DEVELOPING A MACHINE LEARNING MODEL FOR AIR POLLUTION PREDICTION

Aasika ES[1a]
Sangavi G[1b]
Jeeva D[1c]

Keywords:
Air pollution; machine learning; prediction model; Evaluation metrics; web application

A B S T R A C T

In India, air pollution is still a major environmental and public health issue because it can cause a variety of respiratory and cardiovascular conditions. The goal of this project is to create a precise machine learning model that can be used to forecast air pollution levels in India. The model will add pollutants such as PM2.5, PM10, NO2, SO2, and O3, using real-time data from India.gov. We will investigate and assess the different type of algorithms such as random forests, support vector machines, Cat Boost regression, and advanced decision trees, using measures such as mean squared error, root mean square error, and R-squared. The public, academics, and legislators will be able to access trustworthy air pollution forecasts to the implementation of the top-performing model as an intuitive online service or API. This easily navigable tool will help make well-informed decisions and increase awareness of this important environmental issue.

1. INTRODUCTION

Air pollutants poses a substantial danger to public health and the environment. Accurately predicting air pollution levels is vital for implementing powerful mitigation strategies and informing the general public approximately ability health risks. Beyond the health implications, air pollutants may have a long way-reaching results for agriculture, ecosystems, and weather alternate. Certain air pollutants, such as floor-stage ozone and particulate rely, can harm vegetation, lessen yields, and disrupt plant increase, posing a chance to meals security. Moreover, air pollutants contribute to the degradation of natural habitats, affecting biodiversity and ecosystem services which can be critical for human properly-being. In current years, the sector of system getting to know has revolutionized numerous domains via offering effective strategies for studying big and complex datasets. Machine gaining knowledge of fashions have the ability to find problematic styles and relationships inside data, making them well-appropriate for the mission of air pollution prediction.

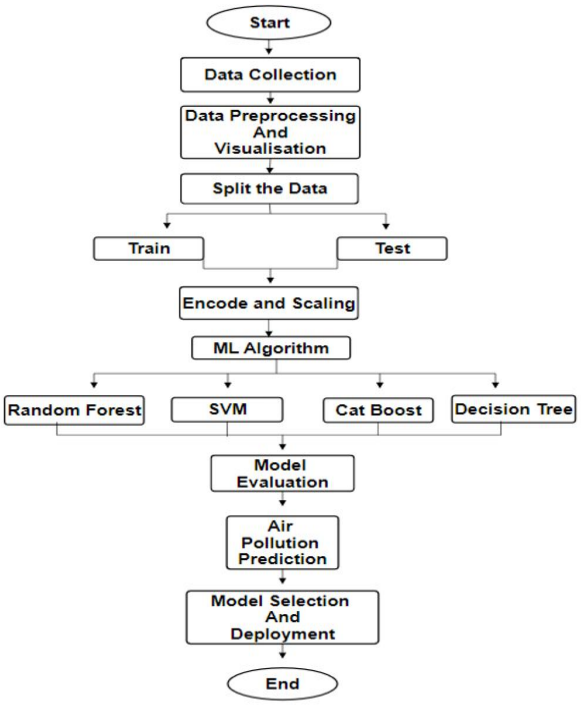
The authors of [1] highlights the urgent need for reliable prediction models because of the serious air quality challenges that are common in metropolitan areas. It also gives a case study on air pollution prediction in Indian cities using machine learning approaches. It highlights how precise forecasting can help decision-makers carry out focused actions to reduce pollution levels and successfully safeguard the public's health. In addition to talking about air quality prediction techniques, the research paper [2] focuses on the drawbacks of conventional monitoring techniques, namely their expensive nature and narrow geographic coverage. The goal of the project is to overcome these obstacles and offer scalable, affordable solutions for forecasting air pollution levels in diverse scenarios by utilizing supervised machine learning techniques. In addition to using machine learning approaches to forecast air quality, the research study in [3] also emphasizes the possibility of combining several data sources, such as meteorological and satellite imaging, to increase prediction accuracy. The study attempts to offer thorough insights into the factors impacting air pollution levels and their temporal and spatial variability by utilizing a holistic approach to data analysis. The authors in [4] highlights the evaluation explores the wider societal costs associated with air pollution, including effects on productivity, healthcare expenditures, and quality of life, in addition to its economic implications. It highlights how important it is for decision-makers to take these wider ramifications into account when developing successful methods to deal with air quality problems. In [5], the study

