FIOCRUZ

Concurso Público Fiocruz 2023

Tecnologista em Saúde Pública

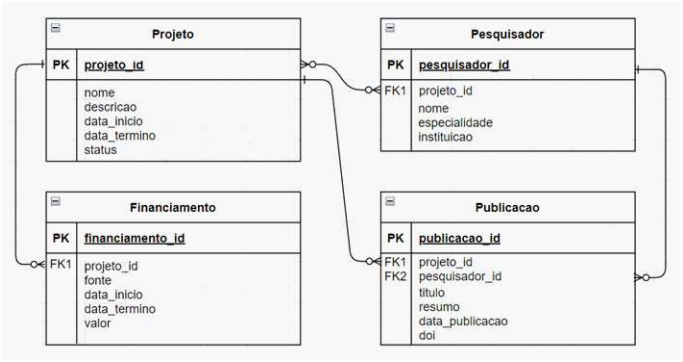
Prova Objetiva e Discursiva

TE56 - Cientista de Dados em Saúde

Prova Objetiva

01. Para a construção de um sistema de apoio à pesquisa e desenvolvimento na área de saúde, um modelo ER associado deve abranger entidades essenciais que facilitam a gestão de dados de pesquisa, desenvolvimento de estudos epidemiológicos e monitoramento de saúde pública. Este sistema poderia auxiliar na análise de tendências, na resposta a emergências de saúde pública e no desenvolvimento de políticas de saúde baseadas em evidências.

Seja o diagrama ER apresentado abaixo, desenhado na notação crow's foot, para um sistema de gestão de pesquisa.



Entre as opções abaixo, a que apresenta corretamente uma consulta SQL para retornar o valor total de financiamento de um projeto chamado "Inovação em Saúde" é:

(A) `SELECT SUM(valor) FROM Financiamento WHERE nome = 'Inovação em Saúde';`

(B) `SELECT valor FROM Financiamento WHERE fonte = 'Inovação em Saúde';`

(C) `SELECT SUM(valor) FROM Projeto WHERE nome = 'Inovação em Saúde';`

(D) `SELECT valor FROM Financiamento WHERE projeto_id = 'Inovação em Saúde';`

(E)
```
SELECT SUM(valor) FROM Financiamento WHERE
projeto_id = (SELECT projeto_id FROM Pro-
jeto WHERE nome = 'Inovação em Saúde');
```

02. Com base no diagrama ER apresentado na questão anterior, a consulta SQL que lista os nomes de todos os projetos que estão associados a menos de 4 pesquisadores e que têm um financiamento total maior que 20.000,00 é:

(A)
```
SELECT Proj.nome
FROM Projeto AS Proj
JOIN Financiamento AS Fin ON Proj.projeto_id
   = Fin.projeto_id
GROUP BY Proj.projeto_id
HAVING COUNT(DISTINCT Fin.financiamento_id)
   > 20000.00
AND COUNT(DISTINCT Pes.pesquisador_id) < 4;
```

(B)
```
SELECT Proj.nome
FROM Projeto AS Proj
JOIN Financiamento AS Fin ON Proj.projeto_id
   = Fin.projeto_id
WHERE SUM(Fin.valor) > 20000.00
GROUP BY Proj.nome
HAVING COUNT(DISTINCT Fin.projeto_id) < 4;
```

(C)
```
SELECT Proj.nome
FROM Projeto AS Proj
WHERE (SELECT COUNT(*)
    FROM Pesquisador AS Pes
    WHERE Pes.projeto_id = Proj.projeto_id)
   < 4)
 AND (SELECT SUM(valor)
    FROM Financiamento AS Fin
    WHERE Fin.projeto_id = Proj.projeto_id)
   > 20000.00);
```

(D)
```
SELECT Proj.nome
FROM Projeto AS Proj
JOIN Pesquisador AS Pes ON Proj.projeto_id
   = Pes.projeto_id
JOIN Financiamento AS Fin ON Proj.projeto_id
   = Fin.projeto_id
GROUP BY Proj.nome
HAVING COUNT(DISTINCT Pes.pesquisador_id) <
   4 AND SUM(Fin.valor) > 20000.00;
```

E)
```
SELECT Proj.nome
FROM Projeto AS Proj
LEFT JOIN Pesquisador AS Pes ON Proj.pro-
   jeto_id = Pes.projeto_id
LEFT JOIN Financiamento AS Fin ON Proj.pro-
   jeto_id = Fin.projeto_id
GROUP BY Proj.nome
HAVING COUNT(Pes.pesquisador_id) < 4 OR
   SUM(Fin.valor) <= 20000.00;
```

03. O campo da Ciência de Dados é dinâmico e está em constante evolução, com o desenvolvimento de tecnologias e ferramentas que tornam a análise de dados mais eficiente e acessível. Uma dessas ferramentas é a biblioteca Pandas para a linguagem de programação Python. Por ser uma biblioteca de análise de dados conhecida principalmente por suas estruturas de dados poderosas que facilitam a manipulação de dados, como dataframes, é amplamente utilizada em processos de ETL (*Extract*, *Transform and Load*) por engenheiros e cientistas de dados que necessitam pré-processar e transferir dados entre plataformas de dados, como, por exemplo, bancos de dados relacionais e *Data Lakes*.

Considere o seguinte código Python que implementa parte de um ETL sobre a tabela `Financiamento`.

```
import pandas as pd
from sqlalchemy import create_engine
from datetime import datetime

engine = create_engine("postgresql://
postgres:postgres@localhost:5432/bd_pes-
quisa")
query = "SELECT * FROM Financiamento"
df = pd.read_sql_query(con=engine.connect(),
sql=sql_text(query))
df['data_inicio'] = pd.to_datetime(df['data_
inicio']).dt.strftime('%d/%m/%Y')
df['data_fim'] = pd.to_datetime(df['data_
fim']).dt.strftime('%d/%m/%Y')
df.to_csv('financiamentos_transformados.csv',
index=False)
```

Observe as afirmativas a seguir sobre a execução do código.

I. O código se conecta a um banco de dados PostgreSQL usando a biblioteca SQLAlchemy e extrai todos os dados da tabela `Financiamento`.
II. As colunas `data_inicio` e `data_fim` são transformadas para o formato DD/MM/AAAA, mas esses dados não são atualizados no banco de dados.
III. O dataframe resultante da transformação é salvo em um arquivo CSV chamado `financiamentos_transformados`.csv na máquina local, incluindo o índice do datadrame como uma coluna adicional.

Sobre as afirmativas acima, pode-se dizer que:

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

04. Quando se trabalha com grandes conjuntos de dados no Pandas, a eficiente alocação de memória torna-se crucial para manter um bom desempenho e evitar o esgotamento dos recursos do sistema. Dado este desafio, analise as opções abaixo para otimizar o uso da memória ao manipular grandes volumes de dados com Pandas.

