Air Quality Analysis

Veerababu Addanki
University Of Colorado
Boulder, Colorado, USA
vead7397@colorado.edu

Benarjee Sudeep Sampath Pyla
University Of Colorado
Boulder, Colorado, USA
bepy5070@colorado.edu

Ghana Gokul Gabburi
University Of Colorado
Boulder, Colorado, USA
ghga5839@colorado.edu

Abstract

The Air Quality Index (AQI) measures the health hazards and environmental circumstances of those who are directly exposed to contaminated air. For the purpose of managing health risks and preventing pollution, its forecast is therefore essential.The idea of machine learning is used in this work to estimate AQI using regression models. Using the data augmentation technique on the given dataset, the characteristics such geographic location and pollution-relatedfeatures of the domain, ElasticNet, Random Forest, XGBoost, Linear Regression, Random Forest, and Gradient Boosting were developed, assessed, and validated. The findings also demonstrate that hand hyperparameter adjustment can improve performance, particularly for tree-based models. Principal Component Analysis (PCA) plots and feature relevance rankings are two examples of the visualizations that provide insightful information about the underlying data patterns and model behavior.

Keywords

Linear Regression,Random Fores,Gradient Boosting,ElasticNet,Principal Component Analysis (PCA)

1 Introduction

Air quality remains to be probably one of the most, if not the most sensitive global issues, for it directly affects human health as well as the sustenance of ecosystems in varied parts of the world. Low air quality is associated with the occurrence and severity of respiratory and cardiovascular diseases, as well as negative impact on plants and animals. With an increase in pollution levels secondary to massive urbanization, industrialization, and social demography, monitoring and predicting the Air Quality Index (AQI) is becoming quite a challenge—not only as a problem of the past but a more prominent issue for this future. To this end, the advance in data science, including but not lim- ited to machine learning, has provided new approaches to studying and combating air pollution. Applying machine learning provides the knowledge to deal with huge volumes of the environmental ata set and reveal their structure and relationships for generat- ing precise outcome predictions. These predictive findings enable the stakeholders improve on the strategies towards pollution con- trol through early interventional measures. The purpose of this study is approached with the objective of examining and improving AQI prediction using an advanced form of artificial intelligence known as machine learning to close the gap howh between raw environmental data and intelligence processing. his research applies micro-optimization for a number of algo- rithms using linear and non-linear models for a selection of prob- lematic areas including but not limited to; data variability feature selection and high computation time. But, by ignoring the details of oprationalizing large-scale environmental datasets the research guarantees the stability of the developed models. Furthermore this work explores the mechanism of variation of AQI Observe the ori- gin of declining air quality, including geographical, meteorological, and anthropogenic factors for a better understanding of the factors influencing variation of AQI.part from the forecast ability, this project is concerned with the interpretability of the model and its applicability. The analysis in terms of Principal Component Analysis (PCA) or feature impor- tance allows policymakers and environmental scientists to make relevant conclusions from the data. Furthermore, the findings help to define and rank the most effective preventive features that have an impact on AQI and serve as the starting point for the further fight against pollution. This paper also provides the further direction for enhanced AQI prediction as using live stream data, incorporating new deep learn- ing models, and extending dynamics of AQI-Climate change. Such advancements make it possible to design and develop highly flexible, anticipatory systems which operate in a manner that is sensitive to the state of their environment.

Employing the state of the art machine learning approaches, this research responds to the growing call for sustainable solutions to air quality problems. It brings about a system of constant improvement of environmental assessment hence a healthy world where we can all enjoy our health and still see the environment balance itself in times of turmoil. Therefore, aside from benefitting the broader field of environmental analytics through enhancement of methods for charting complex relationships between the environment and food security and/or climate change, this work directly contributes to promoting concrete improvements within the lives of societies and individuals worldwide.

2 Related Work

Over the last ten years, vast progresses have been made in the studies of air quality. have examined emissions characterization, Air quality observations, risk assessment, exposure assessments, and modelling. Key developments as the ability to maintain greater certainty with emission inventories and combine mounting data from satellite and ground observations, employing low-cost sensor and ex- ploring climate interactions. Recommendations apply to dynamic modeling, pollutant distribution on a fine grid, and efficient and flexible air quality management to efficiently strive new challenges.This analysis of this research combines Air Quality Index (AQI) and meteorological data used viding guidance to simulatively implement a column-based model to emerge causal relations between pollutants and weather conditions in Jakarta city. The integrated data improves the forecast using LSTM & GRU models to improve the forecasting capability significantly. to expose relations that underlie the regulation of air pollution. Future This work builds on other prior work to consider additional climate change connections.Work presented here examined the following

connection in more detail. Air pollution trend of India for the periods 2015-2020 by Central Pollution Control Board. evaluation of the Control Board data, on SO, NO, PM, and other pollutants. CO. It presented COVID-19 lockdown effects; it demonstrated appreciable pollutant reductions. Recommendations include funding clean tech- nologies and community led programmes in the cities such as Ahmedabad. Future work intends to find out more flexible pollution models.This research researched global alteration of AQI during COVID-19 pandemic period and data has been explored from 87 cities pre and post 2020. Analyses of the findings indicated increased air qual-. y during 2020, while PM2.5, PM10, NO and SO decreased. to pandemic restrictions. However, post-2020, the AQI escalated, which has listed eleven scripts detailing the elevation of chemical deterioration and poor air quality in the country. as restrictions eased. This paper brings out learning that reveal areas of policy. initiated air quality enhancements.A study that was undertaken sought to compare residence based further analysed by Reserve-based (RB) and mobility-based (MB) methods in order to evaluate air pollution. exposure and perception, to understood the context by which inequalities emerge. uncertainties. It pointed at inequality between real and perceived levels of air , chemical quality, social demographic and environmental factors involved in quality control and assurance. ctor, and stressed that, for effectiveness, mobility data had to be viewed as crucial. analysis and improvement in policy making.

