Concepts and Technologies of AI

5CS037

Student Name: Aaditya Acharya

Student ID: 2510333

Abstract

The main goal 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 model Random Forest demonstrated the 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 significant for the health and environment sustainability. Predicting Air Quality requires the complex analysis. The main Goal is to make a model which predicts 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.
Date	Date of measurement
PM2.5	Particulate
PM10	Particulate Matter
NO	Nitric Oxide
NO2	Nitrogen Dioxide
Nox	Nitrogen Oxide
NH3	Ammonia
CO	Carbon Monoxide
SO2	Sulphur Dioxide
O3	Ozone
Benzene	Benzene
Toluene	Toluene
Xylene	Xylene
AQI	Air quality Value
AQI_Bucket	Categorical Classification

This Dataset is aligned 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

- Make Multiple models that can predict the Air Quality Index.
- To Perform data preprocessing.
- Compare model using the standard evaluation metrices.
- To Optimize the models with the hyperparameter tuning.
- Identification of important features affecting the quality of the air.

1.4 Research Questions

The project answers the following questions:

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

2 Methodology

2.1 Data Pre Processing

For preprocessing of data different 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 analysis different EDA techniques was done along with the descriptive statistics.

AQI Distribution

Histogram was generated for the target AQI.

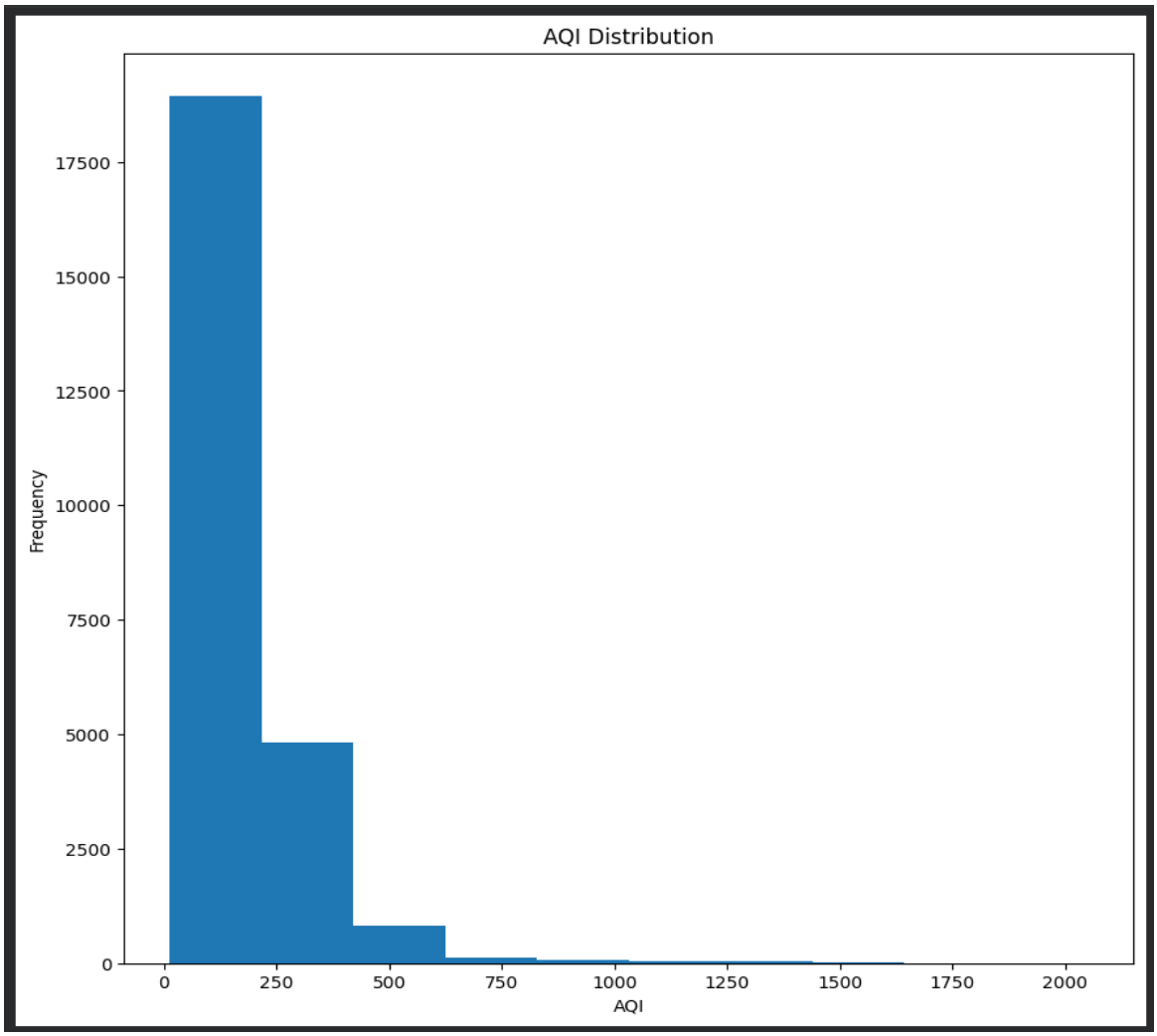


Figure 1 histogram of target variable

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

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

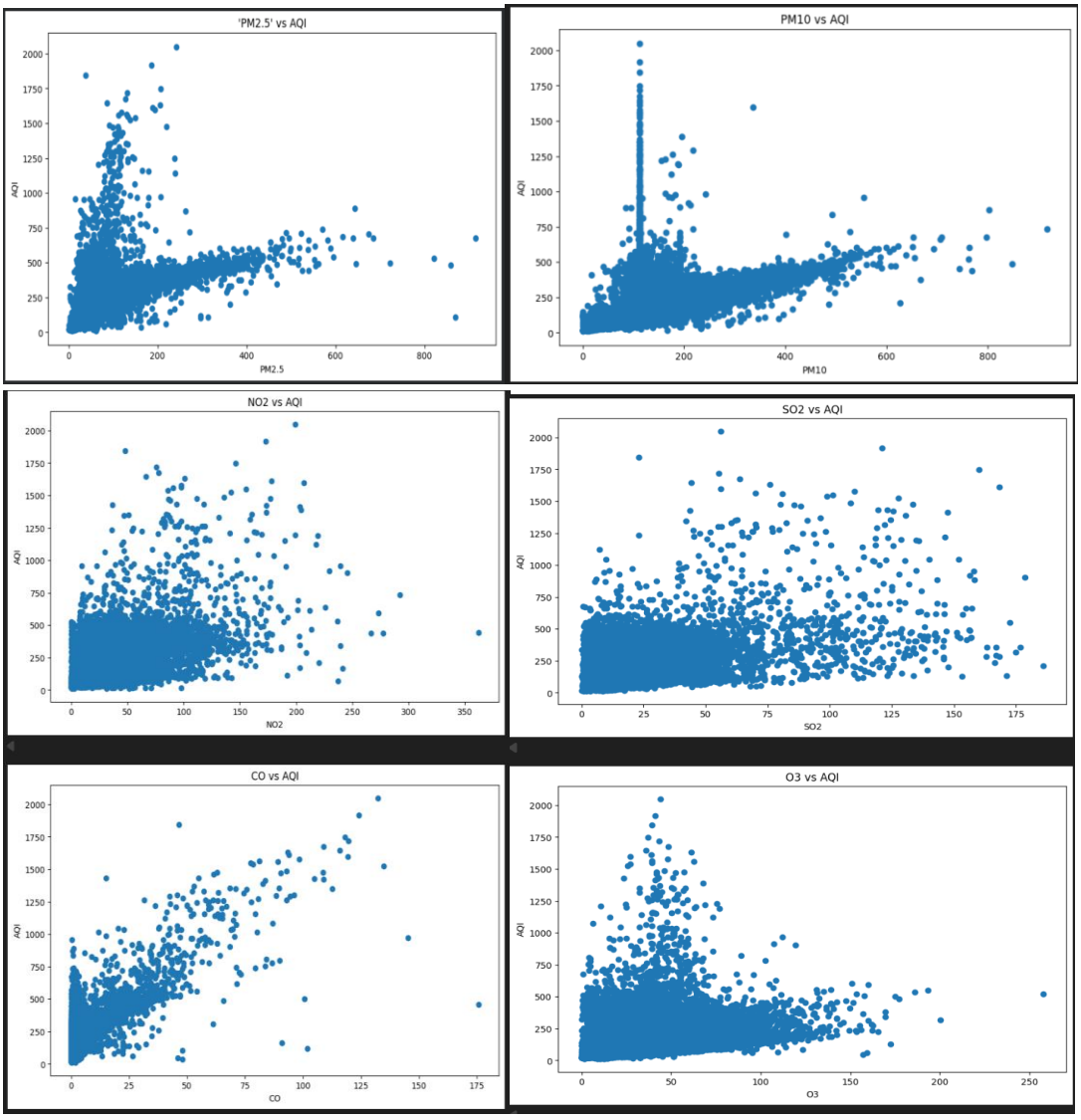


Figure 3 Scatter plot between numerical columns and AQI

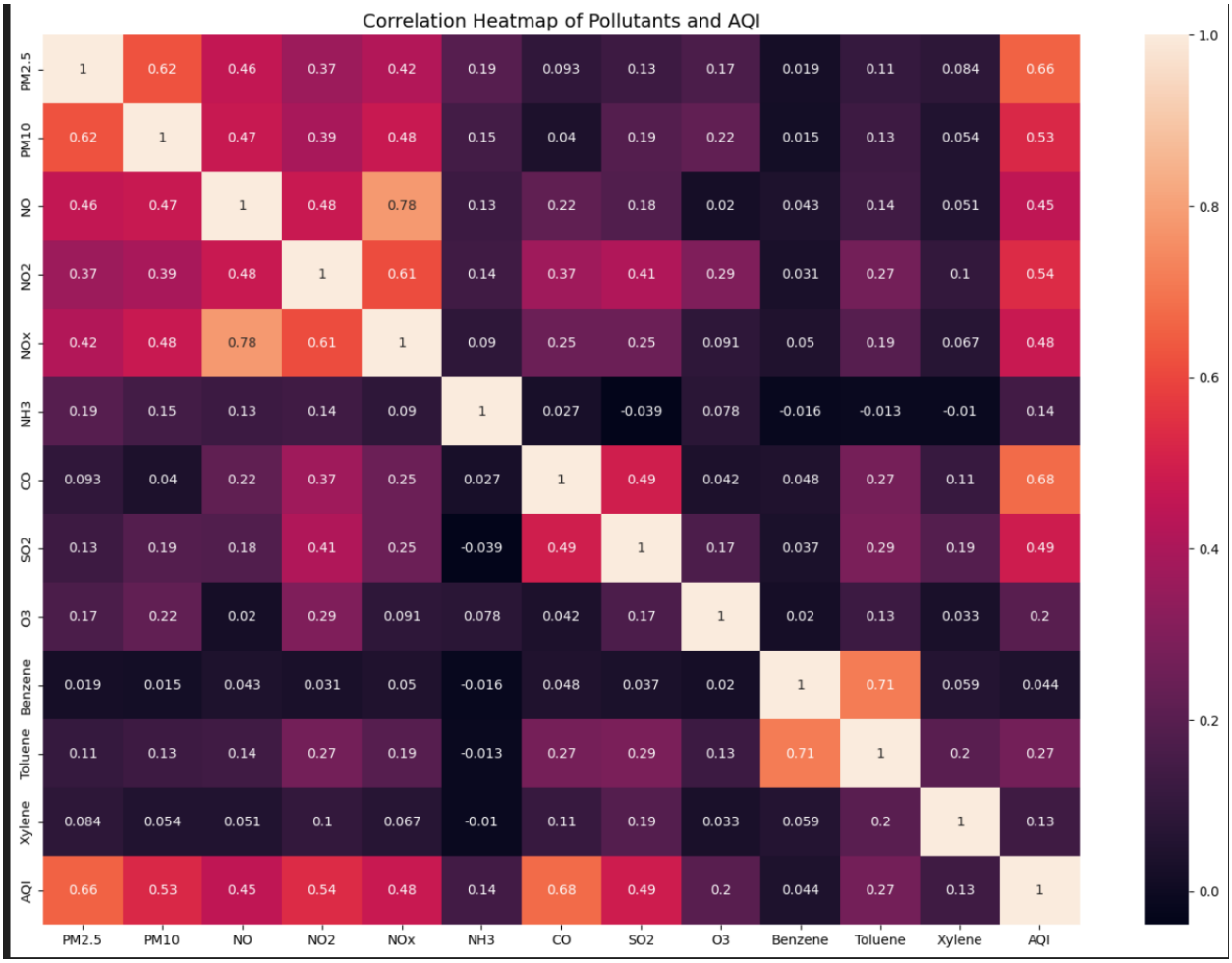
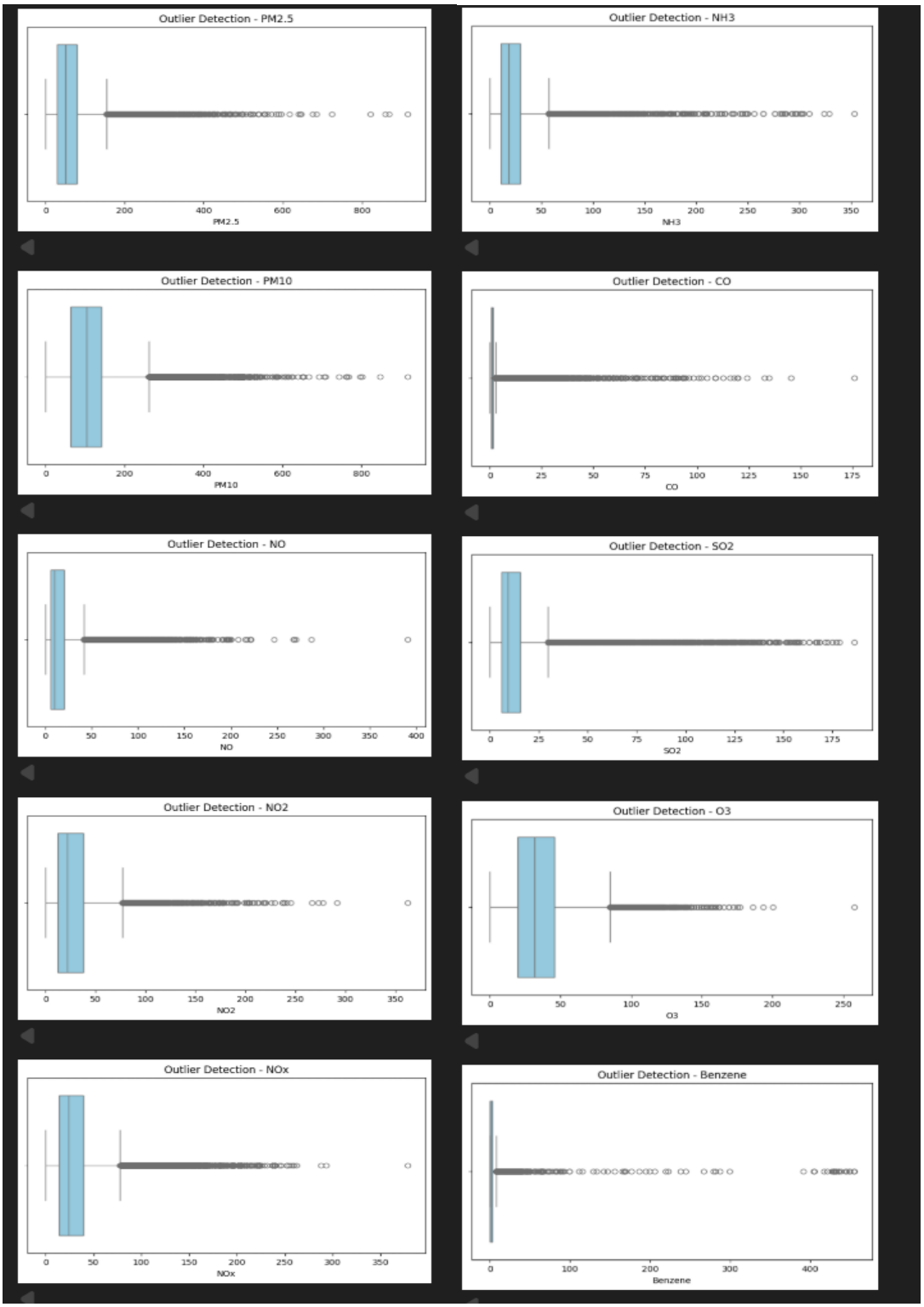


Figure 4 Heatmap

Outlier Detection

Outlier was detected and later it was managed using clip.



