Machine Learning Approaches to Ambient Air Quality Prediction

Sweta

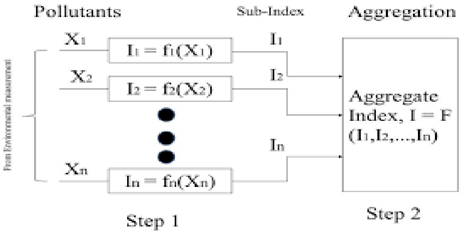
Department of CSE, Galgotias University

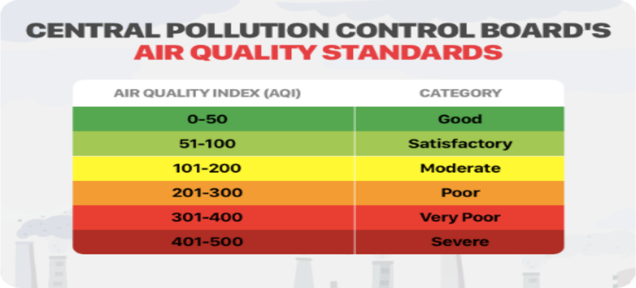
Abstract: ***Air pollution is a significant concern in today's smart city environment. Pollution is monitored in real time by confined authorities who have the authority to assess current traffic conditions and make recommendations based on what they discover.Sensors based on the Internet of Things have been deployed, drastically altering the features of air quality prediction. To gain a better understanding of their processing time for multiple datasets, a comparative analysis of these methodologies is required. To gain a better understanding of their processing time for multiple datasets, a comparative analysis of these methodologies is required. The pollution prediction in this work was done utilizing advanced supervised approaches as well as a comparative evaluation of numerous models to select the optimum model for effectively predicting air quality. Kaggle was used to conduct experiments, and multiple datasets were used to estimate pollution levels.) Aside from these methods, the processing time of each methodology is estimated using standalone learning, and hyperparameter tweaking is fitted using Kaggle. Performance measures were analyzed to obtain the best-fit model in terms of processing time and the lowest error rate.***

Keywords: AirQualityIndex, Nitrogen dioxide(NO_2),Particulate Matters(PM_{10},$PM_{2.5}$),Carbon Monoxide(CO), Sulphur dioxide(SO_2), MAE, RMSE, MSE

1. Introduction

The primary resources for a sustainable lifestyle are air, land, and water. As a result of recent technology breakthroughs, a large amount of data on ambient air quality has been collected and utilised to illustrate the air quality in various places. The vast amount of data collected results in a massive amount of information that neither gives a decision-maker nor a common person a clear picture of air quality. A technique to characterise air quality is to provide the amounts of all contaminants that meet acceptable requirements. Various sample sites and pollutant parameter sampling provide a description of air quality that is both scientific and technical in nature.The general public frequently doesn't often acknowledge actual data, multivariate charts, statistical investigations, and other sophisticated discoveries regarding the air quality. As a result, they get disinterested. In order to address theseissues, the notion of an Air Quality Index (AQI) has been created and successfully implemented in many developed countries over the previous three decades. A method for converting weighted values of individual air pollution-related factors into a single number or group of numbers is known as an Air Quality Index. Prior to the implementation of a modest air quality monitoring scheme in our country in the mid-1980's, there were growing concerns about air pollution. The conversion of complex pharmacogenomics information into straightforward and comprehensive knowledge, as well as engaging with citizens in historical, current, and speculative terms, are the key obstacles to encouraging participation in a coherent form.





