TESTE 1 -

1) Exercício 1

Suponha que você possui uma base de dados rotulada com 10 classes não balanceadas, essa base é formada por 40 features de metadados e mais 3 de dados textuais abertos.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

Bibliotecas usadas:

- **Scikit-learn** para processamento de dados e aprendizado de máquina.
- **Imbalanced-learn** para técnicas de oversampling.
- **Pandas** para manipular os dados.
- **NUMPY** para análise de matrizes

1. *Descreva como faria a modelagem dessas classes.*

R: Realizaria o pré-processamento dos dados para compreender as classes desbalanceadas. Após a análise, empregaria a técnica de oversampling utilizando o SMOTE para gerar instâncias das classes minoritárias, evitando a criação de exemplos artificiais.

2. *Ao finalizar essa modelagem, como iria apresentar essa modelagem para a área contratante?*

R: Apresentaria através de um relatório contendo:

- Descrição detalhada do processo de modelagem;
- Métricas de avaliação do modelo;
- Explicação do impacto das classes não balanceados;
- Como a área contratante pode utilizar a modelagem para tomar decisões;

3. *Como faria a validação desse modelo?*

R: No caso da base de dados com 10 classes não balanceadas, a validação poderia ser feita da seguinte forma:

1. Divida os dados em conjuntos de treinamento e teste. O conjunto de treinamento será usado para treinar o modelo, e o conjunto de teste será usado para avaliar o desempenho do modelo.
2. Treine o modelo no conjunto de treinamento.
3. Avalie o modelo no conjunto de testes.

*Pode ser usada diversas técnicas de avaliação, como a validação cruzada, MEAN, MSE..

4. *Supondo que esses dados são recebidos diariamente, como iria trabalhar com esse desafio?*

R: Criando pipeline de pré-processamento e modelagem que pudesse ser facilmente atualizado com novos dados. A reavaliação periódica do modelo seria necessária para garantir sua eficácia contínua. Um algoritmo que se comporta muito bem com atualizações contínuas é o Random Forest.

5. *Como levaria esse projeto para um ambiente produtivo?*

R: Implementaria um sistema com conteinerização (Docker) para garantir consistência no ambiente de produção, realizando testes com dados reais antes da implementação, configurando monitoramento contínuo para ajustes no desempenho do modelo e integrando-o aos sistemas existentes.

2) Exercício 2:

Suponha que você tenha uma base de dados de vendas de uma loja de varejo que inclui informações sobre produtos, clientes, datas de compra e valores das vendas. A base de dados possui, em média, 10.000 registros diários.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

Bibliotecas usadas:

- **SciPy** usada para análise estatística

1. Como você iria explorar os dados para obter insights sobre o desempenho das vendas.

R: Usando a análise exploratória, podemos separar as features que têm correlação direta com o desempenho. Por exemplo, podemos comparar as vendas, campanhas de marketing, sazonalidade, entre outras métricas.

2. Como você responderia as seguintes questões:
    i. Qual é o desempenho de vendas ao longo do tempo?
    R: Usando gráficos temporais de linhas podemos visualizar o desempenho das vendas com facilidade.
    ii. Quais são os produtos mais vendidos?
    R: Usando uma simples análise de frequência podemos gerar visualizações com gráficos de barras ou tabelas classificadas.
    iii. Como as vendas variam por categoria de produtos?

R: Eu poderia usar um gráfico de linhas para visualizar as vendas por categoria de produtos ao longo do tempo, destacando o volume ou o valor total das vendas de cada categoria..

iv. Qual é a distribuição dos valores de venda?
R: Particularmente para distribuição eu gosto de usar um histograma para visualizar a distribuição dos valores de venda.

v. Como os preços dos produtos afetam as vendas?
R: Usando a análise de regressão podemos estimar o impacto do preço dos produtos nas vendas.

vi. Qual é o perfil dos principais clientes em termos de compras?
R: Eu poderia criar um perfil dos clientes com base em suas características demográficas, comportamentais ou de compra.

3. Como você faria para identificar grupos de clientes nessa base de dados?

R: Posso segmentar os grupos pela idade média dos clientes, o gênero dos clientes, o local de residência dos clientes ou os produtos mais comprados.

4. Qual teste estatístico você usaria para provar uma hipótese referente aos segmentos de clientes? e como iria aplicá-lo?

R: Para identificar a segmentação de dois ou mais clientes, eu uso a biblioteca SciPy para realizar o teste t, que compara métricas entre dois grupos. Ao aplicá-lo, utilizamos a hipótese nula (H0) e a hipótese alternativa (H1), onde H0 representa a igualdade das médias e H1, a diferença de média.