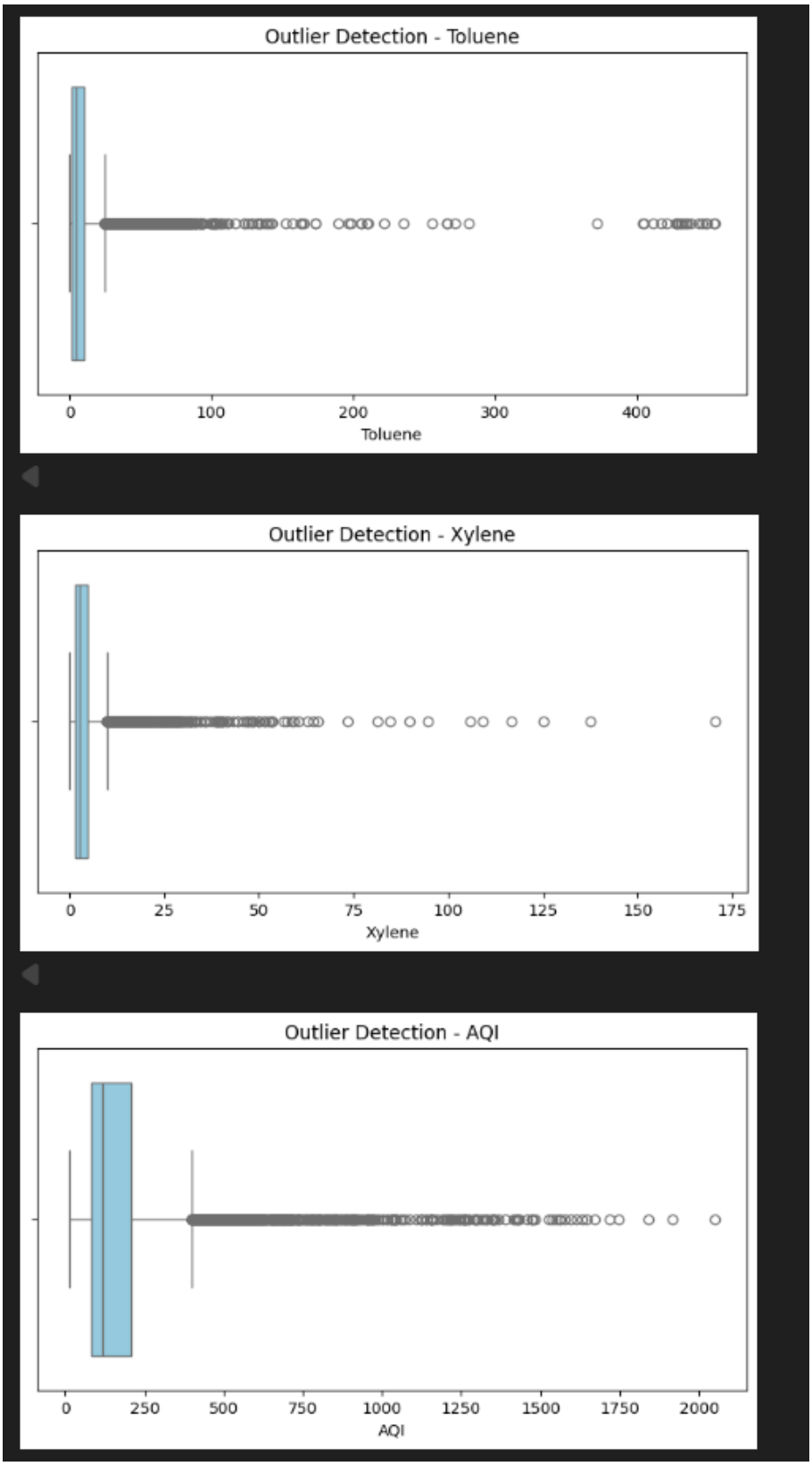


Figure 5 Outliers

2.3 Model Building

Feature Selection

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

Multi-Layer Perception (MLP)

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

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: prevents from overfitting.

