	Cliente_Nome	Cliente_Endereço	Produto_Nome	Quantidade
1	João Silva	Rua A, 123	Produto A	2
1	João Silva	Rua A, 123	Produto B	1
2	Maria Souza	Rua B, 456	Produto A	3
3	Carlos Lima	Rua C, 789	Produto C	1
3	Carlos Lima	Rua C, 789	Produto B	2

(D)

Tabela Clientes

Cliente_ID	Cliente_Nome	Cliente_Endereço
1	João Silva	Rua A, 123
2	Maria Souza	Rua B, 456
3	Carlos Lima	Rua C, 789

Tabela Pedidos

Pedido_ID	Cliente_ID	Produto_Nome	Quantidade
1	1	Produto A	2
1	1	Produto B	1
2	2	Produto A	3
3	3	Produto C	1
3	3	Produto B	2

(E)

Tabela Clientes

Cliente_ID	Cliente_Nome	Cliente_Endereço	Produto_Nome	Preço_Unitário
1	João Silva	Rua A, 123	Produto A	50,00
2	Maria Souza	Rua B, 456	Produto B	30,00
3	Carlos Lima	Rua C, 789	Produto A	50,00

Tabela Pedidos

Pedido_ID	Cliente_Nome	Produto_Nome	Quantidade
1	João Silva	Produto A	2
2	Maria Souza	Produto B	1
3	Carlos Lima	Produto A	3

V

60

As tabelas, a seguir, foram criadas no banco de dados relacional para armazenar informações sobre vendas:

```
CREATE TABLE Clientes (
    Cliente_ID INT PRIMARY KEY,
    Nome VARCHAR(100),
    Cidade VARCHAR(100),
    Estado CHAR(2)
);

CREATE TABLE Pedidos (
    Pedido_ID INT PRIMARY KEY,
    Cliente_ID INT,
    Data_Pedido DATE,
    Valor_Total DECIMAL(10,2),
    FOREIGN KEY (Cliente_ID) REFERENCES Clientes(Cliente_ID)
);
```

A consulta SQL, a seguir, retorna quais resultados?

```
SELECT c.Nome, COUNT(p.Pedido_ID) AS Num_Pedidos, COALESCE(SUM(p.Valor_Total), 0) AS Total_Gasto
FROM Clientes c
LEFT JOIN Pedidos p ON c.Cliente_ID = p.Cliente_ID
WHERE p.Data_Pedido >= '2025-02-01'
GROUP BY c.Nome
HAVING COUNT(p.Pedido_ID) >= 1;
```

(A) Retorna todos os clientes, independentemente de terem feito pedidos, mas exibe NULL no campo Total_Gasto para aqueles que não realizaram compras.
(B) Retorna apenas os clientes que fizeram pedidos antes de 01/02/2025, pois a condição no WHERE exclui pedidos posteriores.
(C) Retorna os clientes que fizeram ao menos um pedido a partir de 01/02/2025, incluindo seu número de pedidos e o total gasto.
(D) Retorna os clientes que fizeram pedidos a partir de 01/02/2025, mas pode incluir clientes sem pedidos devido ao LEFT JOIN.
(E) Retorna erro, pois a cláusula HAVING não pode ser usada com COUNT() dessa maneira.

Questão dissertativa

Uma empresa de monitoramento industrial coleta, diariamente, dados de sensores de temperatura de suas máquinas. O objetivo é identificar medições anômalas que possam indicar falhas nos equipamentos e provocar acidentes. O trecho de código inicial, a seguir, simula a criação de um conjunto de medições realizadas nos últimos 100 dias para uma determinada máquina monitorada. Repare que o conteúdo foi gerado a partir de uma distribuição normal, com média de valores igual a 70 e desvio padrão igual a 5. Note ainda, que foram inseridas temperaturas anômalas, na linha 6.

```
1     import pandas as pd
2     import numpy as np

3     # Gerando dados fictícios
4     np.random.seed(42)
5     dados = np.random.normal(loc=70, scale=5, size=100)
6     dados[::10] = dados[::10] + np.random.randint(10, 20, size=10)

7     # Criando o DataFrame
8     df = pd.DataFrame({"temperatura": dados})
9     print(df)
```

Dê continuidade, em Python, ao código apresentado, implementando as 3 solicitações a seguir:

1. Calcule a média e o desvio padrão da amostra e exiba esses valores de forma clara na saída do programa.
2. Identifique as anomalias da amostra. São consideradas temperaturas anômalas, aquelas que estiverem acima ou abaixo de 2 desvios padrão da média. Crie uma nova coluna chamada “Anomalia”, que deverá conter *True* ou *False*, em função da identificação da anomalia.
3. Utilize a biblioteca Pandas para exibir uma tabela com as temperaturas e suas respectivas anomalias. Indique as anomalias diretamente no DataFrame, utilizando uma formatação condicional para destacar as colunas com valores anômalos (*True*) em vermelho. Utilize Styler.map() e suas variações, para exibir a formatação solicitada.

Exemplo de saída:

```
Média:  71.01076741302954
Desvio Padrão:  6.849046233225284
```

	temperatura	anomalía
0	90.483571	True
1	69.308678	False
2	73.238443	False
3	77.615149	False
4	68.829233	False
5	68.829315	False

Instruções:

- As respostas deverão ser redigidas de acordo com a norma padrão da língua portuguesa.
- Escreva com letra legível e não ultrapasse o espaço de linhas disponíveis da folha de respostas.
- Receberão nota zero textos que desrespeitarem os direitos humanos e textos que permitirem, por qualquer modo, a identificação do candidato(a).

RASCUNHO

NÃO SERÁ CONSIDERADO NA CORREÇÃO

RASCUNHO

NÃO SERÁ CONSIDERADO NA CORREÇÃO

V3

Universidade de São Paulo

Fundação universitária para o vestibular

GABARITO – CONCURSOS DRH USP – ABRIL/2025

Analista de Sistemas (especialidade: Ciência de Dados) – Edital RH Nº 002/2025

PROVA ASD			
1	D	31	A
2	B	32	E
3	B	33	B
4	A	34	D
5	E	35	D
6	E	36	C
7	D	37	E
8	C	38	E
9	C	39	B
10	E	40	D
11	B	41	C
12	C	42	B
13	D	43	B
14	B	44	A
15	E	45	B
16	B	46	D
17	C	47	E
18	D	48	C
19	D	49	A
20	A	50	B
21	E	51	D
22	C	52	C
23	E	53	B
24	B	54	C
25	C	55	A
26	A	56	E
27	C	57	D
28	D	58	E
29	B	59	A
30	C	60	C

Universidade de São Paulo

vencerás pela educação

ASD

RH nº 002/2025

Analista de Sistemas (Ciência de Dados)

QUESTÃO DISSERTATIVA

RESPOSTA ESPERADA

Deve-se apresentar muita familiaridade com as Bibliotecas Pandas e Numpy, que são amplamente utilizadas em aplicações de Ciências de Dados.

Código padrão fornecido:

```
import pandas as pd
import numpy as np

# Gerando dados fictícios
np.random.seed(42)
dados = np.random.normal(loc=70, scale=5, size=100)
dados[::10] = dados[::10] + np.random.randint(10, 20, size=10)

# Criando o DataFrame
df = pd.DataFrame({"temperatura": dados})
print(df)
```

Questão 1 - Calcule e exiba a média e o desvio padrão da amostra

```
# Cálcular e exibir a média e o desvio padrão da amostra
media = df['temperatura'].mean()
desvio_padrao = df['temperatura'].std()
print("Média: " , media)
print("Desvio Padrão: " , desvio_padrao)
```

Questão 2 - Identificar as anomalias da amostra. São consideradas temperaturas anômalas, aquelas que estiverem acima ou abaixo de 2 desvios padrão da média. Crie uma nova coluna chamada “Anomalia”, que deverá conter True ou False, em função da identificação da anomalia.

```
# Detecção de anomalias: Valores acima ou abaixo de 2 desvios padrão da média
limite_inferior = media - 2 * desvio_padrao
limite_superior = media + 2 * desvio_padrao

# Criando uma nova coluna "anomalia" para identificar os valores anômalos
df['anomalía'] = (df['temperatura'] < limite_inferior) | (df['temperatura'] > limite_superior)
```

