INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 9, Issue 2 - V9I2-1158)
Available online at: https://www.ijariit.com

Detection and prediction of air pollution using Machine Learning

Patan Masthan Vali
masthankhan547@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

D. P. Neeha Kousar
dpneehakousar@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

T. Sai Pranathi
saipranathitalari@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

K. Nandini
kogaranandini2002@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

M. Saikanth
msaikanth99@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

A. Ramesh Babu
ambatiramesh84@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

ABSTRACT

The regulation of air pollutant levels is rapidly increasing and its one of the most important tasks for the governments of developing countries, especially India. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The meteorological and traffic factors. burning of fossil fuels, industrial parameters such as powerplant emissions play significant roles in air pollution. Among all the particulate matter (PM) that determine the quality of the air. When its level is high in the air, it causes serious issues on people's health. Hence, controlling it by constantly keeping a check on its level in the air is important.

Keywords: *Decision Tree, PM10, PM2.5, SO2, CO, NO2, O3, Air Pollution, Air Quality Index, Assistive Vector Technology, Regression using Logit, Forest of the chance tree, K-Nearest Neighbor, and Naive Bayes.*

1. INTRODUCTION

From fledgling companies to significant platform suppliers, machine learning has the highest concentration of any industry. Daily, the quantifiable data. Assuming proper public treatment, air quality qualities go much beyond the highest feasible values.

The theory of AI, which believes that the machine gradually makes its own decisions, has an effect on all facets of our society rather than just issuing commands as usual works [4]. Pollution is a major issue in India. Some of the most noteworthy technological advances in recent years include artificial computing and machine learning. A synthetic intelligence system takes input from sensors in the neighborhood of machine learning, where it learns how to act in a given situation. One of the reasons we selected was the capacity to adapt to machine learning (ml) methods. This is a member of the generalized linear classifications family. The geometrical minimization of the empirical classification error is one distinctive feature of SVM. Margin was maximized concurrently. In bigger places like Delhi, where the AQI hit a record high of 999, the situation is significantly worse.

Several of the algorithms that we employ were developed by different researchers. None of them, however, examines the outcomes of all six trials as if they were one research with the same conditions and comparable data. To make things better, the first stage necessitates the forecast of the air quality index [3]. When the dataset is analyzed using the supervised machine learning technique (SMLT), a wealth of information is captured, including variable identification, univariate analysis, bivariate, and analysis of the knowledge, statistical process, and missing pricing treatments. Six entirely unique classifiers that supported various algorithms were developed as part of this project. Six distinct classifiers based on various methods were developed throughout this research.

Table1: Pollutants, AQI Category, and Wellbeing Breakpoints

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

The record revolves around six significant particle contaminations in India, as it does in many other countries. carbon monoxide (CO), ammonia (NH3), gas (NO2), and particulate matter under 10 micrometers in measurement (PM10), particulate matter under 2.5 micrometers wide (PM2.5), and ozone (O3) (O3). By that time, a checking station should be able to provide you with the grouping of a chosen poison. The standard over some period of time for CO and O3 is assumed control for about eight hours, while the standard for the other three is for about 24 hours. Micrograms are the measurement per cubic meter (or milligram, on account of CO).

2. RELATIVE ASSOCIATED WORKS

The proposed technique, according to the authors of the research, uses two different types of deep network systems to absorb a lot of data. Urban air pollution expectation with a focus is handling an abundance of thorough ecological checking information and sophisticated modifications in air toxicity.

In this research, a new contamination fixation forecast method is put forward. It is based on extensive deep learning techniques and vast amounts of natural data. The information contained in data gathered from various sources could be shady. To ensure high-quality data, cleaning the data should be done before putting it in the warehouse. Despite being surrounded by data, we are information-hungry, and knowledge is crucial for decision-makers who want to increase their income in the sector. For forecast accuracy to increase and for natural contamination occurrences to be avoided, credible expectation tactics are required [2]. The Semantic ETL (Extract Transform Load) approach for AQ cloud prediction is discussed in this article. We employ ontology to refocus PM2.5 from different data sources and to mix this data with the same term with separate names are entered into the integral database [15].

We make advantage of the nodes and storage nodes that make up the ETL cloud network. The computational nodes are utilized for data mining and storage for the forecasting, retrieval, and interpretation of the acquired data [1]. We employ residual web providers as the front-end API to receive analyzed data, and we use the browser to display visualized data to present projections and predictions. The analysis of air quality has demonstrated the value of the Big Data Cloud Technology Architecture.