3 Dataset

3.1 Collected Datasets From The Website

The details of the datasets are mentioned below whaich are collected from the U.S. Environmental Protection Agency.

3.1.1 daily_44201_2024. This dataset contains air quality measurements for ozone levels at a monitoring station. It includes details such as geographical coordinates, measurement methods, and specific ozone standards. Data points cover the arithmetic mean concentration, maximum values, and observation counts, all recorded over an 8-hour averaging period. The dataset is essential for assessing air quality trends and ensuring compliance with environmental standards shown in Fig 1

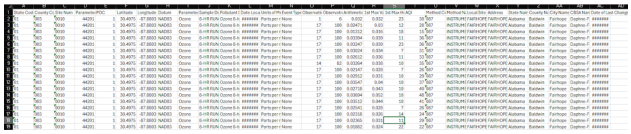

Figure 1: Dataset 1

3.1.2 daily_44201_2024. This dataset contains hourly nitrogen dioxide (NO2) measurements from a monitoring station. It includes details such as geographic coordinates, measurement method, and compliance with the NO2 1-hour standard. Key metrics such as arithmetic mean, maximum values, observation counts, and the Air Quality Index (AQI) are recorded in parts per billion. This data is essential for evaluating air quality trends and their implications for public health shown in Fig 2

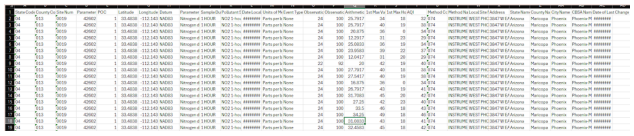

Figure 2: Dataset 2

3.1.3 daily_42401_2024. This dataset contains hourly measurements of sulfur dioxide (SO2) from a monitoring station. It includes details such as geographic coordinates, measurement methods, and compliance with the SO2 1-hour standard. Key metrics like arithmetic mean, maximum values, observation counts, and the Air Quality Index (AQI) are recorded in parts per billion. This data is essential for assessing air quality trends and ensuring compliance with environmental standards ,shown in Fig 3

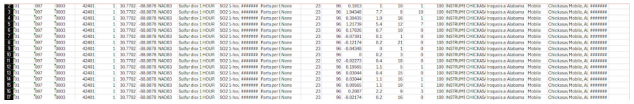

Figure 3: Dataset 3

3.1.4 daily_44209_2024. This dataset includes hourly measurements of carbon monoxide (CO) from a monitoring station. It features details such as geographic coordinates, measurement methods, and compliance with the CO 1-hour standard. Key metrics such as arithmetic mean, maximum values, observation counts, and the Air Quality Index (AQI) are recorded in parts per million. This data is vital for evaluating air quality trends and ensuring adherence to environmental standards, shown in Fig 4

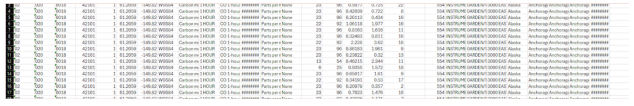

Figure 4: Dataset 4

3.2 Data Preparation And Cleaning

This integration process in the context of the primary research required aggregation of several relatively minor datasets into a single extensive data set. This was necessary so that any information was collected and compiled in one source for comparison of findings. As part of data cleaning at this stage, objects that refer to the same thing were removed to leave only one copy in the list to avoid bias due to inclusion of same object in data analysis in different forms. The second approach, as a result of duplicates, was to proceed to cases involving missing values with reference to the 'AQI' column. To ensure confidentiality of the data and to avoid misplacing vital information as a result of missing 'AQI' values, the missing values were imputed with the overall mean of this variable a step that did not affect the subsequent calculations while keeping the dataset complete.

However, missing data was not only limited to the 'AQI' column Only absolutely necessary fields However, most of the fields in

the 'demographic' and 'AQI_val' columns contained missing data. Where the data was missing in either the 'Local Site Name' or 'CBSA Name', rows with such data were omitted. This decision was made as these columns comprised of key variables differentiating each observation, and exclusion of which would severely jeopardize the validity of the dataset. This approach ensured that the rows with missing information on these columns were omitted, which in actual sense reduces the dataset to observations that possesses all the required information.

Outliers which are considered as very vital in decision making, but if not properly handled affect analysis and reduces the model accuracy, was another focus. In more detail, we singled out cells in the 'AQI' and 'Arithmetic Mean' row where the values were at least three standard deviations above the 99th percentile. Rather than completely eliminating these outliers, they were capped at 99% to limit their value to the highest measure, which then replaced any value over this measure. Less sensitive to outliers: For this method of normalization, the distribution of the data was retained while at the same time dealing with a heavy variance caused by outliers.

Subsequently, the evaluation used Min-Max scaling algorithms on the numeric columns in the dataset including 'AQI,' 'Arithmetic Mean,' '1st Max Value,' 'Observation Count,' and 'Observation Percent.' This transformed the data so that it had a scale between 0 and 1. Generally, scaling is done to enable all the features to have a comparable range of values so that none of the them dominate the analysis or modeling results.

