****

**Desafio - Ciência de Dados**

Dentre as habilidades necessárias ao cientista de dados podemos destacar:

* Conhecimentos sobre estatística.
* Linguagem Python avançada.
* Domínio dos processos de Data ETL (Extraction, Transformation, and Loading).
* Manipulação e organização de dados (Data Wrangling and Data Exploration).
* Domínio das técnicas de Machine Learning e Deep Learning.
* Familiaridade com frameworks para processamento com Big Data.
* Senso crítico para visualização de dados (Data visualization).

O desafio que se segue tem a intenção de avaliar as habilidades essenciais ao cientista de dados, fornecendo um cenário hipotético para que o candidato possa mostrar seus conhecimentos e habilidades propondo uma abordagem capaz de solucionar os problemas propostos. Antes de iniciar o desafio [faça o download dos arquivos](https://drive.google.com/drive/folders/1wDi2LhO9TibEq9t-GWSE9S41Swb7oE5J) que serão utilizados em cada uma das partes.

**obs: Todas as atividades aqui descritas devem ser realizadas através de um notebook do Google Collaboratory.**

**Parte 1 - Data ETL, Data Wrangling e Data Exploration**

Para começar, realize o download dos arquivos necessários no link acima.

**Zip instructions** - Arquivos de coleta de vibração crua:

* Os arquivos se encontram no formato CSV “\*.csv”.
* Em cada arquivo existe uma coleta de vibração realizada pelo sensor “band-aid”.
* O nome do arquivo traz as seguintes informações: “{start}-{interval}-{sensor\_id}.bin”
  + Exemplo: 1623535615-3006-IAJ9206.csv
    - start: 1623535615 [epoch Unix]
    - interval - sampling duration: 3006 [ms]
    - sensor\_id: IAJ9206 [string que identifica um sensor]
* As coletas dizem respeito aos dados de aceleração em g nos eixos X, Y, Z de um acelerômetro.

Utilizando estes arquivos você deve completar as seguintes etapas:

1. Apresentar os dados contidos nos arquivos no domínio do tempo.
2. Apresentar os dados contidos nos arquivos no domínio da frequência (fft).
3. Aplicar filtros, se necessário, para limpar e corrigir os sinais da melhor forma possível.
4. [Bônus] Criar uma função capaz de identificar os harmônicos e picos no domínio da frequência, de maneira a reduzir a quantidade de dados e filtrar a informação relevante.

**Parte 2 - Machine Learning, Deep learning e Data Driven Solutions:**

Assim como na etapa anterior você deve utilizar os arquivos disponibilizados, os quais podem ser encontrados aqui.

**Zip instructions** - Arquivos de coletas de vibração processados:

* Os arquivos se encontram no formato CSV “\*.csv”.
* **collects.csv**: Contém uma lista de objetos que representam as coletas obtidas para diferentes ativos.
* **assets.csv**: Contém as informações sobre os ativos aos quais as coletas fornecidas pertencem.

Considerando os dados contidos nesses arquivos você deve completar as seguintes etapas:

1. Apresentar visualmente os dados contidos em cada arquivo, juntamente com as informações do ativo a que pertencem.
2. Desenvolver um modelo/função capaz de calcular o tempo de downtime e uptime para um ativo qualquer.
3. Desenvolver um modelo/função capaz de identificar mudanças nos padrões de vibração para um ativo qualquer.
4. Identificar possíveis falhas nos ativos utilizando o modelo desenvolvido no item 3 ou um novo modelo (a identificação deve ser autônoma e não uma análise visual).

**Parte 3 - Perguntas:**

Responda às seguintes perguntas com suas próprias palavras, da maneira mais inteligível possível para um leigo da área. Você pode utilizar a internet como ferramenta de pesquisa, entretanto as respostas devem ser de sua autoria.

1. O que é um 'viés de seleção'?
2. O que é uma 'análise do poder'?
3. Qual é melhor ou mais importante, ter bons dados ou construir bons modelos?
4. Quais são as maneiras de definir o número de clusters em um algoritmo de clustering? Como saber se o resultado obtido foi satisfatório?

O cientista de dados deve ser o mais completo dentre os profissionais da área de dados, sendo assim, não se limite de nenhuma forma durante a execução desse desafio, deixe claras todas as suas considerações e conclusões.