CENTRAL POLLUTION CONTROL BOARD'S
AIR QUALITY STANDARDS

AIR QUALITY INDEX (AQI)	CATEGORY
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor
401-500	Severe

The Implications of the Air Quality Index
The AQI's goals are as follows:
1) Utilization of Resources
2) Location based ranking
3) Standards Regulation
4) Analysis of Tendencies
5) Information to the audience
6) Scientific Prosecutions

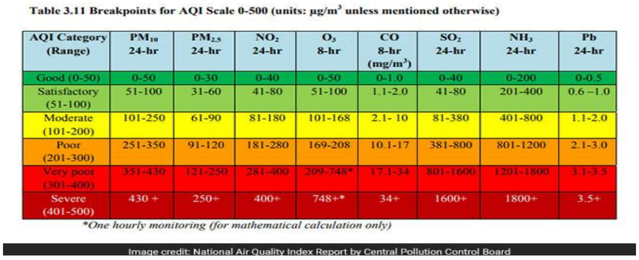


Table 3.11 Breakpoints for AQI Scale 0-500 (units: µg/m³ unless mentioned otherwise)

AQI Category (Range)	PM_{10} 24-hr	$PM_{2.5}$ 24-hr	NO_2 24-hr	O_3 8-hr	CO 8-hr (mg/m³)	SO_2 24-hr	NH_3 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.6 –1.0
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1- 10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748*	17.1-34	801-1600	1201-1800	3.1-3.5
Severe (401-500)	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+

*One hourly monitoring (for mathematical calculation only)

Image credit: National Air Quality Index Report by Central Pollution Control Board

2. Related Work

In this work, machine learning algorithms have been applied. These algorithms are computationally efficient, cost-effective, and accurate. Many machine learning methods have been applied to predict air quality. The ensemble methods Ada Boost and Gradient Boosting were utilised. Multiple Layer Perceptron (MLP) and Random Forest (RF) approaches, as well as Decision Tree Classifier (DCT) and Artificial Neural Network (ANN) techniques, were used. Neural network models, particularly deep neural networks, are becoming increasingly prevalent in today's world.

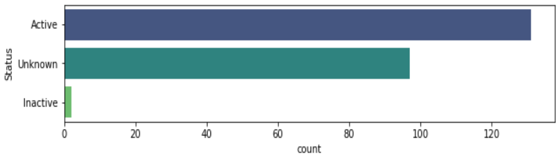
3. Methodology

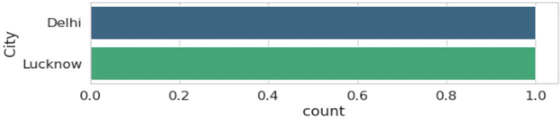
1) **Decision Tree:**A decision tree is a flowchart-like structure in the shape of a tree that is used for categorization and prediction. Internal nodes, branch nodes, and leaf nodes make up a decision tree, with each internal node representing a test on the attribute, each branch implying a test outcome, and each terminal node retaining a class label. The model divides the dataset into smaller subsets and uses simple decision rules to make decisions based on the results. To determine the system's correctness, a cross validation should be undertaken.
2) **Random Forest:**Random Forest is an ensemble methodology that utilizes a vast number of Decision Trees, reminiscent of a forest with several trees. It is a supervised learning model that anticipates improvedaccuracy by taking the weighted average of alldecision trees on wide and varied subcategories of a dataset. In contrast to other approaches, it requires less time to equip. The tree is being devised, and the resulting split is the best among an arbitrary feature selection, but rather the best among all the options available. The accuracy score of the Decision Tree tends to improve as the number of trees in the forest expands, as the problem of overfitting is mitigated to a certain extent. It performs well with larger datasets.
3) **Adaptive Boosting:**Adaptive boosting, often known as Ada boost, is a Machine Learning Algorithm that is used to enhance performance. It combines a number of weak classifiers to create a single powerful classifier. The weights are reassigned to each instance, with the erroneously classified instances receiving higher weights. Boosting is a technique for reducing the model's bias as well as its variance.
4) **Gradient Boosting:** It's a well-known boosting algorithm. The weights of the training instances are not changed in this technique because each consecutive predictor corrects the inaccuracy of its predecessor. The residual errors of its predecessors are used as labels for each successor's training. Gradient Boosting is an important regularisation procedure that aids in the changing of rules. It diminishes the amount of incremental and penalises the value of each subsequent epoch as a result.
5) **Multi-Layer Perceptron:**A feed forward The Artificial Neural Network is a multi-layer perceptron. There are three layers in this network: an input layer, a hidden layer, and an output layer. The input layer is the bottom layer of the network that collects input from datasets and is the portion that is visible. Hidden layers execute all mathematics on the features entered via the input layer, then pass the outcomes to the output layer. The final layer, the output layer, produces a sequence of outputs from a set of inputs. In a neural network, an activation function stipulates how the weighted sum of the input is processed into an output from a node or nodes in a layer of the network, including whether the neuron should be accepted or removed. It solves exceedingly difficult tasks, such as fitness approximation.
6) **Deep Neural Networks:**The deep neural network is a type of machine learning in which the system emerges from input data into high-level functions by using layers upon layers of vertices.It employs neural network models to yield portable solutions. Deep Neural Networks have been used in virtual assistants, face recognition, as well as vision for self-driving cars, along with many other applications.
7) **Ensemble Methods:**These are mechanisms for constructing many models and then consolidating them together to select the best alternative.