Loss Function: Mean Squared Error standard for Regression Task.

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

Linear Regression

To build the Linear Regression model standard scaler data was used in LinearRegression to train and predict the model in which 80% for training whereas 20% for the testing of the data was used.

Random Forest Regressor

To built the Random Forest model RandomForestRegressor was used to fit the training dataset and predictions for the training and testing was done in which 80% for training whereas 20% for the testing of the data was used.

2.4 Model Evaluation

Multi-Layer Perception (MLP)

Metric	Training Performance	Testing Performance
MSE	841.57	982.88
MAE	19.01	20.25
RMSE	29.01	31.35
R^2 Score	0.92	0.91

Linear Regression

Metric	Training Performance	Testing Performance
MSE	1978.01	2053.43
MAE	31.11	31.37
RMSE	44.47	45.31
R^2 Score	0.82	0.81

Random Forest Regressor

Metric	Training Performance	Testing Performance
MSE	114.70	823.49
MAE	6.61	17.68
RMSE	10.71	28.70
R^2 Score	0.99	0.92

Initial Model Comparison

Random Forest have slightly lower test RMSE 28.70 and higher R^2 0.92 compared to Linear Regression RMSE 45.31 and R^2 0.81. However Random forest suggest slightly overfitting as R^2 is 0.99 on training and 0.92 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.

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 has different hyperparameters to improve the performance of the model.

- n_estimators: 200
- max_depth: 20
- min_samples_split: 5
- min_samples_leaf: 1
- max_features: ‘log2’

Cross-Validation Scores

5-Fold Cross-Validation was done during the hyperparameter search. The best cv achieved is 0.92 for random forest and 0.81 for linear regression.

2.6 Feature Selection

Feature Selection includes finding 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)	18.89	0.815	45.47	0.807
Random Forest	Selected (8)	31.38	0.922	29.20	0.921

3.1 Key Findings

The model evaluated on test data shows:

- Random forest achieved 29.20 RMSE which is lower than of linear regression 45.47.
- There is 0.921 R^2 OF Random Forest compared to 0.807.
- CV score of Linear regression is slightly lower 0.815 than Random Forest 0.922.

3.2 Final Model

The final model which was 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.

```
Model Comparison Table:
                          Model  Features Used CV Score    MAE  Test RMSE  \
0                           MLP            12         -  19.16      30.09
1  Linear Regression (Classical)            12         -  31.37      45.31
2      Random Forest (Classical)            12         -  17.76      28.72
3      Linear Regression (Final)             8     0.815  31.38      45.47
4          Random Forest (Final)             8     0.922  18.89      29.20

   Test R²
0    0.916
1    0.809
2    0.923
3    0.807
4    0.921
```

Figure 6 Final Model

Before optimization Test R^2 was 0.923 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 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 0.921 as it reduces the train test performance gap. Also the CV score of 0.922 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

features which was chosen and the model which performed in a consistent manner with the expectations. The Multi-Layer Perceptron (MLP) shows the strong and consistent performance, finding the testing R^2 of 0.916. The Random Forest shows the good testing R^2 of 0.921 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

- on more independent dataset perform the testing.
- 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].