I. Empregar categorias para dados textuais repetitivos ao invés de *strings*.
II. Segmentar os dados em *chunks* menores durante a leitura de arquivos grandes, utilizando o parâmetro `chunksize` no `read_csv`.
III. Fazer uso intensivo de operações `inplace`.

Sobre as afirmativas acima, pode-se dizer que:

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

05. Além do Pandas, NumPy, que é um acrônimo para Numerical Python, é outra biblioteca fundamental para a computação em Python. Ela serve como um dos pilares do ecossistema de ciência de dados e análise numérica, oferecendo suporte para poderosas estruturas de dados de arrays e matrizes multidimensionais.
Seja o dataframe Pandas df carregado da tabela Financiamento e um extrato de seus dados mostrado abaixo.

financiamento_id	projeto_id	fonte	data_inicio	data_termino	valor
0	1	Finep	2023-02-01	2023-05-31	10000.0
1	1	Finep	2023-06-01	2023-06-30	4000.0
2	1	BNDES	2023-08-01	2023-11-30	7000.0
3	2	CNPQ	2023-02-01	2023-09-30	22000.0
4	3	BNDES	2023-02-01	2023-05-31	1000.0
5	3	CNPQ	2023-07-01	2023-12-31	10000.0
6	4	BNDES	2023-02-01	2023-05-31	10000.0
7	4	BNDES	2023-08-01	2023-09-30	5000.0

E seja o seguinte código NumPy, que transforma `df` em matriz e manipula suas linhas e colunas.

```
import numpy as np
matriz = df.values
subconjunto = matriz[matriz[:, 1] == 1, 4:6]
```

Das opções abaixo, a que apresenta corretamente o array extraído pela operação NumPy é:

A) [['2023-05-31', 10000.0], ['2023-06-30', 4000.0], ['2023-11-30', 7000.0]].
B) [['Finep', '2023-02-01'], ['Finep', '2023-06-01'], ['BNDES', '2023-08-01']].
C) [['1', 'Finep'], ['1', 'Finep'], ['1', 'BNDES']].
D) [['2023-02-01', '2023-05-31'], ['2023-06-01', '2023-06-30'], ['2023-08-01', '2023-11-30']].
E) [['10000.0'], ['4000.0'], ['7000.0']].

06. Além da linguagem Python, a linguagem R é uma poderosa ferramenta estatística e gráfica utilizada por cientistas de dados em todo o mundo. Originária do ambiente acadêmico e com forte apoio da comunidade de estatística, R rapidamente se consolidou como uma das linguagens de programação de escolha para análise de dados, pesquisa científica, e qualquer aplicação que exija manipulação intensiva de dados, análise estatística ou visualização gráfica.

Considere o sumário exibido abaixo, saída do comando summary(df) da linguagem R:

```
financiamento_id   projeto_id       fonte           data_inicio
Min.   :1.00     Min.   :1.000   Length:8           Length:8
1st Qu.:2.75     1st Qu.:1.000   Class :character   Class :character
Median :4.50     Median :2.500   Mode  :character   Mode  :character
Mean   :4.50     Mean   :2.375
3rd Qu.:6.25     3rd Qu.:3.250
Max.   :8.00     Max.   :4.000
data_termino          valor
Length:8           Min.   : 1000
Class :character   1st Qu.: 4750
Mode  :character   Median : 8500
                   Mean   : 8625
                   3rd Qu.:10000
                   Max.   :22000
```

Com base nesta informação, a opção que contém uma observação INCORRETA é:

(A) a distribuição da variável `financiamento_id` mostra uma amplitude total de valores que vai de 1 a 8, evidenciando a variação total nos identificadores de financiamento dentro do conjunto de dados.
(B) os indicadores de tendência central para `projeto_id`, com uma média de 2.375 e uma mediana de 2.500, refletem uma distribuição dos dados que tende a ser equilibrada, sem uma inclinação acentuada para valores mais altos ou mais baixos.
(C) as variáveis `fonte`, `data_inicio` e `data_termino` são categorizadas como dados categóricos nominais, dado que representam informações qualitativas sem uma ordem inerente, e são armazenadas como caracteres, indicando o tipo de dado textual.
(D) o terceiro quartil da variável `valor` é 10.000, o que indica que 75% dos valores de financiamento são iguais ou inferiores a 10.000, demonstrando a posição dos valores de financiamento no contexto de dispersão e distribuição de quartis.
(E) a proximidade entre a média e a mediana dos valores de financiamento sugere uma distribuição altamente assimétrica, com uma presença significativa de valores extremos que distorcem a média, como é o caso do valor 22.000.

07. Modelos de *Machine Learning* (ML) são parte fundamental do conhecimento no campo de um cientista de dados, objetivando a compreensão de padrões complexos e a tomada de decisão baseada em dados. Esses modelos permitem que cientistas de dados transformem grandes volumes de dados brutos em insights acionáveis, previsões e recomendações com precisão que frequentemente supera análises tradicionais.
Considerando a base de dados contendo projetos, pesquisadores, publicações e financiamentos, diversos modelos de aprendizado de máquina podem ser criados. Entre as opções abaixo, a que apresenta uma relação INCORRETA entre objetivo, tipo de aprendizado e tipo de algoritmo de aprendizado de máquina é:

(A) previsão de financiamento de projetos com o objetivo de calcular o valor de financiamento que um projeto pode receber, baseando-se em características do projeto, atributos dos pesquisadores envolvidos e dados históricos de financiamento de projetos similares; trata-se um aprendizado supervisionado com algoritmo de regressão, que pode ser implementado por uma regressão polinomial ou regressão com regularização.
(B) detecção de comunidades de pesquisa com o objetivo de identificar grupos dentro de um campo específico, com base na análise de coautoria e citações entre pesquisadores. Trata-se de um aprendizado não supervisionado com algoritmo de clusterização, que pode ser implementado por SVMs – *Support Vector Machines*.
(C) análise de tendências de pesquisa com o objetivo de identificar áreas emergentes de pesquisa e tendências ao longo do tempo com base em análise de tópicos em publicações. Trata-se de um aprendizado não supervisionado com algoritmo de modelagem de tópicos, como LDA – *Latent Dirichlet Allocation*.
(D) análise de sentimentos de publicações com o objetivo de avaliar revisões e comentários e identificar *feedbacks* predominantemente positivos ou negativos; trata-se de um aprendizado supervisionado, que pode ser implementado com Redes Neurais Recorrentes (RNN) e *Long Short Term Memory* (LSTM).
(E) classificação de projetos com o objetivo de categorizar projetos de acordo com critérios relevantes, como disciplina científica, tipo de financiamento, escopo, entre outros; trata-se de um aprendizado supervisionado, que pode ser implementado por árvores de decisão.

