



# SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act, 1956)



THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

# Optimal Machine Learning Techniques to Predict Air Quality Index

## GUIDE

Dr. Sumathi  
Asst. Professor-III  
School of Computing

## TEAM

HEMANTH BABU CHAVA	125003105
MANJUNATH P	125003174
PRAVANTH DEVAKI	125003232

# Contents of this Presentation

- *Base Paper*
- *Abstract*
- *Literature Survey*
- *Problem Statement*
- *Objectives*
- *Methodology*
- *Flow diagram of the work*
- *Modules*
- *Work Plan*
- *References*



# Base Paper

Paper Title: Impact of air pollutants on climate change and prediction of air quality index using machine learning models

Journal Name: Environmental Research Volume 239, Part 1, Article 117354

Year of Publication: 2023

Publisher: Elsevier

Indexing: SCI / Scopus

Paper Link: [prediction of air quality index using machine learning models](#)

# Abstract

- ❖ Air pollution demands better monitoring, rising pollution necessitates accurate air quality prediction, crucial for managing environmental quality.
- ❖ Costly limitations of traditional methods: Manual monitoring stations, though existing, are expensive and limited.
- ❖ Machine learning emerges as a powerful tool: Ensemble methods will be used to predict AQI using open-source CPCB data.
- ❖ Scalability and reusability are key: The research seeks a robust ML framework applicable to diverse cities with the CPCB data.

# Literature Survey



S.No	Paper Title	Methodology	Merits	Limitations
1	Air Quality Index prediction using Machine Learning for Ahmedabad City	<p>These are the methods used for the work:</p> <ul style="list-style-type: none"><li>• SARIMA</li><li>• SVM</li><li>• LSTM</li></ul>	<p>Number of data-preprocessing methods are presented to remove the outliers, normalize the datasets, which are taken from different sources (CPCB boards)</p> <p>Can also be expanded to forecast other pollution indices at different levels.</p>	<p>Lots of missing values are present in the dataset of Ahmedabad city.</p>
2	Air Quality Index prediction using Machine Learning Algorithms ~International Journal of computer Applications Technology and Research	<p>These are the methods used for the work:</p> <ul style="list-style-type: none"><li>• ARIMA</li><li>• Auto Regression</li><li>• Linear Regression</li></ul>	<p>Good prediction and Time Series Analysis was also used for recognition of future data points and air pollution prediction</p>	<p>Not able to show expected output as the data is not in sequence</p> <p>The error is high which they are working to overcome in near future.</p>
3	Indian Air Quality Prediction And Analysis using Machine Learning	<p>The model used for the work is:</p> <ul style="list-style-type: none"><li>• Naïve forest</li><li>• Linear regression</li><li>• Gradient Boosting Algorithms</li></ul>	<p>Parameter reducing formulations for better performance than standard regression models</p>	<p>Low Accuracy</p>

S.No	Paper Title	Methodology	Merits	Limitations
4	Air Pollution Prediction Using ML techniques. An approach to replace existing monitoring stations With virtual motoring stations	The model used for the work is: <ul style="list-style-type: none"> <li>• Ridge regression</li> <li>• SVR</li> <li>• Random Forest</li> <li>• Xtreme Gradient Boosting</li> </ul>	Hyperparameter tuning developed technique can be transferred to any location where pollutant prediction is required  benefit from incorporating other neural networks such as CNN-LSTM (Convolution Neuron Network - Long Short Term Memory) in capturing temporal dependencies and patterns in data	The limitation of this study is that the forecasting of pollutant concentration is not possible as the data from other monitoring stations is required for prediction1
5	Detection and Prediction of air Pollution using Machine Learning models	The model used for the work is: <ul style="list-style-type: none"> <li>• Logistic regression</li> <li>• Auto regression</li> </ul>	Mean accuracy and standard deviation accuracy to be 0.998859 and 0.000612 respectively.	No gaseous pollutants were considered Taken very less data
6	Air Quality prediction by using ML models: a case study on the Indian coastal city Visakhapatnam	The model used for the work is: <ul style="list-style-type: none"> <li>• Random Forest</li> <li>• Light GBM</li> <li>• Cat Boost</li> <li>• Adaptive boosting</li> </ul>	Cat Boost model yielded high prediction accuracy (0.9998) and low RMSE (0.76). Cat Boost incorporates parameters that help mitigate overfitting in datasets	Performance needs to be validated under diverse air quality conditions The covid lockdown has affected the AQI levels

# Problem Statement

- To enhance accuracy in AQI forecasting to facilitate informed decision-making for environmental regulation.
- To provide timely warnings and precautions to the public regarding air quality levels.
- To Enable proactive measures for mitigating the impact of air pollution on public health.

