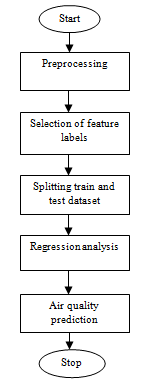
**AIR QUALITY ANALYSIS AND PREDICTIONS IN TAMIL NADU**

** **

**INTRODUCTION:-**

Analyzing air quality and making predictions in Tamil Nadu using advanced data science techniques is a complex and multi-faceted task that involves several steps. Below is an outline of the process, but please note that a full-fledged project would require significant expertise, resources, and access to data sources.

**Data Collection:**

1. Gather historical air quality data from relevant sources. In India, you can obtain this data from organizations like the Central Pollution Control Board (CPCB) or the Tamil Nadu Pollution Control Board (TNPCB).
2. Collect meteorological data, geographical data, and other relevant information that could impact air quality.

**Data Preprocessing:**

1. Clean the data to handle missing values, outliers, and inconsistencies.
2. Aggregate data by location and time to create meaningful time series data.
3. Feature engineering: Create relevant features like weather conditions, traffic density, industrial activities, and population density, which can influence air quality.

**Exploratory Data Analysis (EDA):**

1. Visualize the data to identify trends, patterns, and correlations.

2. Analyze how different pollutants (e.g., PM2.5, PM10, NO2, SO2) vary over time and space.

**Feature Selection and Engineering:**

1. Use domain knowledge to select the most relevant features for modeling.

2. Create new features that may better represent the data and its relationships.

**Model Development:**

1. Choose appropriate machine learning or statistical modeling techniques such as time series forecasting, regression, or deep learning.

2. Split the data into training and testing sets for model evaluation.

3. Train and tune models to predict air quality based on historical data and selected features.

**Model Evaluation:**

1. Use appropriate evaluation metrics (e.g., RMSE, MAE, R-squared) to assess the model's performance.

2. Consider cross-validation to ensure model robustness.

3. Test the model on unseen data to check its predictive capabilities.

Communicate the insights to the stakeholders using data visualization

**Deployment:**

1. Deploy the model in a way that it can make real-time or near-real-time predictions.

2. Set up automated data pipelines to feed new data into the model.

**Monitoring and Maintenance:**

1. Continuously monitor the model's performance and retrain it periodically.

2. Update the model as new data becomes available and the environment changes.

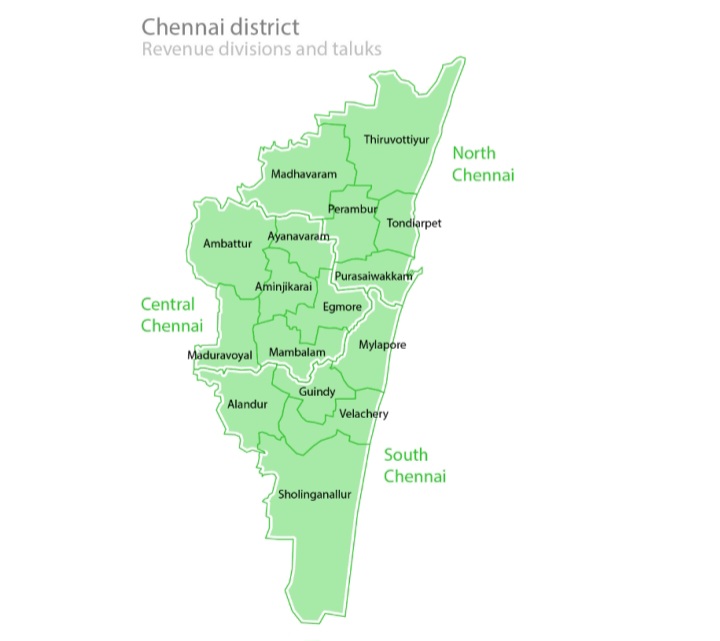
**Communication and Visualization:**

1. Create user-friendly interfaces or dashboards to communicate air quality information to the public and stakeholders.

2. Provide alerts and recommendations when air quality reaches dangerous levels.

**Compliance and Collaboration:**

1 Ensure compliance with environmental regulations and collaborate with government agencies and environmental organizations to improve air quality.



The city's population is 7,088,000. The area is 426 km2

and is the densely populated area. The climatic conditions of

Chennai is dry in summer tropical wet to the months of May

to June and the cool in the month of January with occasional

rainfall. The rivers that flow in Chennai are Kortalaiyar in the

northern part, Cooum rivers and Buckingham canal flows

parallel to the coast and the Otteri Nullah that is east - west

stream.



