JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTING AIR QUALITY INDEX AND COMPARING MODELS USING MACHINE LEARNING ALGORITHMS IN HIMACHAL PRADESH

Mrs.Ruchi Thakur

Department Of Computer Science

Sardar Patel University, Mandi(H.P.)

Abstract: The air we breathe is one of the most important components of life on earth, and the quality of the air is becoming a serious problem for the general population. In many countries Air pollution has become a major public health risk, as a result there are many health risks which include respiratory problems, cardiovascular diseases, and even cancer.
In this paper we analyses the application of two machine learning algorithms, first is time-series forecasting using the Facebook Prophet Algorithm and Regression Analysis using Support Vector Machines (SVM)—for evaluating air quality. The study consists of a comprehensive dataset provided by Himachal Pradesh Pollution Control Board (HPPCB) containing air quality indicators from various cities in Himachal Pradesh. The dataset focuses on ambient concentrations of fine particulate matter (PM10 and PM2.5), gaseous pollutants (SO2, NO2, NH3, and O3), and carbon monoxide (CO).The first step involves data preprocessing, which includes data loading, handling missing values, and ensuring date format consistency to ensure smooth integration with the forecasting and regression models. The pre-processed data is then divided into the training and testing and incorporated into the two machine learning algorithms. The accuracy of the forecasts being assessed using metrics such as mean absolute error (MAE) and root mean squared error (RMSE).On the other hand, Support Vector Regression (SVR), a version of Support Vector Machine (SVM), is being utilized to forecast Air Quality Index (AQI) values by taking into account the concentrations of various pollutants. Radial Basis function (RBF) kernel allowed to SVR for obtaining most accurate prediction. The main cause associated with air pollution are due to rising traffic volumes, extensive infrastructure projects such as tunnel construction that necessitate tree cutting, and the heavy rush of tourists to popular destinations, which increase congestion and fossil fuel emissions. Air quality prediction can assist public authorities and policy makers in developing strategies to reduce air pollution and improve public health. Highlighting their respective strengths and weaknesses in terms of accuracy, scalability and computational efficiency, the performance of both methodologies is thoroughly evaluated and compared. The findings from this comparison study offer important advice to policymakers and researchers when choosing appropriate models for evaluating and controlling air quality. This research improves the field and provides the opportunity for improved monitoring plans and accurate decision-making processes aimed at reducing the adverse impacts of air pollution on the environment and the general population, by highlighting the advantages of innovative methods for machine learning .After comparing these two machine learning algorithm Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel outperforms Prophet in terms of accuracy in predicting air quality index with the maximum coefficient of determination and less mean absolute error.

Keywords: Air Quality Index, Machine Learning, Time Series Forecasting, Support Vector Regression, Facebook Prophet Algorithm

I. INTRODUCTION

Air is the invisible mixture of gases. Most living things rely on oxygen and nitrogen found in the air to survive. Animals wouldn't be able to survive without air; life as we know it would not exist. Without air, there would be no sound at all. Human Beings (Homo sapiens) are one of these species. Air pollution is the contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere (WHO).
Key health harmful air pollutants include following- Particulate Matters (PM2.5 and PM10), Carbon Monoxide (CO), ground level Ozone (O3), Volatile Organic Compounds, metals, Sulphur dioxide (SO2), Ammonia (NH3) and Nitrogen dioxide (NO2).

Air Quality:

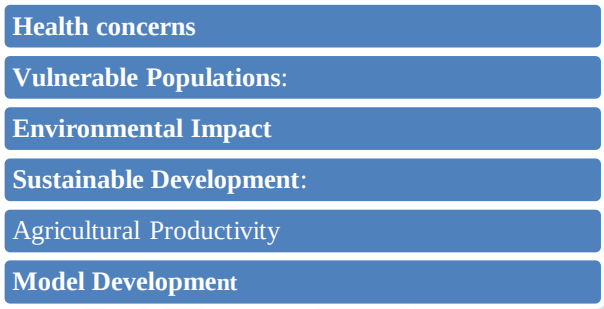
Air quality is the degree of purity or healthiness of the air for humans or the environment. A numerical scale used to communicate how polluted it is forecast to become or how polluted the air currently is. The AQI scale ranges from 0 to 500, where higher values indicate higher levels of air pollution and correspond to greater health concerns. Low air quality can exacerbate respiratory conditions like breathing problems, irritation of the nose, throat, and eyes, cause breathing difficulties, and harm the heart and blood circulation. In the last ten years, studies have found that high levels of air pollution may actually cause depression, impair children's cognitive development, and increase the risk of cognitive ageing in adults.