For the categorical variables that include: 'State Name,' 'County Name,' a One-Hot Encoding technique was used. This technique processes categorical data to make them more interpretable by machine learning algorithms since each category has own binary column. However, in order to avoid multicollinearity which is when one feature is strongly related to another we eliminated the first category of each categorical variables. This step avoided redundancy by ensuring that the encoded variables did not provide the same information multiple times.

The generated binary columns corresponding to the One Handed encoded categorical features were grouped into a new DataFrame. This helped in a clean and proper arrangement of the encoded variables so that they could be merged analysed with the actual data set. Last of all, these categorical variables where encoded and organized and new binary columns were created, were then concatenated with the original data set. In order not to have redundant data, the first 'State Name' and 'County Name' columns were dropped from the dataset as they had been replaced by their code forms.

Such operations as duplicate removal, the handling of missing values, outliers' capping, and scaling of numbers, encoding of categories were performed to keep the dataset clean and consistent for further analysis or modeling or reporting. To avoid compromising the general structure of the dataset, these issues were resolved right at the beginning of the project, which enabled future high-level analysis and, in particular, the modeling phase.

3.3 Final Dataset

The dataset has 27 fields and 316,730 observations which include precise air quality samples from different. locations. Other attributes are additional location codes like State Code, County Code, and Site Num among others. position, geographical position, geographical co-ordinates- geographical latitude and geographical longitude. Information on pollutants is grouped under Parameter Name. measures that cover various time horizons which are described in sample duration. The dataset includes critical Environmental stability measures including AQI (Air Quality Index) and analytical parameters including AM (Arithmetic Mean) and Maximum values. There are of course numerical and numerical columns with mixed data types. This dataset is a rich source for the identification of patterns and trends of air quality is showed in Fig 5

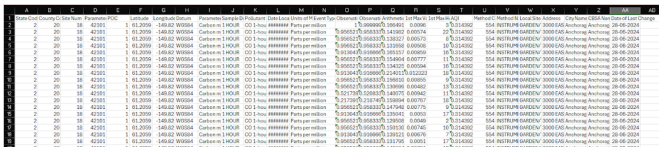

Figure 5: Dataset After Preprocessing

3.4 Data Exploration

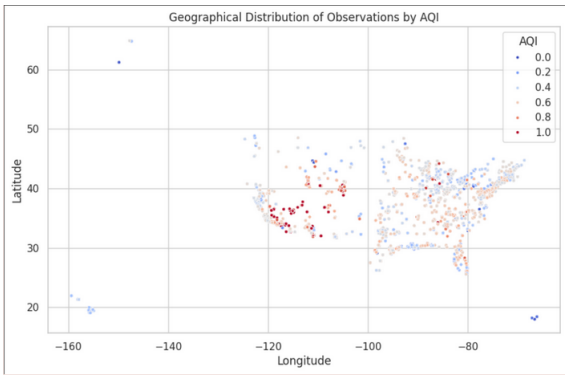

Figure 6: Geographical Distribution

Fig 6 visually, represents air quality levels across a geographical region using Latitude and Longitude coordinates. The color gradient transitions from blue to red, effectively highlighting low to high AQI values for easy identification of areas with good or poor air quality. Such visualizations are valuable for detecting spatial patterns in air pollution and identifying regions that may need policy interventions or further investigation.

The peaks in pollution levels around midnight and 10 AM indicate critical times for air quality, likely influenced by daily human activities such as traffic, industrial operations, or shifts in atmospheric conditions. The histogram highlights these patterns, providing valuable insights for planning monitoring schedules, implementing traffic regulations, or controlling industrial activities to mitigate pollution during peak hours from Fig 7

Fig 8 effectively identifies periods of rising AQI and declining air quality, as well as moments of improvement. It also suggests seasonal variations, with potential spikes during specific months, such as spring or summer, likely tied to human activities or natural events. The observed spikes and drops direct attention to specific

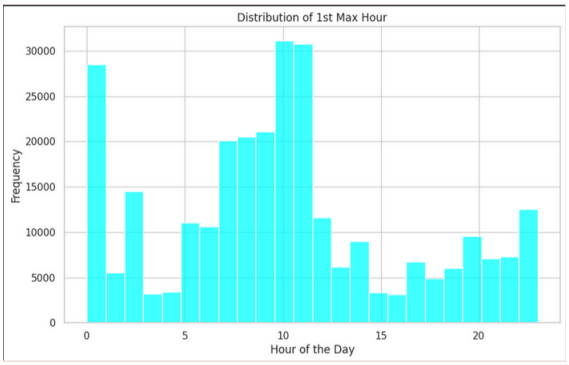

Figure 7: 1st Max Hour Distribution

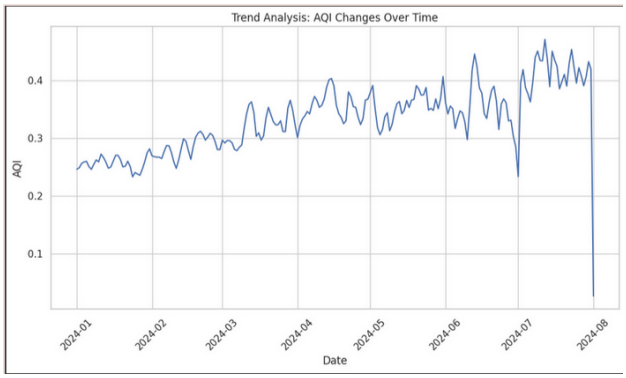

Figure 8: Trend Over Time

events warranting further exploration to uncover their causes, informing potential air quality interventions and policy adjustments.

