FIOCRUZ

Concurso Público Fiocruz 2023

Tecnologista em Saúde Pública

Prova Objetiva e Discursiva

TE49 - Ciência de dados em saúde

Prova Objetiva

01. É correto afirmar que os dois tipos de variáveis são quantitativas:

(A) nominal e ordinal.
(B) nominal e intervalo.
(C) intervalo e razão.
(D) intervalo e ordinal.
(E) razão e intervalo.

02. "Processos de mineração de dados são usualmente aplicados em conjuntos de dados coletados para outros propósitos, para uso futuro ou aplicações diversas. Por essa razão, aplicações de mineração de dados quase nunca podem se beneficiar de estratégias que endereçam a correção de erros na fonte dos dados." Entretanto, a maioria das estatísticas aplicadas em processos de mineração de dados depende da qualidade de dados. Como prevenir problemas na qualidade dos dados na sua geração não é uma opção, o processo de limpeza de dados inclui a seguinte tarefa:

(A) remover anomalias.
(B) agregar redundâncias.
(C) selecionar atributos.
(D) remover ruídos.
(E) amostrar instâncias.

03. Em relação à maldição da dimensionalidade, avalie se são verdadeiras (V) ou falsas (F) as afirmativas a seguir:

I. Refere-se ao fenômeno de que muitos tipos de análises de dados se tornam mais difíceis a medida que a dimensionalidade de dados diminui.
II. Para tarefas de classificação, significa que não há instâncias de dados suficientes para criar um modelo que atribua de forma confiável a classe real das instâncias.
III. Quando a dimensionalidade cresce, os dados se tornam cada vez menos esparsos no espaço.

As afirmativas I, II e III são respectivamente:

(A) V, F e F.
(B) F, V e F.
(C) V, V e F.
(D) F, V e V.
(E) V, V e V.

04. Sobre os bancos de dados relacionais, é possível afirmar que:

(A) a linguagem SQL permite a definição de restrições de integridade dos dados.
(B) caíram em desuso devido à necessidade de processar big data.
(C) são uma ótima opção para armazenar dados não estruturados como texto, áudio e vídeo.
(D) o MongoDB é o SGBD mais utilizado para gerenciar bancos de dados relacionais.
(E) uma desvantagem da linguagem SQL é que não é padronizada, o que dificulta a portabilidade.

05. Sobre a linguagem SQL, é INCORRETO afirmar que:

(A) os comandos CREATE, ALTER, DROP e TRUNCATE são usados para a definição da estrutura do banco de dados.
(B) os comandos INSERT, UPDATE, DELETE e SELECT são usados para modificar os dados.
(C) os comandos GRANT e REVOKE são usados para controlar as opções de acesso aos dados.
(D) os comandos COMMIT, ROLLBACK, SAVEPOINT e SET TRANSACTION são usados para gerenciar transações.
(E) o comando SELECT é usado para selecionar dados.

06. Sobre a o modelo entidade-relacionado, é INCORRETO afirmar que:

(A) modela objetos do mundo real como entidades e seus relacionamentos.
(B) entidades descrevem objetos ou instâncias do mundo real.
(C) entidades fracas não possuem atributos próprios para identificação.
(D) atributos chave são utilizados para identificar de forma única uma entidade.
(E) a cardinalidade é o número máximo de ocorrências de uma entidade associadas.

07. Sobre os relacionamentos dos bancos de dados relacionais, podemos afirmar que:

I. Não precisam ser entre duas entidades distintas, sendo possível a presença de um relacionamento entre apenas uma entidade.
II. Os relacionamentos podem ter atributos.
III. São representados por elipses no modelo entidade-relacionamento.
IV. A cardinalidade especifica o número mínimo e o máximo de instâncias que uma entidade pode participar.

(A) apenas I e II estão corretas.
(B) apenas I, II e III estão corretas.
(C) apenas I, II e IV estão corretas.
(D) apenas II, III e IV estão corretas.
(E) todas estão corretas.

08. Quando analisamos dados visualmente, buscamos encontrar e compreender as partes da informação e como elas se relacionam com outras. Por exemplo, em uma série temporal, visamos analisar como determinadas variáveis se relacionam com a variável tempo. Um análise de parte-todo ilustra como as partes se relacionam entre si e com o todo. Séries temporais e parte-todo são dois exemplos de relacionamentos quantitativos clássicos que podem ser visualizados através de técnicas de visualização.