Air Quality index:
AQI or Air Quality Index, an indicator for daily reporting on AQ is the AQI. It calculates the rate of health damage caused by air pollution. The AQI was created to help people understand how the local AQ affects their health in order to safeguard public health. It uses color-coded categories and provides statements for each category. National Air Quality laws have been established for five primary air contaminants, for which calculates the AQI.
1. Ozone at ground level
2. Particulate matter and particle pollution (PM2.5/PM10)
3. Carbon Monoxide
4. Sulphur Dioxide
5. Nitrogen dioxide

India: National Clean Air Programme (NCAP):
India launched the National Clean Air Program in 2019 with the goal of reducing air pollution levels in various cities. The program focuses on a collaborative, multi-sectorial approach to address pollution from various sources.
A commonly used metric to measure the condition of the air is the Air Quality Index (AQI), which integrates several contaminants into a single statistical number. With the use of algorithms based on machine learning, this research proposal aims to create a prediction model for AQI that will provide accurate and on-time data for strategic decision-making processes.

II. Importance of AQI -

The Ministry of Health and Family Welfare has issued a vital advisory to address and mitigate the adverse health effects of prevailing air pollution conditions under National programme on climate Change and Human Heath (NPCCHH). Some importance of AQI are:



Health Concerns:
Himachal Pradesh, despite its pristine image, faces air pollution challenges in some areas. Accurate AQI predictions allow individuals to take preventive measures like wearing masks or avoiding outdoor activities during high-pollution periods, reducing respiratory illnesses and other health problems.
Vulnerable Populations: Children, pregnant women, and the elderly are particularly vulnerable to air pollution. AQI predictions can help these groups adapt their routines and protect their health.
Environmental Impact:
Understanding the factors influencing AQI helps identify pollution sources and implement targeted control measures.
Sustainable Development: Sustainable development cannot ignore air quality. Economic Benefits:
Himachal Pradesh relies heavily on tourism. Good air quality is a major attraction for tourists..
Agricultural Productivity: Air pollution can harm crops and reduce agricultural yields.
Model Development: Research in this area contributes to the development and improvement of machine learning models for AQI prediction. This knowledge can be applied to other regions and contribute to global air quality improvement.

III. Objectives:

- ✓ To predict Air Quality Index (AQI) and compare two machine learning algorithms.
- ✓ Evaluate the performance of proposed models.

IV. Literature Review

In the study, Researcher has used support vector regression (SVR), a machine learning technique, to forecast pollutant and particle levels and to estimate the air quality index (AQI). Radial basis function (RBF) was the kind of kernel among the investigated choices that gave SVR the most precise predictions. It was shown that employing all of the available variables was more effective than using principal component analysis to choose features. The results show that SVR with RBF kernel enables to precisely forecast hourly pollutant concentrations, such as carbon monoxide, sulphur dioxide, nitrogen dioxide, ground-level ozone, and particulate matter 2.5, as well as the hourly AQI for the state of California. On unseen validation data, classification into the six AQI categories set forth by the US Environmental Protection Agency was carried out with an accuracy of 94.1% (Castelli et al., 2020)

Since air quality is increasingly affecting people's health, the government must take required steps to forecast it. The air's quality is measured by the air quality index. Carbon dioxide, nitrogen dioxide, and carbon monoxide are among the air pollutants that cause air pollution and are released when natural gas, coal, and wood are burned, as well as by industry, cars, and other sources. Serious illnesses including lung cancer, brain sickness, and even death can be brought on by air pollution. The air quality index is determined with the aid of machine learning techniques. So many studies are being conducted in this area, but still the outcomes are still not reliable. Datasets from Kaggle and air quality monitoring stations are accessible and separated into Training and

Testing. Linear Regression, Decision Tree, Random Forest, Artificial Neural Network, and Support Vector Machine are some of the machine learning techniques used for this (Madan et al., 2020)

In five key areas of the Tamil Nadu city of Chennai, this article examined the impact of shutdown on the air quality index and other contaminants. The monitoring stations' data on air pollutants such PM10, PM2.5, SO2, and NO2 from 2018 to 2019 (the pre-Covid period) and 2020 to 2021 were studied (during-Covid period). The findings showed that PM10 and PM2.5 concentrations decreased by roughly 48% and 39%, respectively. The pollutants SO2 (25%) and NO2 (10%) have also been seen to significantly decrease. The AQI level in Chennai city was also found to be satisfactory to moderate both before and during the lockdown. Adyar (50.38%) saw the greatest drop in AQI, followed by Nungambakkam (44.18%), T-Nagar (40.31%), Anna Nagar (39.98%), and Kilpauk (30.74%). According to the overall research, environmental protection and pollution control are achieved by implementing appropriate regulations at the right time and place. (Sangeetha, n.d.)