addresses the significance of model evaluation and validation to guarantee the accuracy of forecasting findings in addition to examining several regression strategies for PM2.5 predictions. The research attempts to give stakeholders confidence in the accuracy of the produced prediction models by using rigorous evaluation techniques, thereby enabling informed decision-making. The authors in [6] demonstrates the difficulties when modeling complex environmental systems in addition to suggesting a deep learning method to predict the air pollution level. It highlights how crucial scalability, interpretability of models, and data quality are to creating workable solutions for real-world applications. In addition to synthesizing the body of literature on data mining and machine learning in air pollution epidemiology, the research study in [7] also points out gaps and restrictions in the state of the field. It urges more multidisciplinary cooperation between epidemiologists, environmental scientists, and data scientists to tackle these issues and improve our knowledge of how air pollution affects human health. In the study [8], the authors not only suggest a supervised machine learning methodology but also highlights the significance of feature selection and data preprocessing in the creation of models. It emphasizes how crucial meticulous data preparation is to guaranteeing prediction models' dependability and applicability in many environmental scenarios.

In addition to presenting a machine learning-based model for PM2.5 estimations, the authors [9] shows how these models may be used to support policy choices and public health initiatives. It highlights how crucial stakeholder participation and cooperation are to guaranteeing the applicability and relevance of prediction tools in practical contexts. This study[10] investigates the integration of sensor networks and Internet of Things (IoT) technologies for real-time monitoring, in addition to talking about the accurate prediction of air pollution. It draws attention to the revolutionary potential of these technologies in promoting community involvement in air quality management and enabling data-driven decision-making. In addition to suggesting a multi-output machine learning model for air pollution forecasting, this paper [11] also addresses the computational efficiency and scalability of such models. It highlights how important it is to have scalable algorithms that can process massive environmental datasets and give decision-makers accurate projections in a timely manner**.** In the research study [12] they demonstrated the application of machine learning methods to improve deterministic air pollution forecasts is a noteworthy development in environmental prediction skills. Furthermore, these algorithms facilitate more informed decision-making and proactive pollution control measures by enhancing the forecasts' accuracy and dependability**.** The authors in [13] used machine learning techniques to enhance air quality forecasting skills is crucial for tackling the more complicated problems that pollution presents. Through the utilization of sophisticated computational techniques, scientists can improve their comprehension of the dynamics of air quality and create more efficient plans for mitigating pollution the research study [14] covers the biological mechanisms and pathways that underlie the effect of air pollution with respect to climate change on respiratory allergies. It highlights how crucial multidisciplinary study is to understanding the intricate relationships between environmental influences and the consequences for human health. In addition to discussing the use of satellite remote sensing data for PM2.5 assessments, the authors [15] also draws attention to ongoing initiatives aimed at enhancing the accuracy and resolution of data. It highlights how cutting-edge remote sensing technologies, such unmanned aerial vehicles and hyperspectral photography, have the ability to offer comprehensive insights into the dynamics of air quality at local and regional scales.

Our main goal is to create a thorough framework that would close the gap in India's air pollution prediction models. While previous research frequently focuses on specific pollutants, like PM2.5 or NO2, my method incorporates a number of categories pollutant characteristics to forecast average pollution levels across various geographic areas. The goal of this comprehensive approach is to offer more precise and nuanced insights for efficient pollution control plans suited to the many environmental circumstances found in India.