Alandur is one of the zones of Chennai corporation, and an

urban node in Guindy division in Chennai district in the state

of Tamil Nadu, India. Alandur is the densely populated urban

area in Chennai. It is located at the latitude and longitude of

13.03°N 80.21°E. It has an average elevation of 12 meters (39

feet) from mean sea level (MSL). Alandur had a population of

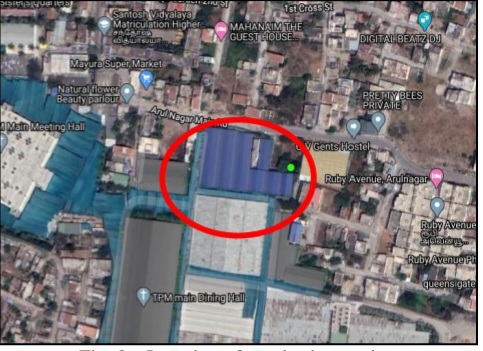
164,430 according to 2011 Census. It has land area of 2

sq.km. It has State highway SH - 48, National highway NH -

45, Kathipara grade fly over and SIDCO industrial estate. This

area was so busy with their vehicular movement and it is one

of the congested areas in Chennai.



**II. METHODS AND MATERIALS**

**Data Sets**

The first and foremost step in modelling is to collect and

group the relevant data, both past data and data from air

quality monitoring. The data is collected from the website of

Central Pollution Control Board. It is very important that the

required data and the factors that cause pollution are collected.

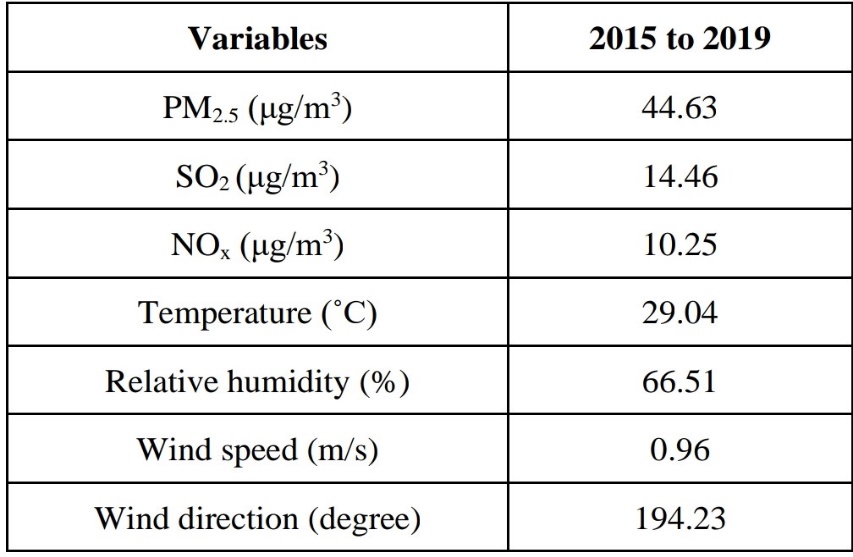
The daily 24-hour average data for five years (2015-2019) is

collected for the following parameters; wind speed, relative

humidity, wind direction, temperature, sulphur dioxide, oxides

of nitrogen, PM2.5. The five year mean of above parameters is

given in table I.

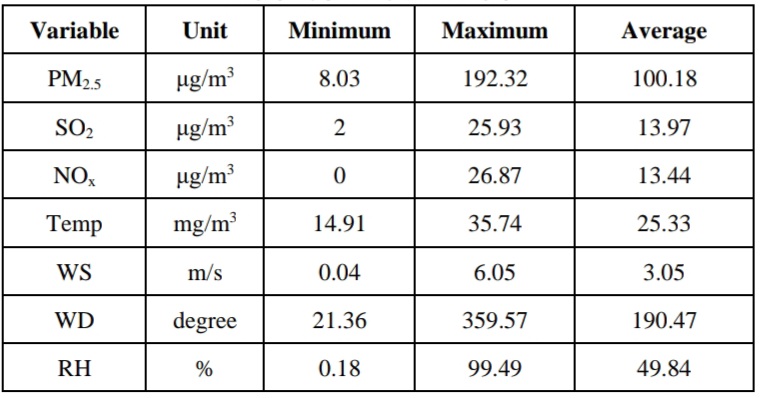


The above table I shows the average value for five years of all

variables and the table II gives the parameters to be used in

prediction, its units, minimum and maximum range of values,

their mean.



**Software**

The Neural Network Toolbox from MATLAB (The

MathWorks Inc. USA) is used for developing prediction

**Performance of Tansig Function**

The following figure 5 appears during the training process.

This graph shows the performance of network versus the

number of epochs. During training the performance of the

network starts from a large value at first and the weights are

altered to have minimum epoch value in the function. In the

graph, the black dashed line represents the best performance

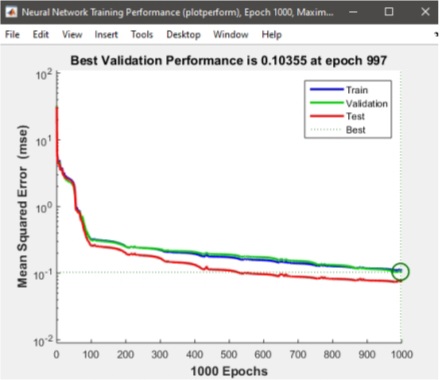
validation of the network. The green line represents the

validation training set, when it intersects with the black line the training process stops. The network performance function

is shown in the figure 5. The best performance is achieved by

the model using tansig transfer function with the minimum

MSE of 0.103.