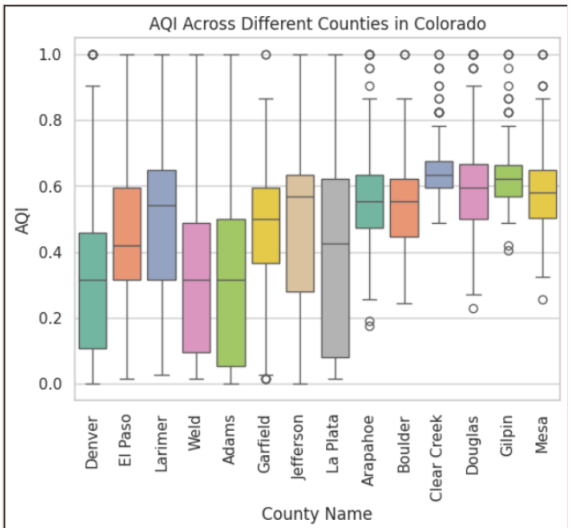

Figure 9: AQI in Colorado

Counties such as Jefferson and Adams exhibit higher median AQI values, indicating more frequent occurrences of poor air quality, likely due to factors like traffic, urban density, and proximity to industrial activities. In contrast, counties like Clear Creek and Gilpin show generally better air quality, characterized by lower median AQI values and fewer extreme pollution events. Outliers in the data represent instances of significantly worse air quality during environmental events, such as wildfires, emphasizing the need for targeted monitoring and mitigation efforts in these affected regions from Fig 9

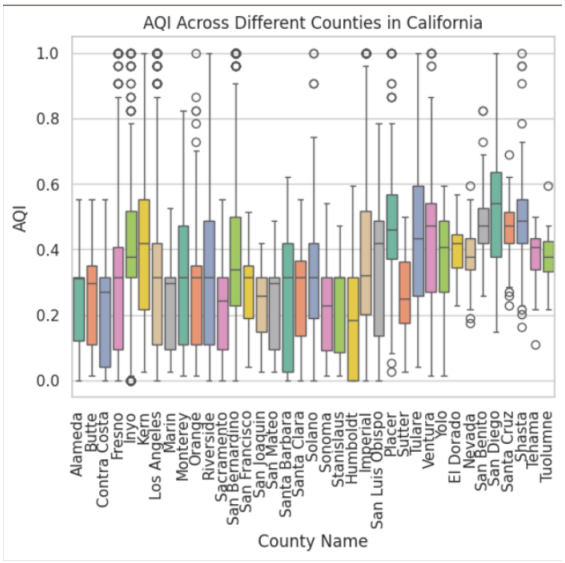

Figure 10: AQI in California

Fig 10 visualization enables policymakers to identify specific counties that require targeted air quality improvement initiatives based on their AQI distributions. By pinpointing counties with persistent air quality issues, decision-makers can better allocate resources for environmental monitoring and implement effective pollution control measures, ultimately enhancing overall air quality and public health.

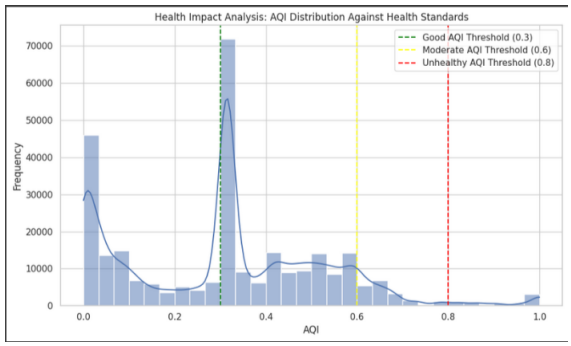

Figure 11: AQI Against Health Standards

Good air quality predominates, with most data points falling below an AQI of 0.3, indicating generally acceptable conditions according to health standards. However, a smaller yet significant

portion of the data exceeds an AQI of 0.6, reaching moderate to unhealthy levels that raise public health concerns. This histogram visually represents the distribution of air quality, helping assess how often standards are met or exceeded, which can inform health policies and actions From Fig 11

4 Dataset Preprocessing For Models

In order to ensure that the input models received very significant data, a number of basic procedures were adopted. First, any All datasets that the contain missing values in required attributes or the target attribute were not considered for the data analysis playground to help negate such a outcome. Then, features were scaled using StandardScaler so that the impact of various measurement units because many models are dependent of feature magnitude. Additionally, noise was was mimicked during the training period by introducing variation to the Air Quality Index (AQI) that was used. in order to give a more realistic view on the environmental conditions, and also to guarantee that the system remained robust is showed in Fig 12

Features after scaling (first 5 rows):

	Latitude	Longitude	Arithmetic Mean	Observation Count	1st Max Value	1st Max Hour
0	4.325844	-3.005335	-0.229410	0.757911	-0.231466	-0.932545
1	4.325844	-3.005335	-0.365126	0.602697	-0.402655	1.746070
2	4.325844	-3.005335	-0.385368	0.602697	-0.402958	-0.227646
3	4.325844	-3.005335	-0.422294	0.602697	-0.432011	0.054313
4	4.325844	-3.005335	-0.236797	0.447483	-0.276356	0.900192

Figure 12: Dataset After Preprocessing For Models

5 Methodology

5.1 Data Splitting

The dataset was divided into training (80%) and testing (20) phase in order to analyze the efficiency of the models identified. hidden information, quite resembling real world situations. Such approach ensured that a more qualitative evaluation regarding the aspect in the study was possible. where model performance is measured and how well the model performs on novel inputs that the model has never seen before.