2. FLOWCHART



3. METHODOLOGY

3.1 DATA COLLECTION

The information was gathered from the daily updated india.gov webpage, which offers real-time statistics. The most recent update to the dataset we are using was made at 10:00 AM on February 22, 2024. 11, characteristics (ID, country, state, city, station, last update, latitude, longitude, pollutant_id, pollutant_min, pollutant_max, and pollutant_avg) and 3,305 data points make up this dataset.

Pollutant Types:

PM10: 0.01 mm-diameter particulate matter, frequently present in smoke and dust. Its small size allows it to enter the lungs deeply and impact respiratory and cardiac health.

PM2.5: Particles that are even smaller, with a size of 2.5 microns. Reduced visibility and foggy air are caused by these particles.

Ozone (O3): A secondary contaminant produced when smog builds.

Nitrogen Dioxide (NO2): A gaseous air pollutant released during the high-temperature combustion of fossil fuels like coal and oil. Allergies and respiratory issues might result from NO2 exposure. **Carbon Monoxide (CO):** It's a poisonous, tasteless, odorless, and colorless air contaminant that is created when incomplete combustion methods are used. **Ammonia (NH3):** By reacting with acidic species in the environment, ammonia (NH3) helps to generate secondary particle dependencies. **Sulfur Dioxide (SO2):** Another gas released during the burning of fossil fuels, which is mostly used in commercial initiatives.

3.2 DATA PREPROCESSING

There are no noisy data points in our dataset, but we did find 247 missing values in each of the pollutant, pollutant, and pollutant columns. To fill in these gaps, we imputated the corresponding column means.

3.3 VISUALIZATION

Representing records visually, using charts, graphs, maps, and other graphical elements, is known as data visualization. It makes it possible for styles, trends, and correlations in the data to be more easily identified and shared. For the purpose of evaluating records, providing information, assisting in decision-making, and communicating insights from complicated datasets, effective statistics visualization is essential.

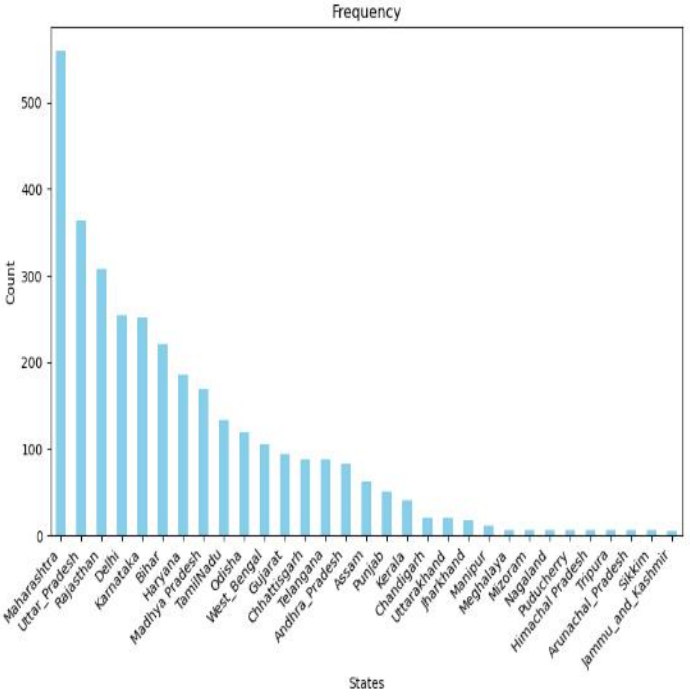


Fig 1. Represent the bar plot of air pollution of various states

From Fig 1. shows the frequency distribution across states is shown in a bar graph; Maharashtra has the highest count, approximately 490, followed by the states namely Uttar Pradesh, Tamil Nadu, Gujarat, and West Bengal, all of which have counts between 300 and 400. A disparate geographical representation in the dataset is indicated by the lower frequencies below 200 in a number of states.