A coluna I mostra os relacionamentos quantitativos e a coluna II as técnicas de visualização. Estabeleça a correta correspondência entre as colunas I e II.

Coluna I

1. Série temporal.
2. Parte-todo.

Coluna II

( ) gráfico de linhas.
( ) gráfico de pizza.
( ) treemap.
( ) gráfico de radar.
( ) gráfico de marimekko.

A sequência correta, de cima para baixo, é:

(A) 2, 1, 1, 2, 1.
(B) 2, 2, 1, 2, 1.
(C) 1, 1, 2, 1, 2.
(D) 1, 2, 2, 1, 1.
(E) 1, 2, 2, 1, 2.

09. Em relação ao nosso sistema de percepção visual e cognição, é INCORRETO afirmar que:

(A) para realizar análise visual de dados, precisamos mais que apenas exibir os dados usando um gráfico.
(B) ferramentas gráficas são instrumentos que potencializam o raciocínio sobre informação quantitativa.
(C) a maioria das análises envolvem entender as relações entre mais de duas variáveis ao mesmo tempo.
(D) a memória tem um papel essencial na análise de dados e nós seres humanos temos uma vasta memória de trabalho.
(E) visualização de dados é sobre como utilizar ferramentas externas a nossa mente para potencializar nossa cognição.

10. Observe as afirmativas a seguir, em relação a técnicas de interação analítica para análise visual de dados:

I. A ordenação e o agrupamento de elementos facilitam a análise dos dados.
II. O uso de escala logarítmica para representação de dados não é recomendada por distorcer os dados.
III. A filtragem dos dados potencializa o processo analítico por trazer foco aos dados de interesse.

Sobre as afirmativas acima, pode-se dizer que:

(A) apenas I está correta.
(B) apenas II está correta.
(C) apenas II e III estão corretas.
(D) apenas I e III estão corretas.
(E) todas estão corretas.

11. Imagine que você está responsável por organizar uma base de dados de vacinação de cidadãos de um determinado município. Tendo em vista que que existem potenciais problemas de duplicidade de registros, você decidiu realizar algumas tarefas de sumarização dos dados. A opção mais adequada para esse processo é:

(A) fazer uma sumarização por campo-chave, unificando os registros com o mesmo CPF. Desta forma, a base terá apenas uma entrada para cada cidadão.
(B) fazer uma sumarização por dois campos-chave, unificando os registros com o mesmo CPF e a mesma vacina. Você terá então vários registros para o mesmo cidadão, cada qual referente a uma vacina aplicada.
(C) fazer uma sumarização por vários campos-chave, unificando os registros com o mesmo CPF, a mesma vacina e a data da vacinação. Você terá então vários registros para o mesmo cidadão, cada qual referente a uma dose de uma vacina aplicada.
(D) fazer uma sumarização por vários campos-chave, unificando os registros com o mesmo CPF, a mesma vacina e a data da vacinação. Você terá então vários registros para o mesmo cidadão, cada qual referente a uma dose de uma vacina aplicada. Usar medidas de sub-totalização para tentar identificar entradas espúrias na base como vacinações registradas em duplicidade.
(E) fazer uma sumarização por campo-chave, unificando os registros com a mesma vacina. Desta forma, a base terá apenas uma entrada para cada vacina. Totalizar o número de cidadãos que receberam doses de cada vacina.

12. Em relação a coleções de valores aleatórios gerados a partir de distribuições de probabilidade:

I. Se selecionamos um valor, em seguida outro e outro formando uma lista, sua média é o valor esperado.
II. Variáveis independentes são aquelas que não dependem das outras variáveis ou seja não se influenciam.
III. Muitos algoritmos de aprendizado de máquina requerem variáveis independentes e identicamente distribuídas ou seja selecionadas da mesma distribuição.

De cima para baixo, a sequência correta é:

(A) V, F e F.
(B) F, V e F.
(C) V, V e F.
(D) F, V e V.
(E) V, V e V.

13. Em muitas situações precisamos trabalhar com dados muito volumosos. Imagine que se queira saber a média de altura de todas as pessoas vivas no mundo e não houvesse uma maneira factível de medir todas as pessoas (população). Usualmente, extraímos um conjunto de dados menor mas representativo e então analisamos este subconjunto (amostra). Medimos alguns milhares de pessoas e esperamos que essa medida possa ser próxima o bastante da medida que obteríamos se medíssemos todo mundo. Para que essa medida seja confiável, precisamos calcular o intervalo de confiança. Para isto, precisamos selecionar diversas amostras da população. Este tipo de técnica é chamada de:

(A) bootstrapping.
(B) seleção de atributo.
(C) covariância.
(D) correlação.
(E) estimativa.

