National College of Ireland

School of Computing
Module: Data Mining And ML
Week: 2 Lab: 2

The purpose of the lab is to load a sample data and perform analysis on it. To complete this lab, students must follow the instruction below.

1. Download the Energy_Consumption.csv dataset provided in Moodle to a location on your PC
2. Open the Jupyter notebook from the Anaconda.
3. Browse to the location where you have downloaded the dataset and create a new notebook file using python 3.

4. Using the code and knowledge form previous lecture, import the Energy_Consumption.csv file.

5. Perform the following Analysis on the dataset in different blocks:
   a) Display summary
   b) Check for column names
   c) Display the top 10 rows using the head command.
   d) Display basic information about the dataset
   e) Using any appropriate visualization (Plot) of your choice, analyse the Renewable Energy and Energy Consumption based on you're the time stamp.

6. Using the sample code from class, display all the data points with missing values

7. Drop the rows with null values

8. Fill the missing values using the mean approach.

9. Display all the duplicate data points in the dataset.

10. Remove all duplicate rows

11. Encode all the HVACUsage, LightingUsage, and Holiday columns with Label Encoding

12. Encode DayOfWeek column with one-host encoding

13. Repeat the instructions in (5) again.

14. Split your dataset into X with the columns (Timestamp Temperature Humidity SquareFootage Occupancy HVACUsage LightingUsage RenewableEnergy DayOfWeek Holiday) and Y with the column (EnergyConsumption).

15. Normalize all the data in the X with the exception of the Timestamp column.

16. Perform feature selection using RandomForest Classifier to use only top 8 features from the X.

17. Repeat the instructions in (5) again.

Note: Submit all your notebook files to Moodle.