08. O scikit-learn é uma biblioteca de aprendizado de máquina para Python que fornece uma ampla variedade de classes e funções para análise de dados e modelagem de *Machine Learning*. Ele inclui algoritmos para classificação, regressão, clusterização, redução de dimensionalidade, seleção de modelos, pré-processamento de dados, entre outros.
Entre as opções abaixo, a que apresenta corretamente a combinação de classes e funções do scikit-learn usadas para implementar regressão do tipo polinomial e classificação com árvores de decisão é:

(A) para regressão polinomial: `linear_model.PolynomialRegression e preprocessing.LinearFeatures;` para árvores de decisão: `tree.DecisionTreeRegressor`.
(B) para regressão polinomial: `preprocessing.PolynomialFeatures` e `linear_model.LinearRegression;` para árvores de decisão: `tree.DecisionTreeClassifier`.
(C) para regressão polinomial: `preprocessing.PolynomialFeatures` e `linear_model.LinearRegression;` para árvores de decisão: `tree.DecisionTreeRegressor`.
(D) para regressão polinomial: `linear_model.PolynomialFeatures e preprocessing.LinearRegression;` para árvores de decisão: `tree.DecisionClassifier`.
(E) para regressão polinomial: `preprocessing.LinearFeatures` e `linear_model.PolynomialRegression;` para árvores de decisão: `tree.TreeDecisionClassifier`.

09. Considere a seguinte implementação de um modelo de regressão linear múltipla utilizando NumPy e scikit-learn, usado para prever o financiamento de projetos com base em características de projetos e pesquisadores. O código abaixo foi executado e algumas métricas de desempenho foram obtidas.

```
import numpy as np
from sklearn.model_selection import train_
test_split
from sklearn.linear_model import LinearRe-
gression
from sklearn.metrics import mean_squared_er-
ror, mean_absolute_error, r2_score

X = np.array([[1, 50], [2, 60], [3, 70], [4,
80], [5, 90], [1, 55], [2, 65], [3, 75], [4,
85], [5, 95]])
y = np.array([100000, 120000, 150000, 200000,
250000, 110000, 130000, 170000, 230000,
290000])
X_train, X_test, y_train, y_test = train_
test_split(X, y, test_size=0.2, random_
state=0)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
mae = mean_absolute_error(y_test, y_pred)

print(f"R-Quadrado: {r2}, MSE: {mse}, RMSE:
{rmse}, MAE: {mae}")
```

Após executar o código, foram obtidas as seguintes métricas de desempenho:

```
R-Quadrado: 0.9020746527777778, MSE:
156680555.5555556,
RMSE: 12517.210374342823, MAE:
10083.333333333343
```

Com base nessas informações, analise as observações abaixo.

I. O valor de R-Quadrado próximo de 1 indica que o modelo explica uma grande proporção da variância dos dados de financiamento. Isso sugere que o modelo tem um bom ajuste aos dados, sendo capaz de capturar uma grande parte da relação entre as variáveis independentes e a variável dependente.

II. Um valor de MSE de aproximadamente 156 milhões sugere que, em média, o quadrado dos erros das previsões do modelo em relação aos valores reais é significativo. Isso indica que o modelo tem um bom ajuste de acordo e não existem erros consideráveis nas previsões.

III. Um MAE de aproximadamente 10083 sugere que, em média, as previsões do modelo desviam cerca de 10083 unidades dos valores reais. Comparado ao RMSE, o MAE não dá um peso tão grande a erros maiores, o que sugere que o modelo pode ter um número relativamente consistente de pequenos a moderados erros de previsão.

IV. A diferença entre o RMSE e o MAE sugere que o modelo pode estar lidando com alguns outliers ou previsões particularmente imprecisas que afetam mais o RMSE, pois o RMSE penaliza mais erros maiores do que erros menores.

Sobre as afirmativas acima, pode-se dizer que:

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

10. As Redes Neurais Recorrentes (RNNs) são projetadas para processar dados sequenciais ou temporais, destacando-se pela sua capacidade de reter memória de entradas anteriores através de loops internos na sua arquitetura. Entre os algoritmos mais utilizados, destacam-se o *Long Short-Term Memory* (LSTM) e o *Gated Recurrent Unit* (GRU), ambos projetados para preservar informações ao longo do tempo e superar o desafio do desaparecimento do gradiente. Além disso, técnicas fundamentais como *softmax*, *backpropagation* e o processo *feedforward* são fundamentais para o treinamento e a eficácia das RNNs. Acerca dessas técnicas, a opção que apresenta uma observação INCORRETA é:

(A) a função *softmax* pode ser usada na camada de saída das RNNs para realizar tarefas de regressão, convertendo os *logits* em valores contínuos que representam diferentes magnitudes.
(B) durante o processo de *feedforward* em redes neurais, incluindo as RNNs, a informação é processada sequencialmente da camada de entrada até a camada de saída, utilizando funções de ativação para introduzir não-linearidade.
(C) o *backpropagation* é o método pelo qual o erro é propagado de volta pela rede para atualizar os pesos, utilizando o gradiente do erro em relação a cada peso para fazer ajustes que minimizem o erro total da rede.
(D) a função *softmax* na camada de saída de uma RNN é crucial para problemas de classificação, onde os *logits* são transformados em probabilidades que somam 1, facilitando a determinação da classe mais provável para a entrada dada.
(E) o processo de *feedforward* e *backpropagation* em RNNs inclui o cálculo de gradientes para cada etapa temporal, ajustando os pesos não apenas com base na saída atual, mas também considerando a influência de entradas anteriores.

11. O Processamento de Linguagem Natural (PLN) busca melhorar a capacidade das máquinas de entender e interagir com a linguagem humana de forma natural e semanticamente adequada. Ao longo dos anos, a evolução dos modelos de Machine Learning tem desempenhado um papel fundamental nesse processo, permitindo avanços significativos em tarefas como tradução automática, análise de sentimentos e assistentes virtuais. Esses modelos dependem de uma série de técnicas de pré-processamento para transformar texto bruto em formas que possam ser eficientemente analisadas e compreendidas. Numere a 2ª coluna pela primeira, considerando as técnicas e as respectivas definições.

COLUNA 1

(1) Tokenização,
(2) POS Tagging,
(3) Stemização,
(4) Lematização e
(5) Chunking.

COLUNA 2

( ) Técnica que transforma uma palavra para sua forma de dicionário, considerando o contexto, a classe gramatical e outras características linguísticas.
( ) Trata de dividir o texto em unidades menores, como palavras ou partes de palavras, transformando o texto bruto e preparando-o para ser manipulado por algoritmos de PLN.
( ) Refere-se a reduzir as palavras para suas formas radicais, facilitando a análise de padrões comuns em diferentes variações da mesma palavra.
( ) Técnica de atribuir a cada palavra em um texto a sua classe morfossintática, como substantivos, verbos, adjetivos, etc.
( ) Trata de dividir um texto em segmentos mais curtos, como conjuntos de palavras ou seções de um texto, que serão tratados separadamente em processos posteriores como, por exemplo, vetorização.

