



TRABALHO PRÁTICO - ANÁLISE DE DADOS FALTANTES (COMPUTAÇÃO EM NUVEM E IOT II)

O presente trabalho pode ser realizado em dupla e tem como objetivo avaliar os conhecimentos adquiridos sobre análise de dados faltantes em IoT. O mesmo deverá ser apresentado até os dias **22 e 23 de Dezembro de 2022**, durante as aulas de Computação em Nuvem e IoT II e tem o valor de 3,5 pontos.

A sua tarefa é eventualmente do (a) seu (sua) colega é acessar uma API que provê dados reais de clima pelo Mundo. Nesta atividade, os dados foram retirados da [plataforma Meteostat](#) e a API foi desenvolvida especificamente para esta atividade. Os dados em questão são da temperatura do ar (em graus celsius) de uma localização (omitida nesta atividade).

A API proporciona a obtenção dos dados originais e dos faltantes. Como você(s) já sabe (m), é inevitável que dados obtidos de sensores sejam faltosos ou que não exista a presença de *outliers*.

Para acessar a API, basta utilizar a seguinte URL: [http://3.145.163.55:5000/dados/mes_inicial/mes_final/ano_inicial/ano_final/formato_saida\(opcional\)](http://3.145.163.55:5000/dados/mes_inicial/mes_final/ano_inicial/ano_final/formato_saida(opcional))

- **mes_inicial:** Um valor inteiro que representa o primeiro mês do critério de busca. Possíveis valores: 1-para Janeiro, 2-para Fevereiro, 3-para Março...12-para Dezembro. **OBS.:** São admitidos todos os dias do mês na consulta (do primeiro ao último dia).
- **mes_final:** Um valor inteiro que representa o último mês do critério de busca. Possíveis valores: 1-para Janeiro, 2-para Fevereiro, 3-para Março...12-para Dezembro. **OBS.:** São admitidos todos os dias do mês na consulta (do primeiro ao último dia).
- **ano_inicial:** Um valor inteiro que representa o primeiro ano do critério de busca. Possíveis valores: de 2018 a 2020
- **ano_final:** Um valor inteiro que representa o último ano do critério de busca. Possíveis valores: de 2018 a 2020
- **formato_saida:** **Valor opcional.** Define o formato de saída (padrão *split*). Possíveis valores: *records*, *split*, *index* e *table*. **OBS.:** Independente do formato de saída escolhido (se escolhido), todos os formatos de saída são no formato JSON.

Você (s) também pode (m) [assistir a esta breve apresentação](#) da utilização da API. Exemplo de uso: <http://3.145.163.55:5000/dados/1/2/2019/2019>, retorna todas as temperaturas do ar registradas em uma determinada localidade, de hora em hora, de primeiro de Janeiro de 2019 a 28 de Fevereiro de 2019, com formato de saída *split* (padrão, portanto, não informado), como pode ser visto abaixo:

[illegible]

Neste outro exemplo, a URL <http://3.145.163.55:5000/dados/4/7/2018/2020/table> retorna todas as temperaturas do ar, registradas em uma determinada localidade, de hora em hora, de primeiro de Abril de 2018 a 31 de Julho de 2020, com formato de saída *table*, como pode ser visto na figura a seguir.

[illegible]

Perceba (m) que sempre é retornado o *dataset* com dados faltantes (chamado de *missing*) e o original, para comparação.

A sua tarefa (e eventualmente do (a) seu (sua) colega) é recuperar estes dados (nas disciplinas de Computação em Nuvem e IoT I e II, fizemos vários exemplos de como recuperar dados em formato JSON de uma API). A principal forma de fazer isso é utilizando a [biblioteca request](#), do Python. O tempo mínimo de análise para este trabalho é de um trimestre de dados (de qualquer um dos 12 trimestres de dados disponíveis pela API). Perceba (m) que o que foi mencionado é o tempo mínimo, podendo ser analisado um período superior a este. É desejável que cada membro (a)/dupla faça análise de tempos distintos. Uma sugestão: fazer esta atividade no Collab, deixando seu (s) notebook organizados e bem explicados para ser apresentado (segunda parte da atividade. Apenas uma sugestão.



Sendo assim, você (e eventualmente sua dupla) utilizarão pelo menos três [das diversas técnicas](#) estudadas e que visam solucionar o problema de dados faltantes, tão comum na IoT (seja imputação univariada, multivariada e seus tipos ou interpolação). Outra tarefa da sua equipe é identificar *outliers* (caso existam) e apresentar (seja por gráficos de dispersão, histograma, boxplots ou outras técnicas mais avançadas). Apresentar soluções que visem tratar *outliers* também é bem-vindo, mas não obrigatório.

Finalmente, você (e eventualmente sua dupla) devem apresentar um seminário até o dia 23 de Dezembro de 2022 (conforme mencionado acima).

Os critérios para avaliação deste trabalho serão:

- Identificação de outliers (se houver) - **0,5 pts**
- Comparação de eficiência entre as técnicas de solução para dados faltantes utilizadas (A função `indice_rmse()` do repositório da disciplina pode ser empregada) - **1,5 pts**
- Qualidade da apresentação do seminário (tempo mínimo de 5 minutos. Você (s) podem mostrar o código que utilizaram para registrar os dados da API e etc. Fica a critério seu (s) determinar a qualidade da apresentação) - **1,5 pts**

Bom trabalho!