MINISTÉRIO DO PLANEJAMENTO E ORÇAMENTO
SECRETARIA DE ORÇAMENTO FEDERAL
SUBSECRETARIA DE TECNOLOGIA E DESENVOLVIMENTO INSTITUCIONAL

CARGO 6: ANALISTA DE PLANEJAMENTO E ORÇAMENTO – ESPECIALIDADE: GESTÃO DE DADOS ORÇAMENTÁRIOS

Prova Discursiva P_4 – Dissertação

Aplicação: 07/07/2024

PADRÃO DE RESPOSTA DEFINITIVO

1. O(A) candidato(a) deve descrever cada uma das quatro principais etapas da análise exploratória de dados, conforme apresentado a seguir, podendo utilizar sinônimos para os termos utilizados.

- Coleta de dados: objetiva encontrar e carregar dados para análise.
- Inspeção ou análise dos dados: objetiva observar o conjunto de dados em um nível mais alto, determinando o tamanho do conjunto de dados, quantas linhas e colunas ele possui, além de realizar verificação básica dos dados para identificar erros aparentes, como valores faltantes, duplicados ou discrepantes (*outliers*); análise univariada — em que se analisam dados de apenas uma variável e análise bivariada — na qual se comparam dados de duas variáveis.
- Pré-processamento de dados: objetiva garantir que os dados estejam prontos para serem utilizados na sumarização estatística; e contempla várias etapas, como limpeza de dados, transformação, padronização, redução de dimensionalidade, normalização, entre outras.
- Sumarização estatística: objetiva realizar cálculos de estatísticas descritivas (média, mediana, desvio-padrão) para que se entenda a distribuição e variabilidade dos dados e se identifiquem relações, correlações, assimetrias, lacunas e tendências do conjunto de dados.
- Visualização de dados: objetiva utilizar gráficos e diagramas para explorar visualmente os dados e identificar padrões, tais como anomalias, distribuições e agrupamentos.

2. O(A) candidato(a) deve descrever, em seu texto, três das técnicas listadas a seguir, a serem adotadas no pré-processamento de dados, podendo utilizar sinônimos para os termos utilizados.

- Limpeza de dados: remoção de ~~valores~~ dados ausentes, duplicados, incompletos, irrelevantes e inconsistentes.
- Tratamento de dados ausentes: remoção de linhas ou colunas ausentes, preenchimento dos dados ausentes com valores estatísticos, imputação de dados ausentes por valores estimados.
- Deduplicação: identificação e remoção de dados duplicados ou redundantes em um conjunto de dados.
- Desidentificação de dados sensíveis: remoção ou ocultação de informações pessoais identificáveis de um conjunto de dados, com objetivo de proteger a privacidade e a segurança dos dados.
- Aprimoramento ou enriquecimento de dados: aprimorar ou enriquecer dados adicionais, como informações geográficas, demográficas ou de contexto, para melhorar a qualidade e a profundidade da análise.
- Tratamento de dados discrepantes (*outliers*): tratamento dos dados discrepantes.
- Normalização e padronização: técnicas para ajustar a escala dos dados.
- Transformação de dados: aplicação de transformações nos dados, como logaritmo, raiz quadrada ou outras transformações para tornar a distribuição dos dados mais normal.
- Redução de dimensionalidade: técnicas para reduzir o número de variáveis nos dados, como análise de componentes principais (PCA) ou seleção de características.
- Tratamento de dados categóricos: conversão de variáveis categóricas em uma forma numérica, como *one-hot encoding* ou codificação de rótulos.
- Discretização de dados: transformação de variáveis contínuas em variáveis discretas.
- Amostragem de dados: seleção de uma amostra representativa dos dados originais (especialmente útil quando se está lidando com grandes volumes de dados).
- Balanceamento de dados: técnicas para lidar com conjuntos de dados desbalanceados, em que uma classe é muito mais comum do que as outras, como sobreamostragem (*oversampling*) ou subamostragem (*undersampling*).
- Tratamento de dados temporais: manipulação de dados que tem uma componente temporal, como interpolação para preencher lacunas temporais ou agregação de dados em diferentes intervalos de tempo.

