iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52086

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Air Quality Prediction using Machine Learning and Cloud Computing

Prof. D.J Chaudhari[1], Monali Sonawane[2], Amit Kumar[3], Aniket Magare[4], Dhara Singh[5], Sanjana Shrivastav[6]
Government College of Engineering, Chandrapur, Maharashtra

***Abstract:** Because humans rely so heavily on air for their survival, air quality has also become a topic of controversy. Today's air quality is quite dangerous for human intake due to the rising urbanization. Many people die each year as a result of respiratory ailments brought on by breathing in poor air. Urbanization cannot be prevented, but there must be a human warning system that allows a person to be informed of dangerous air pollution in the area. A person can choose an alternative route of travel or decide not to go somewhere if the pollution levels are excessive. In an effort to find a cooperative solution to the issue, we will try to research numerous papers written by other authors. So, in order to fulfill our goal of predicting air pollution, we must first gather information about the weather's impact on air quality. After gathering data for a particular location, we will construct a training dataset that will be used to train the XGBoost machine learning algorithm, which will forecast the results of the training data input. Then these trained model will be integrated with the web module. The web module will take input and predict the air quality index based on it.*

I. INTRODUCTION

An instrument for assessing the quality of the air we breathe is the Air Quality Index (AQI). It is an indicator that weighs various air contaminants and translates their concentrations into a single number that symbolizes the general state of the air. Ground-level ozone, particle matter (PM2.5 and PM10), carbon monoxide, Sulphur dioxide, and nitrogen dioxide are the pollutants that make up the AQI.

The AQI gives details on air pollution levels and potential health effects. The range of the air quality index (AQI) is 0 to 500, with higher values indicating higher pollution levels and greater health concerns. Good air quality is defined as an AQI value of 50 or below, whereas hazardous air quality is defined as a value more than 300.

AQI is a tool used by governments, environmental organizations, and health organizations to inform the public about the state of the air in a particular area. During times of low air quality, people can use the AQI to take preventative measures and safeguard their health. To reduce exposure to contaminated air, actions can be done include limiting outside activity, using masks, and closing windows.

Numerous detrimental impacts of air pollution on human health have been documented, including respiratory and cardiovascular issues, cancer, and other chronic illnesses. Particularly susceptible to the impacts of air pollution are young children, the elderly, and those with pre-existing medical issues. Additionally, poor air quality can harm the environment through acid rain, crop and forest damage, and ozone layer erosion.

Additionally, people can help the environment by using less energy at home, travelling by public transportation, commuting by foot or bicycle, and properly disposing of hazardous waste. Campaigns for education and awareness can also serve to increase public understanding of the value of good air quality and motivate individuals to take action to minimize air pollution.

Overall, AQI is a crucial tool for understanding the air we breathe and for assisting us in making decisions that will safeguard both our health and the environment.

II. AIR QUALITY INDEX

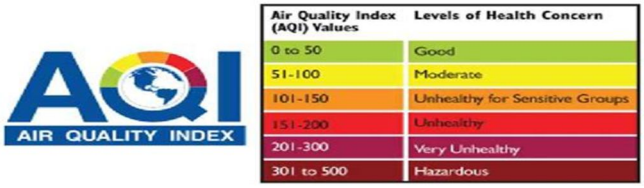


Figure 1. Graphical Representation of AQI.

Governmental organisations and other organisations use the Air Quality Index (AQI) as a measure of air quality to determine the amount of air pollution in a certain location. The concentration of different air pollutants, such as ground-level ozone, particle pollution, carbon monoxide, sulphur dioxide, and nitrogen dioxide, forms the basis of the AQI.

Higher scores on the AQI scale, which runs from 0 to 500, indicate more polluted air and greater health concerns. A score of 100 on the AQI indicates moderate air quality, a score of 150 or higher indicates unhealthy air for sensitive groups, and a score of 200 or higher indicates unhealthy air for the general public. The AQI informs the general population about air quality and how it affects health. People may endure a variety of health impacts, including respiratory issues, heart disease, and stroke, when air pollution levels are high. Particularly susceptible to the impacts of air pollution are young children, the elderly, and those with pre-existing medical issues. Public health experts and government agencies use the AQI to recommend steps to decrease exposure to air pollution and to issue health advisories. For instance, when the AQI is high, people can be instructed to stay indoors, refrain from exercising outside, and use air conditioning to purify the air inside. The AQI can be used to predict future air quality in addition to providing information about present air quality. People can use this knowledge to organise their activities and take precautions for their health. In general, the AQI is a crucial instrument for keeping track of air quality and safeguarding public health. The AQI assists people in making wise decisions about their daily activities and taking action to lessen their exposure to air pollution by giving information about air pollution levels and their effects on health.

