Concepts and Technologies of AI

5CS037

Student Name: Aaditya Acharya

Student ID: 2510333

Abstract

The goal of this report is to predict Air Quality Index (AQI)using the regression techniques. 29531 data from multiple cities with 16 features of Indian Air Quality Dataset which aligns with the United Nation Sustainable Development Goal 11: Sustainable Cities and communities and United Nation Sustainable Development Goal 3: Good Health and Well Being. The methodology includes the EDA, Neural Network (MLP), Classical ML (Linear Regression, Random Fores), Hyperparameter Optimization (Randomized Search CV), Feature Selection using embedded method (Decision Tree) and final model comparison. For the evaluation of the model RMSE R-Squared and MAE were evaluated. The random forest model demonstrated the best performance with test RMSE 40.03 and R-squared 0.912. Key insights include the PM2.5 and PM10 as AQI is the target variable.

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU | ing

Table of Contents

Table of Figures

1 Introduction

1.1 Problem Statement

Air Quality is important for the health and environment sustainability. Predicting Air Quality requires the complex analysis. The Goal of this project is to develop a machine learning model that can predict the AQI value. The main goal is to give the early warning for the air quality which also helps for the planning and policy making of the air quality which can affect the health conditions.

1.2 Dataset

Dataset Name: Indian Air Quality Dataset (2015-2020)

Source: The dataset was obtained from the Kaggle which was made publicly available at Central Pollution Control Board of India.

Time Period: It contains the dataset of 2015- 2020

Records: It contains 29,531 datasets of different cities of India.

Features: 16 features including calculated AQI value.

Features	Description
City	Name of the city from where measurement as taken.
Date	Date of measurement
PM2.5	Particulate Matter
PM10	Particulate Matter
NO	Nitric Oxide concentration
NO2	Nitrogen Dioxide concentration
Nox	Nitrogen Oxide concentration
NH3	Ammonia concentration
CO	Carbon Monoxide concentration
SO2	Sulphur Dioxide concentration
O3	Ozone concentration
Benzene	Benzene concentration
Toluene	Toluene concentration
Xylene	Xylene concentration
AQI	Air Quality Index Value
AQI_Bucket	Categorical Classification

The Dataset aligns with the **SDG3: Good Health and Well Being.** The project directly supports the Sustainable Development Goal 3 as the poor quality of the air can cause the serious disease for the people. Also it aligns with **SDG11: Sustainable Cities And Communities** as predicting the AQI can assist the authorities to make plans as Air Quality can directly affect the quality of the cities**.**

1.3 Objective

The primary objectives of this model are:

- Develop Multiple models that can predict the Air Quality Index (AQI).
- Perform the comprehensive data preprocessing.
- Compare the model using the standard evaluation metrices.
- Optimize the models through the hyperparameter tuning.
- Identification of important features affecting the quality of the air.

1.4 Research Questions

The project seeks to answer the following questions:

- Can Machine Learning models predict the accurate AQI values based on pollutant concentration?
- Which pollutants is the most significant contributor?
- How different regression algorithm differ from each other ?
- What is the impact of hyperparameter optimization and feature selection in model performance?

2 Methodology

2.1 Data Pre Processing

For the preprocessing of the data different preprocessing steps was performed. Initially data contains 29,531 data with different missing values, etc. tor the cleaning the data missing data was filled with the city based mean value and those data which are still missing are managed with date wise mean value.

2.2 Exploratory Data Analysis (EDA)

For the exploratory data analysis different EDA techniques was performed along with the descriptive statistics.

AQI Distribution

Histogram was generated for the target AQI.