14. Quando construímos sistemas para realizar predições e classificações, encontramos padrões nos dados e precisamos ter métricas para nos ajudar a aferir o desempenho desses sistemas. É correto afirmar que:

(A) a revocação é o percentual de instâncias negativas corretamente classificadas.
(B) a precisão é o percentual de instâncias positivas corretamente classificadas.
(C) a precisão é o percentual de instâncias positivas classificadas como positivas.
(D) a acurácia se aproxima de 1 quando os erros diminuem.
(E) o F1 é o percentual de instâncias corretamente classificadas.

15. Sobre os termos apresentados na Lei Geral de Proteção de Dados (LGPD), conforme apresentado na Coluna I. Estabeleça a correta correspondência com as definições da Coluna II.

Coluna I

1. Dado pessoal.
2. Dado pessoal sensível.
3. Dado anonimizado.
4. Banco de dados.

Coluna II

( ) informação relacionada a pessoa natural identificada ou identificável.
( ) conjunto estruturado de dados pessoais, estabelecido em um ou em vários locais, em suporte eletrônico ou físico.
( ) dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural.
( ) dado relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento.

A sequência correta, de cima para baixo, é:

(A) 1, 2, 3, 4.
(B) 1, 2, 4, 3.
(C) 1, 3, 2, 4.
(D) 1, 3, 4, 2.
(E) 1, 4, 2, 3.

16. Sobre a Lei Geral de Proteção de Dados (LGPD), é correto afirmar que:

(A) empresas, profissionais autônomos e pessoas físicas que utilizam dados pessoais devem se adequar à LGPD.
(B) a identificação direta ocorre quando associamos informações, que isoladamente não conseguem identificar um indivíduo, para descobrimos a identidade de uma pessoa.
(C) a LGPD entrou em vigor a partir de agosto de 2021.
(D) órgão de pesquisa é o órgão da administração pública responsável por zelar, implementar e fiscalizar o cumprimento desta Lei em todo o território nacional.
(E) as infrações da LGPD deverão ser aplicadas pela Autoridade Nacional de Proteção de Dados (ANPD).

17. Segundo a LGPD, são direitos dos titulares dos dados pessoais, EXCETO:

(A) correção de dados incompletos, inexatos ou desatualizados.
(B) anonimização, bloqueio ou eliminação de dados tratados em desconformidade com a LGPD.
(C) eliminação dos dados pessoais tratados com o consentimento do titular.
(D) auditoria do tratamento de dados.
(E) revisão de decisões automatizadas.

18. O CRISPR é uma técnica de edição de genes que traz promessas no diagnóstico e tratamento de várias doenças. Entretanto ela levanta uma série de desafios bioéticos. Com relação a isso, é INCORRETO afirmar que:

(A) a edição genética na indústria primária, agricultura e pecuária serão sempre benéficos para a humanidade.
(B) a edição de genes em seres humanos só pode ser cogitada em patologias onde não há tratamentos eficazes ou envolvem efeitos colaterais significativos, mas requer garantias de segurança bastante altas.
(C) é preciso ter cautela, pois qualquer alteração inesperada pode levar a problemas incontroláveis quando se trata da complexidade dos ecossistemas.
(D) a manipulação de espécies vegetais para torná-las resistentes a pragas é de grande benefício para a humanidade mas precisa ser analisada com cautela.
(E) a edição de genes em embriões não é nem científica nem eticamente justificada atualmente.

19. São princípios da ética biomédica, EXCETO:

(A) princípio da intimidade.
(B) princípio da não maleficência.
(C) princípio da beneficência.
(D) princípio da autonomia.
(E) princípio da justiça.

20. A bioética aborda temas delicados, alguns considerados tabus. Dentre os temas abaixo,o que NÃO é relacionado à bioética:

(A) relação cientista x cobaia.
(B) saúde mental.
(C) utilização de células-tronco.
(D) eutanásia.
(E) aborto.