A sequência correta, de cima para baixo, é:

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

12. Observe o código Python abaixo, que utiliza a biblioteca NLTK para tarefas de Processamento de Linguagem Natural.

```
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize

texto = "Fundação Oswaldo Cruz (Fiocruz): Ciência e tecnologia em saúde para a população brasileira."
tokens = word_tokenize(texto)

contador = 0
resultado = 0
while contador < len(tokens):
  for letra in tokens[contador]:
    if letra.upper() in 'FIOCRUZ':
      resultado += 1
  contador += 1
```

O valor da variável resultado, ao final da execução do código, é:

(A) 32.
(B) 33.
(C) 34.
(D) 35.
(E) 36.

13. Entre as observações abaixo sobre a técnica de *Word Embeddings* e sua importância em modelos de Processamento de Linguagem Natural (PLN), a que está correta é:

(A) são algoritmos de criptografia que protegem palavras em um texto para garantir sua privacidade e suas relações semânticas, por isso são essenciais para a segurança e compreensão em aplicações de PLN.
(B) são listas de sinônimos para palavras utilizadas em PLN, permitindo que sistemas de computador compreendam a variedade de vocabulário na linguagem humana e suas relações semânticas.
(C) são representações vetoriais que capturam relações semânticas e sintáticas; são fundamentais para melhorar a precisão de modelos de PLN, ao permitir que computadores interpretem nuances de significado.
(D) são técnicas de compressão de texto que reduzem o tamanho dos dados de linguagem para armazenamento eficiente em bancos de dados de PLN, enquanto permitem compreender suas relações semânticas.
(E) são marcações automáticas de cada palavra em um texto com sua parte correspondente do discurso para análise semântica e sintática em PLN.

14. Você é um cientista de dados trabalhando em um projeto de pesquisa em saúde que envolve a análise de relatórios médicos utilizando técnicas de Processamento de Linguagem Natural (PLN). Parte do seu trabalho é explorar as relações semânticas entre diferentes condições de saúde utilizando um modelo pré-treinado de *word embeddings* em português, focado na área da saúde. Você decide investigar a relação entre diferentes doenças e tratamentos.

Seja o seguinte código Python, que utiliza a biblioteca *gensim* e um modelo hipotético de *word embeddings* denominado *modelo_saude.bin* especializado em termos médicos em português:

```
import numpy as np
from gensim.models import KeyedVectors

def calcular_similaridade(vetor_a, vetor_b):
  numerador = np.dot(vetor_a, vetor_b)
   denominador = np.linalg.norm(vetor_a) *
np.linalg.norm(vetor_b)
  similaridade = numerador / denominador
  return similaridade

model = KeyedVectors.load_word2vec_
format('modelo_saude.bin', binary=True)
vetor_diabetes = model['diabetes']
vetor_hipertensao = model['hipertensão']
vetor_insulina = model['insulina']

vetor_diabetes_ajustado = vetor_diabetes +
vetor_insulina
vetor_hipertensao_ajustado = vetor_hip-
ertensao + vetor_insulina

similaridade = calcular_similaridade(vetor_
diabetes_ajustado, vetor_hipertensao_ajusta-
do)
print(f"Similaridade: {similaridade}")
```

Utilizando o modelo hipotético model_saude.bin, o resultado mostrado pelo código foi de 0.7036085724830627. Baseado no cenário descrito, no código fornecido e no resultado mostrado, a opção que melhor descreve o que está sendo calculado e o significado do resultado é:

(A) a distância euclidiana entre os vetores de "diabetes" e "hipertensão", ambos ajustados pelo vetor de "insulina", sugere que, no espaço semântico do modelo utilizado, as condições de "diabetes" e "hipertensão", quando consideradas no contexto do tratamento com "insulina", possuem uma relação semântica relativamente forte.

(B) a similaridade por cosseno entre os vetores de "diabetes" e "hipertensão", ambos ajustados pelo vetor de "insulina", sugere que, no espaço semântico do modelo utilizado, as condições de "diabetes" e "hipertensão", quando consideradas no contexto do tratamento com "insulina", possuem uma relação semântica relativamente fraca.

(C) a distância euclidiana entre os vetores de "diabetes" e "hipertensão", ambos ajustados pelo vetor de "insulina", sugere que, no espaço semântico do modelo utilizado, as condições de "diabetes" e "hipertensão", quando consideradas no contexto do tratamento com "insulina", possuem uma relação semântica relativamente fraca.

(D) a similaridade por cosseno entre os vetores de "diabetes" e "hipertensão", ambos ajustados pelo vetor de "insulina", sugere que, no espaço semântico do modelo utilizado, as condições de "diabetes" e "hipertensão", quando consideradas no contexto do tratamento com "insulina", possuem uma relação semântica relativamente forte.

(E) a distância euclidiana entre os vetores de "diabetes" e "hipertensão", ambos ajustados pelo vetor de "insulina", sugere que, no espaço semântico do modelo utilizado, as condições de "diabetes" e "hipertensão", quando consideradas no contexto do tratamento com "insulina", possuem uma relação semântica relativamente neutra.

15. A evolução das tecnologias de Inteligência Artificial, especialmente no campo do Processamento de Linguagem Natural (PLN), tem sido marcada por inovações significativas que transformaram a maneira como as máquinas entendem e geram linguagem humana. Uma dessas inovações é a arquitetura de Transformers, introduzida pelo artigo *Attention is All You Need* em 2017, superando as limitações das abordagens anteriores baseadas em Redes Neurais Recorrentes (RNNs) e tornando-se a base fundamental para o surgimento dos *Large Language Models* (LLMs).

Sobre essa arquitetura, pode-se afirmar que:

(A) a arquitetura dos *Transformers* depende exclusivamente de camadas recorrentes para processar sequências de texto, o que melhora a eficiência computacional em comparação com as RNNs.
(B) os *Transformers* introduziram o conceito de atenção seletiva, permitindo que modelos focassem em partes relevantes do texto ao gerar respostas, algo que as RNNs não podem fazer.
(C) os *Transformers* utilizam mecanismos de atenção que permitem a modelagem de dependências de longo alcance sem a necessidade de processamento sequencial, superando as limitações das RNNs.
(D) a arquitetura dos *Transformers* substituiu as unidades de processamento baseadas em regras pelas redes neurais, o que não era possível com as RNNs.
(E) a arquitetura dos Transformers elimina a necessidade de *encoders* e *decoders*, diferenciando-se das RNNs que dependem dessa estrutura para o Processamento de Linguagem Natural.