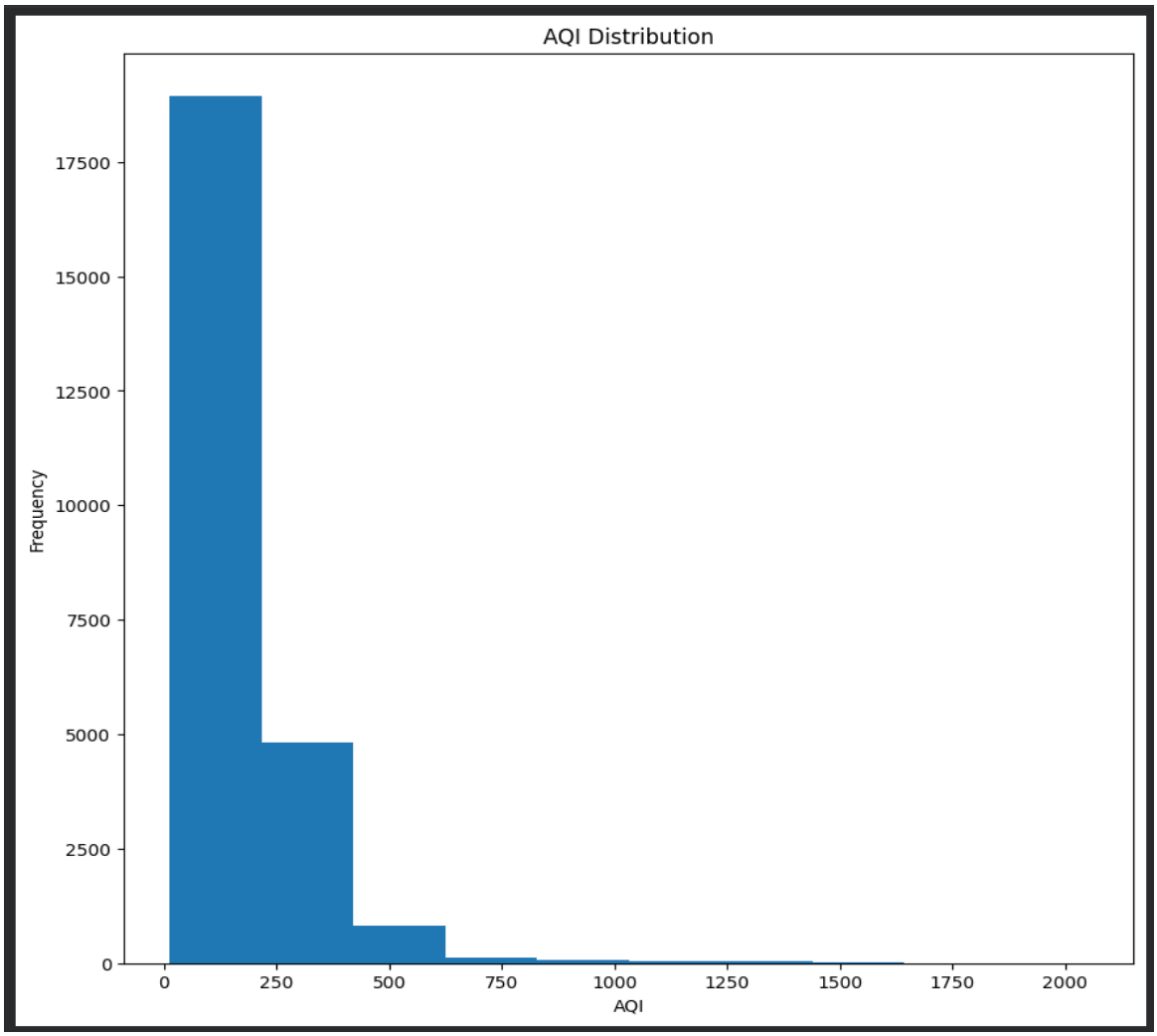

Figure 1 histogram of target variable

Correlation Analysis

Correlation analysis shows the Relation with the target variable.

```
PM2.5    0.656089
PM10     0.528956
NO2      0.535185
CO       0.677250
SO2      0.485447
O3       0.196768
Name: AQI, dtype: float64
```

Figure 2 correlation of target variable with other features

CO shows the strong positive correlation with the target variable whereas O3 shows the weak positive correlation with the target variable.