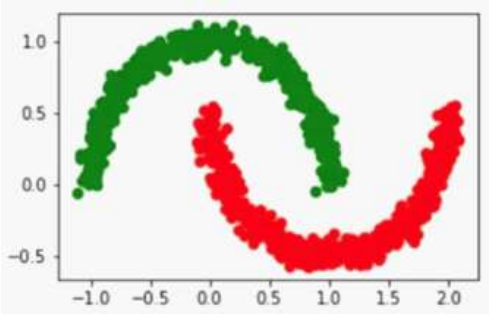
21. Sobre o algoritmo Apriori para mineração de regras de associação, é correto afirmar que:

(A) é eficiente na redução do espaço possíveis padrões, eliminando os pouco frequentes, reduzindo o volume de computação realizado.
(B) é adequado para conjuntos de dados muito grandes, pois a complexidade do algoritmo é exponencial.
(C) é eficaz na identificação de padrões em conjuntos de dados com itens raros ou transações pouco frequentes.
(D) é insensível ao suporte mínimo e aos limites mínimos de confiança.
(E) pode gerar um grande número de regras de associação, dificultando a análise dos resultados.

22. Quando realizamos o agrupamento de objetos de acordo com seus atributos, ou seja, uma tarefa de aprendizado não supervisionado, precisamos avaliar a qualidade deste agrupamento sem informação de supervisão. Como não se tem as classes ou rótulos das instâncias, é preciso avaliar a qualidade dos grupos apenas através de aferições estatísticas de similaridade intra e inter-grupos. Dentre estes índices, podemos citar, EXCETO:

(A) Davies-Boudin.
(B) Dunn.
(C) Coeficiente de silhueta.
(D) Calinski-Harabasz.
(E) Índice de Rand.

23. Considere o problema de calcular agrupamentos dos objetos apresentados na figura abaixo:



Considerando a distribuição dos objetos no espaço de acordo com seus atributos ilustrada na figura, o algoritmo de agrupamento indicado para diferenciar os dois grupos seria:

(A) k-médias.
(B) k-medóides.
(C) neighbor joining.
(D) DBSCAN.
(E) biclustering.

24. Observe as afirmativas a seguir, em relação a seleção de atributos para algoritmos de aprendizado de máquina:

I. Se temos atributos na base de dados que sejam redundantes, irrelevantes ou inúteis, devemos eliminá-los.
II. Podemos eliminar atributos que contribuem muito pouco na construção de um modelo como os que tem um mesmo valor na grande maioria das instâncias.
III. Os atributos removidos do conjunto de treinamento devem ser também removidos do conjunto de testes.

Sobre as afirmativas acima, pode-se dizer que:

(A) apenas I está correta.
(B) apenas I e II estão corretas.
(C) apenas I e III estão corretas.
(D) apenas II e III estão corretas.
(E) todas estão corretas.

25. São algoritmos de classificação, EXCETO:

(A) K-Vizinhos mais próximos.
(B) K-Médias.
(C) Árvores de decisão.
(D) Máquinas de vetores de suporte.
(E) Naive Bayes.

26. Sobre a função de ativação de redes neurais, é CORRETO afirmar que:

(A) Step, ReLU, Adam, Sigmoid, tanh e sigmoid são exemplos de funções de ativação.
(B) é recomendável aplicar funções diferentes a cada neurônio da camada.
(C) é sempre linear, ou seja, pode ser descrita em termos de adições e multiplicações.
(D) é uma função que transforma a saída de um neurônio em um novo valor.
(E) é recomendável usar a função sigmoid em camadas ocultas.

27. Os sistemas computacionais com ou sem o uso de técnicas de aprendizado são apresentados na Coluna I. Estabeleça a correta correspondência com as definições ou exemplos da Coluna II.

Coluna I

1. Sistema especialista.
2. Aprendizado supervisionado.
3. Aprendizado não supervisionado.
4. Aprendizado por reforço.

Coluna II

( ) o sistema recebe um conjunto de registros médicos eletrônicos de pacientes e os agrupa de acordo com as similaridades entre as características presentes nos registros.
( ) o sistema recebe um conjunto de imagens de lâminas referentes a exames de pacientes e assinala com base em sua experiência prévia a patologia associada ou se se trata de um paciente sadio. Um especialista humano então avalia a decisão do sistema retornando para o mesmo uma pontuação que mede o seu desempenho. O sistema evolui de acordo com essa pontuação.
( ) o sistema é programado com regras pré-definidas por alguém treinado no problema.
( ) o sistema recebe um conjunto de imagens de lâminas referentes a exames de pacientes com rótulos indicando a patologia associada ou um paciente sadio.