3. EFFICIANCY OF DATA WORK

• Data Analysis:

The data includes meteorological data for various contaminants for different Indian cities. The strategy is to employ many machine learning techniques to create the optimal machine learning model. To build our model, we used six distinct supervised machine learning methods, including decision trees, k-nearest neighbors, Nave Bayes classifiers, support vector machines, and logistic regression.

The Central Pollution Control Board provided us with information (CPCB). Country, State, City, Location, Last Updated, Minimum, Maximum, Average, and Pollutants are the nine attributes that make up this dataset. Pollutants including SO2, NO2, PM10, PM2.5, CO, and O3 are tracked. To forecast the behaviour of pollutants over the coming days, these data are gathered every day.

Table 2: List of Attributes in dataset

Variable	Description
Country	Home country (India)
State	Indian states name lists
City	City names for each state
Place	Place names for each city
Last update	Date and time (DD/MM/YYYY HH:MM)
Avg.	Average range of pollutants
Max	Maximum range of pollutants
Min	Minimum range of pollutants
Pollutants	Pollutants name

Visualization of Data

The statistics of machine learning rely on data visualization, which is fundamental knowledge in applied statistics. Software analyses and practically on quantitative explanations are useful tools for gaining a qualitative knowledge. Maps and graphs, which are more emotive and consumer-focused than associative or substantive metrics, help to explain what the data means by giving it a visual meaning.

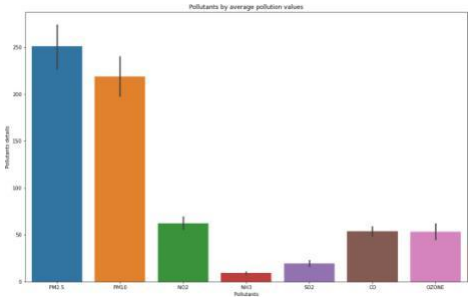

Figure1: Pollutants by average pollution values

As a result, the natural information from the human eye is improved, making it easier to study and comprehend datasets and identify fraudulent data transfer patterns. The spread of pollutants relative to the pollution values is depicted in the image below.

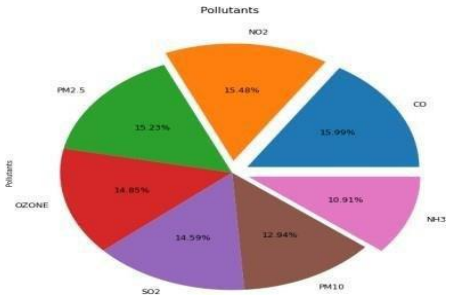

Figure 2: Propagation of Pollutants

Dataset modelling

In addition, we added a new attribute called AQI to our data set when we came to it. This place's air quality is measured as acceptable, satisfactory, or inadequate, depending on how polluted it is. Low-pollution areas are classified as satisfactory, while medium- and high-pollution areas are classified as inadequate. We divided our data collection into features and targets first. Then, we divided the data set for training into two portions: a 70 percent training set for our data model, and a 30 percent testing set for us to use to verify the model's accuracy [13].

In order to distinguish the training, set from the test set, the resulting data set was divided into two series: the training set and the test set. Typically, these two series are distinguished by 7:3 ratios. For the training set, the data model created by combining Random Forest, Logistic Regression, Determined Tree Algorithms, K-Nearest Neighbor (KNN), SVC, and Naive Bayes is utilized, and the prediction for the test set is based on how well the test set prediction performed.

The accuracy of the machine learning model, which was created using decision trees, logistic regression, random forests, and k-nearest neighbors, is based on the training data. The visual representation of the comparison between the air quality indexes (AQI) over cities and locations with pollution vs. No-pollution cities are

