Desafio de ciência de dados para seleção de bolsista pesquisador

Para participar da nossa seleção, você deverá nos enviar a resolução das questões abaixo, elas representam problemas similares a alguns encarados no laboratório, para que possamos avaliar seus conhecimentos. Mesmo que você só consiga resolver uma das questões, não deixe de nos enviar a sua solução.

Um screencast, com duração de aproximadamente 10 minutos, deverá ser gravado explicando a solução proposta. O screencast deverá ser disponibilizado em nuvem via serviço de streaming ou algum outro serviço de compartilhamento de vídeo/arquivo à sua escolha. O código-fonte implementado deverá ser disponibilizado em algum serviço de versionamento de código, como GitHub e GitLab. O link para o screencast deve ser inserido no arquivo README do repositório a ser criado.

Contextualização
No Insight Lab lidamos constantemente com soluções que lidam com grandes massas de dados (muitas vezes sensíveis), seja ingerindo e/ou consultando. Para isso, precisamos pensar em soluções que lidem com dados heterogêneos e de grande volume para serem armazenados e processados.

Questão 01) Faça um programa na linguagem que escolher e/ou usando uma plataforma de processamento de dados de sua escolha, que lê os arquivos com várias mensagens do Instagram (coluna `Conteúdo`) sobre sintomas de COVID. E retorna como saída: as palavras que são únicas e a contagem de repetidas. Nós gostaríamos de uma solução que funcionasse para grandes conjuntos de dados. Os dados podem ser baixados usando o link.

Questão 02) Crie um Jupyter Notebook para que, dado como entrada as features `hora_dia`, `dia_semana`, `bairro` e `cidade`, compute e exiba a quantidade de roubos de carros na próxima hora com base em um modelo preditivo criado a partir de um ou mais algoritmos de regressão. O modelo deve ser treinado usando os dados objetivos no seguinte link. Esses dados referem-se às ocorrências de roubos de carros no estado de São Paulo no ano de 2019.

Inicialmente, importe todos os arquivos em um único dataframe. Em seguida, você deverá realizar agregações de dados para criar o dataset que será utilizado para treinamento.

Deve-se utilizar 80% do dataset para treino e 20% para teste. Compare os algoritmos de regressão e utilize a métrica RMSE para justificar a escolha do algoritmo de regressão

escolhido. Essa comparação deve ser exibida através de um gráfico no Jupyter Notebook.