A sequência correta, de cima para baixo, é:

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

28. “Sua estrutura básica é organizada em camadas. Neurônios em cada camada podem se comunicar com os neurônios da camada anterior e da próxima. É o formato desta estrutura que resulta no nome aprendizado profundo.” (Andrew Glassner)

Segundo Glassner, o que caracteriza uma rede de aprendizado profundo são:

(A) os neurônios artificiais.
(B) as camadas de computação.
(C) as entradas da rede.
(D) as saídas da rede.
(E) os parâmetros da rede.

29. Todos podem cometer erros, inclusive os algoritmos de aprendizado de máquina. Existem técnicas que podem ser usadas para aumentar nossa confiança de que eles farão previsões confiáveis. A ideia é usar uma coleção de classificadores treinados em dados levemente diferentes e usar todos para avaliar cada instância de entrada. Cada um deles realiza a classificação e escolhemos a classe mais votada como resultado. É correto afirmar que contém apenas técnicas que podem ser usadas para aumentar a confiabilidade nas previsões segundo essa ideia:

(A) votação, sampling, random forests.
(B) votação, sampling, boosting.
(C) bagging, boosting e PCA.
(D) random forests, PCA e boosting.
(E) bagging, random forests e boosting.

30. Sobre o algoritmo KNN (K-Vizinhos mais próximos) tradicional, podemos afirmar que:

(A) k é um hiperparâmetro e, portanto, definido ao longo da realização do treinamento.
(B) o KNN é um algoritmo que tem o mesmo propósito que o K-médias e o K-medóides.
(C) é um algoritmo simples e o treinamento é extremamente rápido.
(D) ele cria K agrupamentos das instâncias de entrada.
(E) é um algoritmo que aloca pouca memória, pois não carrega as instâncias todas de uma vez.

31. Recentemente muito tem sido discutido em relação à interpretabilidade dos modelos de aprendizado de máquina. Eles têm sido comparados a caixas-pretas,pois, embora venham apresentando resultados impressionantes com sua acurácia, não se tem muitas vezes ideia do que acontece dentro deles. Em outras palavras, as previsões são úteis e precisas, mas não se sabe como elas foram feitas e quais atributos ou fatores podem ter maior influência nos resultados.
Trata-se de modelos complexos que absorvem relações não lineares e não triviais nos dados. É preciso que o analista de dados tenha uma visão crítica e entendimento dos algoritmos. Suponha que você tenha sido contratado para criar um sistema que utilize modelos de aprendizado de máquina para classificar pacientes segundo a propensão a apresentar uma determinada doença, mas um requisito essencial do sistema é que seja possível explicar claramente como se chegou a essa previsão. Dentre os seguintes algoritmos, é correto afirmar o que se utilizaria é:

(A) Máquinas de vetores de suporte.
(B) Análise de componentes principais.
(C) Floresta aleatória.
(D) Árvore de Decisão.
(E) Naive-Bayes.

32. Dentre as seguintes listas, NÃO contêm apenas algoritmos que podem ser usados para realizar uma regressão, é:

(A) Rede neural, Lasso, Árvore de Decisão.
(B) Random Forest, KNN, SVM.
(C) Rede neural, Ridge, K-médias.
(D) Lasso, KNN, Random Forest.
(E) Ridge, Lasso, Regressão linear.

33. Sobre o algoritmo K-médias, é correto afirmar que:

(A) é usualmente o primeiro algoritmo aplicado para agrupamento inicial de conjuntos de dados.
(B) é um algoritmo bastante simples de ser implementado mas não escala para conjunto de dados grandes.
(C) não existe garantia de convergência do algoritmo K-médias.
(D) consegue, de maneira efetiva, identificar clusters não convexos.
(E) é um algoritmo recomendado para detecção de outliers.

34. Uma das dificuldades de se realizar agrupamentos de dados é a definição do número de grupos. É correto afirmar que contém apenas técnicas ou métricas que podem ser úteis para automatizar a decisão do número K de grupos:

(A) homogeneidade interna e dendrograma.
(B) dendrograma e coeficiente de silhueta.
(C) heterogeneidade externa e método de Ward.
(D) método de Ward e método do cotovelo.
(E) método do cotovelo e coeficiente de silhueta.

