**PROJECT 1: Air Quality Analysis And Prediction**

Phase 4: Development part 2

**INTRODUCTION** **:**

In an era of rising pollution, understanding and predicting air quality is crucial. Discover why accurate analysis and prediction are essential for public health, environmental sustainability, and effective policy-making.

**Necessary steps to follow**:

**Algorithm Selection:**

Choose a suitable machine learning algorithm based on the nature of your project. Common choices include decision trees, support vector machines, neural networks, and more.

**Linear Regression:**

Use for regression problems when you want to predict a continuous target variable.

**Logistic Regression:**

Suitable for binary classification problems (e.g., spam detection).

Z = spam[‘EmailText’]

y = spam[“Label”]

z\_train, z\_test,y\_train, y\_test = train\_test\_split(z,y,)

**Day level**

To get AQI at day level, the AQI values are averaged over the hours of the day.

Df\_station\_hour = df

df\_station\_day = pd.read\_csv(PATH\_STATION\_DAY)

df\_station\_day = df\_station\_day.merge(df.groupby([“StationId”, “Date”])[“AQI\_calculated”].mean().reset\_index(), on = [“StationId”, “Date”])

df\_station\_day.AQI\_calculated = round(df\_station\_day.AQI\_calculated)

**City level**

To get AQI at city level, the AQI values are averaged over stations of the city.

Df\_city\_hour = pd.read\_csv(PATH\_CITY\_HOUR)

df\_city\_day = pd.read\_csv(PATH\_CITY\_DAY)

df\_city\_hour[“Date”] = pd.to\_datetime(df\_city\_hour.Datetime).dt.date.astype(str)

df\_city\_hour = df\_city\_hour.merge(df.groupby([“City”, “Datetime”])[“AQI\_calculated”].mean().reset\_index(), on = [“City”, “Datetime”]) df\_city\_hour.AQI\_calculated = round(df\_city\_hour.AQI\_calculated)

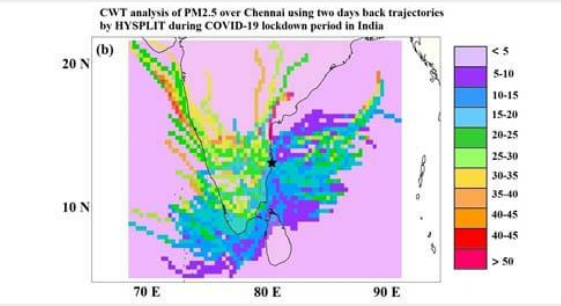
df\_city\_day = df\_city\_day.merge(df\_city\_hour.groupby([“City”, “Date”])[“AQI\_calculated”].mean().reset\_index(), on = [“City”, “Date”])

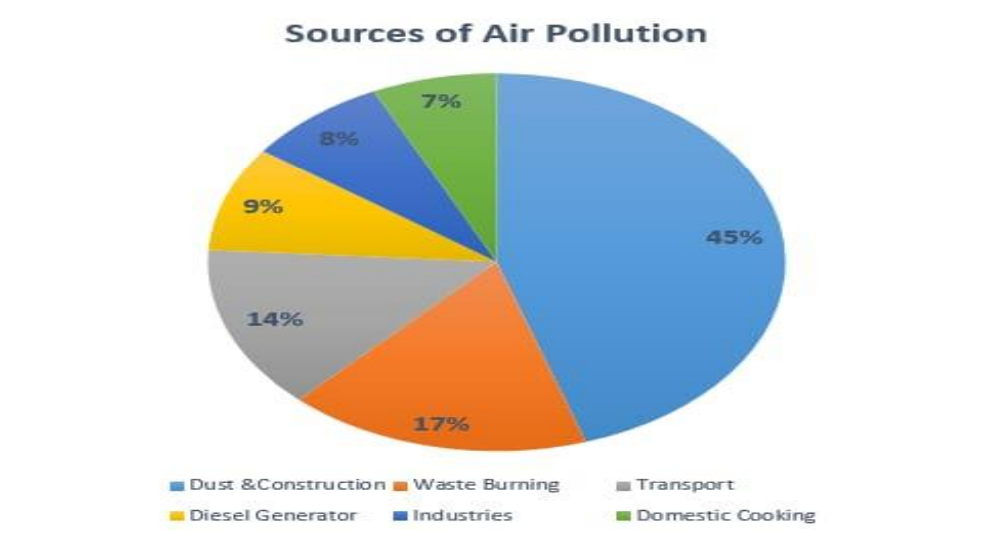
df\_city\_day.AQI\_calculated = round(df\_city\_day.AQI\_calculated)

**Data Visualization**

datacount =sns.countplot(x =“location”,data = tn);

datacount.set\_xticklabels(datacount.get\_xticklabels(), rotation=90);





**Conclusion:**

Air pollution is a global problem; researchers from all around the world are working to discover a solution. To accurately forecast the AQI, machine learning techniques were investigated. The present study assessed the performance of the three best data mining models (SVR, RFR, and CR) for predicting the accurate AQI data in some of India’s most populous and polluted cities. The synthetic minority oversampling technique (SMOTE) was used to equalize the class data to get better and consistent results.