# Objectives

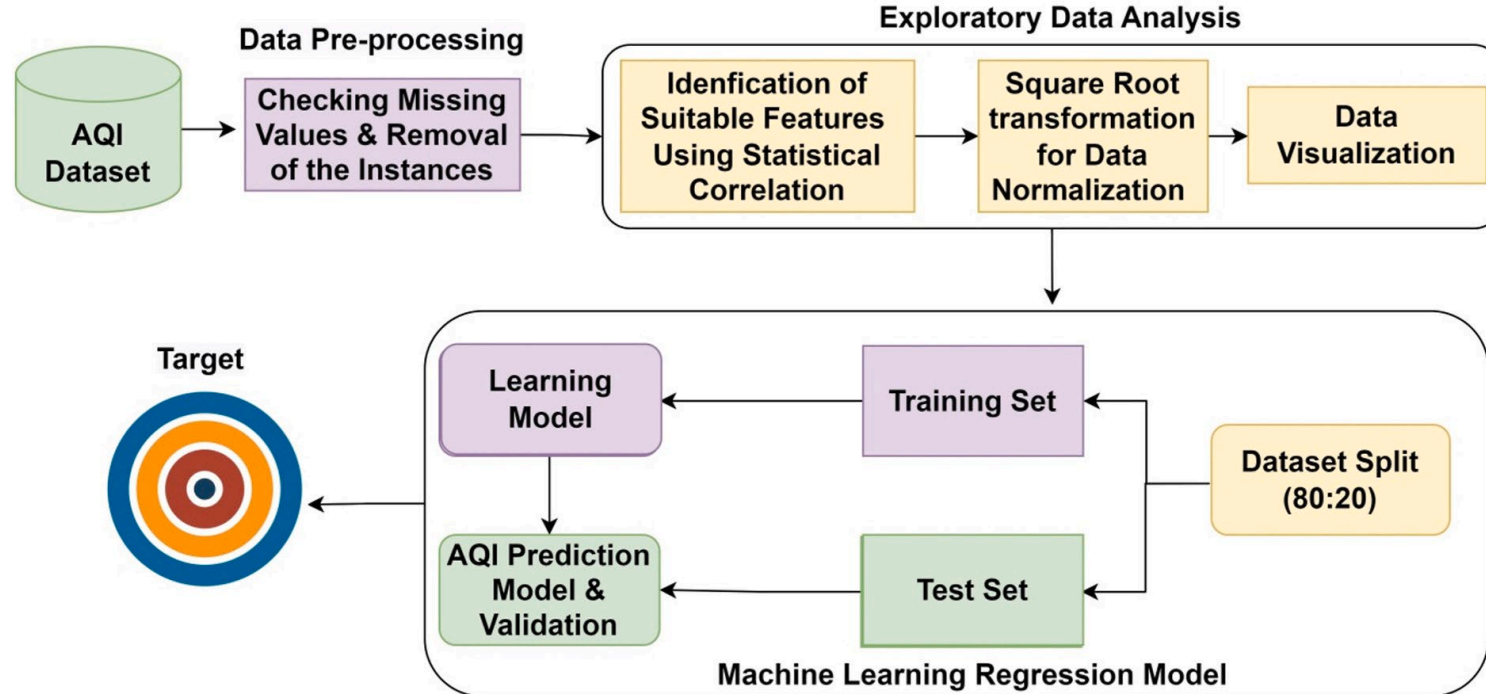
- Implement advanced machine learning techniques on CPCB data to enhance accuracy and reliability in AQI prediction.
- Evaluate the performance of developed models against existing methods to validate effectiveness.
- Demonstrate the potential of the proposed approach to optimize environmental management strategies and safeguard public health against air pollution hazards.