16. O uso de *Large Language Models* (LLMs) na área da saúde, como GPT e BERT, oferece um vasto campo de possibilidades para inovação. Atualmente, é possível criar uma série de aplicações que fazem uso dessas LLMs, variando desde melhorias da qualidade e acessibilidade a conhecimentos até o apoio a novas pesquisas na área.
Entre as opções abaixo, aquela que apresenta uma iniciativa que NÃO pode ser baseada no uso de LLMs é:

(A) LLMs podem ser utilizadas para extrair informações críticas de registros médicos eletrônicos, notas de alta hospitalar e literatura científica, transformando dados não estruturados em insights valiosos para a pesquisa clínica e epidemiológica.
(B) LLMs podem ser utilizadas para processar informações textuais sobre genética e biomarcadores, incluindo sequências de DNA/RNA, realizar análises genéticas complexas e interpretar dados laboratoriais brutos.
(C) LLMs podem automatizar a criação de relatórios de pesquisa, sumários de políticas de saúde e comunicados à imprensa, facilitando a disseminação rápida de informações importantes para o público e a comunidade científica.
(D) LLMs podem analisar extensas bases de dados de literatura científica para identificar padrões, tendências e lacunas no conhecimento, gerando novas hipóteses para pesquisa em saúde pública.
(E) LLMs podem analisar dados de fontes abertas e redes sociais para detectar e monitorar surtos de doenças em tempo real, permitindo respostas rápidas a emergências de saúde pública.

17. Considerando o avanço recente dos modelos de Processamento de Linguagem Natural (PLN) e a necessidade crescente de processar e sumarizar grandes volumes de documentos de forma eficiente, você foi encarregado de desenvolver uma aplicação capaz de sumarizar automaticamente documentos clínicos, proporcionando aos profissionais de saúde acessos mais rápidos e precisos às informações relevantes dos pacientes. Um aspecto primordial no desenvolvimento de aplicações de sumarização é a avaliação dos sumários gerados, na medida em que os usuários passam a confiar nesses sumários para tomada de decisão.
Sobre avaliação de sumários, a opção que NÃO apresenta um modelo adequado para esta tarefa é:

(A) ROUGE.
(B) BLEU.
(C) METEOR.
(D) BERTScore.
(E) PEGASUS.

18. Ao integrar informações provenientes de fontes de dados externas, como documentos ou bancos de dados, com *Large Language Models* (LLMs), é possível empregar uma variedade de técnicas e estratégias para construir aplicações adaptadas às demandas específicas de cada projeto e aos recursos disponíveis.
Das opções abaixo, a que descreve corretamente uma dessas técnicas é:

(A) *Prompt engineering* é a prática de reduzir a complexidade dos modelos de linguagem de IA, simplificando-os para que reconheçam e respondam apenas a comandos básicos de uma palavra, evitando qualquer tipo de prompt contextualizado ou frase mais complexa.
(B) *Few-Shot Learning* é uma abordagem que envolve treinar o modelo com pouco exemplos adicionais e específicos, denominados shots. O modelo usa esses exemplos para entender melhor o contexto ou a tarefa específica solicitada, permitindo que ele generalize a partir de poucos dados e aplique o aprendizado a situações semelhantes.
(C) *Retrieval-Augmented Generation* (RAG) é uma técnica que combina a geração de texto de um LLM com um sistema de recuperação de informações; o modelo original é treinado com uma base de dados ou um conjunto de documentos específico, permitindo a incorporação de novos conhecimentos que não estavam presentes no corpus de treinamento original do modelo.
(D) *Fine-Tuning* é um processo de ajuste fino que consiste em treinar um LLM com um dataset adicional, com o objetivo de personalizar o modelo para tarefas ou domínios específicos. Isso permite que o modelo adapte suas respostas com base no conhecimento ou nos dados contidos nesse dataset adicional.
(E) *Text-to-SQL* é uma funcionalidade padrão integrada em todas as versões do SQL que automaticamente traduz instruções em linguagem natural para comandos SQL, eliminando a necessidade de conhecimento técnico em bancos de dados para a realização de consultas complexas.

19. Acerca dos *frameworks* LangChain e Llamaindex, amplamente utilizados atualmente para construir aplicação integradas a *Large Language Models* (LLMs), a opção que apresenta uma observação correta é:

(A) LlamaIndex é um *framework* projetado para aplicações que utilizam LLMs que se beneficiam de aumento de contexto, fornecendo abstrações que facilitam a ingestão, estruturação e acesso a dados independente de um domínio específico.
(B) no LangChain, *chains* são sequências de chamadas a um LLM, ferramenta ou etapa de pré-processamento de dados, permitindo a criação de pipelines complexos sem necessidade de linguagem específica.
(C) LangChain pode ser utilizado para tarefas como classificação de texto e tradução automática, mas não para extração de entidades nomeadas e geração de texto.
(D) LLamaindex não oferece suporte à busca semântica, não permitindo que os usuários realizem buscas por documentos que contenham conceitos relacionados aos termos de busca.
(E) no LangChain, *chains* implementam uma sequência de ações através de código, ao contrário de agents, onde um modelo de linguagem é utilizado como motor de raciocínio para determinar as ações a serem tomadas.

20. Você é um cientista de dados incumbido de desenvolver uma aplicação de perguntas e respostas para facilitar a extração de informações de documentos PDF contendo artigos científicos na área da saúde. Para construir essa aplicação, as seguintes estratégias foram apresentadas.

I. Utilizar a técnica de *embeddings* de texto para converter documentos PDF em vetores e armazená-los em um *vectorstore*, como ChromaDb ou Pinecone, permitindo buscas semânticas rápidas e eficientes baseadas no conteúdo dos artigos.
II. Desenvolver um sistema de indexação baseado em metadados extraídos dos documentos PDF, como autor, data de publicação e palavras-chave, para facilitar a filtragem e a busca por documentos específicos.
III. Implementar uma abordagem de processamento de linguagem natural (PLN) que empregue a API do modelo de linguagem para gerar respostas precisas às perguntas, utilizando os vetores e metadados armazenados para recuperar informações relevantes dos documentos e inseri-las no contexto do prompt.
IV. Realizar o *fine-tuning* do modelo de linguagem através de um *dataset* que contenha o conhecimento do domínio que se quer adicionar ao modelo, utilizando frameworks como LoRA ou QLoRA para fazer o merge desse *dataset* adicional treinado.
V. Criar uma hierarquia de documentos baseada na classificação dos artigos científicos por tópicos e subtópicos, utilizando algoritmos de *clustering* para organizar automaticamente os documentos em categorias relevantes.

Das estratégias acima:

(A) apenas II e III são válidas.
(B) apenas III, IV e V são válidas.
(C) apenas I, II e III são válidas.
(D) apenas I, III e IV são válidas.
(E) todas são válidas.

21. O Departamento de Informática do Sistema Único de Saúde (DATASUS) disponibiliza inúmeros arquivos para o enriquecimento das bases de dados disponíveis para download. Alguns atributos são preenchidos com informações da classificação estatística internacional de doenças e problemas relacionados com a Saúde (CID-10). São disponibilizados pelo DATASUS arquivos que permitem a agregação das doenças em:

(A) módulo, capítulo e grupo.
(B) capítulo, grupo e categoria.
(C) artigo, grupo e categoria.
(D) capítulo, artigo e inciso.
(E) artigo, inciso e módulo.