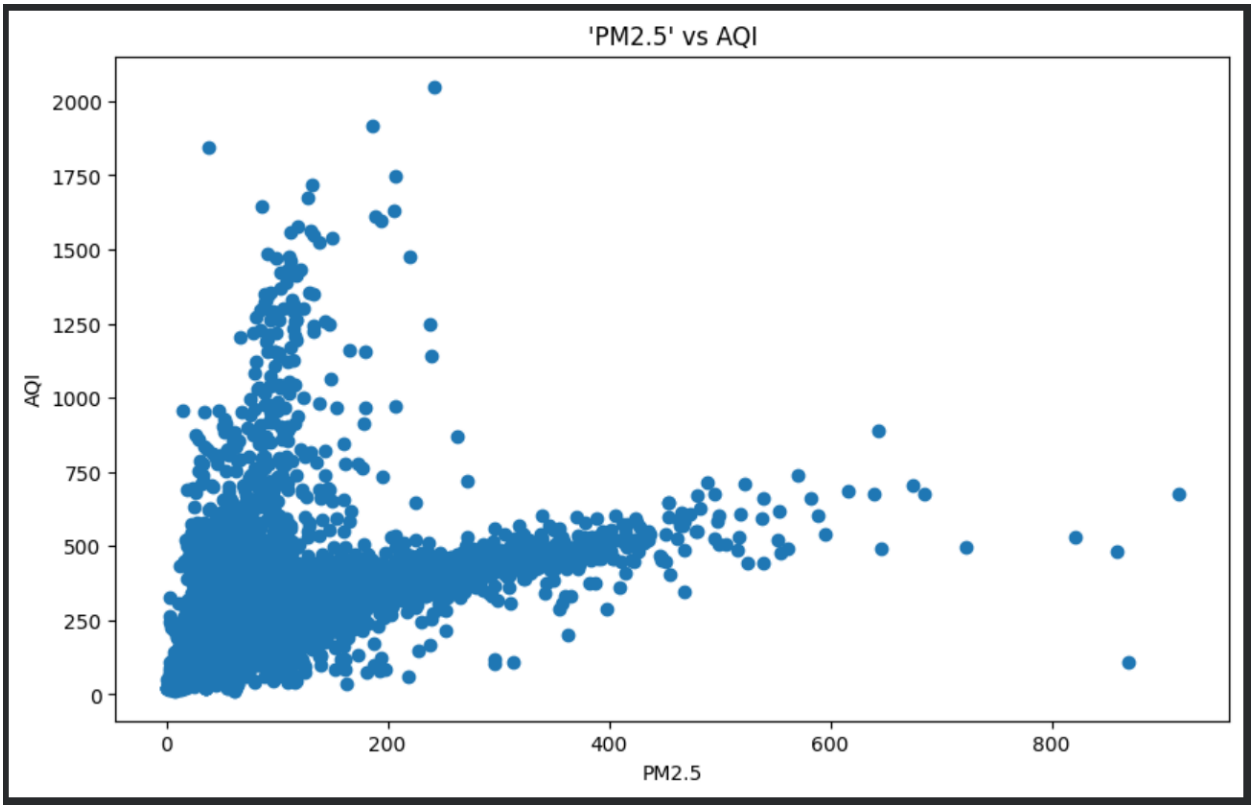

Figure 3 Scatter plot between PM2.5 and AQI

2.3 Model Building

Feature Selection

- The City Date and AQI bucked was removed before training the model as it is categorical data.
- For the training and testing the dataset was split into training(80%) and testing(20%).
- Standard scaler was applied as standardization is crucial for the algorithms like MLP and Linear Regression.

Multi-Layer Perception (MLP)

MlP is a neural network with different hidden layers, activation functions, etc. which is capable of learning complex non-linear patterns in the data.

The model was configured using the following parameters:

- Hidden Layers: (64,32) neurons.
- Activation Function: ReLU
- Solver: Adam Optimizer
- Regularization: alpha(0.001)
- Learning rate: adaptive
- Maximum iterations: 300
- Early stopping: True to prevent the overfitting

Loss Function: Mean Squared Error (MSE) standard for Regression Task.

Optimizer: Adam Optimizer with adaptive learning rate for each parameter.

Linear Regression

Linear Regression is a linear regression model that fits the relationship between the features and target variables. To build the model standard scaler data was used in LinearRegression to train and predict the model in which 80% was used for training whereas 20% was used for the testing of the data.

Random Forest Regressor

Random Forest is a ensemble learning methods that construct the multiple decision trees. To built the model RandomForestRegressor was used to fit the training dataset and predictions for the training and testing was done in which 80% was used for training whereas 20% was used for the testing of the data.

2.4 Model Evaluation

The model performance was evaluated using the following metrices R^2, MAE,

RMSE:

Multi-Layer Perception (MLP)

Metric	Training Performance	Testing Performance
MSE	2099.79	2326.73
MAE	23.60	24.14
RMSE	45.82	48.24
R^2 Score	0.90	0.87

Linear Regression

Metric	Training Performance	Testing Performance
MSE	2799.68	3465.42
MAE	29.42	30.66
RMSE	52.91	58.87
R^2 Score	0.86	0.81

Random Forest Regressor

Metric	Training Performance	Testing Performance
MSE	282.82	1682.23
MAE	7.92	20.90
RMSE	16.82	41.02
R^2 Score	0.99	0.91

Initial Model Comparison

Random Forest have slightly lower test RMSE 41.02 and higher R^2 0.91 compared to Linear Regression RMSE 58.87 and R^2 0.81. However Random forest suggest slightly overfitting as R^2 is 0.99 on training and 0.91 on testing.

2.5 Hyperparameter Optimization

RandomizedSearchCV was used in both linear and Random Forest models. This approach explores the parameters by randomly sampling the combinations, maintaining the computational cost.

Linear Regression

Linear Regression has minimal hyperparameters as the primary consideration is weather to include fit intercept or not.

- fit_intercept : True (it allows model to learn bias term)

Random Forest

Random forest have different hyperparameters to improve the performance of the model, Parameters includes the number of estimators, maximum depth, minimum samples split and leaf. The search identified the best hyperparameters for the model.