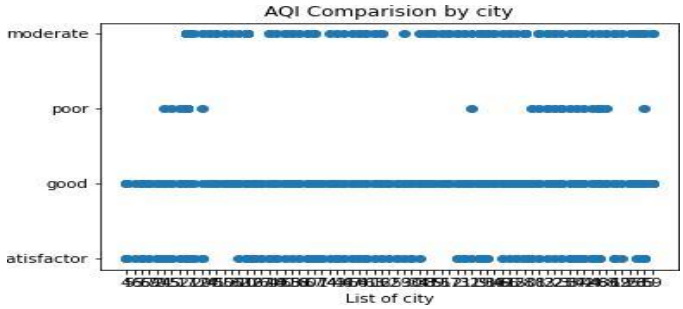

Figure 3: AQI Comparison by city

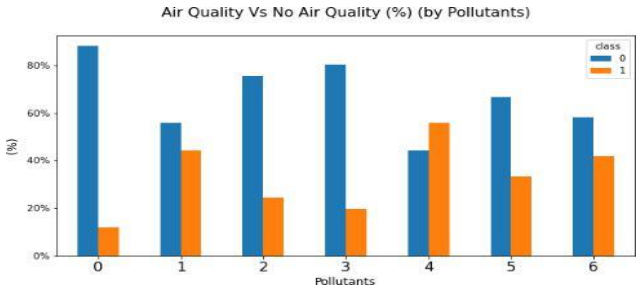

Figure 4: Air Quality Vs. No Air Quality

4. MODELS OF MACHINE LEARNING FOR POLLUTION PREDICTION

Logistic Regression Model Workflow

The most prevalent regression technique appears to be the analysis of logistic regression, which can be applied in modelling of parameters with binary dependence. It frequently depicts the relationships between the independent variables through the use of the computational process known as logistic regression.

The two different types of discrete dependent variables, variables x1 and x2, are each coded in either 0 or 1. Discrete, binary, continuous, or mixed variables can all be independent. Meaning that the logistic regression model predicts P(Y=1) as a function of Logistic regression Assumptions: Binary logistic regression requires that the dependent variable be binary [3]. It should be the level 1 component that is desired as a result of the binary regression of the dependent variable. Consideration should only be given to the most crucial variables. Independent of one another, the variables should be. Alternatively, the model should have very little. Linear correlation exists between the log odds.

Broad sample sizes are required for logistic regression. Our machine learning model's accuracy is 97.46876470588235, and the process by which Logistic Regression works—that is, how our model is trained and tested using Logistic Regression—is graphically depicted.

```
Classification report of Logistic Regression Results:

              precision    recall  f1-score   support

           0       0.97      0.97      0.97        73
           1       0.99      0.99      0.99       175

    accuracy                           0.98       248
   macro avg       0.98      0.98      0.98       248
weighted avg       0.98      0.98      0.98       248

Cross validation test results of accuracy:
[0.94117647 1.         1.         1.         1.         1.
 1.         0.94117647 1.         1.         1.         1.
 0.94117647 1.         1.         1.         1.         0.94117647
 0.94117647 1.         1.         1.         1.         0.94117647
 0.94117647 0.88235294 0.94117647 1.         1.         0.82352941
 1.         0.94117647 1.         0.9375     0.9375     0.9375
 0.8125     1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         0.93333333
 1.         1.        ]

Accuracy result of Logistic Regression is: 97.46960784313725

Confusion Matrix result of Logistic Regression is:
 [[ 71   2]
 [  2 173]]

Sensitivity :  0.9726027397260274

Specificity :  0.9885714285714285
```