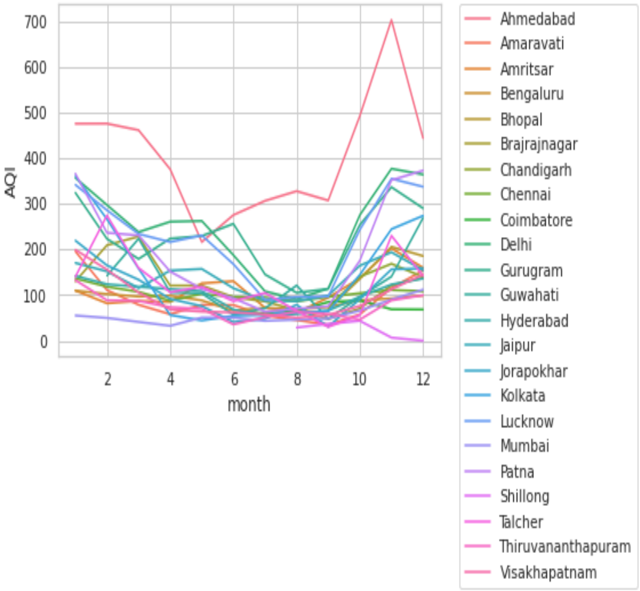
4. Dataset

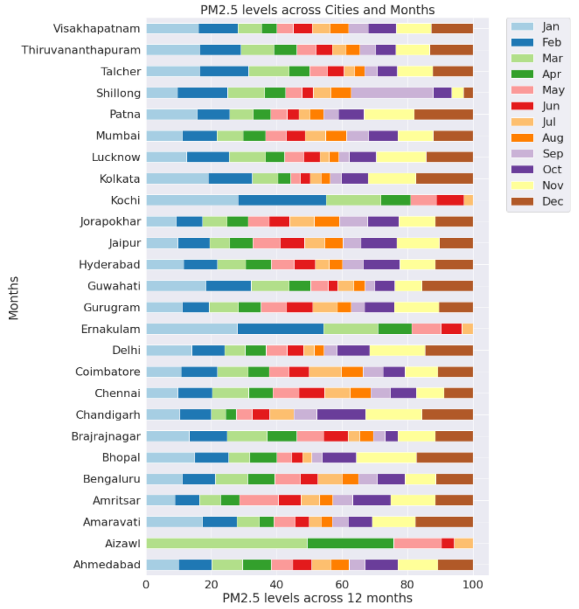
Source: Kaggle (AQI DATA IN INDIA [2015-2020]), AQI and hourly data across stations and city in India.