35. São desafios do processo de agrupamento de dados, EXCETO:

(A) visualização do agrupamento resultante.
(B) sensibilidade do resultados às condições iniciais.
(C) escalabilidade para grandes conjuntos de dados.
(D) determinação do número ideal de clusters.
(E) avaliando o desempenho conhecimento dos rótulos das instâncias.

36. O ChatGPT é um modelo de linguagem desenvolvido pela OpenAI, baseado na arquitetura GPT (Generative Pre-trained Transformer). Sobre o Chat-GPT, é correto afirmar que:

(A) é uma abordagem de aprendizado de máquina que utiliza uma rede neural transformadora treinada sob demanda em uma grande quantidade de dados textuais para realizar tarefas diversas, como geração de texto, tradução, questionamento e resposta, entre outras.
(B) ele aprende a prever a próxima palavra em uma sequência, mas não tem uma compreensão profunda da estrutura e do conteúdo da linguagem.
(C) apesar de ele entender contextos, gerar respostas coesas e manter uma conversa significativa, ele não foi afinado especificamente para tarefas de conversação.
(D) é uma ferramenta poderosa, mas pode gerar respostas que não são necessariamente precisas ou contextualmente apropriadas, pois sua geração é baseada em padrões aprendidos nos dados de treinamento.
(E) lista uma bibliografia segura de onde as informações apresentadas foram retiradas.

37. Sobre os autoencoders, podemos dizer que:

(A) não podem ser utilizados em conjuntos de dados ruidosos.
(B) autoencoders variacionais são determinísticos.
(C) aprendem a representar um conjunto de entradas em variáveis latentes.
(D) há mais variáveis latentes que entradas.
(E) não há perda de informação na codificação.

38. Sobre as redes neurais convolucionais, é correto afirmar que:

(A) a rede VGG16 foi originalmente treinada no famoso conjunto de imagens chamado MNIST que contém 60.000 dígitos de 0 a 9 escritos a mão.
(B) quando usadas como classificadores de imagens são suscetíveis de serem enganadas ajustando os valores dos pixels em pequenas quantidades que são imperceptíveis para um observador humano.
(C) os filtros são operações que permitem que as convoluções sejam invariantes translacionalmente ou invariantes ao deslocamento.
(D) o processo de cercar as imagens com anéis de zeros para que o filtro possa ser deslizado sobre cada elemento de entrada é chamado de pooling.
(E) podemos reduzir as dimensões de um tensor usando downsampling através de pooling, mas é impossível aumentar as dimensões de um tensor.

39. Analise as afirmativas a seguir, em relação à mineração de padrões frequentes:

I. Seu objetivo é extrair conjuntos de itens frequentes de um banco de dados.
II. Um exemplo de padrão frequente são as regras de associação.
III. Dado um conjunto de itens X = {x1, x2,…,xm} e um conjunto de transações T = {t1, t2, …, tn}, um subconjunto de X, S, é chamado de conjunto de itens frequentes se S ocorre em uma porcentagem de todas as transações em T que excede um limite, denominado suporte.
IV. O suporte de um conjunto de itens Y, suporte(Y), é definido como o número de transações em T que contêm o conjunto de itens Y.

Das afirmativas acima, é correto afirmar que:

(A) apenas I está correta.
(B) apenas III está correta.
(C) apenas II e III estão corretas.
(D) apenas III e IV estão corretas.
(E) todas estão corretas.

40. Em relação ao processamento de linguagem natural, NÃO é correto afirmar que:

(A) os modelos baseados em n-gramas recuperam uma quantidade imensa de informação em um idioma e podem ter bom desempenho em identificação de idioma e correção ortográfica.
(B) é importante a seleção de características e o pré-processamento para eliminar anomalias.
(C) a classificação de texto pode ser feita com modelos de n-gramas com qualquer algoritmo de classificação tradicional.
(D) sistemas de recuperação de informação utilizam um modelo de linguagem simples baseado em saco de palavras e conseguem bons desempenhos em termos de cobertura e precisão com corpora muito grandes de texto.
(E) sistemas de extração de informação utilizam um modelo mais complexo que inclui noções limitadas de sintaxe e semântica e podem ser construídos a partir de autômatos de estado finito.

Prova Discursiva

QUESTÃO

Você foi contratado para projetar um sistema para descoberta de conhecimento em uma base de dados de vacinação de cidadãos de um município usando modelos de inteligência artificial. Tendo em vista que você receberá uma base de dados estruturada, mas passível de apresentar diversos tipos de problemas e inconsistências, descreva as etapas que você deverá considerar no desenvolvimento do sistema. Não é preciso mencionar softwares específicos a serem utilizados, mas sim as tarefas a serem implementadas.