# Methodology

- We chose Tirupati monitoring station to work with and apply the base paper.
- Initially, data is processed to remove noises and handle missing values. Then the data is explored for patterns and a comprehensive analysis is made using apt tools like correlation matrix.
- Data is transformed and standardized before fitting it to the model for training and testing.
- To ensure the robustness of the models, we used 5-fold cross-validation, computed MAE, MSE, RMSE, and R2 scores for each fold, and reported the mean and standard deviation of each metric across all folds. When a model has a higher R2 score and lower MAE and MSE scores, it is generally regarded as performing better
- Models and methods - Random Forest, XGBoost, Bagging Regressor, LGBM Regressor
- Python libraries and modules - numpy, pandas, seaborn, sklearn.model\_selection, GridSearchCV

# System Architecture/Flow diagram of the work



# Dataset

1. Source: Central Pollution Control Board (CPCB), India
2. Content: Continuous ambient air quality measurements
3. Parameters: PM2.5, PM10, NO, NO2, NO<sub>x</sub>, NH3, CO, SO2, Benzene, Toluene, Ozone, RH, Xylene, BP,AT, RF, Temp, SR
4. Time period: From 01-01-2017 to 31-12-2022
5. Location: Tirumala, Tirupati, Andhra Pradesh, India
6. Data format: A comma-separated value (CSV) file
7. Potential uses: The dataset can be used to train machine learning models to predict air quality index (AQI), analyze air quality trends, and assess the impact of air pollution on public health.

# Modules

- Module 1: Data Preprocessing, Exploratory Data Analysis, Data transformation
- Module 2: Data spilt and fitting the models
- Module 3: Comparative Analysis, Performance valuation, Result and Discussion

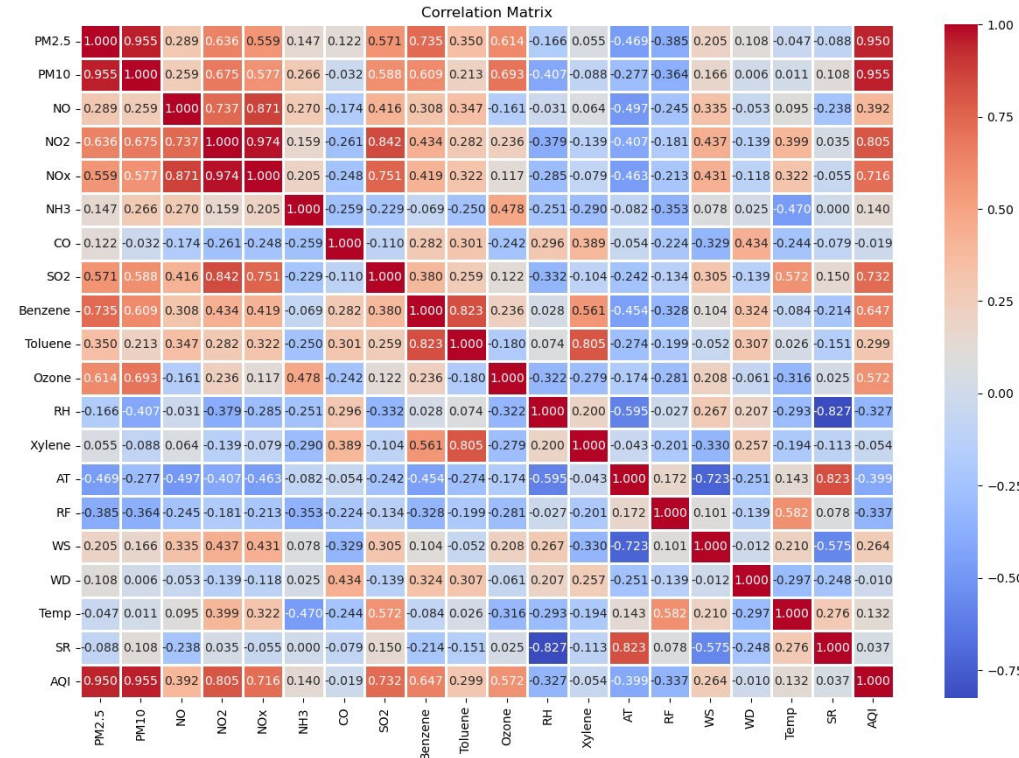
# Module 1 – Data Preprocessing

- The datasets were obtained from the Central Pollution Control Board's - Central Control Room for Air (CPCB-CRR). They include measurements of air pollutants and meteorological parameters.
- Raw dataset contained a total of 2191 observations spanning from 01-01-2017 to 31-12-2022, which included 20 variables ( 13 Air pollutant attributes, 06 Meteorological factors, and 1 AQI )
- Removed rows of instances with null values for the target variable ( AQI ) and BP (Barometric Pressure) parameter was removed since it had only 47 instances of recorded data and not helpful.
- **Handling Missing Values:** Used forward fill (previous day values) for null and 'None' values

