FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Master Thesis: Addressing Missing Values in Madrid Air Quality Dataset Using Transformers

Missing data is one of the major concerns in studies related to environmental air pollutants. This is often occurred due to the equipment failure or data corruption. Various methods have been introduced to estimate the missing data, such as Bayesian Framework and Neural Networks. In this thesis, we aim to estimate the missing values in Madrid air quality dataset using Transformers. First, we will have a literature review over the Machine Learning techniques that have been proposed in missing values estimation for air quality indexes. The dataset that we will work on contains information related to air pollutants in Madrid since year 2000 until now. This information is recorded in 157 stations that are located in different parts of the city. Some stations are newer than other stations and contain the air pollutants of recent years. This is the link to the dataset:

Due to different locations, all the stations do not have the same behavior. Therefore, we aim to apply clustering techniques to find similar stations. Then we use transformers to predict missing values in each station using the model built for the cluster. There are design choices how to cluster similar stations and how to build a model for each cluster to estimate the missing values. These design choices will be investigated in this thesis.

Prerequisites:

- Knowledge of Machine Learning Techniques and AI
- Programming with Python

Contacts:

- Anahita Farhang, M.Sc
  Gebäude 1, Raum 231
  anahita.farhang-ghahfarokhi@fb2.fra-uas.de
- Prof. Dr. Jörg Schäfer
  Gebäude 1, Raum 217
  jschaefer@fb2.fra-uas.de