|  |  |  |  |
| --- | --- | --- | --- |
| **Synopsis Report**  **on**  **“PREDICTON OF AIR QUALITY INDEX”**  **using Machine Learning**  **Submitted as requirement for the**  **Second Year Mini Project Lab**  **Session 2022-23**  **By:**  **Janhavi Dwivedi (2100321530083)**  **Khushi Sharma (2100321530088)**  **Manisha Tayal (2100321530097)**  **Prachi Sharma (2100321530119)**  **Under the guidance of:**  **Mr. Dhyanendra Jain**  **Assistant Professor**  **DEPARTMENT OF CSE-AIML**  **ABES ENGINEERING COLLEGE, GHAZIABAD** | | | |
| m |  |  | naac A Grade Engineering College |
| **AFFILIATED TO**  **DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW**  **(Formerly UPTU)** | | | |

# Student’s Declaration

I / we hereby declare that the work being presented in this report entitled **“PREDICTON OF AIR QUALITY INDEX using Machine Learning”** is an authentic record of my/our own work carried out under the supervision of Mr. Dhyanendra Jain, Assistant Professor, CSE-AIML. The matter embrodied in this report has not been submitted by us anywhere else.

**Date:**

**Signature of student Signature of student**

**(Name: Janhavi Dwivedi) (Name: Khushi Sharma)**

**(Roll No: 2100321530083) (Roll No: 2100321530088)**

**Signature of student Signature of student**

**(Name: Manisha Tayal) (Name: Prachi Sharma)**

**(Roll No: 2100321530097) (Roll No: 2100321530119)**

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

**Signature of HOD Signature of Supervisor**

**…………………… Mr. Dhyanendra Jain**

**CSE-AIML Assistant Professor**

**Date:** **CSE-AIML**

**Acknowledgement**

We would like to convey our sincere thanks to **Mr. Dhyanendra Jain** for giving us the motivation, knowledge and support throughout the course of the project. The continuous support helps in a successful completion of the project. The knowledge provided is very useful for us.

We also like to give a special thanks to the department of CSE-AIML for giving us the continuous support and opportunities for fulfilling our mini project.

**Signature of student** **Signature of student**

**(Name: Janhavi Dwivedi) (Name: Khushi Sharma)**

**(Roll No: 2100321530083) (Roll No: 2100321530088)**

**Signature of student Signature of student**

**(Name: Manisha Tayal) (Name: Prachi Sharma)**

**(Roll No: 2100321530097) (Roll No: 2100321530119)**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Contents** | **Page No.** |
|  | Student’s Declaration | I |
|  | Acknowledgement | II |
|  | List of Figures | III |
|  | Abstract | Iv |
|  |  |  |
| Chapter 1: | Introduction | 7 |
| Chapter 2: | Related Work/Methodology | 8-9 |
| 2.1: | Existing Approaches | 8 |
| 2.2: | Comparative Analysis of Existing Works | 9 |
| Chapter 3: | Project Objective | 10 |
| Chapter 4: | Proposed Methodology | 11 |
| Chapter 5: | Design and Implementation | 12 |
| 5.1: | Work Flow Diagram | 12 |
| Chapter 6: | Results and Discussion | 13 |
| Chapter 7: | Conclusion and Future Scope | 14-15 |
|  | References | 16 |

**List of Figures**

|  |  |
| --- | --- |
| **Figure No and name** | **Page No.** |
| **Fig.1.** Proposed Approach | 4 |
| **Fig.2.** Work Flow Diagram | 5 |
|  |  |

**ABSTRACT**

A crucial natural resource, air quality has been degraded by economic activity. Governments in both developing and populous nations view the regulation of air as a crucial issue. The combustion of fossil fuels, transportation patterns, and industrial variables like power plant emissions all have a substantial impact on air pollution. To maintain excellent air quality, the air quality monitoring framework calculates various air pollutants at various locations. In the present context, it is the main problem. The introduction of hazardous gases into the environment contaminates the air. Prediction of air pollution is not exempt from the sectors where machine learning technologies are having a significant increase in influence and penetration. This study examines how machine learning algorithms based on sensor data might be used to predict air pollution in the setting of smart cities. From businesses, vehicle exits, and so forth. Nowadays, basic levels of air contamination have been reached, and the air contamination level in many important urban areas have exceeded the standards for air quality set by the public authority. It has a huge impact on a person's soundness. Using the progress in ML technology, it is now possible to predict the poisons based on information from the past. The influence of machine learning technology in nearly every business is expanding and infiltrating, and air pollution prediction is no exception. This study updates prior work on forecasting the air quality index using machine learning approaches. The most renowned databases were used to select the papers that were the most significant. After carefully reading those publications, the salient features were condensed, providing a framework for relating and contrasting them.