Regression analysis was performed to investigate the

correlation between the actual and predicted results based on

the value of correlation coefficient, R. The R value of 1

indicates perfect fit between the training data and the

produced results. The regression analysis plots of the network

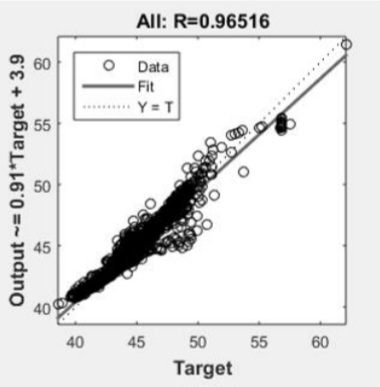
structure were shown in the figure 6. In regression plot, the

solid line indicates the perfect fit, which shows the good

correlation between predicted and target values. The dashed

line in the regression plot shows the best fit produced by the

algorithm.



Form the above figure 6, the regression value is R = 0.965.

The relevancy of the target and the ANN output is given by

the regression plot. R= 0.965 shows that the output of ANN

matches with the target. The relevance between the outputs

and targets were indicated by the Regression (R) value. When

R is 1, precise linear relevance is achieved between targets

and outputs. Similarly, when R is zero, there is no linear

relevance is achieved between targets and outputs. In this

study, the training data show proper relevance between targets

and outputs. Also, the validation and checked outcome gives

R values greater than 0.965.

B. Performance of Purelin Transfer Function

The following figure 7 shows that the best performance is

achieved by the model using purelin transfer function with the

minimum MSE of 0.094.

air quality analysis and predictions in tamil nadu in advanced development datascience with python

Form the above figure 8, the regression value is R = 0.964.

The relevancy of the target and the ANN output is given by

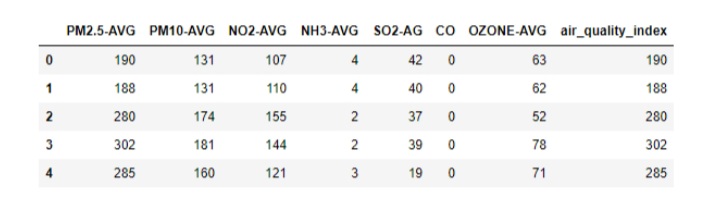
the regression plot. R= 0.964 shows that the output of ANN

matches with the target.

Program

#importing pandas module for data frame  
  
import pandas as pd  
   
# loading dataset and storing in train variable  
  
train=pd.read\_csv('AQI.csv')  
   
# display top 5 data  
train.head()"

Output



Program:

# importing Randomforest

from sklearn.ensemble import AdaBoostRegressor

from sklearn.ensemble import RandomForestRegressor

# creating model

m1 = RandomForestRegressor()

# separating class label and other attributes

train1 = train.drop(['air\_quality\_index'], axis=1)

target = train['air\_quality\_index']

# Fitting the model

m1.fit(train1, target)

'''RandomForestRegressor(bootstrap=True, ccp\_alpha=0.0, criterion='mse',

                      max\_depth=None, max\_features='auto', max\_leaf\_nodes=None,

                      max\_samples=None, min\_impurity\_decrease=0.0,

                      min\_impurity\_split=None, min\_samples\_leaf=1,

                      min\_samples\_split=2, min\_weight\_fraction\_leaf=0.0,

                      n\_estimators=100, n\_jobs=None, oob\_score=False,

                      random\_state=None, verbose=0, warm\_start=False)'''

# calculating the score and the score is 97.96360799890066%

m1.score(train1, target) \* 100

# predicting the model with other values (testing the data)

# so AQI is 123.71

m1.predict([[123, 45, 67, 34, 5, 0, 23]])

# Adaboost model

# importing module

# defining model

m2 = AdaBoostRegressor()

# Fitting the model

m2.fit(train1, target)

'''AdaBoostRegressor(base\_estimator=None, learning\_rate=1.0, loss='linear',

                  n\_estimators=50, random\_state=None)'''

# calculating the score and the score is 96.15377360010211%

m2.score(train1, target)\*100

# predicting the model with other values (testing the data)

# so AQI is 94.42105263

m2.predict([[123, 45, 67, 34, 5, 0, 23]])

Output

RandomForestRegressor score: 97.96%

Predicted AQI using RandomForestRegressor: 123.71

AdaBoostRegressor score: 96.15%

Predicted AQI using AdaBoostRegressor: 94.42

Conclusion:

This helps the stakeholders to understand about the data insights.