In this research, a brand-new hybrid interpretable predictive machine learning model with two innovations is put forth for the prediction of Particulate Matter 2.5. Deep neural networks and a nonlinear auto regressive moving average with exogenous input model are used to build a hybrid model structure first. Second, this hybrid model incorporates automatic feature generation and feature selection processes. The experimental findings show that the model outperforms other models in terms of peak value prediction accuracy and model interpretability. The suggested model demonstrates how historical PM2.5, weather, and season data are used to estimate PM2.5 forecast. The correlation coefficients for the accuracy of the 1, 3, and 6-hour predictions are 0.9870, 0.9332, and 0.8587, respectively. More significantly, the suggested method offers a fresh interpretable machine learning framework for time series data, allowing for the development of accurate predictive models and the explanation of complicated interdependencies across multimode inputs. (Gu Y, 2022)

In this study, different error-prone methodologies, including R-Squared (R2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) methods, are listed in order to estimate the AQI value for Particulate Matter (PM2.5) m at a specific area of Delhi using a range of data forecasting ways. The suggested method combines the Long-Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) deep learning models to forecast the AQI of the surrounding environment. In order to compare their performances with the proposed hybrid (LSTM-GRU) model, several stand-alone machine learning (ML) and deep learning (DL) models are trained on the same dataset, including LSTM, Linear-Regression (LR), GRU, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). It is discovered that the proposed hybrid model outperforms all others, with MAE values of 36.11 and R2 values of 0.84. (Sarkar N, 2022)

Researcher has used Data Mining and Machine Learning models to forecast the AQI and classify the AQI into buckets, in this research project. Principal Component, Partial Least Square, Principal Component with Leave One Out CV, Partial Least Square with Leave One Out CV, and Multiple Regression AQI Data of Multiple Indian Cities are the five regression models we used to predict AQI. Based on the value of the AQI, the AQI Index is further divided into six different categories termed buckets: "Good, Satisfactory, Moderate, Poor, Very Poor, and Severe." Researchers have created three classification models, including Multinomial Logistic Regression, K Nearest Neighbour, and K Nearest Neighbours with Repeat CV Classification method, to forecast the AQI bucket. When just taking into account the fifth component from all the models, the PLS model with Leave One Out Cross Validation performed the best at dimension reduction from the Air Quality Dataset of Different Indian Cities. PLS model was the most accurate and had the lowest RMSE. In terms of accuracy and AUC, the KNN Model with Repeated CV and Tune Length 10 performed the best from Station Wise Data of Indian Cities. (Mahalingam et al., 2019)

In this researcher collected dataset from Central pollution Control Board (CPCB) India of 23 different cities. Five machine learning Models were employed to predict the air quality and found that the Gaussian naive Bayes model achieves the highest accuracy. The XGBoost model performed the best and gets the highest linearity between the predicted and actual data. (Kumar & Pande, 2023)

In this study focus was on Data set of New Delhi for predicting ambient air pollution and quality using several Machine learning algorithms. Several ML algorithms were widely used such as Random Forese, SVM, regression, classification. (Sinha & Singh, 2020)

V. Problem Statement and Data

(a) Statement of the problem:

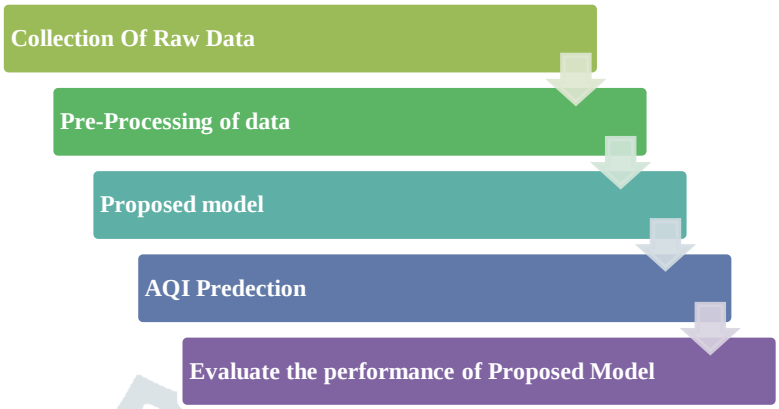
Various researches are being done in this field but still results are not accurate. Despite extensive researches, I haven't been able to locate any research specific to Himachal Pradesh up to this point until now; previous studies have mainly adopted a traditional machine learning method to analyze the air pollution disaster risk. Moreover, these models are not interpretable. The Studies are not able to explain the interpretation between determining factor and their impact on polluted air. More research is required for the accurate prediction of AQI. In eight districts, PM2.5 levels exceed 50, indicating a cause for environmental concern. Air pollution in the region is a health hazard and represents the third-highest risk for premature death.