## 

## Chapter 1

**Introduction**

Economic activities have impacted the quality of air, a crucial natural resource. Air quality, transportation, and health care are only a few of the challenges that the rise of urbanization is responsible for. Because it significantly affects both human health and biological equilibrium, air pollution is a topic of increased public awareness today. Air pollution affects health, job productivity, and energy efficiency in addition to the environmental effects of harmful pollutants. Air pollution has an effect on both the natural equilibrium and human health. Human health is significantly impacted by gas concentrations in the air, which can also have harmful effects. The volume of rain is also altered by an increase in air pollution. Since air pollution can have a number of harmful consequences on people, it needs to be continuously studied in order to be properly controlled. One way to manage air pollution is to understand its origin, driving force, and entry point. The volume of rain is also altered by an increase in air pollution. Since air pollution can have a number of harmful consequences on people, it needs to be continuously studied in order to be properly controlled. One way to manage air pollution is to understand its origin, driving force, and entry point. The primary contributors to air pollution are ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO), sulfur dioxide (SO2), and particulate matter PM2.5, PM10. These invisible gases are produced by burning fossil fuels, wood, commercial boilers, and volcanic eruptions. They have the ability to cause harm to individuals and are primarily to blame for illnesses including cancer, birth defects, and problems with breathing. Due to COVID-19's recent rapid spread, several researchers are examining the underlying pollution-related issues that contribute to pandemics in many nations. Evidence that has been made public by the WHO raises questions about how widespread the contamination is. We realize that we should screen the air because the opportunity to do so has already passed. The Air Quality Prediction model seeks to work on the concentration of various pollutants, such as PM2.5, PM10, S02, N02, C0, benzene, toluene, and xylene, as well as on the weather conditions that also affect the AQI, or the Air Quality Index, of a region scaling them to a range and defining whether it is healthy, satisfactory, moderate, or unhealthy for the region. Various machine learning techniques are applied after the data has been pre-processed and correctly scaled. The creation of AQI prediction models for upcoming acute air pollution episodes is the main goal of the study addressed in this paper. The machine learning (ML) methods decision tree regression, lasso regression, and linear regression models are all being looked into. This study examines the reduction in forecasting precision over a longer time frame. Even after it has been forecasted, a layperson would not be able to interpret the numerical data for the air quality index, which can range from 0 to 300+. As a result, after being forecasted, the AQI will be split into 5 categories using a Quality Check tool. They are classified as "Healthy, moderate, "Unhealthy," "Very Unhealthy," and "Hazardous". Several machine learning approaches, some of which are listed below, are used to achieve the research's objectives.

## 

## 

## Chapter2

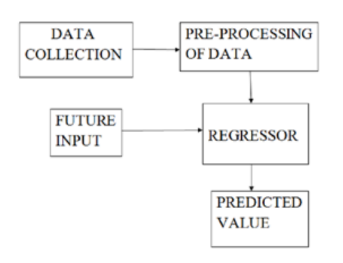
## Related Work

The related work associated with our project is given below:

### Existing Approaches

There are mainly two approaches viz. single pollutant index and multi-pollutant index to determine the air quality index. Every index has its own characteristic strengths and weaknesses that affect its suitability for particular applications. This paper attempts to present a review of all the major air quality indices developed worldwide.

The air quality file is anticipated by utilizing Machine Learning calculations for the recognition of PM2.5 level utilizing strategic relapse . There are applications that show the steady PM2.5 levels, while some show the gauge of a particular day. This structure mishandles ML models to perceive and figure PM2.5 levels reliant upon an educational assortment comprising of meteorological conditions in a specific city. The informational index utilized in discovery of PM2.5 level comprises of Temperature, Wind speed, Pressure, PM2.5, Concentration(ug/m^3).



### FIG.1.

### Central pollution control board built many pollution monitoring stations in heavily polluted areas, we collect the data from those monitoring stations.

### Implementation of software

### In Software specifications, used IDE is Anaconda python, Operating systems must be Windows 7/10 and we used the Coding language as Python.

### 

### FIG.2

### Comparative Analysis of Existing Works

In those papers, they only implemented the prediction of PM2.5. In this project they want to implement prediction of all the pollutants [CO, O3, NO2, SO2, PM2.5, PM10,toluene, xylene] with the help of meteorological data for better prediction.