- Tratamento de dados textuais: pré-processamento específico para dados textuais, incluindo-se *tokenização*, remoção de pontuação, *stemming*/*lemmatization* e vetorização (transformação de texto em vetores numéricos), *lowercsing* (conversão de textos em letras minúsculas), remoção de *stopwords*, *parts of speech tagging*.

3. O(A) candidato(a) deve mencionar, em seu texto, três dos gráficos descritos a seguir, utilizados na visualização de dados, a fim de facilitar a compreensão dos padrões e das tendências nos dados, podendo utilizar sinônimos para os termos utilizados.

- Histograma: utilizado para mostrar a distribuição de uma variável numérica, permitindo a identificação de padrões de frequência e *outliers*.
- Gráfico de dispersão (*scatter plot*): utilizado para visualizar a relação entre duas variáveis numéricas, sendo útil para identificar correlações e tendências.
- Gráfico de caixas (*boxplot*): utilizado para mostrar a distribuição estatística de uma variável, incluindo-se mediana, quartis e *outliers*; útil para identificar a dispersão e a presença de *outliers* nos dados.
- Gráfico de barras: ideal para comparar categorias ou grupos de dados; podem ser horizontais ou verticais.
- Gráfico de linhas: mostra a mudança no valor de uma variável ao longo do tempo ou de outra variável contínua, sendo útil para identificar tendências.
- Gráfico de *pizza*: mostra a distribuição proporcional de uma variável categórica como partes de um todo, sendo útil para mostrar a composição de um conjunto de dados.
- Diagrama de setores (radial): uma variação do gráfico de *pizza* em que os setores são dispostos radialmente em vez de circularmente.
- Gráfico de área: mostra a mudança na proporção de uma ou mais variáveis ao longo do tempo, sendo útil para visualizar tendências e comparações ao longo do tempo.
- Gráfico de bolhas (*bubble plot*): uma variação do diagrama de dispersão em que o tamanho dos pontos é usado para representar uma terceira dimensão de dados.
- Mapas de calor (*heatmaps*): representam dados em uma matriz de cores, em que cores mais intensas indicam valores maiores; é útil para visualizar padrões em grandes conjuntos de dados.

4. O(A) candidato(a) deve mencionar as sete etapas do ciclo de vida para desenvolvimento de uma solução de inteligência artificial (IA), descritas a seguir, podendo utilizar sinônimos para os termos utilizados.

- Definição do problema: clarificação dos objetivos do projeto e identificação das questões que a solução de IA deverá resolver.
- Coleta e preparação de dados: reunião de dados relevantes e execução das técnicas de pré-processamento para garantir dados de qualidade.
- Análise exploratória de dados: investigação inicial dos dados para entender suas características e identificar padrões preliminares.
- Desenvolvimento do modelo: seleção e treinamento de algoritmos de IA, utilizando parte dos dados para construir modelos preditivos ou de classificação.
- Avaliação do modelo: validação do modelo com um conjunto separado de dados para medir sua eficácia, a partir de métricas como precisão, revocação (*recall)* e F1-*score*.
- Implantação: integração do modelo em um ambiente de produção, onde ele pode ser utilizado para fazer previsões ou identificar fraudes em tempo real.
- Monitoramento e manutenção: avaliação contínua do desempenho do modelo, além de ajustes, conforme o necessário, para garantir sua eficácia.

5. O(A) candidato(a) deverá descrever duas das seguintes abordagens e(ou) algoritmos de IA, que podem ser utilizados para detecção de anomalias e fraudes em dados, podendo utilizar sinônimos para os termos utilizados.