(b) Significance of Study

The significance of this research is, it will develop an interpretable predictive machine learning model which can be used to predict accurate AQI and can be used as a strong measurement tool. If successful, this machine learning model can be replicated in the other states also.

VI. Methodology

The sequences of steps for the prediction of AQI are explained in this section.



(a) Data Collection:

The dataset used in this study was collected from Himachal Pradesh State Pollution Control Board (HPSPCB) from March 2, 2020 to January 31, 2024.The dataset focuses on ambient concentrations of fine particulate matter (PM10 and PM2.5), gaseous pollutants (SO2, NO2, NH3, and O3), and carbon monoxide (CO). The events were collected between March 2, 2020 and January 31, 2024. Total number of records used 11005.

This data contains the pollutant concentration data , AQI values. With the respective information on the breakpoint for each pollutant concentrations. Table 1. Shows the detailed description of the variables used in the study.

Table 1. Summary of measurement location and variable observed:

Measurement location	Type	Variables
Himachal Pradesh	Particulates	$PM_{2.5}$
		PM_{10}
	Gaseous pollutants	Sulfur dioxide (SO_2)
		Nitrogen dioxide (NO_2)
		Ammonia(NH_3)
		Ozone (O_3)
		Carbon monoxide(CO)

The pollutant data are presented in mass concentration units, specifically (µg/m³) for micrograms per cubic meter except for CO which is in (mg/m³) or milligrams per cubic meter.

Table 2. Description of dataset variables used:

Name of Variable	Explanation
State Name	Name of the state where the monitor is located
Parameter code	The CPCB code corresponding to the parameter being measured.
Parameter name	The designation or label given in CPCB for the parameter being monitored
Date GMT	The calendar date of the average in Greenwich Mean Time.
Units of measure	The units of measured parameter.
Qualifier	Sample values may include qualifiers that explain their absence such as null data, natural occurrences, and quality assurance issues.

Table 3. A concise summary of pollutant levels, particulate data, and environmental event metrics, including minimum, maximum, mean, standard deviation, quantiles, kurtosis, and skewness.

	PM10 (µg/m³)	PM2.5 (µg/m³)	SO? (µg/m³)	NO? (µg/m³)	NH? (µg/m³)	O? (µg/m³)	CO (mg/m³)	AQI
Mean	64.12	64.12	64 .12	64.12	64.12	64.12	64.12	64.12
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Max	549.0	341.0	128.0	66.4	370.0	981.0	20.0	508.0
Std	37.69	15.42	3.87	4.99	5.22	10.50	0.23	33.86
25%	40.0	18.70	2.0	4.5	0.30	1.50	0.38	41.0
50%	54.0	26.76	2.0	6.1	1.50	3.10	0.38	54.0
75%	77.2	28.00	2.4	12.0	2.29	3.95	0.38	77.0
Kurtosis	8.93	49.56	426.76	11.69	2260.03	6835.20	4416.18	9.88
Skew	2.09	4.17	19.32	2.05	34.42	73.78	53.43	1.96

Both metrics point to heavy tails and asymmetry in the data distribution. High kurtosis values for pollutants like PM2.5 and NH_3 suggest the presence of extreme outliers. PM2.5 exhibits a sharp increase between the 75th percentile and the maximum, indicating the presence of outliers. M10 displays high variability, whereas CO shows much less fluctuation. Significant pollution events are highlighted, particularly with PM10 reaching 549 µg/m³ and O_3 at 981 µg/m³. The mean concentrations across all pollutants are consistent but significantly lower than the maximum values, indicating high variability and potential pollution spikes.

The Prophet model is used to forecast Air Quality Index (AQI) based on two key variables:

ds: The date or timestamp of the observation.

y: The AQI values you want to predict.

Figure 1 shows the forecast using prophet model. Where ds and y are two variables.