Questão 3 - Utilize a biblioteca Pandas para exibir uma tabela com as temperaturas e suas respectivas anomalias. Indique as anomalias diretamente no DataFrame, utilizando uma formatação condicional para destacar as colunas com valores anômalos (True) em vermelho. Você deverá usar Styler.map() e suas variações, para exibir a formatação solicitada.

```
# Definindo a função de estilo para colorir anomalias
def colorir_anomalia(val):
    if val:
        return 'color: red'  # Colorir as anomalias de vermelho
    else:
        return 'color: black'  # Manter as normais com a cor preta

# Aplicando o estilo com Styler.map
styled_df = df.style.map(colorir_anomalia, subset=['anomalía'])

# Exibir o DataFrame com a formatação condicional
styled_df
```

Uma empresa de monitoramento industrial coleta diariamente, dados de sensores de temperatura de suas máquinas. O objetivo é identificar medições anômalas que possam indicar falhas nos equipamentos e provocar acidentes. O trecho de código abaixo simula a criação de um conjunto de medições realizadas nos últimos 100 dias para uma determinada máquina monitorada. Repare que o conteúdo foi gerado a partir de uma distribuição normal, com média de valores igual a 70 e desvio padrão igual a 5. Note ainda, que foram inseridas temperaturas anômalas, na linha 6.

Código inicial fornecido (o candidato deve dar continuidade a ele em Python):

```
1   import pandas as pd
2   import numpy as np

3   # Gerando dados fictícios
4   np.random.seed(42)
5   dados = np.random.normal(loc=70, scale=5, size=100)
6   dados[::10] = dados[::10] + np.random.randint(10, 20, size=10)

7   # Criando o DataFrame
8   df = pd.DataFrame({"temperatura": dados})
9   print(df)
```

Solicitações

Dê continuidade ao código acima, implementando as 3 solicitações a seguir em Python:

1- Calcule a média e o desvio padrão da amostra e exiba esses valores de forma clara na saída do programa.

1) Exemplo de Saída Esperada no Console (Os valores podem variar por conta da semente aleatória)

```
Média: 71.01
Desvio Padrão: 6.85
```

2) Exemplo de Saída Esperada no Console

	temperatura	Anomalia
0	90.48	True
1	69.31	False
2	73.23	False
3	77.61	False
4	68.82	False

3)Exemplo de Saída Esperada no DataFrame Formatado

temperatura	Anomalia
90.48	**True**
69.31	**False**
73.23	**False**
77.61	**False**
68.82	**False**

CRITÉRIOS DE CORREÇÃO

• **Critério 1:** Completude e abrangência dos conceitos (0 a 3 pontos):

Faixa de nota	Critério
3	Os conceitos principais são abordados com profundidade e detalhamento.
2	A maioria dos conceitos principais é abordada, mas pode faltar algum detalhe ou profundidade.
1	Alguns conceitos principais são abordados, mas a explicação é superficial ou incompleta.
0	Pouco ou nenhum conceito relevante é abordado.

• **Critério 2:** Domínio e aprofundamento dos conceitos (0 a 3 pontos):

Faixa de nota	Critério
3	A resposta é precisa, com informações corretas e bem explicadas.
2	A resposta é em sua maioria precisa, mas pode conter alguns pequenos erros ou imprecisões.
1	A resposta contém várias imprecisões ou erros conceituais, mas a ideia geral é compreensível.
0	A resposta está incorreta e confusa.

• **Critério 3:** Aplicação prática / exemplificação dos conceitos (0 a 3 pontos):

Faixa de nota	Critério
3	A resposta faz uma excelente conexão entre os conceitos teóricos e suas aplicações práticas.
2	A resposta faz boas conexões entre teoria e prática, mas pode ser aprimorada com mais exemplos ou detalhes.
1	A conexão entre teoria e prática é mencionada, mas é superficial ou pouco clara.
0	A resposta não aborda a aplicação prática e não apresenta exemplos dos conceitos.

• **Critério 4:** Clareza e Coerência (0 a 1 ponto):

Faixa de nota	Critério
1	O texto é extremamente claro e coerente, apresentando uma explicação lógica e bem estruturada dos conceitos.
0,5	O texto é claro e coerente, com algumas pequenas falhas na estrutura ou na explicação.
0	O texto é compreensível, mas apresenta várias falhas na clareza ou na coerência que dificultam a compreensão total.