Assessment Report

on

“Air Quality Index (AQI) Prediction”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY

DEGREE

SESSION 2024-25

in

CSE(AIML)

By

Rishi Patwa (202401100400156)

Sneha Sahu (202401100400188)

Tanya Yadav (202401100400197)

Vanshika Aggarwal (202401100400207)

Vikas Kumar Singh (202401100400210)

Section: C

Under the supervision of

“Abhishek Shukla”

KIET Group of Institutions, Ghaziabad

May, 2025

Introduction

What is AQI and Why It Matters?

The **Air Quality Index (AQI)** is a numerical scale used to communicate how polluted the air currently is or how polluted it is forecast to become. A high AQI indicates greater air pollution and higher health risks, especially for vulnerable populations.

In India, increasing urbanization and industrial activity have made AQI monitoring crucial. Predicting AQI allows authorities and the public to take proactive measures to safeguard health and manage pollution sources.

  Figure: AQI Levels and Associated Health Impacts

Objective of the Project:

This project aims to:

- Build a regression model to predict AQI based on environmental factors like PM2.5, PM10, NO2, SO2, CO, O3, etc.
- Visualize pollution levels in various Indian regions.
- Provide insight into features most responsible for changes in AQI.

Methodology

1. Data Collection & Exploration:

- Loaded the dataset using Pandas.
- Performed data cleaning: removed null values, handled missing AQI levels.
- Checked for outliers and inconsistencies.

2. Feature Selection:

Selected relevant pollutants as features for AQI prediction:

- PM2.5
- PM10
- NO2
- SO2
- CO
- O3

3. Preprocessing:

- Handled missing values using forward fill/backward fill.
- Standardized the data using StandardScaler.

4. Model Building:

- Applied Linear Regression, Random Forest Regression, and XGBoost Regressor.
- Compared performance using MAE, MSE, and R^2 Score.

5. Visualization:

- Used Seaborn and Matplotlib for:
    - Feature correlation heatmap
    - AQI distribution across cities
    - Predicted vs actual AQI values

6. Tools Used:

- Python, Jupyter Notebook
- Libraries: Pandas, NumPy, Sklearn, Matplotlib, Seaborn, XGBoost

Code

```
# Imports and File Upload
import pandas as pd
import numpy as np
from google.colab import files
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import zscore
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, r2_score

# Upload dataset
uploaded = files.upload()

# Load dataset
data = pd.read_csv('city_day.csv')

# Report missing values before cleaning
print("Missing Values Before Cleaning:\n")
print(data.isnull().sum())
```

```
# Data Preprocessing
# Drop rows with missing AQI values
data = data.dropna(subset=['AQI'])

# Fill remaining missing numeric values with column means
data = data.fillna(data.mean(numeric_only=True))

# Define pollutant features
pollutants = ['PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'CO',
              'SO2', 'O3', 'NH3', 'Benzene', 'Toluene', 'Xylene']

# Feature & Target Separation
X = data[pollutants]
y = data['AQI']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

# Scale features for KNN
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
```

```
X_test_scaled = scaler.transform(X_test)

# KNN Model
knn = KNeighborsRegressor(n_neighbors=5)
knn.fit(X_train_scaled, y_train)
y_pred = knn.predict(X_test_scaled)

# Evaluation
r2_test = r2_score(y_test, y_pred)
mse_test = mean_squared_error(y_test, y_pred)

# Train performance
y_train_pred = knn.predict(X_train_scaled)
r2_train = r2_score(y_train, y_train_pred)

print("\n✅ Model Evaluation:")
print(f"Train R² Score: {r2_train:.4f}")
print(f"Test R² Score: {r2_test:.4f}")
print(f"Mean Squared Error (Test): {mse_test:.2f}")

# Boxplot of AQI across Cities
plt.figure(figsize=(12,6))
sns.boxplot(x='City', y='AQI', data=data)
```

```
plt.xticks(rotation=90)
plt.title("AQI Distribution Across Indian Cities")
plt.tight_layout()
plt.show()

# Residual plot
residuals = y_test - y_pred
plt.figure(figsize=(8,5))
sns.histplot(residuals, kde=True, bins=30)
plt.title("Distribution of Residuals (Actual - Predicted AQI)")
plt.xlabel("Residual")
plt.grid(True)
plt.tight_layout()
plt.show()

# Actual vs Predicted AQI Plot
plt.figure(figsize=(8,6))
sns.scatterplot(x=y_test, y=y_pred, alpha=0.6)
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--')
plt.xlabel("Actual AQI")
plt.ylabel("Predicted AQI")
plt.title("KNN Regression: Actual vs Predicted AQI")
plt.grid(True)
```

plt.tight_layout()

plt.show()