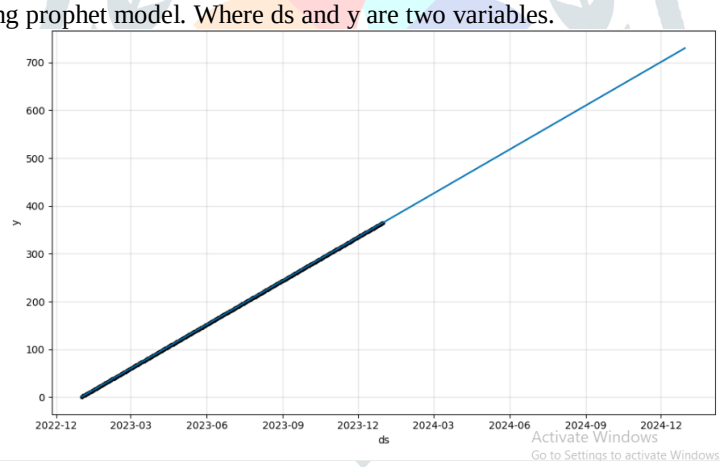
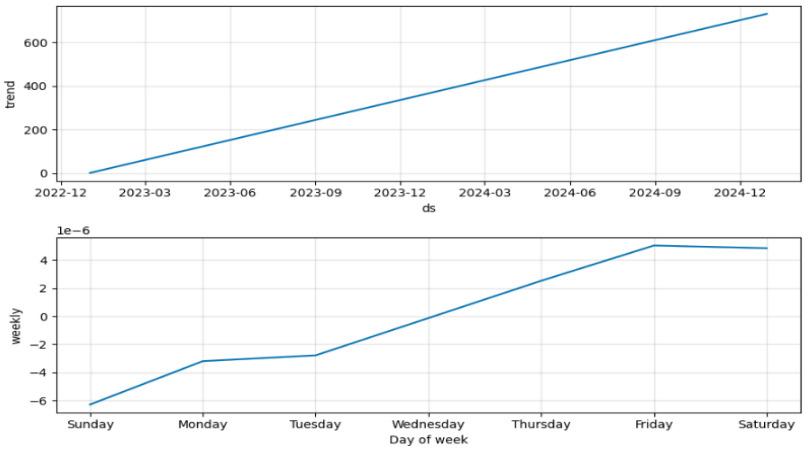
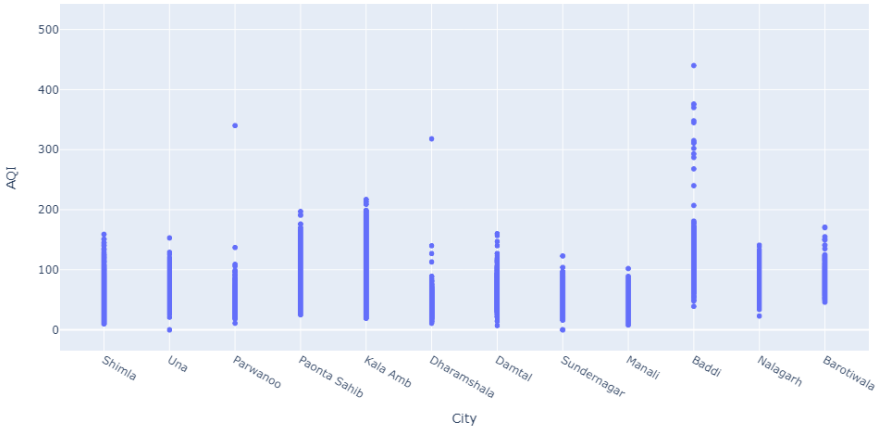


Figure 2 generates a plot that visualizes the different components of a forecast produced by a statistical or machine learning model. This typically includes trends, seasonal effects, and holiday effects, helping to understand how each component contributes to the overall forecast.

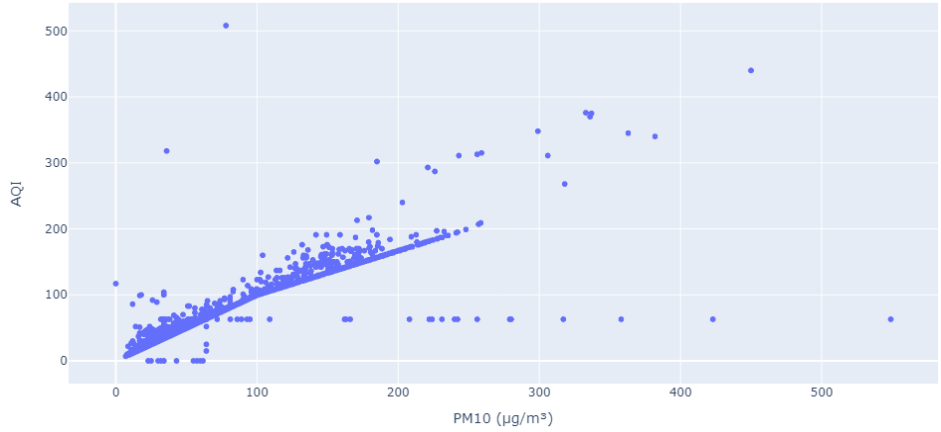


(b) Scatter plot showing the Air Quality Index (AQI) for different cities and then displays the plot.