# Module 1 – EDA

The pollutants with the highest heat map values had a significant impact on AQI predictions

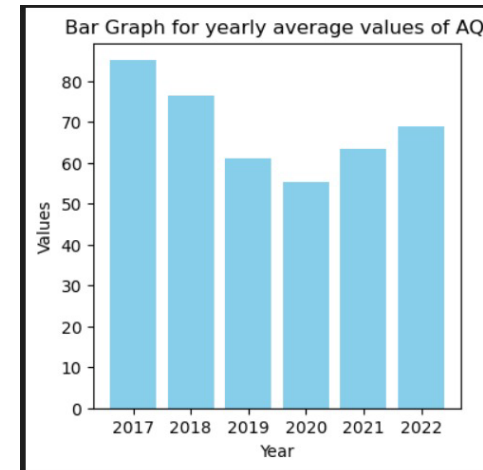
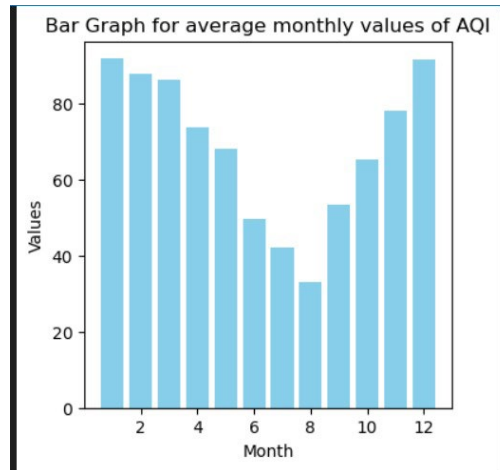
- When compared to the negative or inversion correlation, the positive correlation plays a major role in AQI prediction.
- With a correlation coefficient of 0.95, the highest among all parameters, the relationship between PM10 and AQI is significant, indicating that PM10 and PM2.5 plays a crucial role in determining the AQI.



Heat Map – Input parameters with AQI

# Module 1 – EDA

- The results show that AQI levels were extremely high in Dec and Jan. Seasonal variation may be the reason for the highest level of AQI in these months. These months fall under the winter season when temperatures are lower and more mist formation is observed. Because of the presence of moist air in the atmosphere, this could result in the formation of a temperature inversion and the pollutants emitted from the source get retained in the atmosphere.



Seasonal and annual variation of AQI

# Module 1 – Data Transformation

- The most commonly used data transformation techniques are Box-Cox transformation, log transformation and square root transformation.
- Log transformation is a powerful tool to reduce skewness in data. It works by compressing larger values and spreading out smaller ones.
- Log transformation was used to change the data to make it more normal.

## Before transformation

## After transformation

Skew Kurtosis

Skew Kurtosis

PM2.5	1.08	0.96
PM10	0.8	1.43
NO	1.49	2.74
NO2	2.06	9.4
NOx	1.59	5.04
NH3	1.11	0.62
CO	5.88	58.01
SO2	1.81	2.9
Benzene	2.87	19.29
Toluene	3.36	18.35
Ozone	1.12	1.17
RH	-0.66	-0.01
Xylene	6.09	61.63
AT	-0.01	-0.74
RF	10.13	103.69
WS	1.03	1.34
WD	-0.06	-0.54
Temp	4.31	27.93
SR	1.04	2.66
AQI	1.53	3.96

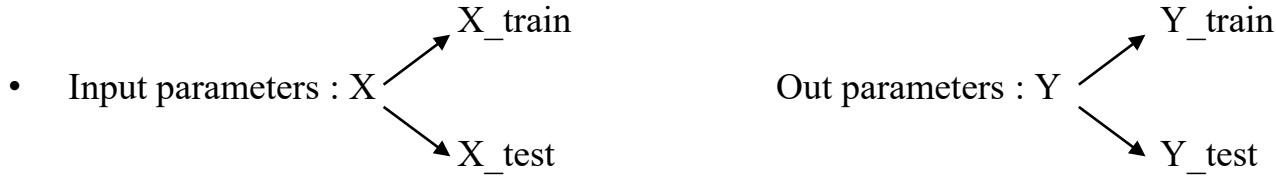
PM2.5	-0.27	0.07
PM10	-0.36	-0.38
NO	-0.44	1.05
NO2	-0.21	-0.13
NOx	-0.97	5.09
NH3	-0.99	6.17
CO	-4.66	26.7
SO2	0.26	0.35
Benzene	-4.01	22.12
Toluene	-3.78	26.74
Ozone	-0.32	0.75
RH	-1.08	1.08
Xylene	-2.24	6.13
AT	-0.21	-0.75
RF	1.13	-0.48
WS	-0.31	-0.29
WD	-0.98	1.68
Temp	3.21	16.23
SR	-1.4	2.0
AQI	0.29	-0.59