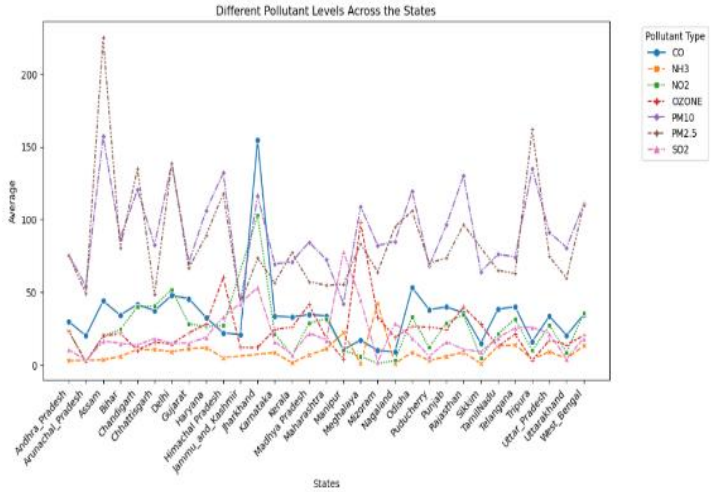


Fig 2. Represent the line plot of various pollutant across the states

From Fig 2. shows the state-by-state variations in PM10, NO2, NH3, OZONE, PM2.5, CO, and SO2 concentrations are shown in the line graph, with some states having higher concentrations of particular pollutants than others. This diversity emphasizes the necessity of implementing focused pollution

management strategies that are adapted to the unique pollutant profiles of each state.

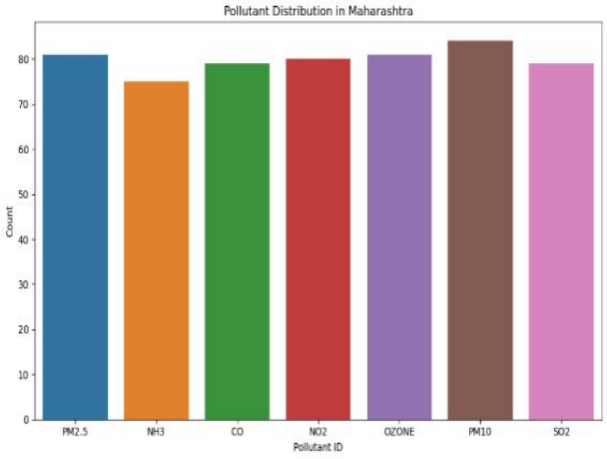


Fig 3. Represent the bar plot of pollutant distribution in Maharashtra

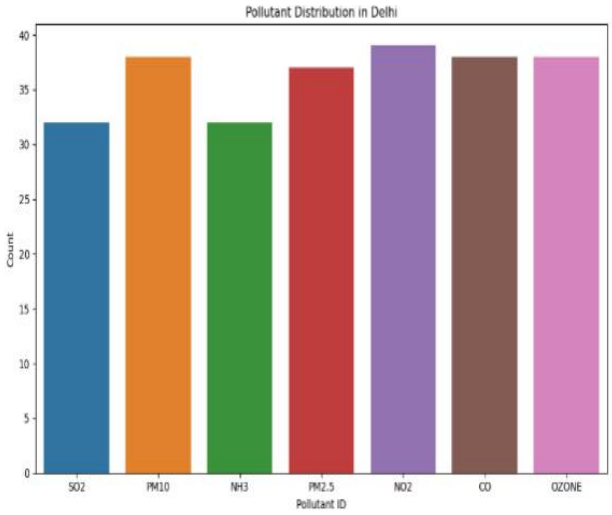


Fig 4. Represent the bar plot of pollutant distribution in Delhi

From Fig 3. and Fig 4. shows Delhi has comparatively greater levels of NO3 (nitrates) and PM10 (coarse particles), whereas Maharashtra has higher levels of PM2.5 (fine particulate matter) pollution, according to the bar plots. The divergent patterns of pollutants indicate distinct primary causes and air quality issues in these two states.