## Chapter 3

**Project Objective**

* This project will predict the air quality index using machine learning.
* AQI informs the public about environmental conditions. It is especially useful for people suffering from illnesses aggravated or caused by air pollution.
* The air quality observing framework estimates different air toxins in different areas to keep up great air quality.
* It also helps in identifying faulty standards and inadequate monitoring programmes.
* We anticipate the air quality list by utilizing distinctive calculations like direct relapse, Decision Tree and Random Forest.
* This project will analyze different Machine Learning Algorithms and finds the one with best accuracy.
* We implemented the different machine learning algorithms in Python using Jupiter notebook.
* The air quality file is anticipated by utilizing ML calculations for the recognition of PM2.5 level utilizing strategic relapse.
* The major objective of this paper was to provide a snapshot of the vast research work and useful review on the current state of-the-art on applicable big data approaches and machine learning techniques for air quality evaluation and predication.
* Another prominent machine learning method, random forest, which can overcome the prediction variance that each decision tree has, is used.

## Chapter 4

## Methodology

Data collection, pre-processing, feature selection, time windowing, and model construction are the steps in this study's approach. All the machine learning models used in this study will be constructed using the open-source data mining platform. Data was gathered to teach the algorithm to recognize air quality. The data set was meant to have the following attribute sets: CO, SO2, O3, NO2, NO, benzene, toluene, and xylene. The meteorological data information set parameters of temperature, wind speed, humidity, and wind direction are used to train the system. Several machine learning approaches are used to achieve the goals of the study, but random forest regression is the most effective machine learning methodology and produces the most accurate findings.

**Random Forest Regression**: Natural forest is a type of bagging as opposed to boosting. The trees in a random wood are parallel to one another. The trees are not near one another when being installed. It completes its tasks by training many decision trees and creating the class that corresponds to each tree's categorization mode or average prediction (regression). A met estimator called a random forest combines multiple decision trees and makes some beneficial adjustments. It incorporates the outcomes of numerous projections. The hyper parameter (also known as the hyper parameter) is a specific proportion of the total number of functions at each node that can be divided on. To further randomie the data and prevent overfitting, each tree employs a random sample from the data set it divides. The phrase "Random Forest" refers to ensemble approaches that employ tree-type classifiers, where "h (x, k), k=1...," denotes a created classifier and "k" denotes an independent random vector with an identical distribution, and "x" denotes an input pattern. Recursive partitioning is used to create multiple trees, and the results are then combined. Each tree is separately constructed by first splitting the parameter set into numerous parts based on one of the parameters, then repeating the process for each part using a bootstrap sample of the training data.

Each decision tree is formed and then the unlabeled data are added. The ultimate probability of the AQI level I in the random forest is defined by p'(ci), where T is the number of decision trees as previously mentioned. The estimated probability of the AQI level I for each tree is given by p(ci).

P'(Ci) = 1/T K=1 Σ T P(Ci)

The equation which determines the result is given by:

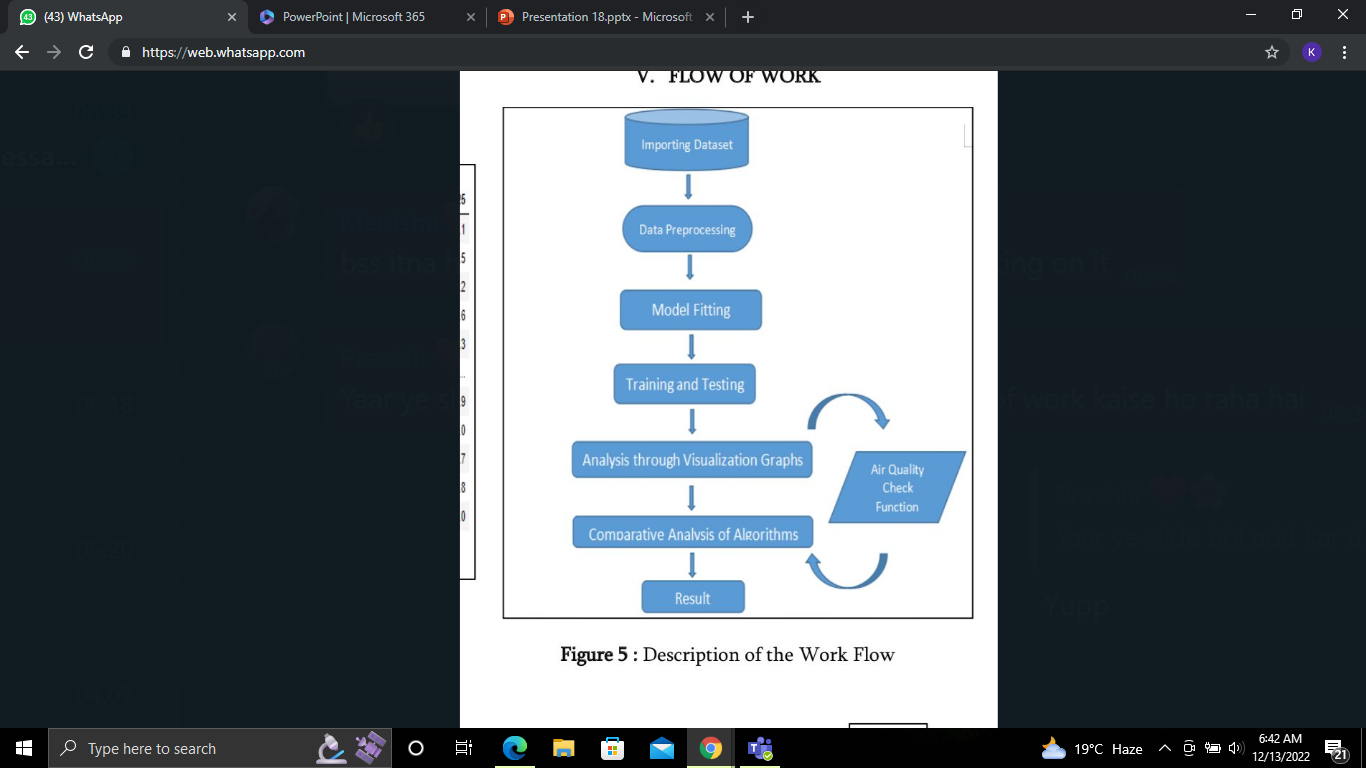
C′(i) = Max(p′(ci))