- Detecção de anomalias por agrupamento (*clustering*): algoritmos de agrupamento, como K-*means*, podem ser usados para identificar *clusters* normais nos dados; em seguida, pontos que não se encaixam nesses *clusters* são considerados anomalias.
- Detecção de anomalias estatísticas: essa abordagem identifica anomalias com base em desvios estatísticos dos padrões normais nos dados. Algoritmos como Z-*score*, desvio-padrão, e métodos baseados em quartis, como gráfico *boxplot*, são comuns nessa categoria.
- Detecção de anomalias por classificação supervisionada: algoritmos de classificação, como árvores de decisão, SVM (*support vector machines*) e redes neurais, podem ser treinados com dados rotulados para identificar anomalias como classes separadas.
- *Isolation forest*: este é um algoritmo de detecção de anomalias baseado em árvores de decisão, que funciona pela divisão repetida dos dados em subconjuntos aleatórios até que as anomalias acabem isoladas em folhas individuais da árvore.
- LOF *(local outlier factor*): este algoritmo calcula a "localização" de cada ponto de dados em relação aos seus vizinhos; pontos que têm uma densidade de vizinhos significativamente menor que os seus arredores são considerados anomalias.
- Redes neurais *autoencoder*: os *autoencoders* são redes neurais que tentam reconstruir a entrada original a partir de uma representação comprimida (codificação); pontos de dados que têm uma alta diferença entre a entrada original e a reconstrução podem ser considerados anomalias.
- *One-class* SVM (*one-class support vector machines*): algoritmo de aprendizado de máquina usado para aprender a fronteira de decisão em torno da maioria dos dados; pontos localizados fora dessa fronteira são considerados anomalias.

- DBSCAN (*density-based spatial clustering of applications with noise*): algoritmo de agrupamento baseado em densidade que é capaz de identificar *clusters* de alta densidade de pontos e separar ruídos (anomalias) como pontos que não estão dentro de nenhum *cluster* denso, e, portanto, são candidatos a serem considerados anomalias.
- *A priori*: algoritmo baseado em regras de associação, o qual busca a relação entre itens ou elementos. Ele utiliza uma abordagem de geração e teste para encontrar itens frequentes em transações, ou seja, emprega a busca por profundidade e gera um conjunto de itens candidatos de k elementos a partir de um conjunto de itens k-1 elementos, sendo eliminado dos padrões menos frequentes.
- *K-Means Clustering*: algoritmo de aprendizado não supervisionado amplamente utilizado para a análise de agrupamentos (clustering) em que o objetivo é dividir um conjunto de dados em *K clusters* distintos. A ideia básica é que pontos de dados que estão longe dos centroides dos *clusters* (ou seja, possuem alta distância) podem ser considerados anômalos. Esses pontos não se encaixam bem em nenhum dos *clusters* formados e, portanto, são candidatos a serem considerados anomalias.
- Regressão logística: não é tradicionalmente utilizado para detecção de anomalia, mas pode ser utilizado para modelar a probabilidade de um evento binário ocorrer. Para isso, deve-se treinar o modelo com um conjunto de dados rotulados com exemplos de comportamentos normais e anômalos, e, depois, fazer a *scoragem* do modelo treinado para prever a probabilidade de anomalias em novos dados.
- Regressão Linear: não é tradicionalmente utilizado para detecção de anomalia, mas pode ser utilizado ao identificar desvios significativos entre valores previstos pelo modelo e os valores reais, desde que o contexto aplicado apresente uma relação entre variáveis independentes e a variável dependente linear. A ideia é que, após ajustar um modelo de regressão linear aos dados, os resíduos (diferenças entre os valores observados e previstos) podem ser analisados para identificar anomalias.
- *k-Nearest Neighbors* (KNN): é usado para classificação ou regressão, em que um novo ponto de dados é classificado com base na maioria dos k vizinhos mais próximos. Para ser utilizado como detecção de anomalias, o princípio deve ser invertido, ao contrário de encontrar os vizinhos mais próximos, identifica-se os pontos que são mais distantes dos demais pontos no conjunto de dados.
- Redes Neurais Recorrentes (RNNs) com *Long Short-Term Memory* (LSTM): são amplamente utilizadas em tarefas de sequência e são capazes de capturar dependências temporais complexas nos dados. Embora não sejam comumente empregadas especificamente para detecção de anomalias, elas podem ser adaptadas, por exemplo, anomalias podem ser identificadas quando há desvios significativos das sequências temporais normais, resíduos são significativamente diferentes do esperado ou dados com padrões que não se encaixam nos padrões normais aprendidos.