22. Disseminados pelo DATASUS para download (ftp.datasus.gov.br), os dados desagregados sobre a declaração de óbito do Sistema de Informação sobre Mortalidade (SIM) estão disponíveis com a extensão:

(A) parquet.
(B) tsv.
(C) pdf.
(D) dbc.
(E) xslx.

23. Considerando a definição, pilares e objetivos da Saúde Coletiva, avalie se são verdadeiras (V) ou falsas (F) as afirmativas a seguir:

I. A saúde é definida como ausência de doenças.
II. Tem como característica ações isoladas da Vigilância Epidemiológica e Sanitária.
II. É considerada a influência de fatores sociais, econômicos e culturais na saúde das comunidades.

As afirmativas I, II e III são, respectivamente:

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

24. Segundo a Lei Orgânica da Saúde (Lei nº 8080/1990), os serviços públicos de saúde e os serviços privados contratados ou conveniados que integram o Sistema Único de Saúde (SUS) devem obedecer aos princípios abaixo, EXCETO:

(A) direito à informação, às pessoas assistidas, sobre sua saúde.
(B) universalidade de acesso aos serviços de saúde em todos os níveis de assistência.
(C) igualdade da assistência à saúde, sem preconceitos ou privilégios de qualquer espécie.
(D) encaminhamento do paciente para inscrição em programas de complementação de renda, caso necessário.
(E) descentralização político-administrativa, com direção única em cada esfera de governo.

25. Sobre o direito à saúde previsto na Lei Orgânica da Saúde (Lei nº 8080/1990) e na Constituição Federal (1988), avalie se são verdadeiras (V) ou falsas (F) as afirmativas a seguir:

I. A saúde é um direito fundamental do ser humano, devendo o Estado, sempre que possível, prover as condições indispensáveis ao seu pleno exercício.
II. O dever do Estado não exclui o das pessoas, da família, das empresas e da sociedade.
III. A saúde é direito de todos e dever do Estado, garantido mediante políticas sociais e econômicas que visem à redução do risco de doença.

As afirmativas I, II e III são, respectivamente:

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

26. Um grupo de pesquisadores deseja acompanhar o histórico de internações hospitalares de mães nascidas após o ano 1997 e que tiveram filhos com baixo peso ao nascer. A ideia central é identificar agravos de saúde que podem contribuir para o baixo peso das crianças no momento do parto. Para isso, os pesquisadores pretendem utilizar duas bases de dados disponíveis para download no DATASUS em acesso aberto: o Sistema de Informações sobre Nascidos Vivos (SINASC) e o Sistema de Informações Hospitalares (SIH/SUS). A pesquisa analisará os dados de nascimentos e internações hospitalar entre 2012 e 2022.

Das opções abaixo, o real motivo que impede o desenvolvimento desse projeto é:

(A) não há dados no SINASC sobre o nascimento das mães em 1997, pois o seu funcionamento se inicia no ano 2000.
(B) não é possível vincular as internações da mãe ao nascimento da criança com baixo peso, pois não há atributos que possam fazer essa vinculação nas bases citadas.
(C) o SINASC é um sistema com informações sobre nascidos vivos, não sobre suas mães.
(D) o SIH/SUS coleta, reúne e publica dados a cada dois anos (somente em anos pares), o que impede uma análise completa no intervalo 2012 e 2022.
(E) o SINASC não registra informações sobre peso ao nascer.

27. Dataframes da biblioteca Pandas no Python são muito versáteis. Com eles é possível ler, processar, transformar e exportar dados tabulares com grande eficiência. Considere um dataframe criado a partir da leitura de um arquivo do tipo csv (*comma separated value*). Só devem ser carregadas as primeiras mil linhas das colunas A, B e C. Além disso, todos os valores devem ser convertidos para o tipo *string*. Os parâmetros e valores do método read_csv() que possibilitam isso são:

(A) nrows=1000, usecols=['A', 'B', 'C'] e dtype=str.
(B) nrows=1k, usecols=['A', 'B', 'C'] e type=String.
(C) lines=1000, columns=['A', 'B', 'C'] e dtype=String.
(D) nrows=1000, names=['A', 'B', 'C'] e type=str.
(E) lines=1k, columns=['A', 'B', 'C'] e dtype=str.

28. Para reproduzir a transformação ilustrada na figura abaixo, o código Python que faz uso da bilblioteca Pandas (pd) e pode ser utilizado para para unir dois dataframes (df1 e df2), criando o dataframe (df3), é:

(A) df3 = pd.join([df1, df2]).
(B) df3 = pd.union([df1, df2]).
(C) df3 = pd.concat([df1, df2]).
(D) df3 = pd.merge([df1, df2]).
(E) df3 = pd.sum([df1, df2]).

29. No campo da saúde, é comum a adoção de métodos para a reduzir a dimensionalidade dos dados, como a segmentação de idades em faixas etárias. O comando Python, com o uso da biblioteca Pandas (pd), que pode ser utilizado para segmentar os valores de uma lista de idades (tipo inteiro) em 10 faixas etárias, é:

(A) pd.qcut(idades, chunks=10).
(B) pd.cut(idades, chunks=10).
(C) pd.qcut(idades, bins=10).
(D) pd.split(idades, bins=10).
(E) pd.cut(idades, bins=10).

30. A biblioteca Pandas do Python possui diversas formas para selecionar partes de um objeto dataframe. Utilizando os dados disponíveis no dataframe df (imagem abaixo), um programador deseja criar um dataframe (df_novo) contendo somente as colunas CODUFMUN e COMPETEN. Das opções abaixo, a única INCORRETA é:

df

	CNES	CODUFMUN	DT_ATUAL	COMPETEN
0	2002183	120033	200610	201012
1	6411312	120038	201007	201012
2	3449300	120040	201011	201012
3	2002833	120040	201005	201012
4	2000148	120060	200904	201012

df_novo

	CODUFMUN	COMPETEN
0	120040	201012
1	120040	201012
2	120001	201012
3	120040	201012
4	120040	201012

(A) df_novo = df.sliceframe(: , [1,3]).
(B) df_novo = df.loc[: , ['CODUFMUN', 'COMPETEN']].
(C) df_novo = df.iloc[0:40 , [1,3]].
(D) df_novo = df.drop(columns = ['CNES','DT_ATUAL']).
(E) df_novo = df[['CODUFMUN', 'COMPETEN']].

