**Assignment 2**

**CS430-01**

**Machine Learning on Cloud**

**Fall 2022**

**Exploratory Data Analysis and Regression on Air Quality Data (100 points)**

**Goal:** The goal of this assignment is to use Pandas/Matplotlib/Seaborn to explore the dataset, use Sciket-learn libraries to fill up missing data, encode categorical data, normalize the data, and split the data and train machine learning model on training data and evaluate the results on test data .

**Instructions:** For this assignment, you work on a Google Colab notebook. First, create a new notebook titled **Assignement2\_XXX**, where **XXX** are your initials. Also create a GitHub repository titled **Assignement2\_XXX** to which you can push your code. Then complete the following:

1. In this assignment, you will need to work on the “Beijing Multi-Site Air-Quality Data Data Set” from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data>).
2. (PM2.5) is an air pollutant that is a concern for people's health when levels in air are high. PM2.5 are tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated.
3. You will need to train regression models to find out the concentration of PM2.5 in the air.
4. Do necessary preprocessing (cleaning, missing value imputation, encoding categorical values, split the data into training and test set).
5. Perform regression analysis:
   1. Linear Regression
   2. Stochastic Gradient Descent
   3. Ridge Regression
   4. Lasso Regression
   5. ElasticNet
6. Analyze the results (Score, RMSE) using pretty table and put comments on your findings.

Ensure your notebook is organized and has proper **Markdown comments,** etc. You can assume that after someone see the raw notebook, so it should be clear.

**Assignment Submission:** Upload a link to your GitHub repository for the project in the area provided in Moodle by the deadline specified.