INFO411: Data Mining and Knowledge Discovery

Instructions:
This task is a real-world data mining problem. You are required to prepare a set of presentation slides that must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) your proposed data mining approach and methodology; (3) the strengths and weaknesses of your proposed approach; (4) the performance measures that can evaluate your data mining results; (5) the results and a brief discussion. Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Gas Turbine CO and NOx Emission

Background:
The dataset contains 36733 instances of 11 sensor measures aggregated over one hour from a gas turbine located in Turkey's north western region for the purpose of studying flue gas emissions, namely CO and NOx (i.e., NO and NO2). This data is collected in a date range from 01.01.2011 to 31.12.2015, includes gas turbine parameters (such as Turbine Inlet Temperature and Compressor Discharge pressure) in addition to the ambient variables. Note that the dates are not given in the instances but the data are sorted in chronological order. See the attribute information and the relevant paper for details. Please follow the protocol mentioned in the paper (using the first three years' data for training/cross-validation and the last two for testing) for reproducibility and comparability of works.

The dataset "pp_gas_emission.zip" can be downloaded from https://archive.ics.uci.edu/ml/machine-learning-databases/00551/ . Unzipping this file will show five files: gt_2011.csv, …, gt_2015.csv. The first row of the .csv file contains the information related to each attribute. In short, the task is to use the first 9 attributes to predict the 10^{th} (CO) and 11^{th} (NOX) attributes, respectively. Information on this data set, including the relevant paper, can be found at https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set# .

Requirements:
1. Download the dataset from the above website. Present a general description of the dataset and present the general properties of the dataset.
2. You are required to implement at least two (more will be better) prediction methods for this prediction task. You shall follow that relevant paper to partition the data, that is, "*split the dataset into train (data from 2011-2012), validation (year 2013 data), and test sets (data from 2014 and 2015)*." Also, you need to tune the hyperparameters of your prediction model in a principled way.
3. Discuss any data preprocessing which have been applied.
4. Provide the performance measures of your prediction results. Also, compare your performance with those obtained in that relevant paper and discuss the potential causes of the discrepancy.
5. Conduct the analysis of attribute importance and attribute selection. Discuss your selection methods and results by comparing with those in that relevant paper.