Task - 2: Implement a Machine Learning Model for the Regression problem

Dataset Link: <https://www.kaggle.com/datasets/ineubytes/pm25-airpolution-dataset>

This dataset describes about air pollution, specifically the levels of PM2.5 (particulate matter) in the air. PM2.5 is tiny particles that can be harmful when inhaled. This provides data over time, including factors like temperature, humidity, and wind speed, which can influence pollution levels. It also includes the location or city where the measurements were taken. This can be used to study air quality, identify trends, and understand how various factors impact PM2.5 levels, helping to address air pollution and its effects on health.

These are the columns of dataset:

1. No represents row numbers in the dataset.
2. year indicates the year of the data in the row.
3. month specifies the month of the data in the row.
4. day denotes the day of the data in the row.
5. hour indicates the hour of the data in the row.
6. pm2.5 represents PM2.5 concentration in micrograms per cubic meter (ug/m^3).
7. DEWP corresponds to Dew Point temperature in degrees Celsius (°C).
8. TEMP refers to Temperature in degrees Celsius (°C).
9. PRES signifies atmospheric Pressure in hectopascals (hPa).
10. cbwd describes Combined wind direction.
11. Iws represents Cumulated wind speed in meters per second (m/s).
12. Is signifies Cumulated hours of snow.
13. Ir indicates Cumulated hours of rain.

Q/A for Task-2

1. What challenges did you encounter while preparing the dataset for preprocessing and EDA analysis?

Ans: While I’m preparing the dataset for preprocessing and EDA I’ve encountered challenges like Null values, unstructured data. For overcoming these, I’ve used the functions by which I can get the right path.

1. Describe the difference observed while modeling with linear and random forest regressors.

Ans: Modeling with a Random Forest Regressor made me to work easier on complex, non-linear relationships and handle various data types effectively. Whereas Ridge Regressor is a linear method that prevents overfitting but gives linear relationships in the data.

1. How did you evaluate the performance of the linear and logistic regressor?

Ans: The performance of the Ridge Regressor and Random Forest Regressor was done using standard regression evaluation metrics. For both models, common metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) were calculated.

1. Did you observe any signs of overfitting during training? If so, how did you handle it?

Ans: I didn’t observe any overfitting of data because the accuracy of test data is more compared to training data.

1. How well did your model perform on the testing set compared to the training set?

Ans: In both the models, testing set of data performed well. As in the both of the models overfitting isn’t present I didn’t had any challenges during evaluating the performance of both of the data.

1. How did you choose the number of clusters for the K-means algorithm? explain why?

Ans: Here, in my evaluation of models using metrics I didn’t use any K-means algorithm because data is good enough to evaluate without any issues. If I had any issue, I would have used cross-validation of K-means to evaluate the data using models.