- n_estimators: 200 (number of trees in the forest)
- max_depth: 20 (maximum depth of the tree)
- min_samples_split: 5 (minimum sample required to split internal node)
- min_samples_leaf: 1 (minimum sample required at leaf node)
- max_features: ‘log2’ (number of features required for each split)

Cross-Validation Scores

5-Fold Cross-Validation was done during the hyperparameter search. The best cv achieved is 0.90 for random forest and 0.85 for linear regression.

2.6 Feature Selection

Feature Selection includes selecting the top features that best suits predicting the target variable. Embedded method Decision Tree was used to select the features in which built in method feature_importances_ was used. Features were ranked by importance and top 8 features were selected for model training.

- 'PM2.5'
- 'CO'
- 'NO'
- 'PM10'
- 'NOx'
- 'O3'
- 'Toluene'
- 'NH3'

3 Results And Conclusion

Model	Features	MAE	CV Score	RMSE	R^2
Linear Regression	Selected (8)	20.52	0.854	60.51	0.800
Random Forest	Selected (8)	30.94	0.908	40.03	0.912

3.1 Key Findings

The model evaluated on test data shows:

- Random forest achieved 40.03 RMSE which is lower than of linear regression 60.51.
- There is 0.912 R^2 OF Random Forest compared to 0.800.
- CV score of Linear regression is slightly lower 0.854 than Random Forest 0.908.

3.2 Final Model

The final model based on Random forest was more effective predicting the target variable. Random forest is good on non-linear dataset than the linear regression. In Random Forest MAE, CV score and R^2 is higher and RMSE is lower which is good in case of CV Score, RMSE and R^2 but in case of MAE linear regression is good as random forest has higher value.

```
Final Model Comparison Table:
               Model  Features Used  CV Score    MAE  Test RMSE  Test R²
0      Random Forest              8     0.908  30.94      40.10    0.912
1  Linear Regression              8     0.854  20.52      60.51    0.800
```

Figure 4 Final Model

Before optimization Test R^2 was 0.91 and in final model test R^2 was 0.912 the improvement is not seen but there is successful reduction of overfitting as initial train R^2 was 0.99.

3.3 Challenges

The challenges we faced are:

- Quality of the dataset is not good as it contains missing value in target variable and there is the increase of potential bias.
- There is overfitting in the complex model
- There is lack of external validation

3.4 Future Work

To make the model better:

- More advanced ML algorithms can be used to make model more better.
- Testing on external dataset can be done.

4 Discussion

4.1 Model Performance

A random Forest model's Test R^2 shows the excellent prediction capability. While Linear Regression is less accurate which still shows the good linear relationship for the AQI value. This shows that while non-linear effects are more important, entirely the problem will not be non-linear. The model could be more efficient but not perfect.

4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning and feature selection helped lot in model performance.

Hyperparameter helps in the reduce of the overfitting as the initial train R^2 OF Random Forest is 0.99 after the optimization the final model R^2 was 0.91 as it reduces the train test performance gap. Also the CV score of 0.908n shows the consistent performance across different data indicating generalization.

Feature selection was performed using the Embedded Method (Decision tree) to find the most important features affecting the quality of the air. The feature was reduced form 12 to 8 which gives different advantages as focusing on the important features simplifies the model and gives more clear AQI values. With the less feature there is less chance of model overfitting and it reduces computational requirement for the model fit and prediction.

4.3 Interpretation of results

The chosen features and model performed in a consistent manner with expectations. The Multi-Layer Perceptron (MLP) demonstrated strong and consistent performance, achieving a testing R^2 of 0.912. The Random Forest shows the good testing R^2 of 0.8 in the final model as earlier model shows 0.99 R^2 on Training dataset which shows the prevention of overfitting as Hyperparameter Tuning and Feature selection helped to prevent the overfitting.

4.4 Limitations

- Initially the data contains 29,531 AQI data which is good but later the size decreases to 24850 after the cleaning of the as target variable is missing in many rows.
- The model was only trained on Indian cities, on other region performance is untested.

4.5 Suggestions for Future Research

- Perform testing on more independent dataset.
- Use of larger dataset and increase the quality of the dataset.
- Apply more better feature selection techniques.
- More advanced model training can be done for better non-linear relationship

5 References

Nation, U., n.d. *Sustainable Cities and communities*. [Online]
Available at: https://sdgs.un.org/goals/goal11
[Accessed 2026].

United Nations, n.d. *Good Health and Well Being*. [Online]
Available at: https://sdgs.un.org/goals/goal3
[Accessed 2026].

Vopani, 2020. *Air Quality Data in India (2015 - 2020)*. [Online]
Available at: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india/data
[Accessed 2026].