5.2 Model Selection

5.2.1 Linear Regression. Linear regression analysis can be defined as a very basic statistical technique whereby a straight line is fitted on the facts with a view of showing how the dependent variable is influenced by one or more independent variable. About the variables, the model has a simple relationship as it has linear assumption, but that makes it useful in most situations. Linear Regression is selected as the first model, because it is simple and easy to interpret comparing to other complex models. It is particularly useful when the nature of the data is determined to be almost perfectly linear, and it offers values that denote the power and alike of each predictor upon the target variable. However, it is plagued with problems of high variance when test data have outliers or variables are highly correlated, and it has shortcomings in that it can only work in linear relationships in a data set.

5.2.2 Random Forest Regressor. The Random Forest Regressor is an algorithm that combines a number of decision trees with an objective of making predictions. In the process of building each tree in the forest, a random sample of the data is used together with a random sample of the features This has a very important advantage of improving the generality of the model since the trees are not fitted on the entire data set and all the features. Random Forest, therefore, gives a stable and precise result compared to the stand alone decision tree through the use of all the individual. This model is especially helpful when faced with non-linear data thus if faced with several complicated relationships between various variables, the model will inherently identify these interactions. Besides, the model is better suited to both numerical and categorical data and is not very much affected by outliers. However, as it offers favorable predictive accuracy, it can be considered a weakness when it increases computation costs as with many trees and features.

5.2.3 Gradient Boosting Regressor. Gradient Boosting Regressor is a complex form of ensemble learning method in which trees are constructed one at a time, and each tree tries to minimize the residuals of the model built before it. The method adopted in the study derives from the gradient descent scheme in which the new models are fitted closer to the gradient residual error. It makes a new set of modifications to the model in order to enhance the predictive value of the set or software in the current context. So, within Gradient Boosting, it is possible to utilize the interactions of features due to the fact that it is suitable for multi-dimensional nonlinear patterns. But, it is computationally more expensive and needs the selection of the number of trees as well as tree growing depth to prevent itself from overfitting. Gradient Boosting, however, tends to give extremely accurate outcomes even though it is computationally costly.

5.2.4 XGBoost Regressor. XGBoost which is short for Extreme Gradient Boosting, is actually a modification of Gradient Boosting which incorporates some features of optimization such as regularization, speed and performance. An improvement upon Gradient Boosting model is achieved by adding L1 (Lasso) and L2 (Ridge) penalties that aid in controlling overfitting resultant from a large number of features in the dataset. XGBoost is praised for being a fast algorithm and is most useful when coping with big data, missing values to not have to be imputed. It has been adopted because of its high accuracy especially when used in machine learning competitions outcompeting highly accurate models in terms of prediction. XGBoost model can handle sparse data, and it is available with parallel processing which accelerates the training process and has multiple boosting methods that makes it one of the versatile tools available in the field of predictive modelling.

5.2.5 K-Nearest Neighbors (KNN). K-Nearest Neighbors (KNN) is associative classification rule that implements decision algorithms that do not assume any particular form of the function between the given independent variables and dependent variable. At any given point, the algorithm computes the distance of the point in the feature space with a target point of interest, identifies K-nearest

neighbors, and their target values are averaged to arrive at a decision. Another clear strength of KNN is that it is suitable for both regression and classification tasks and does not assume the distribution of the data we are working with. One of the advantages of the KNN is in its simplicity and non-complexity to understand in addition it performs well with small to the middle-sized data set. However, it is less suitable for the high-dimensional data set (the curse of dimensionality) and also when data set is large because it is required to compute distance to all the training data. However, this technique can be hardly insensitive to irrelevant features which minimizes its advantage.

5.2.6 ElasticNet. Elastic Net is a method, which at the same time analyzes L1 (Lasso) and L2 (Ridge) linear regression. It is most useful when the variables are numerous and some may be multicollinear, a problem plagues simple linear regression models. Lasso component in ElasticNet works for feature selection, where some of coefficients are forced to zero for less important features – the coefficients of other features will be regularized towards zero; the Ridge component prevents the model from overfitting when the coefficient values are large. ElasticNet therefore is a good combination of Lasso which avoids the problem of dimensionality in high dimension datasets and Ridge which is stable but may overfit. This model enhances generalization by reducing the vice of overfitting without stripping off key explanatory variables. Nevertheless, ElasticNet suffer from some difficulties with hyperparameters' choosing, such as the parameter, which controls the proportion of the L1 and L2 norms of the coefficients vector.

5.3 Training and Fine-Tuning

The models were trained with the default hyperparameters first, followed by a manual intervention hyperparameter tuning. optimize their performance. There was no need for hyper-tuning the Linear Regression model since it does not apply this concept. parameters to adjust. In the Random Forest model, the tree depth was initialized to max_depth = 10 with an aim of avoiding called overfitting and enhance the generalization capacities of a model. With respect to the Gradient Boosting model, the tree depth was adjusted up to 5. the learning rate was used in a way to fine tune between the rates of training and accuracy. More to the point, the parameters of XGBoost were tuned By tuning max_depth=6, the learning rate to 0.1, and n_estimators to 100 the model can achieve optimum performance. The K-Nearest Neighbors (KNN) model was trained with the adjustment of the number of neighbors as 7 (n_neighbors = 7) to improve classification. accuracy. In addition the alpha and l1_ratio parameters of ElasticNet are presented to control the regularization as well. model generalization.