Figure 5: Classification report for Logistic regression

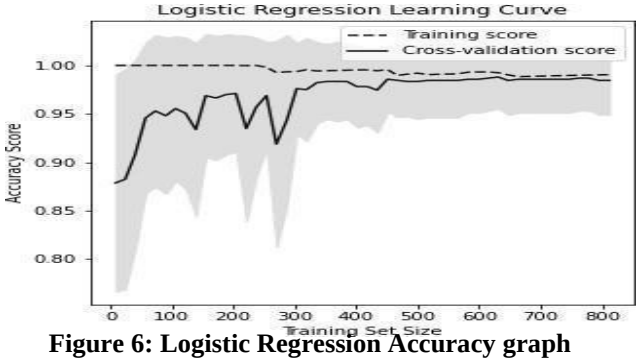

Figure 6: Logistic Regression Accuracy graph

Naive Bayes Model Supervision

The Bayes Theorem is a foundational concept for the Naive Bayes statistical classification approach. This is among the most fundamental methods for supervised learning that are accessible. A trustworthy, quick, and precise algorithm is the Naive Bayes classifier. Classification Theorem Bayes is the foundation for the statistical method known as Naive Bayes. One of the most straightforward learning algorithms is this one. The Naive Bayes classifier has an efficient and quick method. Classifications The most extensive results sets and excellent precision are found in naive bayes. Naive Bayes assumes that an influence of clothing functioning in a class is independent of other attributes. Depending on factors including income, the history of prior applications and the transaction, age, and location, a loan applicant's suitability or unsuitability might be determined.
For instance, a bond applicant's eligibility depends on his income, upbringing, age, job, and loan history [5]. These traits are autonomous, despite the fact that they are linked [4]. It is referred to be naive since this assumption makes equations simpler. Class equality is assumed under this presupposition. Our machine learning model's accuracy is 97.38352941176471, and the workings of Naive Bayes, or the training and testing of our model using Naive Bayes, are graphically depicted below using a learning curve mechanism.

```
Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

           0       0.94      0.99      0.96        73
           1       0.99      0.97      0.98       175

    accuracy                           0.98       248
   macro avg       0.96      0.98      0.97       248
weighted avg       0.98      0.98      0.98       248

Cross validation test results of accuracy:
[1.         1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         0.94117647 1.
 0.94117647 1.         1.         0.94117647 1.         0.94117647
 0.88235294 0.94117647 1.         1.         0.94117647 0.94117647
 1.         0.9375     0.875      0.9375     0.9375     0.9375
 1.         0.875      1.         1.         0.9375     1.
 0.9375     1.         0.9375     1.         0.9375     1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]

Accuracy result of Naive Bayes is: 97.32352941176471
```

Figure 7: Classification Report of Naïve Bayes

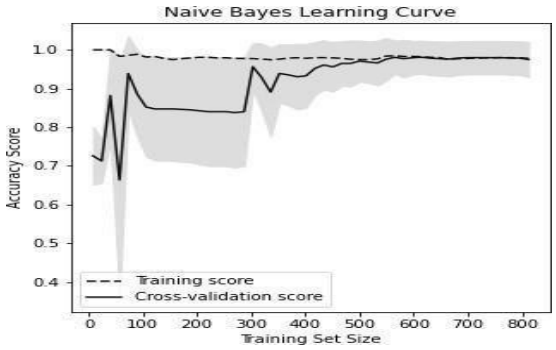

Figure 8: Naïve Bayes Accuracy graph

The algorithm Develop the Support Vector Machines Model versus the Random Forests Model:
Support vector machine model control

Support For the purposes of regression and categorization, Vector Machine is a group of related supervised learning techniques. The generalized linear classifications family includes this. The empirical classification error and geometrical margin are minimized and maximized concurrently in SVM, which is a distinctive feature [10].

Random forests were a hybrid of trees that forecasted errors in general forest statements since each tree functions in its own unique way. With respect to each tree's random variable theory. Because the proportion of forest trees is so important, the mistake of generalizing forests depends on the strength of each individual tree and their relationship to one another. Error levels are comparable to those of the Ad boost algorithm but are noisier due to the random selection of the attributes for dividing each node. For the random forest algorithm to be used, the following procedures must be followed: From of the given data set, choose N random data. Build a decision tree using these N records as the basis. Repeat step 1 of the algorithm using the number of trees you've selected.

when the model is saved on the disk, the representation that SVM will utilize. The accuracy of our machine learning model is 70.6505882352942, and the Support Vector Machine classifier's operation, which entails training and evaluating it using the Support Vector Classifier, is graphically depicted.


```
Classification report of Support Vector Machines Results:

              precision    recall  f1-score   support

           0       0.97      0.97      0.97        73
           1       0.99      0.99      0.99       175

    accuracy                           0.98       248
   macro avg       0.98      0.98      0.98       248
weighted avg       0.98      0.98      0.98       248

Cross validation test results of accuracy:
[1.         1.         1.         1.         0.94117647 0.94117647
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         0.94117647 1.
 0.9375     0.9375     0.9375     0.9375     0.9375     1.
 1.         1.         1.         0.9375     0.9375     1.
 0.9375     1.         0.8125     1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]

Accuracy result of Support Vector Machines is: 98.27205882352942
```

