**Questão 1 :**

Um dos processos mais desafiadores na área de Inteligência Artificial é o processamento de linguagem natural (NLP - Natural Language Process em inglês).

Geralmente são utilizados algoritmos de classificação para executar a tarefa, inclusive aqueles para classificação multi-label. Diante desse tipo de desafio, necessitamos executar um processo minucioso de pré-processamento da base de dados que envolvem diversos processos até o treinamento do modelo.

a. Atualmente temos um trabalho onde necessitamos classificar uma base de dados que contém textos de ligações telefônicas em português. Você como um cientista de dados necessita classificá-lo. Sendo assim, indique todo o processo de classificação, desde a seleção de uma base para o treinamento, passando pelo pré-processamento até a seleção do melhor algoritmo.

Também haverá alguns problemas como uma alta utilização de memória e processamento, caso esses problemas venha a ocorrer, indique sua estratégia para solucioná-los.

b. Você deverá descrever todo o processo e após concluir deverá compartilhar esse processo com o time de Cientistas de Dados da Callink.

c. Você será avaliado pela criatividade utilizada na solução proposta.

**Questão 2:**

**Descrição da Atividade**

Desenvolva um modelo preditivo capaz de classificar a cada dia a probabilidade de que a ação da empresa AMZN negociada na Nasdaq alcance nos próximos 10 dias uma valorização de no mínimo 4% e que não desvalorize mais que 2,5%.

A variável dependente já está calculada para facilitar o desenvolvimento e ela representa a condição de atingimento de resultados citada acima.

Para essa atividade disponibilizaremos um dataset com os dados diários de 10 anos de negociação segmentado em Treino (80%) e Teste (20%).

Utilize o dataset de teste para calcular as principais métricas de um modelo de classificação binário, considerando também que as classes são desbalanceadas, portanto deve ser escolhida além das métricas tradicionais métricas que considerem a condição de classes desbalanceadas.

O objetivo dessa atividade não é encontrar respostas certas, mas sim avaliar o pipeline de desenvolvimento de um modelo de Machine Learning e sua criatividade em achar caminho para solucionar um problema, portanto é importante que você descreva cada etapa do desenvolvimento explicando o porquê dos métodos utilizados.

Utilize o notebook do GOOGLE COLAB (<https://colab.research.google.com/>) para desenvolver a atividade e ao terminar compartilhe conosco:

[Juliano.alves@callink.com.br](mailto:Julio.amaral@callink.com.br)

Documente no notebook cada etapa executada de forma que facilite a sua apresentação na etapa de avaliação e apresentação.

**Sobre o DataSet**

O dataset é formado por dados referente aos níveis de preços em que a ação foi negociada diariamente, sendo composto pelo candle (diário) e demais fields fornecidas no terminal Bloomberg.