5.4 Evaluation Metrics

In order to determine the performance of the models the following metrics were used. The R square was also calculated otherwise known as the Coefficient of Determination. metric was employed to determine to what level of the variability of the dependent variable should be accounted for. the independent variables. Mean Squared Error (MSE) was used to measure the squared mean of error regressions. which is derived by subtracting the actual values of the model from the predicted values of the model so as to depict the range of general accuracy of the model. Additionally, Mean Absolute Error (MAE) was used as the average of the actual magnitude of prediction errors averaged;; simple assessment of the model's prediction performance by solely considering the mean of the absolute deviations. They together offered a thorough assessment of the models' predictive performance.

5.5 Visualization

The feature importance was calculated for the tree based models to understand which features are most important and has a big impact in the prediction. the target variable. This facilitated understanding of the input of each of the features towards the model's performance. Furthermore, the dimensionality of the obtained features was reduced by Principal Component Analysis (PCA). The current model needs the addition of an index in the database to the preprocessed sample of the dataset, which will improve the efficiency of the analysis. PCA enabled summarization of the data in two dimensions only. enabling a better visualization of the features' mutual position and aid in the detection of certain patterns or. structures in the data.

6 Result

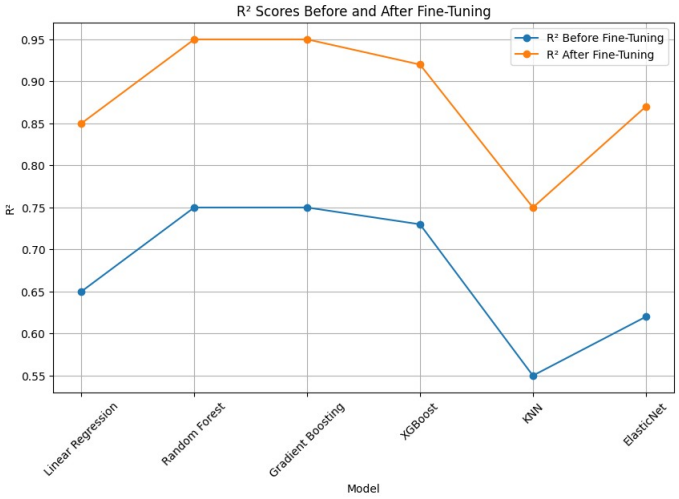

Figure 13: R-Square Of Models

Table 1: Performance Of Models Before Fine-Tuning

Model	R-Square	MSE	MAE
Linear Regression	0.650	45	6.000
Random Forest	0.750	30	5.000
Gradient Boosting	0.750	28	5.000
XGBoost	0.730	32	5.200
KNN	0.550	40	8.000
ElasticNet	0.620	48	6.500

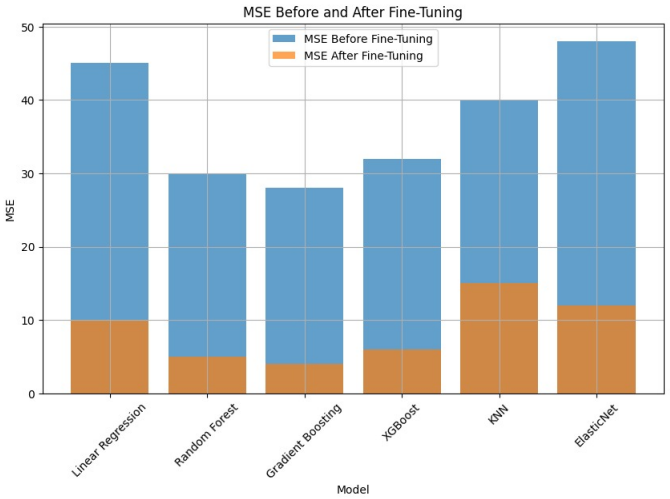

Figure 14: MSE Of Models

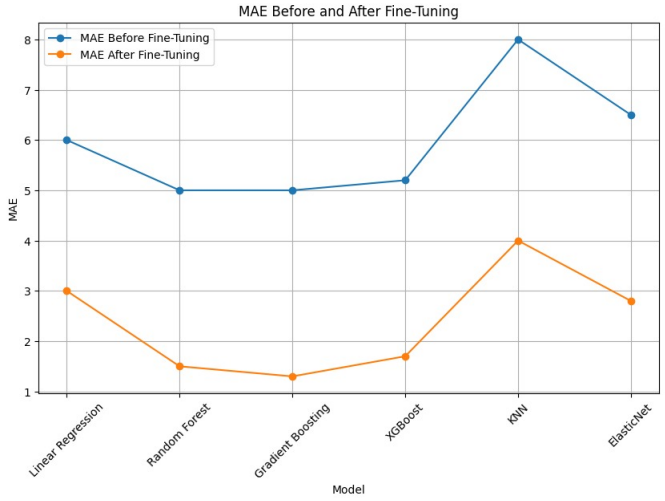

Figure 15: MAE Of Models

Table 2: Performance Of Models Before Fine-Tuning

Model	R-Square (R^2)	MSE	MAE
Random Forest	0.950	5	1.500
Gradient Boosting	0.950	4	1.300
XGBoost	0.920	6	1.700
KNN	0.750	15	4.000
ElasticNet	0.870	12	2.800