Output/Result

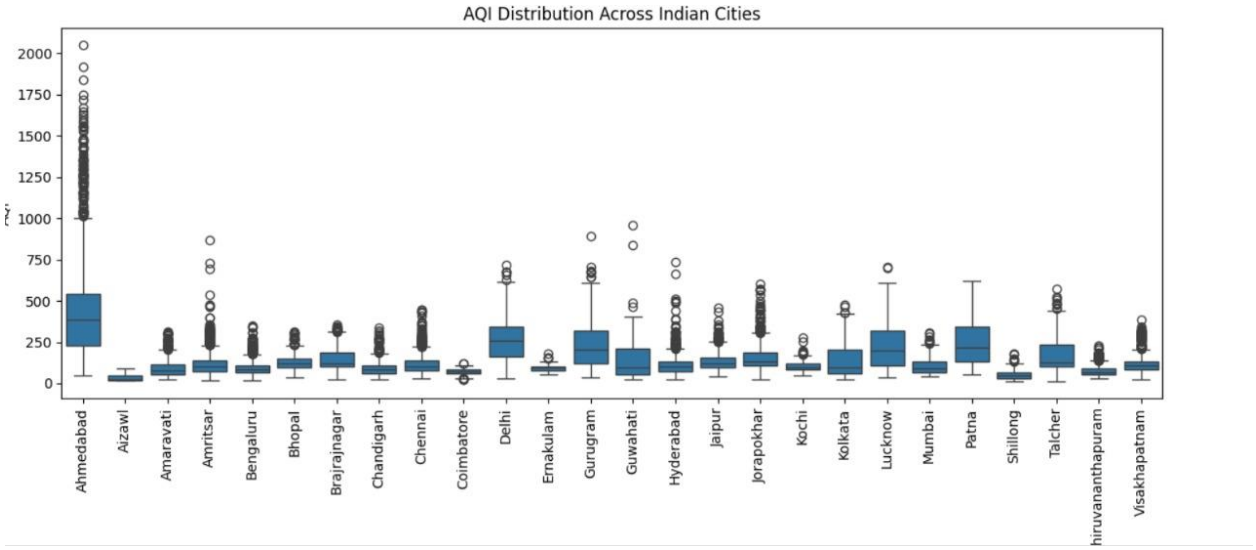
```
Choose Files city_day.csv
```

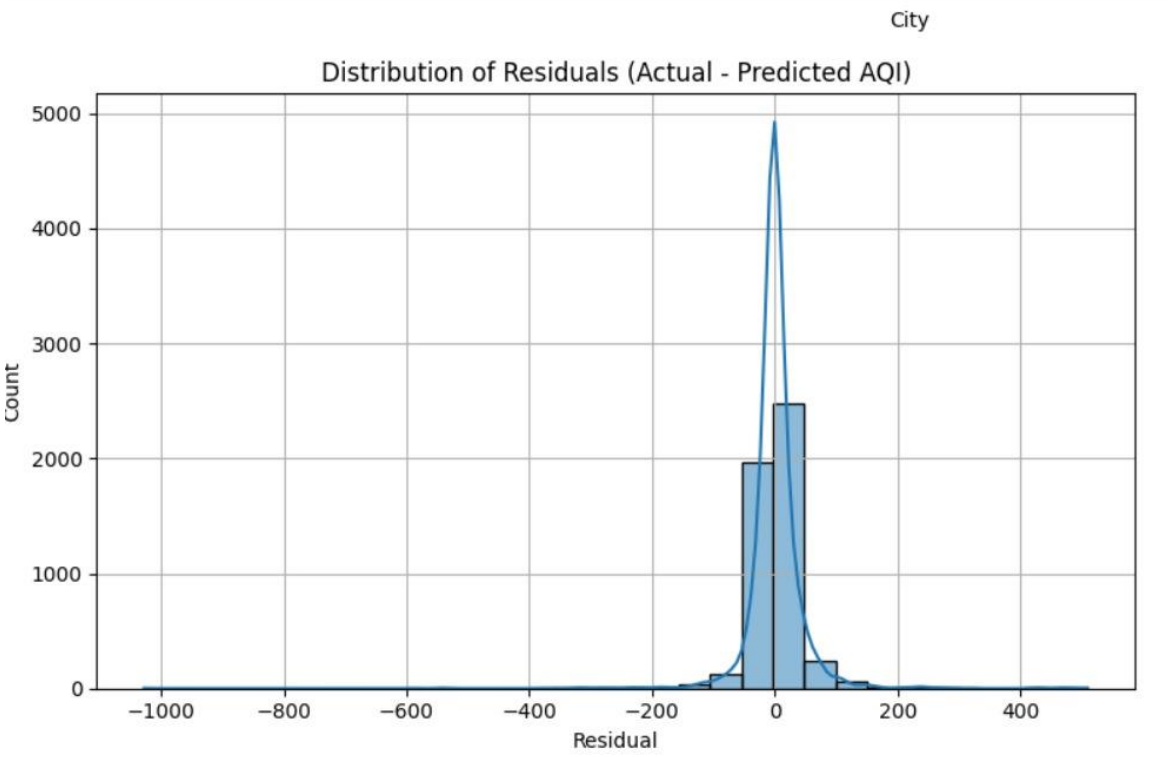
- **city_day.csv**(text/csv) - 2574056 bytes, last modified: 5/27/2025 - 100% done