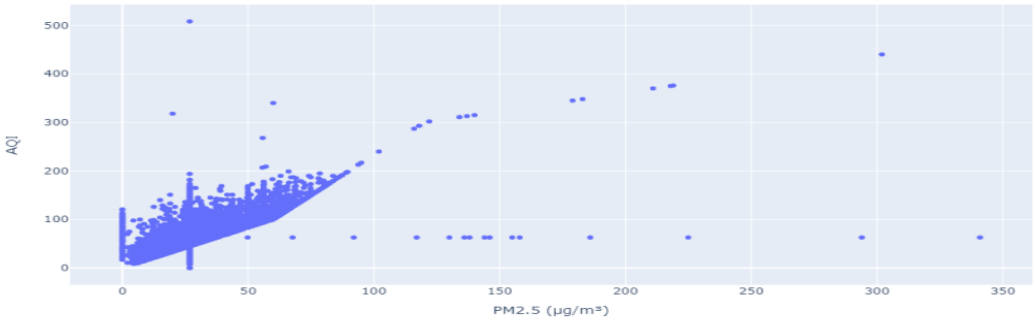
❖ Figure 3. Measurement of city with AQI for the state of Himachal Pradesh



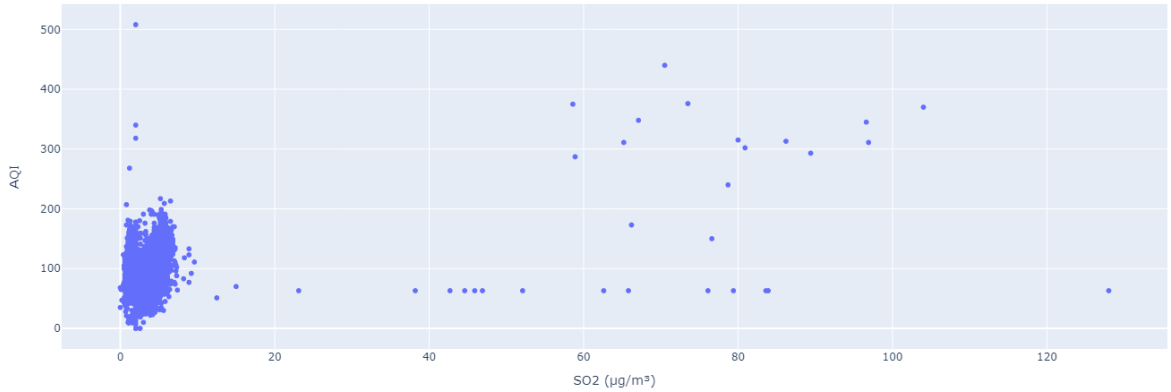
❖ Figure 4 Measurement of AQI with PM10 for the state of Himachal Pradesh



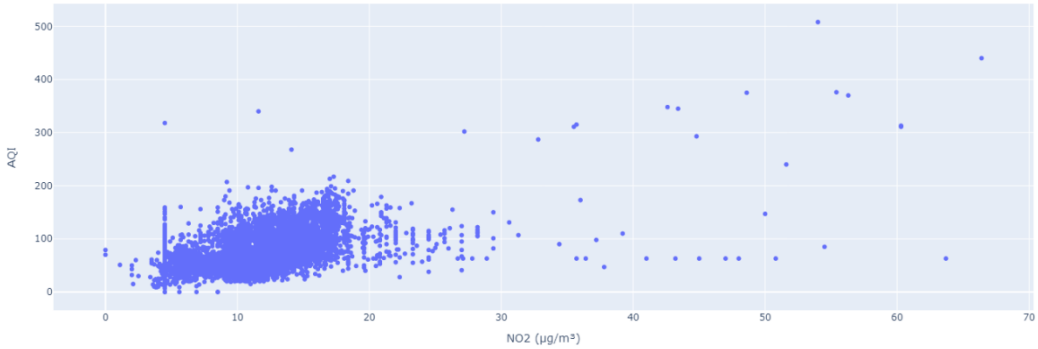
❖ Figure 5 .Measurement of AQI withPM2.5 I for the state fo Himachal Pradesh



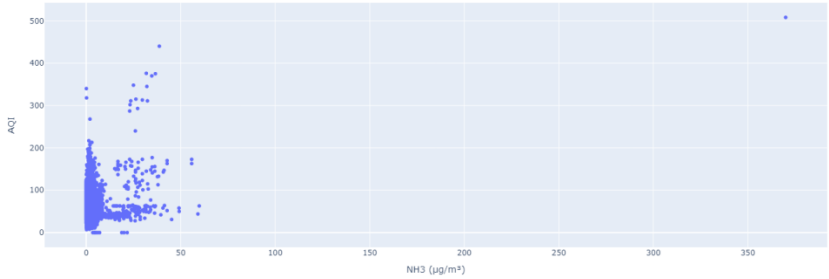
❖ Figure 6 Measurement of SO2 with AQI for the state of Himachal Pradesh