*Caso tenha mais de dois grupos podemos usar o método de comparação múltipla, como a análise de variância.

3) Exercício 3

Suponha que você tenha uma base de dados contendo textos jurídicos, como decisões judiciais, petições e documentos legais. A base de dados inclui informações sobre o conteúdo do texto, data, jurisdição e outras informações relevantes. Seu objetivo é criar um sistema de recomendação que sugira textos jurídicos semelhantes a um texto de referência.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

Bibliotecas utilizadas:

- NLTK: Para o pré-processamento dos dados e a representação dos textos.

- Scikit-learn: Para o cálculo da similaridade e a seleção dos resultados.

1. *Descreva como você desenvolveria o sistema de recomendação que recebe um texto de referência e sugere os textos mais semelhantes a ele na base de dados.*

- Usaria a biblioteca NLTK para processamento de linguagem natural, ele oferece diversas ferramentas para tokenização, stemming, lematização e outras tarefas essenciais de pré-processamento de texto.
- O segundo passo é fazer o pre-processamento dos dados retirando pontuações, caracteres especiais, stopwords e stemming.
- Realizar tokenização para facilitar o processamento.
- Fazer a vetorização do texto usando TF-IDF para converter texto em números.
- O último passo é descobrir a similaridade de cosseno para achar textos semelhantes.

2. *Como você avaliaria esse sistema de recomendação?*

R: Podemos usar métricas como precisão como o F1-score para comparar as recomendações de textos similares.

c) *Suponha que novos textos jurídicos sejam adicionados diariamente. Como você manteria o sistema de recomendação atualizado e garantiria que ele continue a fornecer recomendações relevantes?*

R: Podemos incluir uma verificação diária dos scores relevantes no script. Se os scores estiverem abaixo de um valor mínimo pré definido, isso indicaria a necessidade de revisar e possivelmente atualizar o modelo. Isso garantirá que o sistema se ajuste às mudanças na base de dados e continue a fornecer recomendações relevantes ao longo do tempo.

TESTE 2 –

1) Como funciona o teste de hipóteses e qual é a sua finalidade na análise estatística?

R: O teste de hipóteses avalia amostras, utilizando estatísticas para determinar se os resultados são significativos ou devido ao acaso. A finalidade é fornecer evidências estatísticas para aceitar ou rejeitar uma hipótese nula, contribuindo para decisões informadas na análise de dados, onde a hipótese alternativa pode ser considerada se a nula for rejeitada.

2) O que são redes generativas adversárias (GANs) e quais são os possíveis usos dessas redes?

   R: GANs são modelos de inteligência artificial em que dois componentes competem para criar dados realistas (gerador e discriminador). Usadas para gerar imagens, vídeos e músicas, as GANs fazem a melhoria de qualidade de imagens e aprendem a partir de exemplos.

3) O que são modelos de linguagem? Qual a diferença entre LLMs e modelos de linguagem tradicionais?

   R: Modelos de linguagem são projetados para gerar linguagem natural através de treinamento, ele aprende padrões e relações semânticas entre as palavras para gerar textos, tradução automática entre outras tarefas. LLMs são treinados para interpretar textos complexos e nuances em grandes conjuntos de dados enquanto o modelo tradicional é segmentado por estruturas fixas.

4) Suponha que você tenha um conjunto de dados com três ou mais grupos para comparar e deseja determinar se há diferenças significativas entre eles. Descreva como você escolheria entre o teste ou outras técnicas estatísticas

   R: Escolhi a análise de variância multivariada (MANOVA), porque ela lida com muitas variáveis simultaneamente, dentro da manova existem alguns testes estatísticos como à matriz de covariância e à matriz de médias multivariadas. Além desses testes, ainda existem os testes de correlação e o teste de regressão que eu uso frequentemente.

5) Qual é a importância do pré-processamento de texto em tarefas de NLP? Quais são as etapas comuns no pré-processamento de texto?

   R: O pré-processamento de texto é crucial em tarefas de Processamento de Linguagem Natural (NLP) porque melhora a qualidade dos dados, reduz o ruído e facilita a extração de padrões significativos. Existem muitas etapas, mas vou citar as mais importantes.

   - **Tokenização:** Dividir o texto em unidades menores para facilitar o processamento.
   - **Remoção de Stopwords:** Eliminar palavras comuns.
   - **Stemming e Lematização:** Reduzir palavras às suas formas básicas.
   - **Codificação de Texto:** Converter o texto em uma representação numérica, como embeddings word2vec ou TF-IDF, para uso de machine learning.