Figure 9: Classification Report of SVM

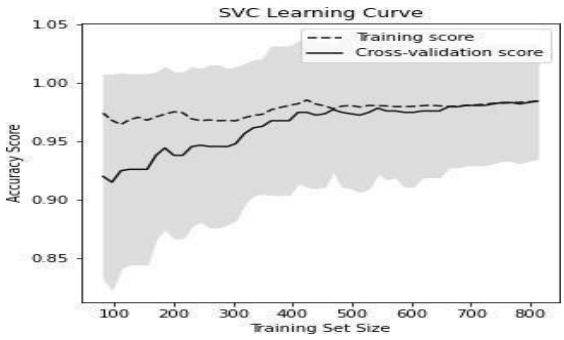

Figure 10: Support Vector Machines Accuracy graph

K Nearest Neighbor Model vs. Decision Tree Model Algorithm Work:
Overseeing the decision tree model

Decision Trees are among the most effective algorithm. As a controlled learning algorithm, the decision array algorithm fits this description. Both categorical and continuous output variables are compatible.

Constraints of a decision tree: Beginning with the root, consider the complete training. Given that they are continuous, qualities for gaining knowledge are regarded as categorical. Based on the attributes' values, recursive distributions are created. Statistical techniques using internal or root node properties. Decision The categorization and analysis of data using trees is a method [6]. The core recursive building block of classification sequencing is decision trees.

Random Forests Model Oversight: Since each tree runs under its own internal random variable theory, random forests are a hybrid of trees that foresee errors in general forest statements. Because the proportion of forest trees is so important, the mistake of generalizing forests depends on the power of each individual tree and their relationship to one another. The error levels are similar to those of the Ad boost technique but are noisier because each node's splitting characteristics are chosen at random.

For the random forest algorithm to be used, the following procedures must be followed: Out of the given data set, choose N random data. Build a decision tree using these N records as the basis. Repeat the algorithm steps with the number of trees you've selected.
1

```
Classification report of Random Forest Results:

              precision    recall  f1-score   support

           0       0.99      0.99      0.99        73
           1       0.99      0.99      0.99       175

    accuracy                           0.99       248
   macro avg       0.99      0.99      0.99       248
weighted avg       0.99      0.99      0.99       248

Cross validation test results of accuracy:
[0.94117647 1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         1.         1.
 0.94117647 1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         0.9375     1.         0.9375
 1.         1.         1.         1.         0.9375     1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]

Accuracy result of Random Forest is: 99.15441176470588
```

Figure 11: Classification Report for Random Forests

K-Nearest Neighbor Model Under Supervision

The supervised learning technique K-Nearest Neighbor retains all n-dimensional occurrences in which they train the data points. The most previewed class is finally returned, and the average of K-similar Neighbors is returned for real value results. Only the closest k number of stored instances (near neighbours) are assessed throughout this process.

Each of the k neighbours in the nearest distance algorithm evaluates the contribution of distance using the following question and gives more weight to the nearby neighbours. When the closest neighbours are pooled, KNN is often robust to noisy data. The K-dearest algorithm is a supervised classification technique that uses a set of defined points to teach other points how to be labeled[15].

There should be categories for all of the traits. In the alternative, they would be pre-discretized. The knowledge gain principle is applied because top-of-the-tree traits have a bigger impact on classification. It is simple to overfit a tree with too many branches, which can reflect noise or bolt-related disturbances. It is illustrated graphically below using a learning curve method how the Decision Tree classifier functions, which involves training and evaluating our machine learning model using Decision Tree. 99.88235294117646 is the accuracy of our model.

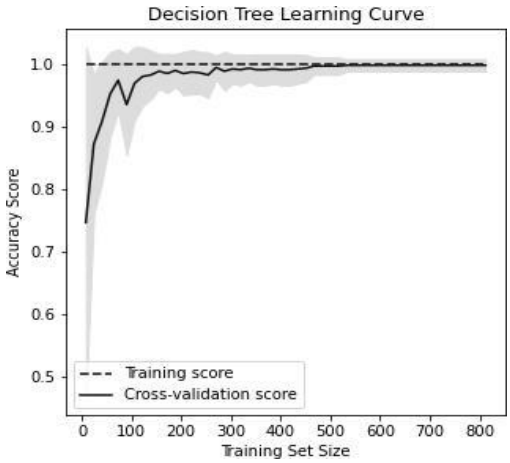

Figure 12: Decision Tree Accuracy graph

K-Nearest Neighbor Model Oversight:

The supervised learning technique K-Nearest Neighbor stores every n-dimensional instance in which a data point is taught. To get to the most previewed class and the average of K-similar Neighbors for real value outcomes, just the closest k number of stored instances (near neighborhoods) are analyzed.