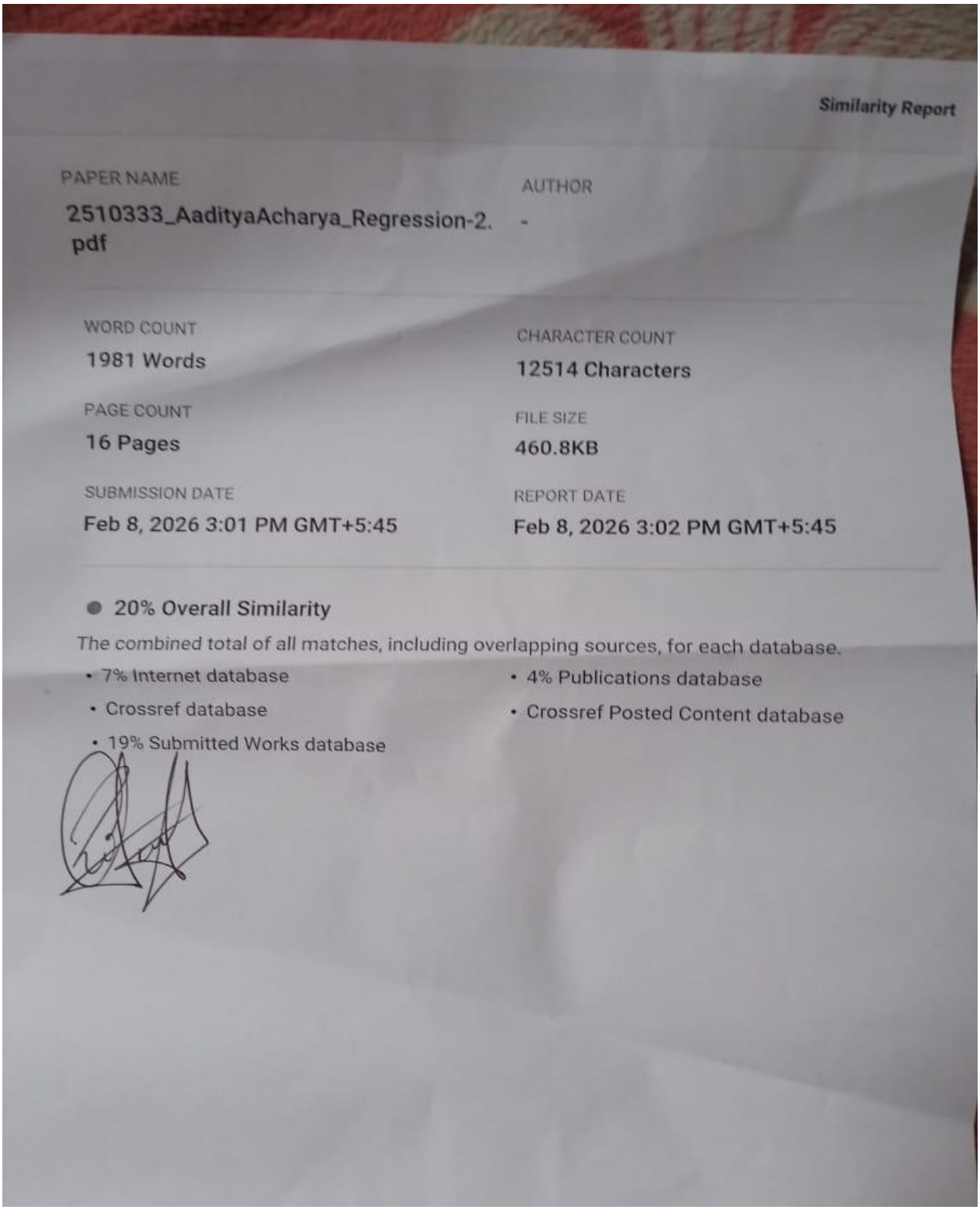
United Nations, n.d. *Good Health and Well Being*. [Online]
Available at: https://sdgs.un.org/goals/goal3
[Accessed 2026].

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].

6 GitHub Link

https://github.com/aaditya6882/2510333_AadityaAcharya_Final

7 Appendix



Similarity Report

PAPER NAME
2510333_AadityaAcharya_Regression-2.pdf

AUTHOR
-

WORD COUNT
1981 Words

CHARACTER COUNT
12514 Characters

PAGE COUNT
16 Pages

FILE SIZE
460.8KB

SUBMISSION DATE
Feb 8, 2026 3:01 PM GMT+5:45

REPORT DATE
Feb 8, 2026 3:02 PM GMT+5:45

● 20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 7% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 19% Submitted Works database

Figure 7 Plagiarism Report