3.4 ENCODING

One-hot encoding is a technique used to convert categorical variables right into a format that can be without difficulty understood and processed by using device learning algorithms. The columns ‘city’ and ‘pollutant_id’ need to be one-hot encoded. This approach that every unique fee in those columns might be represented as a separate binary column inside the transformed facts. One-hot encoding is useful when running with express variables as it allows device learning algorithms to treat these variables as numerical inputs even as retaining their authentic meaning.

3.5 SCALING

Scaling is crucial because many gadget gaining knowledge of algorithms are sensitive to the dimensions of the enter functions. Features on large scales can dominate those on smaller scales, main to bias inside the model. By scaling the functions, we ensure that all capabilities make contributions equally to the version's predictions.
The MinMaxScaler, specifically, scales the capabilities through subtracting the minimal value and dividing by means of the variety (most value - minimum fee). This transforms the features to a not unusual scale, usually between zero and 1.

3.6 SPLIT THE DATA

In system studying, splitting the available records into separate training and testing out sets is an essential step. To train the model, the training set of data has to be applied, allowing it to study styles from the information, while the checking out set evaluates the model's overall performance on unseen statistics. In our data we split training is to allocate 70% and testing is to allocate 30%.

3.7 ML ALGORITHM

a. Random Forest:

An ensemble learning method called Random forest, that combines more than one decision tree to improve predictive accuracy and manipulate over fitting. It works through building more than one choice bushes on distinctive subsets of the information and aggregating their predictions.

$$f(x) = 1/M * \Sigma(m=1 \text{ to } M)\ fm(x) \qquad (1)$$

From equation (1):
(a) x denotes the input feature vector
(b) M represents the total number of decision trees in the ensemble
(c) fm(x) denotes the prediction of the m-th decision tree

b. Support Vector Machine (SVM):

A supervised learning algorithm that can be applied to any type of regression problem is called the Support Vector Machine(SVM). It unearths the foremost hyperplane that separates instructions with the maximum margin, making it powerful for high-dimensional areas.

$$W^T X + b = 0 \qquad (2)$$

From equation (2):
(a) W denotes the weight vector, perpendicular to the hyperplane
(b) X represents the input feature Vector
(c) b denotes the bias term
For non-linear instances, the statistics is mapped into a higher-dimensional space the use of kernel functions, and the hyperplane is discovered in that space.

c. Cat Boost:

Cat Boost (Categorical Boosting) is a gradient boosting framework advanced with the aid of Yandex. It is designed to address express features efficiently and can robotically manage missing values. Cat Boost uses an ordered goal encoding scheme for specific features, that is based at the target mean price for every category.

$$F(x) = F\theta(x) + \Sigma(m=1 \text{ to } M)\ \gamma m * fm(x; \theta m) \qquad (3)$$

From equation (3):

(a) Fθ(x) is the initial prediction model, often a constant or the mean of the target variable.
(b) M is the total number of iterations (Weak learners)
(c) fm(x;θm) is the m-th weak learner with parameters Θm
(d) γm is the step size (learning rate) for the m-th iteration

The algorithm works as follows:
1. Initialize the model with Fθ(x)
2. For each iteration m from 1 to M
3. Compute the negative gradient (or residuals) of the loss function with respect to the current model as:
rm = -∂L(y,F(x)) / ∂F(x)
4. Fit a weak learner fm(x;θm) to the negative gradients rm.
Update the model by adding the new weak learner with a step size γm:

From equation (3)

The key idea is to iteratively minimize the loss function by adding new weak learners that correct the errors made by the previous models.

d. Decision Tree:

Decision Trees are a kind of supervised gaining knowledge of set of rules that may be employed to each classification and regression problems. They work with respect to recursively partitioning the enter area primarily based on the feature values, developing a tree-like shape of selections and their feasible results.

$$f(x) = \Sigma(I=1 \text{ to } k)\ ci * I(x \in Ri) \qquad (4)$$

From equation (4):
(a) x denotes the input feature vector
(b) Ci represents the value associated with the i-th leaf node
(c) Ri denotes the region of the input space corresponding to the i-th leaf node
(d) I (x ∈ Ri) is an indicator function that returns 1 if x falls in region Ri, and 0 otherwise.