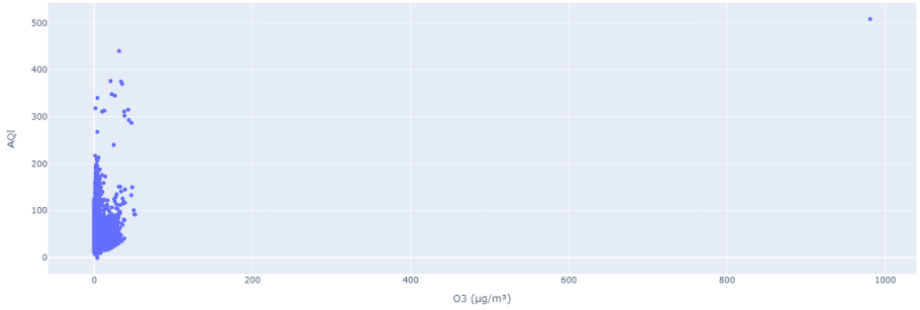
❖ Figure7 Measurement of NO2 with AQI for the state of Himachal Pradesh



❖ Figure 8 Measurement of NH3 with AQI for the state of Himachal Pradesh



❖ Figure 9. Measurement of O3with AQI for the state of Himachal Pradesh



Figures showing non-linear relationships for various pollutants and particulates. Non-linear plots can indicate complex interactions between the pollutants or varying impacts on health and the environment.

VII. PERFORMANCE EVALUATION

The Prophet model demonstrates strong predictive performance, with both MAE and RMSE values indicating minimal errors in predictions. The average absolute error of approximately 9.82188076 units and the RMSE of around 0.00011624 units highlight that the model's forecasts are very close to actual values. These low error metrics suggest that the Prophet model is reliable and effective for this dataset, reinforcing its suitability for accurate time series forecasting.
On the other hand where Support Vector Regression Model gave Mean Absolute Error around 4.959813362154864 units and Root Mean Square Error around 0.00011624 units.
We compare the performance of the Facebook Prophet Model and Support Vector Regression (SVR) models.
The SVR model's MAE of around 4.959813362154864 units indicates smaller average prediction errors compared to the Prophet model's MAE of approximately 9.82188076 units.
Prophet Model -MAE measures the average absolute difference between actual and predicted values. In this case, it indicates that, on average, our Prophet model's predictions are very close to the actual values, with an error of approximately 0.00009822 units.
RMSE measures the square root of the average of squared differences between predicted and actual values. It penalizes larger errors more than MAE. RMSE suggests that the typical error in predictions is around 0.00011624 units.

- **Accuracy (MAE)**: Support Vector Regression (SVR) performs better with a lower Mean Absolute Error (MAE) (4.95) compared to Prophet (9.821), indicating that SVR's predictions are closer to the actual values on average.

  Calculates as –MAE $= \sum_{i=1}^{n} | \text{yi} - \hat{\text{y}}\text{i} |$

  Where {yi} represent 'y_true' contains actual values and { ŷi} represent 'forecasted' contains predicted values.

- **Precision Root Mean Squared Error (RMSE) Calculation (RMSE)**: Both models have identical RMSE values (0.0001162), suggesting that they predict the magnitude of errors similarly.

  Calculates as $-RMSE = \sqrt{\sum_{i=1}^{n}(\text{yi} - \hat{\text{y}}\text{i})^2}$

  Where, n is the number of observations {yi} represent 'y_true' contains actual values and { ŷi} represent 'forecasted' contains predicted values.

In conclusion, while both models have the same RMSE, indicating similar precision, the SVR model outperforms the Prophet model in terms of accuracy (MAE), as it has a significantly lower MAE score. This suggests that for this particular dataset or problem, the SVR model might be providing more accurate predictions relative to the Prophet model. Both MAE and RMSE values in this case are very small, which generally indicates that your Prophet model's predictions are quite accurate relative to the scale of the data. Lower values of MAE and RMSE imply better model performance in terms of prediction accuracy