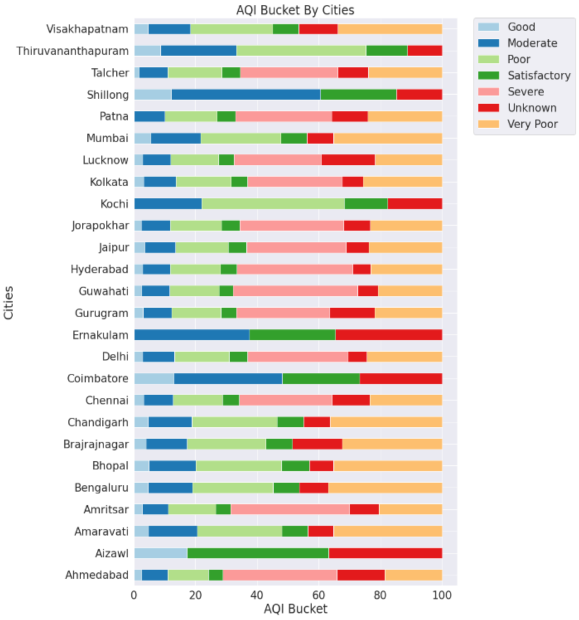
Dataset Description:
PM2.5, PM10, NO2, NH3, SO2, OZONE are independent dimensions in the aforementioned dataset, which contains eight attributes, seven of which are pollution accumulators and one of which is the Air Quality Index. An AQI is an implicit component.Although this dataset contains outliers, pre-processing is required. While our goal is to anticipate the AQI, we can use either classification or regression to achieve its goals and objectives. Since our class label is supervised, we need to use the Regression analysis. Regression is a statistical method of supervised learning that matches data into some kind of specific range.

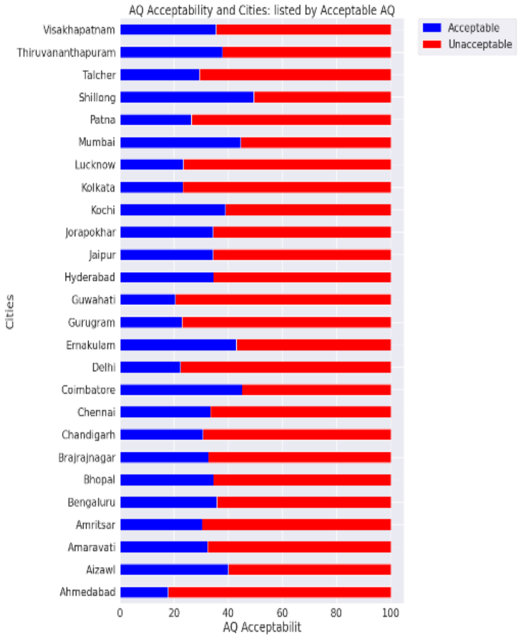












5. Conclusion

The Decision Tree, Random Forest Classifier, Multi-Layer Perceptron, and Ada Boosting and Gradient Boosting algorithms were deployed as ensemble approaches in this work. The error rate and time consumption of these algorithms were compared. As per the outcome, Random Forest excelled in all of the techniques and performed well on datasets of varied sizes and features based on the processing, as well as possessing the minimum error rate of all of these techniques. The Decision Tree had the fastest processing time, but the greatest error rate. In a nutshell, we can conclude that, among all the algorithms investigated in our research project, Random Forest was the right choice.

6. Future Work

We seek to investigate more factors that may influence environmental pollution in the future.

References

[1] 7 Million Premature Deaths Annually Linked to Air Pollution. Accessed: Apr. 27, 2019. [Online]. Available: https://www.who.int/ phe/eNews_63.pdf

[2] F. C. Moore, ‘‘Climate change and air pollution: Exploring the synergies and potential for mitigation in industrializing countries,’’ Sustainability, vol. 1, no. 1, pp. 43–54, 2009.

[3] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, ‘‘Inferring air quality for station location recommendation based on urban big data,’’ in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 437–446.

[4] M. Johnson, V. Isakov, J. S. Touma, S. Mukerjee, and H. Özkaynak, ‘‘Evaluation of land-use regression models used to predict air quality concentrations in an urban area,’’ Atmos. Environ., vol. 44, no. 30, pp. 3660–3668, 2010.