6) Descreva o processo de vetorização de texto e como modelos de linguagem como o Word2Vec ou o TF-IDF podem ser usados para representar palavras e documentos.

R: Modelos como Word2Vec capturam relações semânticas, gerando embeddings distribuídos, enquanto o TF-IDF destaca a importância relativa de termos. Ambos transformam o texto em vetores, facilitando a análise em tarefas de Processamento de Linguagem Natural (NLP).

7) O que é a análise de sentimento em NLP e quais são os principais métodos para realizar essa tarefa? Como você avaliaria a eficácia de um modelo de análise de sentimento?

R: O processamento de NLP envolve a identificação e classificação de sentimentos, podem ser positivos, negativos ou neutros. Os melhores métodos são:

- **Abordagens baseadas em Regras:** Usam regras e listas de palavras para atribuir polaridades.
- **Aprendizado Supervisionado:** Modelos de classificação, como SVM, treinados em conjuntos rotulados.
- **Redes Neurais:** Modelos como LSTM e redes neurais convolucionais, capturando contextos mais complexos.

8) Qual é a diferença entre a classificação de texto e o agrupamento (clustering) de texto em NLP? Em que situações cada um é mais apropriado?

R: A classificação de texto atribui rótulos predefinidos, enquanto agrupamento organiza documentos sem rótulos em grupos com base em semelhanças. Classificação é para categorias específicas, agrupamento explora estruturas desconhecidas. Classificação tem rótulos definidos, agrupamento não.

9) Explique o conceito de reconhecimento de entidades nomeadas (NER) em NLP e suas aplicações práticas.

R: O NER classifica entidades, como nomes de pessoas, locais e organizações, em textos. Executa a extração de informações, análise de sentimentos e aprimoramento da busca de informações.

10) Como você lidaria com problemas de desequilíbrio de classe em tarefas de classificação de texto em NLP? Quais estratégias seriam eficazes?

R: O desequilíbrio pode causar um problema de viés de classe, isso ocorre porque o aprendizado de máquina tende a aprender com a classe mais frequente que nesse caso é classificado como “SPAM”, por esse motivo temos que usar o balanceamento de

amostragem, equilibrando as classes na amostra de treinamento. Atribuir pesos diferentes. Também é possível combinar estratégias de amostragem e aprendizado. Por exemplo, a sobreamostragem pode ser usada para aumentar o tamanho da classe minoritária, e os pesos de classe podem ser usados para ajustar o modelo para lidar com o aumento do tamanho da classe minoritária.

TESTE 3 - CASE

canada_amostra.csv

Amostra:

Contextualização:

O Base de dados canada_amostra em formato CSV representa um conjunto de empresas do Canadá com a respectiva descrição de seus produtos, dados econômicos e localização.

Assim, podemos caracterizar cada variável:

name: nome da empresa;

description: descrição do produto da empresa;

employees: número de empregados da empresa;

total_funding: Total de investimento já recebido pela empresa;

city: cidade;

subcountry: estado;

lat: latitude da cidade;

lng: Longitude da cidade.

1) Problema:

Deseja-se prospectar empresas que possuam soluções em **tratamento de água** , principalmente, elativas à : **solutions on waste and water, Improve water quality and water efficiency use, water contamination, water for human consumption, water resources** .

a) EXERCÍCIO 1 - Aplique um algoritmo de ML (ou um conjunto deles) capaz de selecionar as principais empresas indicadas para desenvolver a solução de acordo com seu alinhamento com o tema (Justifique a escolha do algoritmo).

b) EXERCÍCIO 2 - Faça uma análise exploratória dos resultados acrescentando as demais variáveis contidas no dataset. Quais insights você pode obter a partir desses dados? Quais são as principais cidades (pólos de desenvolvimento) para essa solução?

c) EXERCÍCIO 3 - EXTRA - Se você terminou o desafio de forma rápida, temos mais algumas perguntas para serem respondidas. Elas, como dito, não são obrigatórias, então sinta-se à vontade em não as responder ou até mesmo respondê-las parcialmente. Essa parte visa observar seu entendimento de um ambiente real de produção.

i) a) organize seus códigos em pacotes garantindo seu versionamento e documentação (bibliotecas auxiliares, etc.).

ii) b) construa testes automatizados para validação do seu pacote.

iii) c) crie uma imagem Docker capaz de executar suas análises em um ambiente de produção.

iv) d) crie um GitHub público e suba todo o código do Teste 3, e disponibilize para avaliação.