QUESITOS AVALIADOS

QUESITO 2.1 Descrição de cada uma das quatro principais etapas da análise exploratória de dados
Conceito 0 – Não descreveu nenhuma das quatro principais etapas da análise exploratória de dados ou o fez de forma totalmente equivocada.
Conceito 1 – Descreveu corretamente apenas uma das quatro principais etapas solicitadas.
Conceito 2 – Descreveu corretamente duas das quatro principais etapas solicitadas.
Conceito 3 – Descreveu corretamente três das quatro principais etapas solicitadas.
Conceito 4 – Descreveu corretamente as quatro principais etapas solicitadas.

QUESITO 2.2 Descrição de três técnicas a serem adotadas no pré-processamento de dados
Conceito 0 – Não descreveu nenhuma técnica para o pré-processamento dos dados ou o fez de forma totalmente equivocada.
Conceito 1 – Descreveu corretamente apenas uma técnica para o pré-processamento dos dados.
Conceito 2 – Descreveu corretamente apenas duas técnicas para o pré-processamento dos dados.
Conceito 3 – Descreveu corretamente três técnicas para o pré-processamento dos dados.

QUESITO 2.3 Descrição de três tipos de gráficos utilizados na visualização de dados
Conceito 0 – Não descreveu nenhum gráfico utilizado na visualização dos dados ou o fez de forma totalmente equivocada.
Conceito 1 – Descreveu corretamente apenas um tipo de gráfico.
Conceito 2 – Descreveu corretamente apenas dois tipos de gráfico.
Conceito 3 – Descreveu corretamente três tipos de gráfico.

QUESITO 2.4 Descrição de cada uma das etapas do ciclo de vida para o desenvolvimento de uma solução de inteligência artificial
Conceito 0 – Não descreveu nenhuma etapa do ciclo de vida do desenvolvimento da solução de IA ou o fez de forma totalmente equivocada.
Conceito 1 – Descreveu corretamente apenas uma das sete etapas do ciclo de vida do desenvolvimento da solução de IA.
Conceito 2 – Descreveu corretamente apenas duas das sete etapas do ciclo de vida do desenvolvimento da solução de IA.
Conceito 3 – Descreveu corretamente apenas três das sete etapas do ciclo de vida do desenvolvimento da solução de IA.
Conceito 4 – Descreveu corretamente apenas quatro das sete etapas do ciclo de vida do desenvolvimento da solução de IA.
Conceito 5 – Descreveu corretamente apenas cinco das sete etapas do ciclo de vida do desenvolvimento da solução de IA.
Conceito 6 – Descreveu corretamente apenas seis das sete etapas do ciclo de vida do desenvolvimento da solução de IA.

Conceito 7 – Descreveu corretamente as sete etapas do ciclo de vida do desenvolvimento da solução de IA.

QUESITO 2.5 Menção a dois algoritmos e(ou) abordagens de inteligência artificial que podem ser utilizados para a detecção de anomalias e fraudes em dados e explicação de sua utilização

Conceito 0 – Não mencionou nenhum algoritmo e(ou) abordagem de IA na detecção de anomalias e fraudes em dados nem explicou sua utilização, ou o fez de forma totalmente equivocada.

Conceito 1 – Mencionou corretamente apenas um algoritmo/abordagem de IA e não explicou ou explicou incorretamente sua utilização na detecção de anomalias e fraudes em dados.

Conceito 2 – Mencionou corretamente apenas um algoritmo/abordagem de IA e explicou corretamente sua utilização na detecção de anomalias e fraudes em dados ou mencionou corretamente dois algoritmos/abordagens de IA, mas não explicou ou explicou incorretamente sua utilização na detecção de anomalias e fraudes em dados.