3.8 EVALUATION METRICS

a. R Squared:

R-squared is a statistical measure that represents the proportion of the variance in the dependent variable this is explained by the independent variables in a regression version. It ranges from 0 to 1, with higher values intimate a higher fit of the model to the information.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2} \qquad (5)$$

From equation (5):
(a) Y_i denotes the observed value of the dependent variable
(b) $\hat{Y}$ denotes the predicted value of the dependent variable
(c) $\bar{Y}$ denotes the mean of the true values of the dependent variable

b. Mean absolute error:

Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |yi - \hat{y}| \qquad (6)$$

From equation (6):

(a) N is the total number of instances
(b) Yi is the observed value of the dependent variable
(c) Ŷ is the predicted value of the dependent variable

c. Mean squared error:

The Mean squared error defines the average of the squared difference between the original and predicted values. It measures the variance of residuals.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(yi - \hat{y})^2 \qquad (7)$$

From equation (7):
(a) N is the total number of instances
(b) Yi is the observed value of the dependent variable
(c) Ŷ is the predicted value of the dependent variable

3.9 DEPLOYMENT

An application that has been extensively tested is launched for use in a production setting. It entails transferring the program from the testing or development environment to the operational servers or systems. During deployment, meticulous planning and execution are essential to guarantee a smooth transition and reduce downtime or interruptions. To properly address possible problems or rollbacks, appropriate backup plans, rollback schedules, and monitoring systems must be in place. A well-managed deployment procedure is necessary to provide end users with a high-caliber software product on time.

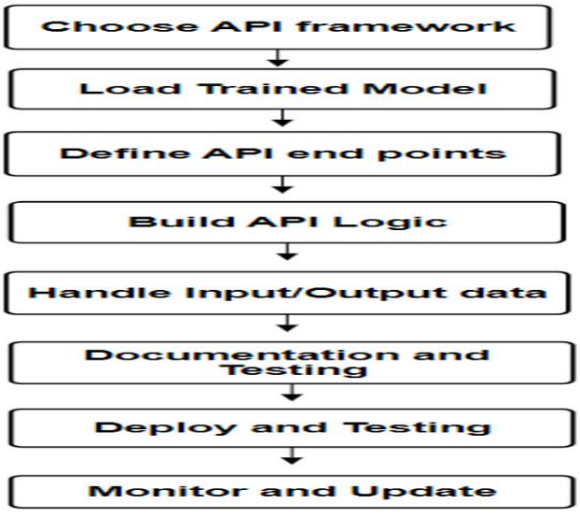


Fig 5. Represent the work flow to create the website

4. FINDINGS AND DISCUSSIONS

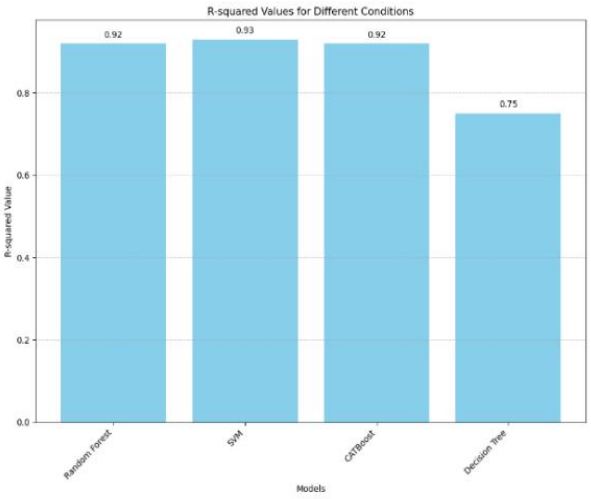


Fig 6. Represent the bar plot of model accuracy

From Fig 5. says the R-squared values, a gauge of how well models match the dataset, are shown for various situations or models in a bar plot. Out of all the models shown, the SVM model has the greatest R-squared value of 0.93, suggesting the best fit. The R-squared values for Cat Boost and Random Forest are both 0.92, but the second Random Forest model's value is 0.75, indicating a worse match than the other two.