# Module 2 – Model Development:

## Data Split (80:20) and Standardization

- We assign the target variable AQI to the variable  $y$  and all other features except AQI to the variable  $X$  as explanatory variables.



- We utilize the `sklearn.model_selection` module to split the data into two distinct sets: training and testing, using the `train_test_split()` function. The dataset is split into 80% for training and 20% for testing.
- An inadequately selected split can lead to the model being overfit or underfit, which can result in inadequate predictions on unseen data. Therefore, it is crucial to select the split carefully and assess the model's performance on the test data to ensure that it can generalize well to new data.
- To normalize the features of the training and testing sets, we use the `StandardScaler()` function from the `sklearn`

# Module 2–RandomForest algorithm

- Random forest regression employs a collection of decision trees trained on random feature subsets and data instances. Implemented via scikit-learn, we use classes like RandomForestRegressor for model creation, GridSearchCV for hyperparameter optimization, and k-fold for cross-validation.
- The model is fitted to the training data using the fit() method of the GridSearchCV object.
- The RandomForestRegressor object is initialized with preset hyperparameters: max\_depth of 15, max\_features set to 'auto', min\_samples\_leaf at 1, min\_samples\_split of 2, and n\_estimators of 100, alongside a designated random\_state.

```
▶ # creating model: # defined hyperparameters are 'max_depth': 15, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2
randFor = RandomForestRegressor(n_estimators=100,max_depth=15,max_features='auto',min_samples_leaf=1,min_samples_split=2)
# Fitting the model
randFor.fit(X_train,Y_train)
```

```
] RandomForestRegressor(max_depth=15)
```

```
▶ randFor.score(X_train,Y_train) * 100
```

```
] 98.4699814378327
```

# Module 3



- Comparative analysis on performance of the ensemble methods and implemented models.
- AQI predictions made using machine learning models
- Normal distribution of datasets, Residual plots, Residual Histograms of different predictive models
- Comparison of performance by different ML models in Training and Validation/Testing
- Result discussion and Conclusion

## WORK PLAN

Review Period	Work Particulars	% of Work Completed
Zeroth Review	<ul style="list-style-type: none"><li>• Base paper Confirmation</li><li>• Problem identification</li><li>• Literature Review</li></ul>	-
First Review	<ul style="list-style-type: none"><li>• System architecture design</li><li>• Module identification</li><li>• Proposed Algorithm implementation</li></ul>	40%
Second Review	<ul style="list-style-type: none"><li>• Proposed algorithm implementation</li><li>• Comparative Analysis</li><li>• Providing security features</li></ul>	100%

# References



- [Gokulan Ravindiran, Sivarethinamohan Rajamanickam, Karthick Kanagarathinam, Gasim Hayder, Gorti Janardhan, Priya Arunkumar, Sivakumar Arunachalam, Abeer A. AlObaid, Ismail Warad, Senthil Kumar Muniasamy: “ Impact of air pollutants on climate change and prediction of air quality index using machine learning models” ,Environmental Research, Volume 239, Part 1, 2023, 117354](#)
- [Gokulan Ravindiran, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, Christian Sonne,” Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam”, Chemosphere, Volume 338, 2023, 139518.](#)
- [C R, Aditya & Deshmukh, Chandana & K, Nayana & Gandhi, Praveen & astu, Vidyav. \(2018\).” Detection and Prediction of Air Pollution using Machine Learning Models”. International Journal of Engineering Trends and Technology. 59. 204-207.](#)
- [A. Samad, S. Garuda, U. Vogt, B. Yang, “Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations”, Atmospheric Environment, Volume 310, 2023, 119987](#)
- [Nilesh N. Maltare, Safvan Vahora, “Air Quality Index prediction using machine learning for Ahmedabad city”, Digital Chemical Engineering, Volume 7, 2023, 100093](#)

Thank you