Conceito 3 – Mencionou corretamente dois algoritmos/abordagens de IA, mas explicou corretamente a utilização de apenas um deles na detecção de anomalias e fraudes em dados.

Conceito 4 – Mencionou corretamente dois algoritmos/abordagens de IA e explicou corretamente a utilização de ambos na detecção de anomalias e fraudes em dados.

Cebraspe
Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos

MINISTÉRIO DO PLANEJAMENTO E ORÇAMENTO
SECRETARIA DE ORÇAMENTO FEDERAL
SUBSECRETARIA DE TECNOLOGIA E DESENVOLVIMENTO INSTITUCIONAL

CARGO 6: ANALISTA DE PLANEJAMENTO E ORÇAMENTO – ESPECIALIDADE: GESTÃO DE DADOS ORÇAMENTÁRIOS

Prova Discursiva P_4 – Questão

Aplicação: 07/07/2024

PADRÃO DE RESPOSTA DEFINITIVO

O modelo de fundação assemelha-se a uma grande enciclopédia digital que foi lida e compreendida por uma inteligência artificial. Esses modelos são treinados com uma quantidade enorme de dados, abrangendo uma variedade de tópicos muito grande. Eles aprendem padrões e informações gerais que podem ser aplicados a uma ampla gama de tarefas sem a necessidade de grandes ajustes. Isso os torna extremamente flexíveis e poderosos, capazes de entender e gerar linguagem, resolver problemas e até criar arte de modo semelhante ao trabalho humano.

RAG (*retrieval-augmented generation*) é como um assistente inteligente que, ao receber uma pergunta, rapidamente consulta uma biblioteca imensa para buscar a informação mais relevante antes de responder. Ele combina a capacidade de geração de respostas do modelo de fundação com um mecanismo de busca, trazendo informações precisas e atualizadas. Isso é particularmente útil em tarefas que exigem respostas detalhadas e baseadas em evidências, como responder perguntas complexas ou fornecer recomendações detalhadas.

O modelo de fundação customizado é uma adaptação mais personalizada desses grandes modelos enciclopédicos. Caso uma empresa tenha necessidades muito específicas, como entender jargões técnicos de uma área específica como engenharia, tecnologia da informação, medicina, direito, entre outras áreas, ou responder a perguntas sobre leis de patentes, um modelo de fundação customizado pode ser ajustado para se especializar nesses tópicos, proporcionando resultados mais precisos e relevantes para a empresa. Ele é customizado para absorver e refletir o conhecimento e as necessidades específicas de seu treinamento, oferecendo uma ferramenta poderosa e personalizada para tarefas específicas.

Esses modelos não são apenas ferramentas, sendo também considerados recursos de busca contínua para expansão das capacidades humanas por meio da tecnologia.

QUESITOS AVALIADOS

QUESITO 2.1
Conceito 0 – Não abordou o quesito ou o fez de forma totalmente equivocada.
Conceito 1 – Descreveu o modelo de forma apenas superficial, sem desenvolvimento.
Conceito 2 – Desenvolveu uma descrição do modelo de forma parcialmente correta ou insuficiente.
Conceito 3 – Desenvolveu uma descrição do modelo de forma correta e completa.

QUESITO 2.2
Conceito 0 – Não abordou o quesito ou o fez de forma totalmente equivocada.
Conceito 1 – Descreveu o modelo de forma apenas superficial, sem desenvolvimento.
Conceito 2 – Desenvolveu uma descrição do modelo de forma parcialmente correta ou insuficiente.
Conceito 3 – Desenvolveu uma descrição do modelo de forma correta e completa.

QUESITO 2.3
Conceito 0 – Não abordou o quesito ou o fez de forma totalmente equivocada.
Conceito 1 – Descreveu o modelo de forma apenas superficial, sem desenvolvimento.
Conceito 2 – Desenvolveu uma descrição do modelo de forma parcialmente correta ou insuficiente.
Conceito 3 – Desenvolveu uma descrição do modelo de forma correta e completa.