Each of the k neighbors evaluates the distance-related contribution using the following question and gives the nearby neighbors greater weight in the nearest distance method.

KNN often handles noisy data well since the closest neighbours are pooled. A collection of defined points is required in order to utilise K-dearest's algorithm, which is a classification algorithm that is supervised [15]. an is sought in this

```
Classification report of K-Nearest Neighbor Results:

              precision    recall  f1-score   support

           0       0.96      0.97      0.97        73
           1       0.99      0.98      0.99       175

    accuracy                           0.98       248
   macro avg       0.97      0.98      0.98       248
weighted avg       0.98      0.98      0.98       248

Cross validation test results of accuracy:
[1.         1.         1.         0.94117647 0.94117647 0.94117647
 1.         0.94117647 1.         0.94117647 1.         1.
 0.94117647 1.         1.         1.         1.         1.
 1.         1.         1.         1.         0.82352941 1.
 0.9375     1.         0.9375     0.9375     0.9375     0.9375
 1.         1.         0.9375     0.9375     0.9375     1.
 0.9375     1.         0.875      1.         0.9375     1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]

Accuracy result of K-Nearest Neighbor is: 97.44117647058823
```

Figure 13: Classification Report for KNN

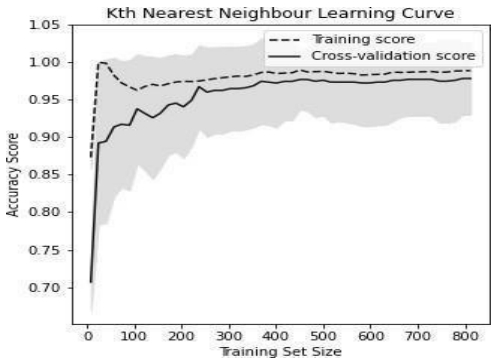

Figure 14: KNN Accuracy graph

5. DESCRIPTION AND RESULTS

The classification reports generated by each of the six algorithms on the data set have been examined and reviewed. The accompanying table shows our comparison of the accuracy of each method.

Table 3: Accuracy result of each algorithm

Algorithm	Accuracy
Logistic Regression	97.46%
Navies Bayes	97.38%
Random Forest	99.16%
Support Vector Machine	70.65%
K-Nearest Neighbors	97.61%
Decision Tree	99.88%

The decision tree, which has a 99.88% accuracy rate, is arguably the most efficient system, as shown in the above table. Contrarily, the Support Vector Machine Classifier, which has an accuracy of 70.65%, is the least accurate algorithm.

Contrarily, we observe that Random Forest has accuracy of 99.16%, which is nearly identical to that of Decision Tree. This is due to the fact that Random Forest is a subtype of Decision Tree and as such, has accuracy that is similar to Decision Tree. We can also see that the accuracy for the logistic regression, Naive Bayes, and KNN is roughly 97%.

Regardless of whether the application process is successful or not, India's meteorological service needs to automate air quality detection. (real time). Governments at all levels, including local and state, make a lot of efforts to forecast and take into account the air quality index for deemed public health. The artificial intelligence environment was made simpler by implementing the prediction framework. Advances in IOT infrastructure, big data applications, and machine learning techniques make real-time monitoring and assessment of air quality appealing in the context of future smart cities [11].

As part of future study, it may be possible to include and combine in an MTL context the combined features of neighboring weather stations, which can provide the forecast an additional boost. It is possible to extend this work by including.

6. CONCLUSION

To reduce air pollution, we have created effective machine training methods. We talked about the application of machine learning techniques for pollution estimation and the Indian air quality index. (AQI).

The importance of pollution event prediction has increased in India's main cities as a result of rising urbanization and the impact of increased traffic. Data was collected and cleaned from a variety of heterogeneous sources for use in machine learning techniques. By using a structure regularize to optimize outputs and reduce the number of model parameters. Our main focus has been on reducing model complexity. With an overall accuracy of 99.8%, we found that the Decision Tree Algorithm produced the greatest results out of all the methods. The importance of air pollution prevention

7. REFERENCES

[1] "Adaptive Algorithms for Signature Wavelet identification in the Musical Sounds," by M. Duraipandian. Review of Soft2020: 120–129 Computing Paradigm (JSCP), 2, no. 2.