Figure 10 Showing PI chart for Distribution of AQI Conditions by City

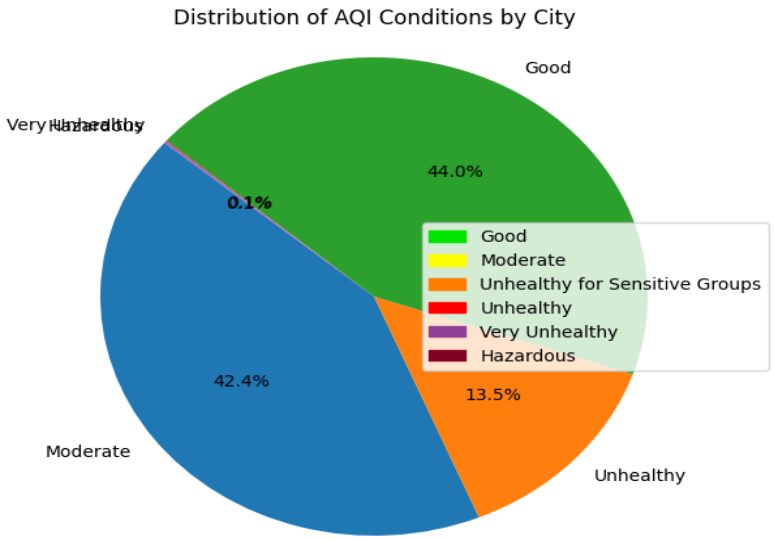
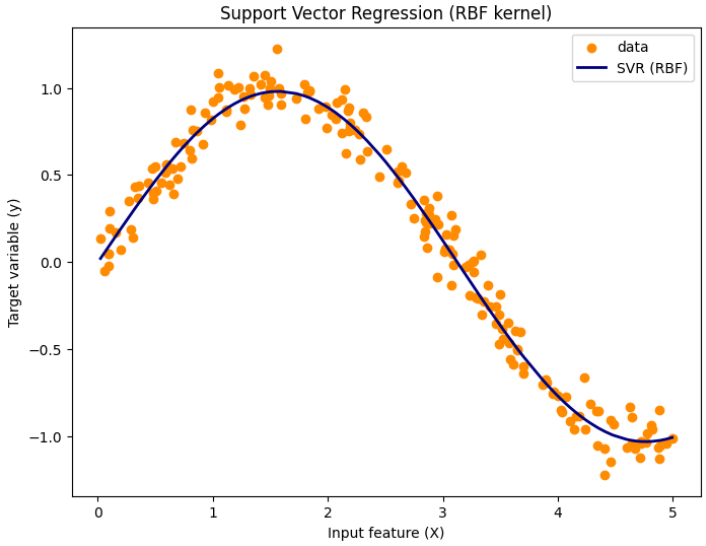


Figure 11, Shows the Support Vector Regression with RBF Kernal



VIII. CONCLUSION

The air quality problem poses significant health risks, so there needs to be effective monitoring and predictions. This study employed two machine learning approaches—Facebook Prophet for time-series forecasting and Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel for regression analysis—to evaluate air quality based on historical data from Himachal Pradesh. Through in-depth data pre-processing and model evaluation using metrics like mean absolute error (MAE) and root mean squared error (RMSE), our findings shows that SVR achieved superior performance, when compared to Prophet in terms of RMSE and MAE .

This research underscores the critical role of advanced machine learning techniques in informing public authorities and policymakers, facilitating the development of targeted strategies to mitigate air pollution and safeguard public health. By providing insights into model performance, we contribute valuable guidance for future research and decision-making processes aimed at addressing air quality management challenges. The results indicate that the predictive machine learning model offers a promising solution for air pollution prediction in Himachal Pradesh making it a valuable tool for sustainable air quality management and environmental protection in the region in the future. Overall, studying AQI prediction using machine learning in Himachal Pradesh is necessary for protecting public health, the environment, and the local economic growth.

As future work, we will try to improve SVR to forecast air quality through more parameters and dataset.

VIII. BIBLIOGRAPHY

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A Machine Learning Approach to Predict Air Quality in California. *Complexity*, *2020*. https://doi.org/10.1155/2020/8049504

Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. *International Journal of Environmental Science and Technology*, *20*(5), 5333–5348. https://doi.org/10.1007/s13762-022-04241-5

Madan, T., Sagar, S., & Virmani, D. (2020). Air Quality Prediction using Machine Learning Algorithms-A Review. *Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020*, 140–145. https://doi.org/10.1109/ICACCCN51052.2020.9362912

Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019). *"A Machine Learning Model for Air Quality Prediction for Smart Cities*. 452–457. https://doi.org/10.1109/WiSPNET45539.2019.9032734.Abstract

Sangeetha, P. (n.d.). Assessment of air quality index and pollutants in Chennai city, India: Pre-Covid and during-Covid period. In *Proceedings of the International Academy of Ecology and Environmental Sciences* (Vol. 2022, Issue 4). www.iaees.org

Sinha, A., & Singh, S. (2020). Review on air pollution of Delhi zone using machine learning algorithm. *Journal of Air Pollution and Health*, *5*(4). https://doi.org/10.18502/japh.v5i4.6446