## Chapter 5

**Design and Implementation**

The design and implementation of our project is as follows:

### 5.1. Work Flow Diagram



**Fig.2.** Work Flow Diagram

## Chapter6

## Results and Discussion

All of the algorithms—random forest, decision tree, linear regression, AdaBoost, and ANN—perform admirably, but random forest regression slightly outperforms the competition in terms of accuracy. In order to increase the projected accuracy of the input dataset, the Random Forest classifier averages the results from multiple decision trees applied to various subsets of the input dataset. Random Forest is a machine learning method for classification and regression problems that is highly well supervised. One of the important features of this technique is its capacity to handle a dataset containing continuous variables, such as in the case of regression. It takes shorter training time than other algorithms. It functions well and makes accurate predictions of the outcome even with the massive dataset. Accuracy can be kept even when a sizable amount of data is missing.

The precise methodology and dataset utilised will have an impact on the outcomes and discussion of an air quality prediction using machine learning project. But generally speaking, the outcomes and conversations below can be anticipated:

**Results:**

One of the project's main outcomes will be the machine learning model's prediction accuracy. Metrics like mean squared error, root mean squared error, and correlation coefficient will be used to quantify this.

The machine learning algorithm will find the most significant factors that cause air pollution. This can aid in prioritising interventions and locating probable pollution sources.

On the basis of recent data, the machine learning model can generate real-time predictions of air quality. This can provide early warning systems for high pollution episodes and aid to inform public health advisories.

**Discussion:**

Comparison with current models: The outcomes of the machine learning model can be contrasted with those of current models like statistics and chemical transport models. This can be used to assess how well the machine learning strategy works.

Sources of pollution can be located using a machine learning model; examples include wildfires, industry, and traffic. This can help to target interventions and informing policy decisions.

Predictions made by the machine learning model may have an impact on public health, especially for vulnerable groups including young children, the elderly, and persons who have respiratory ailments. The possible health effects of air pollution and the significance of lowering emissions can both be discussed.

Overall, the findings and analysis of an air quality prediction using machine learning project can offer insightful information about the causes and effects of air pollution, and they can help guide the formulation of policies targeted at cutting emissions and enhancing public health.