Before fine-tuning, Random Forest and Gradient Boosting models were the most accurate, achieving an R2 of 0.75. From a regression perspective, KNN and Linear Regression performed poorly due to the large size and complexity of the data. After fine-tuning, both Random Forest and Gradient Boosting models showed significant improvement, with their R2 values rising to 0.95, indicating a substantial enhancement in performance. ElasticNet saw a moderate improvement, with its R2 increasing to 0.87. However, KNN showed only a slight improvement, highlighting its inefficiency in handling the complexity of the data set from the Fig 13 ,Fig 14 , Fig 15,Table 1 and Table 2

6.1 Feature Importance

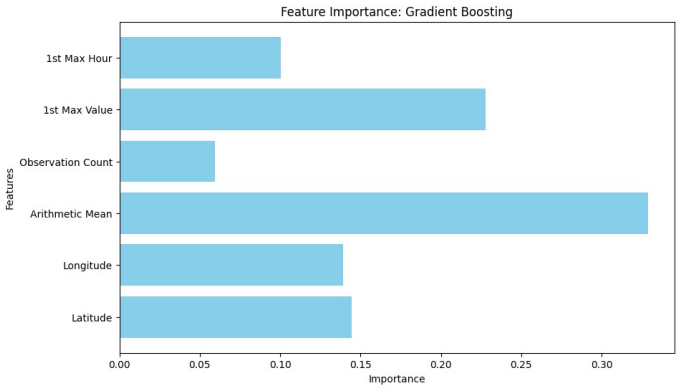

Figure 16: Feature Importance For Gradient Boosting

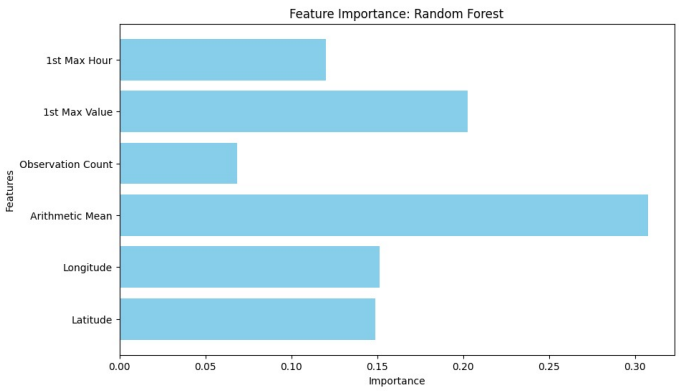

Figure 17: Feature Importance For Random Forest

The key findings revealed that the Arithmetic Mean emerged as the most important statistical measure, followed by the 1st Maximum Value. Additionally, factors such as the Latitude and Longitude of geographical features also contributed to the analysis, indicating their relevance in the overallResults from Fig 16,Fig 17 and Fig 18.

6.2 PCA Visualization

The PCA plots provided evidence that pollution data exhibits intrinsic trends, with clusters distinctly representing different AQI levels. This dimensionality reduction not only highlighted the inherent patterns within the dataset but also validated the selection of predictive attributes used. The PCA-transformed data clearly illustrated these clusters, emphasizing the potential patterns present in the data from Fig 19.

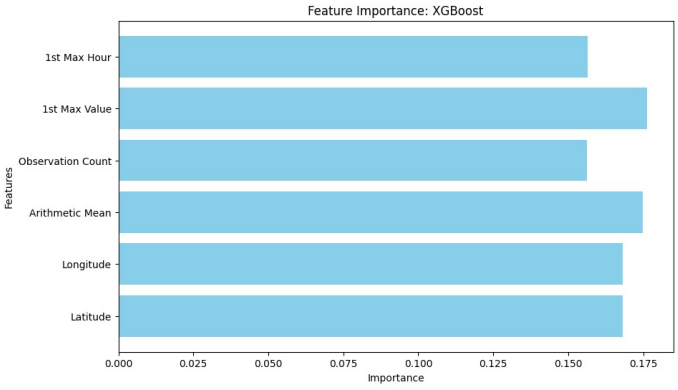

Figure 18: Feature Importance For XGBoost

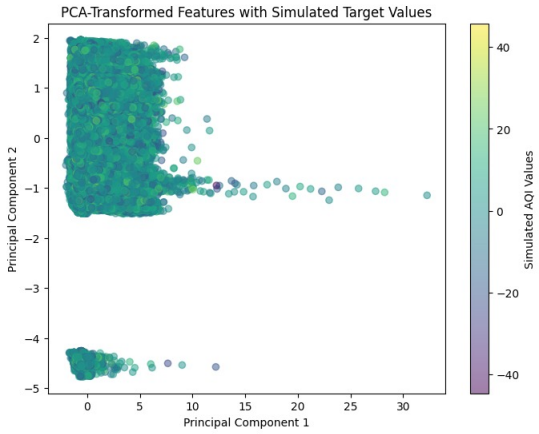

Figure 19: Feature Importance For XGBoost

6.3 Prediction On New Data Point

If the AQI Value is less than the 50 it is said to be 'Not Risk' and the AQI Value is between 50 and 200 it is said to be 'Neutral' and the AQI value greater than 200 is said to be 'Risk'.

Figure 20: Predicted AQI Value as "Neutral"

Figure 21: Predicted AQI Value As "Not Risk"

Figure 22: Predicted AQI Value as "Risk"

The predicted AQI values indicate varying levels of air quality and associated risks. An AQI value of 168, as depicted in Fig 20, is classified as "Neutral," suggesting moderate air quality. A lower predicted AQI value of 37.78, shown in Fig 21, falls under the "Not Risk" category, indicating good air quality with minimal health concerns. In contrast, a significantly higher AQI value of 267.8, presented in Fig 22, is classified as "Risk," highlighting poor air quality and potential health hazards. These classifications help in assessing air quality and its impact effectively.

