Project by Shantanu Sharma

Abstract

Air pollution is a critical environmental concern, it impacts human health and climate. Accurate prediction of the AQI (Air quality index) is essential for informing public health policies and raising community awareness. This study aims to compare the performance of the various machine learning models for AQI prediction by evaluating them against different pollutant data (such as PM2.5, PM10, NO_2, CO, SO_2, and O_3) collected from air quality monitoring stations. We will explore SVM (support vector machine), polynomial regression, decision tree, and linear regression. Model evaluation is based on metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 score values and Root mean squared logarithmic error.. The finding reveals the advantages and disadvantages of each model.

Contents

1. Introduction

Air is essential for the sustenance of life.Air pollution is one of the growing environmental concern globally; it affects not only the quality of the atmosphere but have effect on the health of humans as well. With the rapid increase of industrialization and urbanization, the air pollution has increased rapidly resulting in the degradation of the air quality. Generally air quality is measured in terms of AQI(Air quality index),It is defined as the measurement of air pollutant concentrations in ambient air pollution and the health risks associated with it An AQI number is given based on the air pollutant with the highest concentration at the instant of the air quality is known. AQI ranges from 0 to 500 where good air quality lies between 0 to 50, while measurements over 300 are considered highly dangerous. Normally while measuring the AQI amoung PM2.5, PM10,CO ,sulphur dioxide, nitrogen dioxide, ground-level ozone ,PM2.5 is considered for the AQI as these are found in huge quantity in the atmosphere is considered the most harmful air pollutant that adversely affect the human health.

The below image show the air quality index chart

2. Methodology

The ML model is designed to predict the AQI based on the data available for PM2.5, PM10, etc, it uses the following technique to achieve the result

a. Multiple Linear Regression

- Multiple linear regression, often known as multiple regression, is a statistical method that integrates several explanatory factors to predict the result of a response variable. One type of linear regression is multiple regression.

b. Polynomial Regression

- An nth-degree polyn models the relationship between the independent variable x and the dependent variable y, hence extending Linear regression. This type of regression is used when a higher-order poisson more precisely represents the problem than a linear connection.

c. Decision Tree

- It is a potent tool in ML algorithms. They offer an understandable method for decision-making grounded on evidence by modelling the interrelationships among many factors. A decision tree can be considered as the diagrammatic representation that is utilized for decision-making or forecasting.There are nodes, branches, and leaf nodes in the structure.

d. Support vector machine(SVM)

- A Support Vector Machine (SVM) is a commonly used algorithm in case of both linear and nonlinear classification, as well as regression. It is a type of adaptable algorithm,it offers various type kernel functions but the most commonly used is linear.

3. Library Used

Implementation for the model the following library were used

- from sklearn.model_selection import train_test_split-This is used to divide the data in to training and the testing set
- from sklearn.linear_model import LinearRegression – This library was used to implement the multiple variable Linear Regression.
- from sklearn.tree import DecisionTreeRegressor This library was used to implement the decision tree for the given data
- from sklearn.svm import SVR – This library was used to implement the suppor vector regression
- from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score -It is used to calculate the mean square error , mean absolute error and R^2 error.
- from math import sqrt -This is used to calculate the square root
- import numpy as np and import pandas as pd- These two library are used to ease the task of handling the data
- import matplotlib.pyplot as plt – This is used to plot the graph for predicted and the actual value of the AQI