[2] Haoxiang, S. Smys, and Wang. Using Named Entity Identification, MC-SVM Based Work Flow Preparation on the Cloud.2020's Journal of Soft Computing Paradigm 2, issue 02:130-139 Big data validation and quality assurance: difficulties, challenges, and needs.

[3] J. Gao, C.-L. Xie, and C.-Q. Tao, IEEE Symposium on Service-Oriented System and Engineering, IEEE Computer Society, Oxford, UK, April 2016.

[4] "HazeEst: Machine learning based metropolitan air pollution prediction using fixed and mobile sensors," IEEE Sens. J, vol. 17, no. 11, pp. 3517-3525, 2017. K. Hu, A. Rahman, H. Bhrugubanda, and V. Sivaraman.

[5] "SVR based dense air pollution assessment model employing static and wireless sensor network," IEEE SENS J Orlando FL, pp. 1-3, by K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang, and A. Rahman.

[6] A new air quality forecasting model employing data mining and artificial neural networks is presented by M. Huang, T. Zhang, J. Wang, and L. Zhu in the 6th IEEE International Conference on Program A.

[7] Li, S. Song, and X. Fei (2011). "Spatial characteristics of air pollution in Chengdu, China's central business district." Geoinformatics 2011: 19th International Conference, 1-4.

[8] Zhang, B., Yu, J., Zou, G., Yong, R., and Zhao Q. (2019). "A novel combination prediction strategy for urban PM2.5 concentration based on CNN and lstm." IEEE Access, 7(2005):20050–20059.

[9]] Rajesh Sharma, Akey, and Sungheetha. Edge Detection in CT and MRI Using GTIKF-GaborTransform Integrated K-Means and Fuzzy C Means Clustering, Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 111-119.

[10] "Data Mining based Forecast of Demand in Indian Market for Refurbished Electronics," Journal of Soft Computing Paradigm (JSCP) 2, Suma, V. and Shavige Malleshwara Hills no. 02 (2020): 101-110.

[11] "Classification model for air quality using machine learning approaches," International Journal of Software Engineering and Its Applications, pp. 45–52, 2015. S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, and A. A. Jamal.

[12] T. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatiotemporal data from air quality monitoring stations," IEEE, 2016's 14th International Conference on Industrial Informatics (INDIN).

[13] "Least squares support vector prediction for daily atmospheric pollution level," in Proc. 2010 IEEE/ACIS, by W. F. Ip, C. M. Vong, J. Y. Yang, and P. K. Wong

[14] "A Comprehensive Evaluation of Air Pollution Prediction Improvement by a Machine Learning Method," IEEE International Conference, 2015; Xia Xi, Zhao Wei, Rui Xiaoguang, Wang Yijie, Bai Xinxin, Yin Wenjun, et al.

[15] "Wavelet basedContent Based Image Retrieval using Colour and Texture featureExtraction by GLCM and CCM" in Journal of Computer Science, Science Publications, ISSN: 1549-3636(c),Volume 10 No. 1 (2014) by Jeyanthi P. and Jawahar Senthil Kumar V.

[16] "Content Based Image Retrieval based on Colour and Texture Feature Extraction in Android" International Conference on Information Communication and Embedded systems (ICICES), ISBN:978-1-4799-3834- 6/14, IEEE. Jeyanthi P., Jawahar Senthil Kumar V., and Vel Murugan.

[17] Exemplary Content Based Image Retrieval Using Visual Contents & Genetic Approach, 2015 International Conference on Communication and Signal Processing, ICCSP, Kavitha, N., Jeyanthi, P.

[18] Xi X, Wei Z, Xiaoguang R, et al. [18] (2016) a thorough analysis of machine learning's improved ability to predict air pollution. In: Service Operations & Logistics International IEEE Conference. IEEE

[19] Yue, G., Qiao, J., and Gu, K. (2019). "Effective and efficient estimation of PM2.5 concentration from photos." Journal of Instrumentation and Measurement, PP, 1–10.

[20] "Machine Learning Approach to Forecasting Urban Pollution," published in the IEEE journal in 2016, by Yves Rybarczyk and Rasa Zalakeviciute.

[21] 2018. Zhu, Dixian, Cai, Changjie, Yang, Tianbao, and Zhou, Xun. Big Data Cognitive Computing, Volume 2, Number 1, Article 5, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization." https://doi.org/10.3390/bdcc2010005.