31. Atributos numéricos diferentes podem possuir enorme discrepância de amplitude em um mesmo conjunto de dados. Por exemplo, enquanto a idade de uma pessoa tende a estar entre 0 e 130 anos, a altura em metros costuma variar entre 0,5 e 2,5. Em casos assim, alguns modelos de análise podem dar uma importância muito maior para a variável de maior amplitude (idade). Para lidar com esse efeito, é comum o uso de métodos de feature scaling disponíveis em pacotes *Python* como o *Scikit Learn*. Das opções a seguir, a única que NÃO representa um método para *feature scaling* é:

(A) *Min-Max Scaling*.
(B) padronização *(Z-Score)*.
(C) *Maximum Absolute Scaling*.
(D) Distância de *Levenshtein*.
(E) *Robust Scaling*.

32. Bases de dados desbalanceadas podem afetar os resultados de muitos algoritmos que tentam identificar padrões nesses dados. Essa é uma realidade para muitas bases da saúde, pois a prevalência de uma doença na população pode ser algo raro. Sobre o processo de rebalanceamento de bases de dados, avalie se são verdadeiras (V) ou falsas (F) as afirmativas a seguir.

I. A técnica de *oversampling* envolve aumentar o número de instâncias da classe minoritária (menos frequente) para equilibrar a distribuição das classes.
II. A técnica de *undersampling* envolve reduzir o número de instâncias da classe majoritária (mais frequente) para equilibrar a distribuição das classes.
III. Antes de aplicar a técnica de *oversampling*, é importante dividir os dados em conjuntos de treino e teste. A técnica de *oversampling* só deve ser aplicada ao conjunto de testes.

As afirmativas I, II e III são respectivamente:

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

33. Na análise de dados textuais, é muito comum o uso de medidas de similaridade para agrupamento de documentos. Sobre a similaridade por cosseno, das afirmações utilizadas abaixo está correta:

(A) a similaridade por cosseno é uma medida de similaridade que considera apenas a frequência absoluta das palavras nos textos, por isso não permite a comparação entre textos longos e curtos.
(B) a similaridade por cosseno é uma medida de similaridade que ignora completamente a ordem das palavras nos textos e só pode ser aplicada a textos informais.
(C) quanto maior o valor da similaridade por cosseno entre dois textos, maior é a similaridade entre eles.
(D) a similaridade por cosseno não pode ser usada para calcular a similaridade de textos, somente de sequências genéticas.
(E) a similaridade por cosseno é uma medida de similaridade que não pode ser aplicada a textos longos.

34. A análise visual de dados, por meio de gráficos e dashboards, por exemplo, tem papel central na análise exploratória de dados. Sobre o papel da análise visual na descoberta de padrões em dados, é possível afirmar que a análise visual:

(A) é usada apenas para tornar os gráficos e visualizações de dados mais atraentes esteticamente, especialmente com a construção de dashboards.
(B) não desempenha um papel significativo na descoberta de padrões em dados, sendo mais útil para fins de apresentação de resultados.
(C) é limitada na descoberta de padrões em dados e não oferece nenhuma vantagem sobre métodos quantitativos.
(D) é útil apenas para apresentar resultados em dashboards após a descoberta de padrões por métodos estatísticos.
(E) pode permitir a identificação de tendências, anomalias e padrões nos dados de forma intuitiva e rápida.

35. Ao analisar dados do campo da saúde, é comum encontrar atributos com dados faltantes. Sobre as estratégias para lidar com essa situação em pesquisas da saúde, avalie se são verdadeiras (V) ou falsas (F) as afirmativas a seguir:

I. É importante compreender se os dados estão faltando de forma aleatória ou sistemática, o que pode influenciar a escolha da técnica apropriada para lidar com eles.
II. Uma abordagem comum é a imputação de dados, onde os valores faltantes são estimados com base em informações disponíveis.
III. A exclusão de observações com dados faltantes não é capaz de introduzir viés significativo.

As afirmativas I, II e III são, respectivamente:

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

36. Modelos de IA nem sempre são transparentes sobre quais fatores mais influenciam suas decisões. Para mitigar esse efeito, uma abordagem é usar soluções do campo de pesquisa chamado Inteligência Artificial Explicável, ou em inglês: *Explainable Artificial Intelligence* (XAI). O objetivo é ajudar a entender como um modelo complexo funciona, fornecendo alguma explicabilidade e/ou interpretabilidade sobre suas decisões. Sobre o uso de XAI, avalie se são verdadeiras (V) ou falsas (F) as afirmativas a seguir:

I. Métodos de interpretabilidade robustos e consistentes são elementos fundamentais para a construção de confiança e para viabilizar a responsabilização (*accountability*) de decisões algorítmicas.
II. No campo da saúde, a busca por modelos de IA interpretáveis é fundamental não só para dar transparência para médicos e pacientes, mas para diversas outras partes interessadas, inclusive aos órgãos reguladores.
III. Na pesquisa científica, muitas aplicações utilizam modelos baseados em redes neurais profundas, que são por natureza pouco transparentes. Neste caso, a XAI pode desempenhar um papel importante ao dar acesso aos padrões identificados durante o processo de treinamento do modelo, podendo subsidiar a geração de hipóteses de pesquisa.

As afirmativas I, II e III são, respectivamente:

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

37. Modelos de IA que apresentam vieses podem levar a um tratamento desigual e discriminatório contra indivíduos e grupos específicos. Imagine um modelo usado para a seleção de candidatos a vagas de emprego que privilegia homens em detrimento de mulheres, mesmo que elas sejam igualmente qualificadas. Esse tipo de viés de gênero pode perpetuar desigualdades e prejudicar a carreira de muitas mulheres. Dentre os possíveis elementos que podem mitigar esse efeito está:

(A) viés nos dados utilizados para treinar os modelos de IA.
(B) ausência diversidade na equipe de desenvolvimento.
(C) presença de mecanismos de auditoria e avaliação de viés nos modelos de IA.
(D) seleção inadequada de dados.
(E) dados desatualizados ou não representativos.

38. Sobre os impactos e riscos do uso de inteligência artificial (IA) e *machine learning* na saúde, é INCORRETO afirmar que:

(A) a falta de explicabilidade dos modelos de IA e *machine learning* impede que os profissionais de saúde compreendam as decisões tomadas por eles, o que pode levar a erros médicos.
(B) a utilização de IA e *machine learning* na saúde pode levar à discriminação de grupos minoritários, como mulheres e pessoas com deficiência, se os dados utilizados para treinar os modelos forem enviesados.
(C) a falta de transparência nos algoritmos de *machine learning* pode dificultar a compreensão das decisões clínicas tomadas pelas máquinas, comprometendo a autonomia do paciente e a relação médico-paciente.
(D) a privacidade e segurança dos dados dos pacientes não são um problema importante no uso de IA e *machine learning* na saúde, pois sempre são utilizados dados anonimizados previamente.
(E) modelos de IA podem ser enviesados se os dados utilizados para treinamento forem desbalanceados, o que pode levar à discriminação de grupos específicos de pacientes, como minorias étnicas ou pessoas com doenças raras.