**Chapter 7**

**Conclusion and Future Scope**

**Conclusion**

The purpose of this study is to provide an overview of the existing approaches to the topic of air quality prediction by looking at prior works. The most important finding was that, in order to increase the precision of air quality prediction, a dataset of other factors that affect air quality should be provided in addition to air quality data. It might be quite difficult to compare the findings because of the various data sets and temporal granularities used in the analysis of the studies. Future work is recommended to be exhaustive. The results might then be compared on an equal and comparable basis. We forecast the air quality index using specialized calculations such as linear, decision trees, and random forests. Our analysis of the data led us to the conclusion that the Random Forest method predicts the air quality Index more precisely. When using Random Forest Regression Training requires less time. It functions properly and provides the most precise predictions of the outcome. Even in the absence of a significant amount of data, accuracy can be maintained. To sum up, predicting air quality using machine learning is a potential method for predicting and tracking air pollution levels in various locations. Machine learning models may be trained to recognise patterns and produce precise forecasts of future air pollution levels by using historical data on weather and air quality. Machine learning-based air quality prediction has the potential to offer useful insights for policymakers, city planners, and public health officials in order to better understand the impact of air pollution and create practical solutions to reduce its detrimental effects on both human health and the environment. Using machine learning to predict air quality is a promising strategy that can deliver precise and timely information to assist reduce the harmful effects of air pollution on the environment and public health. Overall, using machine learning to anticipate air quality can considerably improve environmental management and safeguard human health. To provide accurate and precise predictions, nevertheless, ongoing updates and enhancements to the models' precision and accuracy are required.

**Future scope**

A potential future lies in the fast developing field of air quality prediction utilising machine learning. Here are a few suggested directions for improvement:

**Accuracy gains:** Machine learning models can be trained to produce predictions about air quality that are more accurate as more data becomes available. Longer-term forecasts as well as more precise real-time forecasts may fall under this category.

**Integration with smart devices**: Machine learning models may be integrated with smart devices to deliver real-time data regarding air quality as more of them become accessible.

**Public awareness:** By utilising machine learning to predict air quality, the public may become more knowledgeable about these issues, enabling both individuals and governments to make more informed decisions.

**Policy creation:** The data produced by machine learning models could assist in the creation of laws and emission reduction objectives for the improvement of air quality.

**Extension to additional contaminants:** At the moment, the majority of machine learning algorithms concentrate on forecasting particulate matter and other typical pollutants. These models might, however, be expanded to forecast other contaminants like ozone and carbon monoxide.

**Integration with other environmental data:** To provide more thorough information regarding air quality, machine learning models may be combined with other environmental data such as weather patterns and satellite imagery.

Overall, the use of machine learning to predict air quality has a promising future, with the possibility of ongoing accuracy increases and wider applications in public health and policy creation.

## References

1. Kennedy Okokpujie, Etinosa Noma-Osaghae, Odusami Modupe, Samuel John, and Oluga Oluwatosin, “A SMART AIR POLLUTION MONITORING SYSTEM,” International Journal of Civil Engineering and Technology (IJCIET), vol. 9, no. 9, pp. 799–809, Sep. 2018.

2. D. Zhu, C. Cai, T. Yang, and X. Zhou, “A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization,” Big Data and Cognitive Computing, vol. 2, no. 1, p. 5, Mar. 2018.

3.A. Masih, “Machine learning algorithms in air quality modeling,” Global Journal of Environmental Science and Management, vol. 5, no. 4, pp. 515–534, 2019.

4. Campbell-Lendrum, D., & Prüss-Ustün, A. (2018). Climate change, air pollution and noncommunicable diseases. Bulletin of The World Health Organization, 97(2), 160-161. <https://doi.org/10.2471/blt.18.224295>

5. Random Forest Regression in Python - GeeksforGeeks. GeeksforGeeks. (2020). Retrieved 22 July 2020, from https://www.geeksforgeeks.org/random-forest-regression-in

6. Decision Tree Regression — scikit-learn 0.23.1 documentation. Scikit-learn.org. (2020). Retrieved 22 July 2020, from <http://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html>.

7. Kostandina Veljanovska and Angel Dimoski, “Air Quality Index Prediction Using Simple Machine Learning Algorithms,” International Journal of Emerging Trends & Technology in Computer Science, vol. 7, no. 1, 2018.

8. Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

9. Yi, X.; Zhang, J.; Wang, Z.; Li, T.; Zheng, Y. Deep Distributed Fusion Network for Air Quality Prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, UK, 19–23 August 2018; pp. 965–973.

10. Veljanovska, K.; Dimoski, A. Air Quality Index Prediction Using Simple Machine Learning Algorithms. Int. J. Emerg. Trends Technol. Comput. Sci. 2018, 7, 25–30.

11. Ghorani-Azam, A.; Riahi-Zanjani, B.; Balali-Mood, M. Effects of Air Pollution on Human Health and Practical Measures for Prevention in Iran. J. Res. Med. Sci. 2016, 21, 1–12.

12. Rocca, J. Ensemble Methods: Bagging, Boosting and Stacking. Available online: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (accessed on 23 April 2019).

13. Awad, M.; Khanna, R. Support Vector Regression. In Ecient Learning Machines; Apress: Berkeley, CA, USA, 2015.