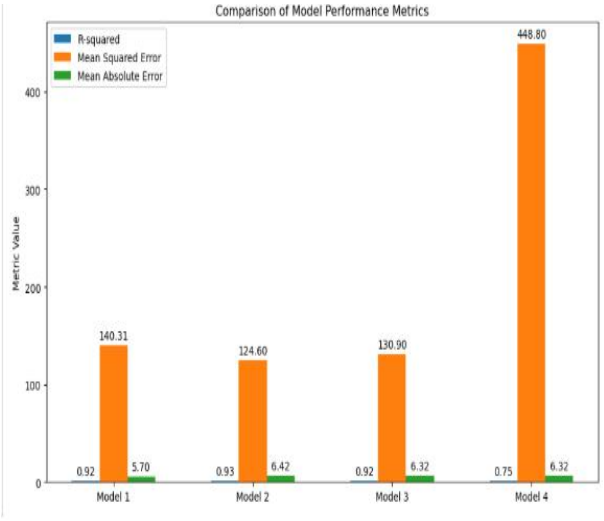


Fig 6. shows A comparison of various performance measures amongst four models is shown in the bar plot. The model with the greatest fit, Model 2, has the highest R-squared value of 0.93 out of all the models. On the other hand, Model 4 exhibits the worst performance, with the lowest R-squared value of 0.75 and the largest Mean Absolute Error of 348.80. A well-informed decision may be made when selecting a model

thanks to the displayed metrics, which include R-squared, Mean Squared Error, and Mean Absolute Error.

These metrics shed light on the models' accuracy and predictive power. A comparison of various performance measures amongst four models is shown in the bar plot. The model with the greatest fit, Model 2, has the highest R-squared value of 0.93 out of all the models. On the other hand, Model 4 displays the greatest Mean Absolute Error of 348.80 and the lowest R-squared value of 0.75, indicating the least successful performance.

A well-informed decision may be made when selecting a model to the displayed metrics, which include R-squared, Mean Squared Error, and Mean Absolute Error. These metrics shed light on the models' accuracy and predictive power.

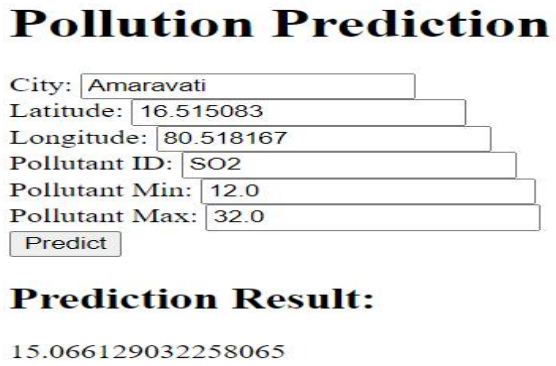


Fig 7. Represent the integrating using Flask as a web application

From Fig 7. says by tracking and predicting pollution levels in certain areas, this interface can help the public and government take the necessary action to reduce environmental problems. The prediction outcome, which comes from underlying models or algorithms, can help with public health and air quality management decision-making.

5. CONCLUSION

In this study, we created and assessed a number of machine learning models to reliably forecast India's air pollution levels. Our models concentrated on important pollutants including PM2.5, PM10, NO2, SO2, and O3 by using real-time data from the India.Gov portal. After extensive testing, the Support Vector Machine (SVM) algorithm was found to be the best-performing model. Its greatest R-squared value of 0.93 indicated that it was the model that fits the data the best out of all the models that were examined. The effective execution of this project demonstrates the potential of machine learning methods in decision assistance and environmental monitoring systems. Policymakers, researchers, and the general public can easily receive trustworthy air pollution forecasts with our user-friendly online application or API. This easily available platform can help people make well-informed decisions, increase public awareness, and aid in the creation of successful mitigation techniques.