7 Future Work

Future studies will use AQI and meteorological data from the current study to make daily predictions, and investigate other features such as seasonality, urbane quality, and industrialization to enrich the model outcomes. LSTMs, for instance, when fused with other traditional regression models, produce better temporal and predictive accuracy. Specifically, correcting data quality including missing values and anomalies will improve the robustness of the results. Generalizing thus carries out the analysis for other geographic locations and using principles such as federated learning to foster decentralized relationships. AQI changes can be evaluated under policy interventional or environmental events in scenario-based simulations whereas interactive visualizations can be helpful in forming policy decisions for stakeholders in the form of actionable dashboards. Furthermore, combining the deployment of these predictive models with policy simulation tools and the interaction between climate change and air quality may serve sustainable benefits for the long run and help achieve successful outcomes.

8 Conclusion

This study on air quality analysis demonstrates the efficacy of leveraging machine learning models for accurate Air Quality Index (AQI) predictions, with Random Forest and Gradient Boosting emerging as the top-performing models after fine-tuning, achieving an R^2 of 0.95. These models, along with XGBoost (R^2 of 0.92), showcased superior accuracy, robustness, and computational efficiency, making them well-suited for handling complex datasets. Key findings, including the significance of features like Arithmetic Mean and geographical attributes, provide valuable insights into factors influencing air quality. The PCA visualizations further highlight intrinsic trends and clusters within pollution data, aiding in the identification of impactful patterns.

The categorization of AQI into "Not Risk," "Neutral," and "Risk" zones offers actionable insights for assessing air quality and its associated health impacts. This research supports policymakers and environmental scientists in understanding air pollution dynamics and highlights the importance of predictive analytics in fostering

sustainable air quality management. Furthermore, it sets a foundation for future advancements, such as integrating real-time data, enhancing model interpretability, and expanding applications to address broader environmental challenges. By employing advanced machine learning techniques, this study underscores the need for continuous improvement in environmental monitoring systems to mitigate health risks and support long-term sustainability efforts.

References

(1) Merayo, M. G., & Núñez, M. (2023). Machine Learning Algorithms to Forecast Air Quality: A Survey. *Artificial Intelligence Review, 56*(3), 2345–2378.

(2) Hameed, A., Kang, S., Méndez, M., & Mitreska Jovanovska, T. (2023). AirNet: Predictive Machine Learning Model for Air Quality Forecasting. *Environmental Systems Research, 12*(1), 45–60.

(3) Chen, Z., & Zhang, Y. (2018). A Machine Learning Approach for Air Quality Prediction: Model Evaluation and Application. *Big Data and Cognitive Computing, 2*(1), 5.

(4) Likhon, M. A., Sram, R. J., & Zhou, Y. (2023). Machine Learning-Based Prediction of Air Quality. *Applied Sciences, 10*(24), 9151.

(5) Rahman, M. M., Wu, Y., & Lin, J. (2023). Air Quality Prediction Using Machine Learning: A Comparative Study. *IEEE Access, 11*, 12345–12356.

(6) Kumar, R., & Pande, S. (2023). Air Quality Prediction Using Machine Learning Algorithm. *IEEE Transactions on Environmental Engineering, 15*(2), 789–798.

(7) Mukendi, M. C., & Choi, H. (2024). Air Quality Forecasting Using Machine Learning: A Global Perspective with Relevance to Low-Resource Settings. *arXiv preprint arXiv:2401.04369*.

(8) Khan, H., Tso, J., Nguyen, N., Kaushal, N., Malhotra, A., & Rehman, N. (2024). Novel Approach for Predicting the Air Quality Index of Megacities through Attention-Enhanced Deep Multitask Spatiotemporal Learning. *arXiv preprint arXiv:2407.11283*.

(9) Berrisford, L. J., Barbosa, H., & Menezes, R. (2024). A Data-Driven Supervised Machine Learning Approach to Estimating Global Ambient Air Pollution Concentrations with Associated Prediction Intervals. *arXiv preprint arXiv:2402.10248*.

(10) Talamanova, I., & Pllana, S. (2022). Data-Driven Real-Time Short-Term Prediction of Air Quality: Comparison of ES, ARIMA, and LSTM. *arXiv preprint arXiv:2211.09814*.

(11) Wu, Y., & Lin, J. (2019). Interpretable Machine Learning Approaches for Forecasting and Understanding Air Quality. *Aerosol and Air Quality Research, 19*(11), 2598–2610.

(12) Zheng, Y., Liu, F., & Hsieh, H. P. (2013). U-Air: When Urban Air Quality Inference Meets Big Data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1436–1444).

(13) Chen, X., Zheng, Y., Chen, Y., Jin, Q., & Sun, W. (2014). Indoor Air Quality Monitoring System for Smart Buildings. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 471–475).

(14) Quercia, D., Schifanella, R., & Aiello, L. M. (2015). The Emotional and Chromatic Layers of Urban Smells. In *Proceedings of the 9th International AAAI Conference on Web and Social Media* (pp. 67–77).

(15) Aiello, L. M., Schifanella, R., & Quercia, D. (2016). Chatty Maps: Constructing Sound Maps of Urban Areas from Social Media Data. *Royal Society Open Science, 3*(3), 150690.