4. Implementation

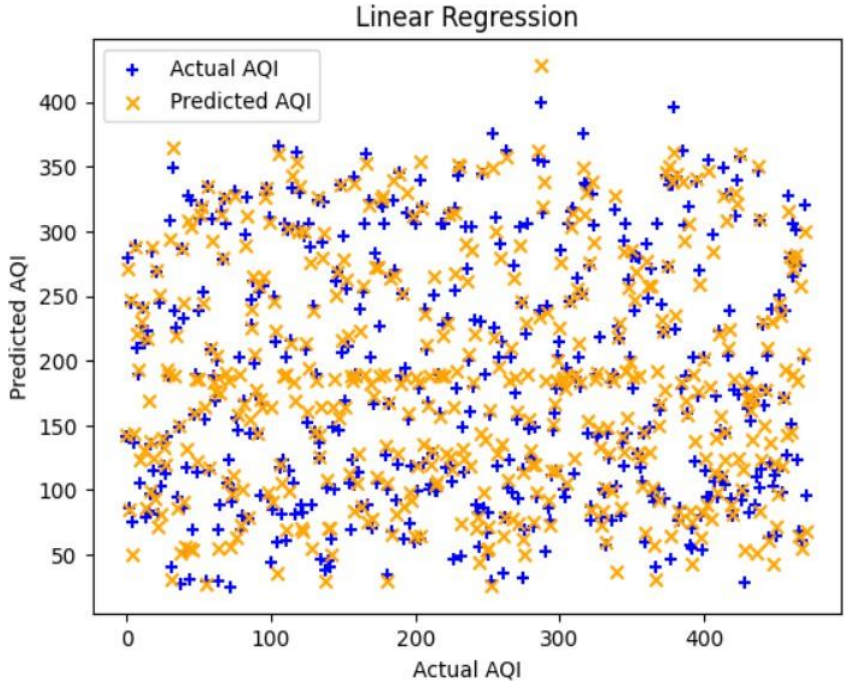
Here first the data set consisting of the PM2.5, PM10 etc value is uploaded , if the data contains any 'NA' Value it is either dropped or replaced with the mean of the value above and the below. Now finally the data is split into the training and testing set. Now various techniques are applied to train the model with the help of a respected function. With the help of predicted and actual value ,root mean square error , mean square logarithm error , mean absolute error ,and R^2 error is calculated.Finally with the help of the matplot the graph for the actual and predicted value is plotted.

