TEST 3 – CASE

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).

Inicialmente foi realizada uma análise exploratória sobre os dados para entender como eles estavam apresentados. Ao analisar a coluna 'description' e o alinhamento das empresas a soluções em tratamento de água, optou-se por excluir o item 'water resources', uma vez que ele trata de recursos hídricos, recursos esses que estão disponíveis para ambiente para uso geral. Sendo assim, não estando associado ao tratamento de água.

Foi criado a classe CarregaDados.py, para fazer a leitura da planilha .xlsm e organização dos dados em colunas para facilitar a leitura via biblioteca pandas.

A etapa de pré-processamento (PreProcessor.py) consistiu na exclusão de caracteres especiais, passagem de todas as letras para minúsculas e o processamento de tokenização, vetorização e aplicação de stopwords. Foi aplicado também a correção de balanceamento das classes.

Foi aplicado o treinamento de 4 modelos (TreinaModelo.py): K-NN, Random Forest, Logistic Regression e Gradient Boost.

Analisou (AvaliaModelo.py) as métricas relativas a esses modelos como: accuracy, precision, recall, F1-Score, curvas ROC e matriz de confusão, a fim de ajustar os parâmetros e definir qual o melhor modelo (SelecionaModelo.py).

A partir dessa análise, optou-se por utilizar o modelo GradientBoost para a análise e para fazer as previsões, prospectando as empresas que poderiam estar alinhadas ao tratamento de água.

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?

Após a seleção das empresas e avaliando as demais informações da planilha, é possível observar que a planilha possui a maior parte das informações como ausentes, não sendo possível dessa forma, chegar a conclusão de qualque insights sobre os pólos de desenvolvimento, localização, número de empresas e total de investimento recebido pela empresa.