39. A Lei Geral de Proteção de Dados Pessoais (LGPD) prevê diversos requisitos para o tratamento de dados pessoais. Avalie se são verdadeiras (V) ou falsas (F) as afirmativas I, II e III a seguir:

São hipóteses previstas na LGPD para o tratamento de dados pessoais:

I. a realização de estudos por órgão de pesquisa, não se aplicando a este a necessidade de anonimização dos dados pessoais que serão tratados, mesmo que possível.
II. o cumprimento de obrigação legal ou regulatória pelo controlador.
III. a tutela da saúde, exclusivamente, em procedimento realizado por profissionais de saúde, serviços de saúde ou autoridade sanitária.

As afirmativas I, II e III são, respectivamente:

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

40. Segundo a LGPD, o controlador deve "comunicar à autoridade nacional e ao titular a ocorrência de incidente de segurança que possa acarretar risco ou dano relevante aos titulares". NÃO é obrigação do controlador comunicar:

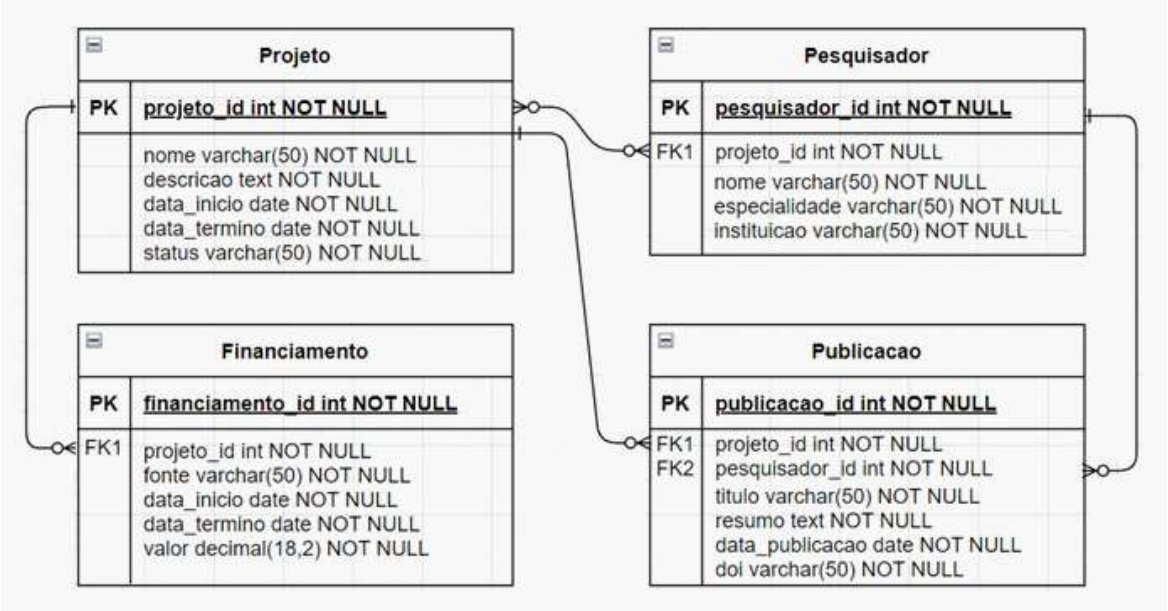
(A) a descrição da natureza dos dados pessoais afetados.
(B) a indicação das medidas técnicas e de segurança utilizadas para a proteção dos dados, mesmo que envolvam segredos comercial e industrial.
(C) os motivos da demora, no caso de a comunicação não ter sido imediata.
(D) as medidas que foram ou que serão adotadas para reverter ou mitigar os efeitos do prejuízo.
(E) os riscos relacionados ao incidente.

Prova Discursiva

QUESTÃO

Entre os desenvolvimentos mais recentes no campo da Inteligência Artificial e do Processamento de Linguagem Natural está o ChatGPT, uma implementação avançada de Large Language Model (LLM) desenvolvida pela OpenAI, que rapidamente ganhou popularidade e se tornou um marco na interação homem-máquina através da linguagem natural. A API da OpenAI permite a criação de interfaces com os modelos de linguagem da linha GPT para uma vasta gama de aplicações, desde a automação de tarefas de escrita até o desenvolvimento de sistemas avançados de conversação e assistentes virtuais personalizados.

Seja o banco de dados PostgreSQL implementado a partir do diagrama ER apresentado a seguir, desenhado na notação *crow's foot* e complementado com os tipos dos dados, que contêm as tabelas Projeto, Pesquisador, Publicacao e Financiamento.



Ele pode ser apresentado a um modelo da linha OpenAI GPT através do framework LlamaIndex através do código abaixo:

```
from llama_index.core import SQLDatabase
from llama_index.core.query_engine import (
    NLSQLTableQueryEngine,
    SQLTableRetrieverQueryEngine,
)
sql_database = SQLDatabase(engine, include_tables=["projeto", "financiamento", "pesquisador", "publicacao"])
query_engine = NLSQLTableQueryEngine(sql_database)
```

Onde `engine` é uma conexão estabelecida corretamente com o banco de dados. Quando utilizamos `query_engine` para perguntar "Quais os projetos em andamento?", o modelo responde corretamente que "*Os projetos em andamento são Horto Escola, Manejo da Fauna Silvestre e Controle de Zoonoses e Produção Sustentável no Território do CFMA*.", e apresenta os metadados da resposta conforme abaixo:

```
{'17064b17-05a3-4e51-9692-4257561285c2': {},
 'sql_query': "SELECT nome FROM projeto WHERE status = 'em andamento';",
 'result': [('Horto Escola',),
  ('Manejo da Fauna Silvestre e Controle de Zoonoses',),
  ('Produção Sustentável no Território do CFMA',)],
 'col_keys': ['nome']}
```

De onde podemos concluir que a pergunta enviada via `query_engine` foi transformada no código SQL contido em `sql_query` e que os dados resultantes foram utilizados para responder à pergunta.

Considere agora que a seguinte pergunta seja enviada:

Qual foi o valor total dos financiamentos dos projetos com dois ou mais pesquisadores envolvidos em 2023?

A resposta correta dada por `query_engine` foi:

O valor total dos financiamentos dos projetos com dois ou mais pesquisadores envolvidos em 2023 foi de R$ 138.000,00.

Considerando a técnica apresentada, e a pergunta e a resposta informadas, em texto com o mínimo de 50 linhas e o máximo de 150, responda às seguintes perguntas:

1 – Como se dá a transformação da pergunta em linguagem natural para o código SQL correspondente?
2 – Apresente um código SQL que pode ser gerado corretamente a partir da pergunta.
3 – Os dados do banco de dados são conhecidos pelo modelo de linguagem no momento da transformação da pergunta em linguagem natural para código SQL?
4 – Como se dá o processo de execução e apresentação da resposta final ao usuário?

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