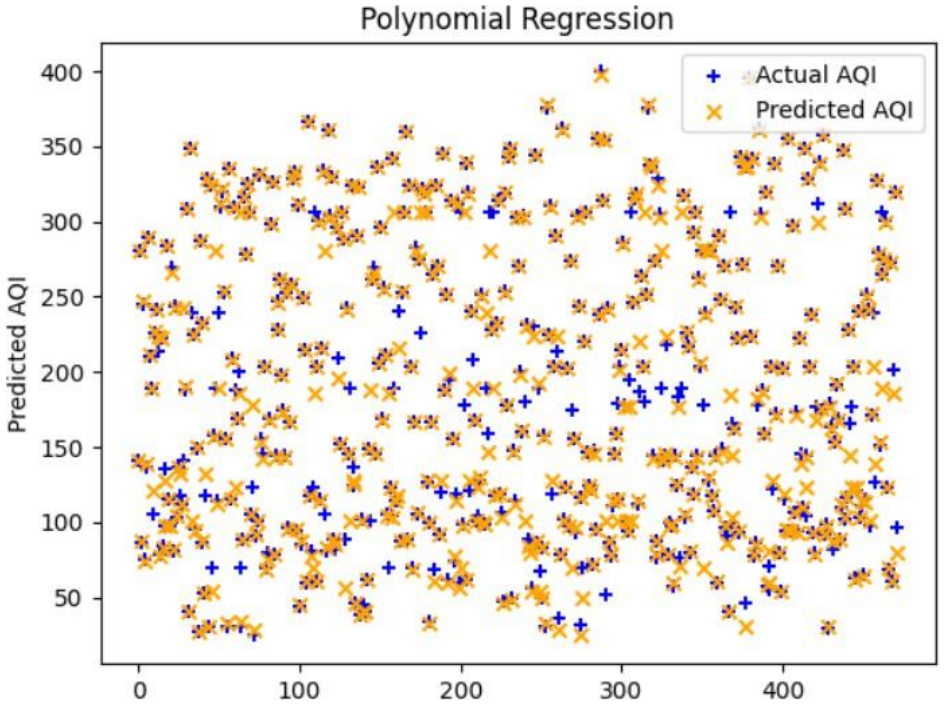
5. Graphs and Result

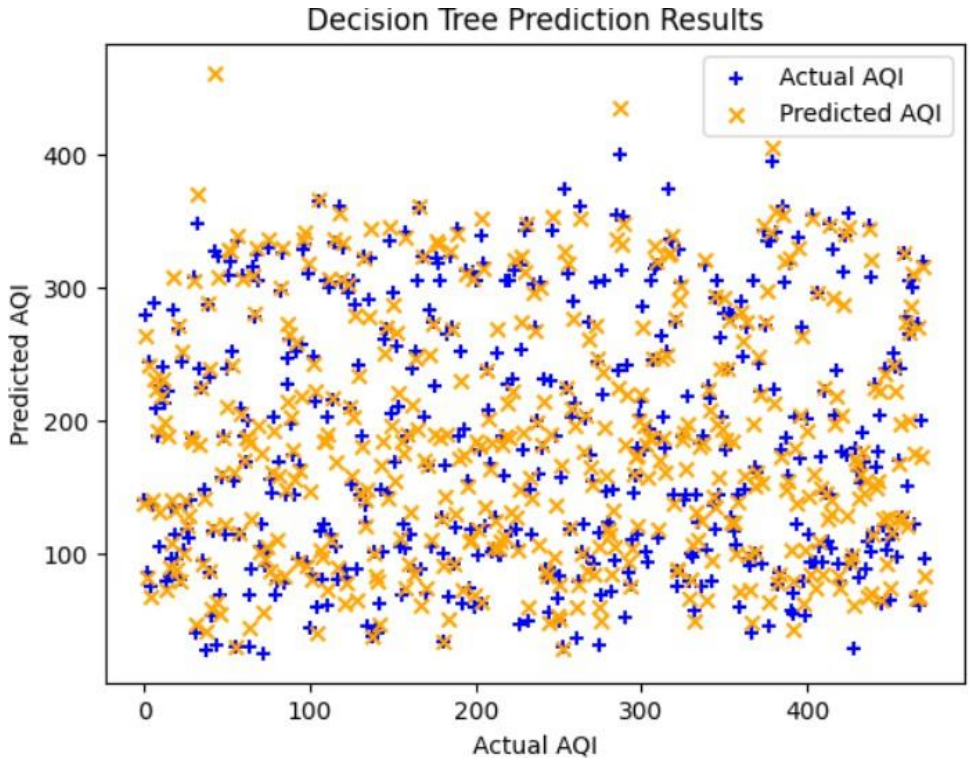
The below image show the various error for the different models