6. FUTURE WORKS

a. **Include other data sources:** Including information from satellite observations, crowd sourcing projects, and regional monitoring stations could help the model's geographical resolution and prediction accuracy even further.

b. **Examine cutting-edge methods for deep learning:** Deep neural network developments in recent years and their capacity to represent intricate non-linear interactions may be used to forecast air pollution even more precisely.

c. **Create regionalized models:** More contextualized predictions might be obtained by customizing models to particular cities or regions by taking into consideration distinct geographic, meteorological, and emission factors.

d. **Include other toxins and environmental elements:** A comprehensive environmental monitoring system might be created by expanding the project's scope to include additional pollutants like heavy metals and volatile organic compounds (VOCs), as well as environmental elements like noise pollution.

e. **Combining early warning systems:** Forecasts of air pollution could be included into early warning systems to allow for timely advisories and notifications to reduce health risks, especially for vulnerable populations.

References:

[1] Aditya, C. R., Deshmukh, C. R., Nayana, D. K., & Vidyavastu, P. G. (2018). Detection and prediction of air pollution using machine learning models. International journal of engineering trends and technology (IJETT), 59(4), 204-207.

[2] Babu, K. M., & Beulah, J. R. (2019). Air quality prediction based on supervised machine learning methods. International Journal of Innovative Technology and Exploring Engineering, 8(9), 206-212.

[3] Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. Journal of big Data, 8(1), 1-21.

[4] Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health, 17, 1-19.

[5] D'amato, G., Pawankar, R., Vitale, C., Lanza, M., Molino, A., Stanziola, A., ... & D'amato, M. (2016). Climate change and air pollution: effects on respiratory allergy. Allergy, asthma & immunology research, 8(5), 391.

[6] Gladkova, E., & Saychenko, L. (2022). Applying machine learning techniques in air quality prediction. Transportation Research Procedia, 63, 1999-2006.

[7] Halsana, S. (2020). Air quality prediction model using supervised machine learning algorithms. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8, 190-201.

[8] Harishkumar, K. S., Yogesh, K. M., & Gad, I. (2020). Forecasting air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Computer Science, 171, 2057-2066.

[9] 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.

[10] Kumar, S., Mishra, S., & Singh, S. K. (2020). A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere. Heliyon, 6(11)

[11] Lin, C., Li, Y., Yuan, Z., Lau, A. K., Li, C., & Fung, J. C. (2015). Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM2. 5. Remote Sensing of Environment, 156, 117-128.

[12] Rakholia, R., Le, Q., Ho, B. Q., Vu, K., & Carbajo, R. S. (2023). Multi-output machine learning model for regional air pollution forecasting in Ho Chi Minh City, Vietnam. Environment International, 173, 107848.

[13] Xu, X., Yang, H., & Li, C. (2022). Theoretical model and actual characteristics of air pollution affecting health cost: a review. International Journal of Environmental Research and Public Health, 19(6), 3532.

[14] Zhang, Z., Johansson, C., Engardt, M., Stafoggia, M., & Ma, X. (2024). Improving 3-day deterministic air pollution forecasts using machine learning algorithms. Atmospheric Chemistry and Physics, 24(2), 807-851.

[15] Zhang, Zhiguo, et al. "Improving 3-day deterministic air pollution forecasts using machine learning algorithms." Atmospheric Chemistry and Physics 24.2 (2024): 807-851.

Aasika ES
Vellore Institute of Technology,
Chennai,
India
ashika.2023@vitstudent.ac.in

Sangavi G
Vellore Institute of Technology,
Chennai,
India
sangavi.g2023@vitstudent.ac.in

Jeeva D
Vellore Institute of Technology,
Chennai,
India
jeeva.2023@vitstudent.ac.in

Dr. David Maxim Gururaj
Vellore Institute of Technology,
Chennai,
India
davidmaxim.gururaj@vit.ac.in