Em texto com o mínimo 50 linhas e o máximo de 150, descreva as principais tarefas a serem realizadas na preparação dos dados, a escolha dos modelos de IA e avaliação e apresentação dos resultados finais.

RASCUNHO

RASCUNHO

RASCUNHO

RASCUNHO

RASCUNHO

INSTRUÇÕES

1. Por motivo de segurança, a Fiocruz solicita que o candidato transcreva em letra cursiva, em espaço próprio no Cartão de Respostas da Prova Objetiva, a frase abaixo apresentada:

"As melhores coisas da vida não podem ser vistas nem tocadas, mas sim sentidas pelo coração." (Dalai Lama)

2. Para cada uma das questões da prova objetiva são apresentadas 5 (cinco) alternativas classificadas com as letras (A), (B), (C), (D) e (E), e só uma responde da melhor forma possível ao quesito proposto. Você só deve assinalar UMA RESPOSTA. A marcação de nenhuma ou de mais de uma alternativa anula a questão, MESMO QUE UMA DAS RESPOSTAS SEJA A CORRETA.
3. A duração da prova é de 4 (quatro) horas, considerando, inclusive, a marcação do Cartão de Respostas e a Prova Discursiva. Faça-a com tranquilidade, mas controle o seu tempo.
4. Verifique se a prova é para o **PERFIL** para o qual concorre.
5. Somente após autorizado o início da prova, verifique se este Caderno de Questões está completo e em ordem. Folhear o Caderno de Questões antes do início da prova implica na eliminação do candidato.
6. Verifique, no **Cartão de Respostas da Prova Objetiva**, se seu nome, número de inscrição, identidade e data de nascimento estão corretos. Caso contrário, comunique ao fiscal de sala.
7. O **Caderno de Questões** poderá ser utilizado para anotações, mas somente as respostas assinaladas no **Cartão de Respostas da Prova Objetiva** e no **Caderno de Respostas da Prova Discursiva** serão objeto de correção.
8. Observe as seguintes recomendações relativas ao **Cartão de Respostas da Prova Objetiva**:
. não haverá substituição por erro do candidato;
. não deixar de assinar no campo próprio;
. não pode ser dobrado, amassado, rasurado, manchado ou conter qualquer registro fora dos locais destinados às respostas;
. a maneira correta de marcação das respostas é cobrir, fortemente, com esferográfica de tinta azul ou preta, o espaço correspondente à letra a ser assinalada;
. outras formas de marcação diferentes da que foi determinada acima implicarão a rejeição do **Cartão de Respostas**;
9. O fiscal não está autorizado a alterar quaisquer dessas instruções.
10. Você só poderá retirar-se da sala após 60 minutos do início da prova.
11. Quaisquer anotações só serão permitidas se feitas no caderno de questões.
12. Você poderá anotar suas respostas da prova objetiva em área específica do Caderno de Questões, destacá-la e levar consigo.
13. Os três últimos candidatos deverão permanecer na sala até que o último candidato entregue ao fiscal todo o seu material de prova.
14. Ao terminar a prova, entregue ao fiscal de sala, obrigatoriamente, o **Cartão de Respostas da Prova Objetiva, o Caderno de Respostas da Prova Discursiva e o Caderno de Questões**.

15. Prova Discursiva:

- A questão discursiva deverá ter um limite mínimo de 50 linhas e máximo de 150 linhas.

- Transcreva sua resposta para a parte pautada do **Caderno de Respostas da Prova Discursiva**. Não assine, rubrique ou coloque qualquer marca que o identifique, sob pena de ser anulado. Assim, a detecção de qualquer marca identificadora no espaço destinado à transcrição do texto definitivo acarretará nota ZERO na respectiva prova discursiva.

- O tempo total de duração das provas será de 4 (quatro) horas, incluindo o tempo para o preenchimento da Resposta Definitiva da Questão Discursiva. Nenhum rascunho SERÁ LEVADO EM CONTA.

Boa Prova!

Ao término da prova, anote aqui suas respostas e destaque na linha pontilhada.

01	09	17	25	33
02	10	18	26	34
03	11	19	27	35
04	12	20	28	36
05	13	21	29	37
06	14	22	30	38
07	15	23	31	39
08	16	24	32	40