[5] C. Malalgoda, D. Amaratunga, and R. Haigh, ‘‘Local governments and disaster risk reduction: A conceptual framework,’’ in Proc. 6th Int. Building Resilience Conf., Building Resilience Address Unexpected. Palmerston North, New Zealand: Massey Univ., 2016, pp. 699–709.

[6] M.-A. Kioumourtzoglou, J. D. Schwartz, M. G. Weisskopf, S. J. Melly, Y. Wang, F. Dominici, and A. Zanobetti, ‘‘Long-term PM2.5 exposure and neurological hospital admissions in the northeastern United States,’’ Environ. Health Perspect., vol. 124, no. 1, pp. 23–29, 2015.

[7] World Health Organization, and UNAIDS Air Quality Guidelines: Global Update 2005, World Health Org., Geneva, Switzerland, 2006.

[8] K.-H. Kim, E. Kabir, and S. Kabir, ‘‘A review on the human health impact of airborne particulate matter,’’ Environ. Int., vol. 74, pp. 136–143, Jan. 2015.

[9] En.wikipedia.org. (2018). Air quality index. Accessed: Dec. 2, 2018. [Online]. Available: https:==en:Wikipedia:org=wiki=Airqualityindex

[10] W. Yi, K. Lo, T. Mak, K. Leung, Y. Leung, and M. Meng, ‘‘A survey of wireless sensor network based air

pollution monitoring systems,'' Sensors, vol. 15, no. 12, pp. 31392–31427, 2015.

[11] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian, ''The impact of PM2.5 on the human respiratory system,'' J. Thoracic Disease, vol. 8, pp. 69–74, Jan. 2016.

[12] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, ''Urban planning and building smart cities based on the Internet of Things using big data analytics,'' Comput. Netw, vol. 101, pp. 63–80, 2016.

[13] M. Asgari, M. Farnaghi, and Z. Ghaemi, ''Predictive mapping of urban air pollution using apache spark on a Hadoop cluster,'' in Proc. Int. Conf. Cloud Big Data Comput., 2017, pp. 89–93.

[14] D. Zhu, C. Cai, T. Yang, and X. Zhou, ''A machine learning approach for air quality prediction: Model regularization and optimization,'' Big Data Cogn. Comput., vol. 2, no. 1, p. 5, 2018.

[15] R. W. Gore and D. S. Deshpande, ''An approach for classification of health risks based on air quality levels,'' in Proc. Int. Conf. Intell. Syst. Inf. Manage. (ICISIM), Oct. 2017, pp. 58–61.

[16] G. R. Kingsy, R. Manimegalai, D. M. S. Geetha, S. Rajathi, K. Usha, and B. N. Raabiathul, "Air pollution analysis using enhanced K-means clustering algorithm for real-time sensor data," in Proc. IEEE Region Conf. (TENCON), Nov. 2016, pp. 1945–1949. http://www.cpcb.nic.in/national-air-quality-index/

[17] I. Bougoudis, K. Demertzis, and L. Iliadis, ''HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens,'' Neural Comput. Appl., vol. 27, no. 5, pp. 1191–1206, 2016.

[18] C. Yan, S. Xu, Y. Huang, Y. Huang, and Z. Zhang, ''Two-phase neural network model for pollution concentrations forecasting,'' in Proc. 5th Int. Conf. Adv. Cloud Big Data (CBD), 2017, pp. 385–390.

[19] C. A. Keller, M. J. Evans, J. N. Kutz, and S. Pawson, ''Machine learning and air quality modeling,'' in Proc. IEEE Int. Conf. Big Data (Big Data), Dec. 2017, pp. 4570–4576.

[20] A. B. Ishak, M. B. Daoud, and A. Trabelsi, ''Ozone concentration forecasting using statistical learning approaches,'' J. Mater. Environ. Sci., vol. 8, no. 12, pp. 4532–4543, 2017.

[21] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, ''Promises and challenges of big data computing in health sciences,'' Big Data Res., vol. 2, no. 1, pp. 2–11, 2015.

[22] https://www.geeksforgeeks.org/predicting-air-quality-index