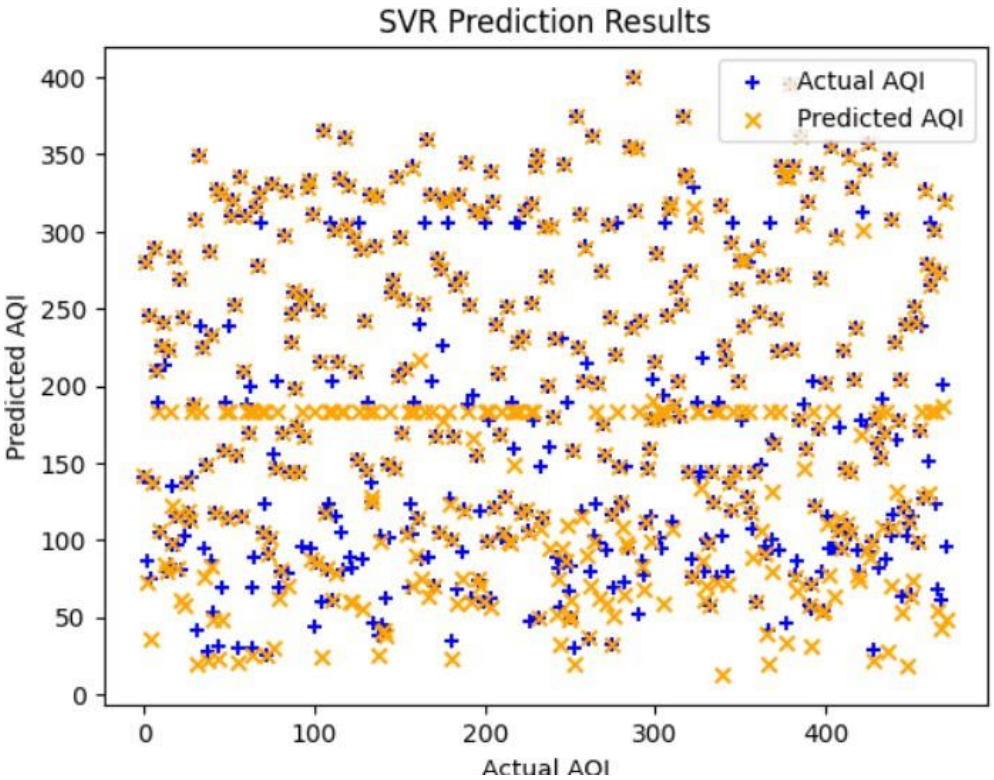
	Model	RMSE	MAE	MSLE	R2
0	Linear Regression	33.230023	18.366167	0.064673	0.878508
1	Decision Tree	30.884650	9.087945	0.034996	0.895052
2	Polynomial Regression	31.112567	18.096346	0.051294	0.893498
3	Support Vector Regression	37.097277	16.403695	0.100917	0.848584

The below images shows the graph for the actual and predicted value of the AQI




Polynomial Regression
Actual AQI
Predicted AQI
Predicted AQI
400
350
300
250
200
150
100
50
0
100
200
300
400


Decision Tree Prediction Results
Actual AQI
Predicted AQI
Predicted AQI
400
300
200
100
0
100
200
300
400
Actual AQI



6. Results Analysis

Here from the above images we can see that decision Tree achieved the lowest RMSE (30.88) and MAE (9.08), indicating that it had the best performance in terms of error metrics among the models also MSLE of 0.034996 and R^2 score of 0.895 shows its accuracy and ability to explain variance in AQI values effectively and form the actual and predicted value graph we can see that it most accurate among all models

Polynomial regression shows similar performance with slightly higher RMSE (31.11) and MAE (18.09) but a competitive R^2 score of 0.893. This suggests that it is capable of capturing non-linear relationships, although it may be more sensitive to overfitting.

Linear regression had a higher error rate than the Decision Tree, its simplicity and interpretability might make it suitable for cases where model transparency is prioritized.

Support Vector Regression (SVR) had the highest RMSE (37.09) and the lowest R^2 score (0.848), indicating that it may not be as suitable for AQI prediction in this dataset compared to the other models.

7 Conclusion

The above shows the machine learning model, particularly Decision Tree and Polynomial Regression models, can effectively predict AQI levels by analysing key pollutant data, the decision achieved the highest accuracy and shows the ability to handle the complex data. Somewhat accuracy is shown by the linear and polynomial regression but could handle only the simple data.

8 Future Scope

In future, this model can be incorporated with Deep Learning Models Experimenting with deep learning models like LSTMs and CNNs could capture temporal patterns in AQI data if time-series data is available. This model can also be deployed as the real time system with continuous update of the data.

Reference

1. S. A. Aram *et al.*, "Machine learning-based prediction of air quality index and air quality grade: a comparative analysis," *International Journal of Environmental Science and Technology*, vol. 21, no. 2, pp. 1345–1360, Jun. 2023, doi: 10.1007/s13762-023-05016-2.
2. "IQAIR | First in air quality." https://www.iqair.com/newsroom/what-is-aqi
3. I. Ayus, N. Natarajan, and D. Gupta, "Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China," *Asian Journal of Atmospheric Environment*, vol. 17, no. 1, May 2023, doi: 10.1007/s44273-023-00005-w.
4. GeeksforGeeks, "GeeksforGeeks," *GeeksforGeeks*. https://www.geeksforgeeks.org/