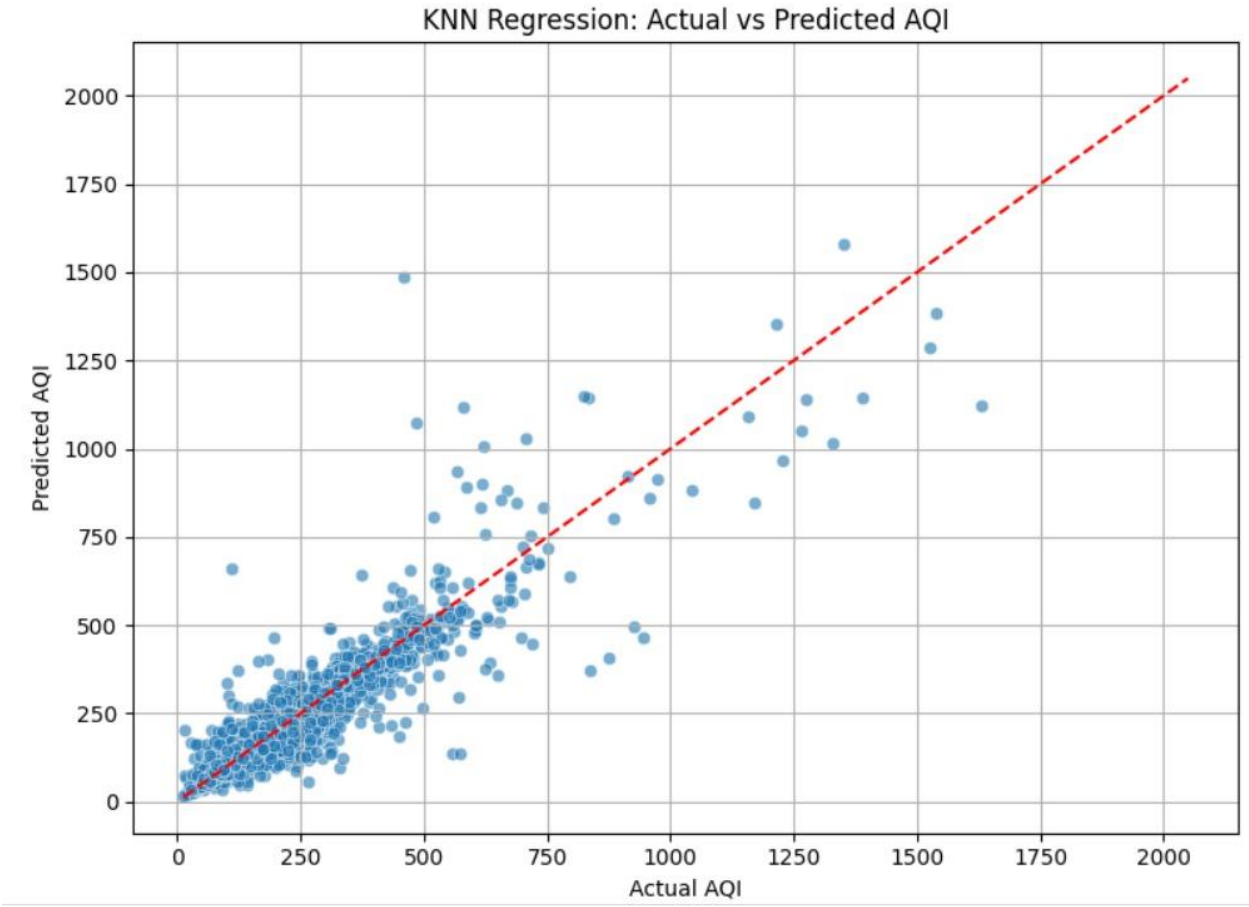
```
Saving city_day.csv to city_day.csv
Missing Values Before Cleaning:

City             0
Date             0
PM2.5         4598
PM10         11140
NO            3582
NO2           3585
NOx           4185
NH3          10328
CO            2059
SO2           3854
O3            4022
Benzene       5623
Toluene       8041
Xylene       18109
AQI           4681
AQI_Bucket    4681
dtype: int64

✅ Model Evaluation:
Train R² Score: 0.9237
Test R² Score: 0.8752
Mean Squared Error (Test): 2284.86
```




City
Distribution of Residuals (Actual - Predicted AQI)
Count
Residual
5000
4000
3000
2000
1000
0
−1000
−800
−600
−400
−200
0
200
400


KNN Regression: Actual vs Predicted AQI
Predicted AQI
Actual AQI
2000
1750
1500
1250
1000
750
500
250
0
0
250
500
750
1000
1250
1500
1750
2000

References

- Dataset:

Rohan Rao (Kaggle Dataset) –

https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india

- AQI Info Image:

U.S. Environmental Protection Agency – AQI Brochure

https://www.epa.gov/air-trends/air-quality-index-aqi

- Libraries Used:

Python (Pandas, NumPy, Sklearn, Matplotlib, Seaborn, XGBoost)