III. SYSTEM ARCHITECTURE

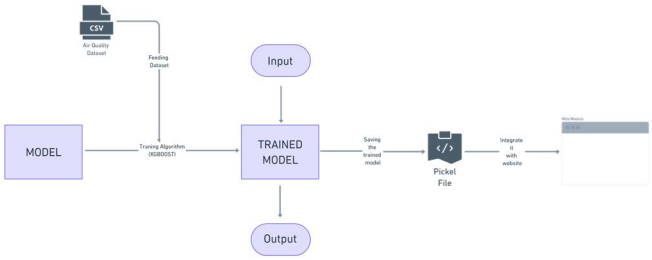


Figure 2: System architecture of project

1) *Trained Model:* A machine learning algorithm that has learned to recognise patterns in a given dataset produces a trained machine learning (ML) model. A set of input data and the matching output values or labels are provided to the algorithm during training. The method tweaks the model's parameters until it successfully predicts the outcome given the input data. Once trained, the model can be used to make predictions based on fresh, unexplored data. In other words, the trained ML model may apply what it has learned to predict outcomes from previously unknown data. The calibre of the training data, the algorithm of choice, and the parameter settings applied throughout the training process all affect how accurate a trained model is.
2) *Pickel File:* A pickle file is a binary file format used to serialise and deserialize Python objects in the context of Python programming. It is a technique for saving a Python object's current state in a file that may be accessed later, even if the Python programme is no longer active. Pickling a Python object turns it into a series of bytes that may be written to a file or sent over the internet. When the object needs to be used again, it can be "unpickled" to revert to the original Python object by reading it from the file or receiving it over the network. Machine learning models that have been trained are frequently saved in pickle files since it can be time-consuming to retrain them each time they are required. A trained model can be quickly loaded into memory and used to make predictions on new data by saving it as a pickle file.
3) *Flask based Website:* A website that was created with Flask, a well-liked Python web application framework, is referred to as being Flask-based. Flask is a compact and adaptable framework that makes it simple to create web applications fast and effectively. Python can be used to create web applications with Flask, making it simple to interface with other Python libraries and tools. URL routing, templating, and request handling are just a few of the capabilities and tools that Flask offers to make online development easier.

An HTML template, static files (like CSS and JavaScript files), and other components make up a Flask-based website. The Python modules that define the routes and views of the website are often included. The Flask framework manages incoming HTTP requests and responses while running on a web server like Apache or Nginx.

IV. WORKING

The working of the model are as follows:

1) Model is prepared using Google Collaboratory.
2) Training Algorithm is applied on model.
3) Dataset is fed into the model.
4) Finally, after training we test the trained model on a set of data
5) The trained model is saved in the form of a pickle file.
6) Then this pickle file is integrated into the website.
7) The website will predict the AQI based on the input parameters
8) Fed the input into the real_2018.csv file
9) On pressing the predict the button the AQI will be predicted.

The described procedure entails building a machine learning model to forecast the Air Quality Index (AQI). Using Google Collaboratory, a training algorithm is used to generate the model. The model is given a dataset, and its performance is assessed by testing it on a set of data. A pickle file is used to store the model after training. The website that uses this file to estimate the AQI based on input parameters is then integrated with it. The website will utilise the trained model to predict the AQI after the user inputs these parameters. The real_2018.csv file receives the input when the predict button is clicked, and the website then displays the projected AQI number.

V. METHODOLOGY

The model was trained on three algorithms which are

1) Linear Regression
2) XGBoost
3) Random Forest

Let's see some information on the proposed models

A. *Linear Regression*

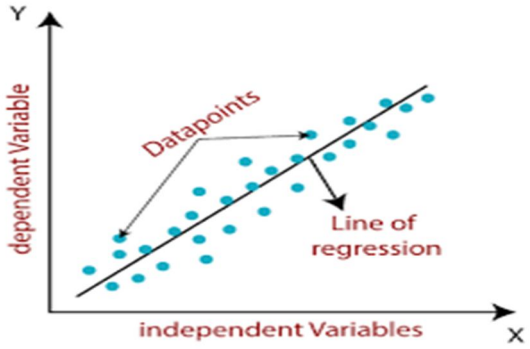


Figure 3. Graphical Representation of Linear Regression

A statistical technique called linear regression seeks to prove the existence of a correlation between a dependent variable and one or more independent variables. Developing a linear equation that can forecast the value of the dependent variable from the values of the independent variables is the objective. Y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept, can be used to represent a simple linear regression model with one independent variable.
In numerous disciplines, including finance, economics, engineering, and science, linear regression is frequently utilized. It can be applied to estimate trends, create predictions, and find correlations between different variables. The validity of hypotheses and the strength of the link between variables can both be assessed using linear regression. However, it makes the assumption that the variables have a linear connection, which may not necessarily be the case in real-world situations.

B. XGBoost

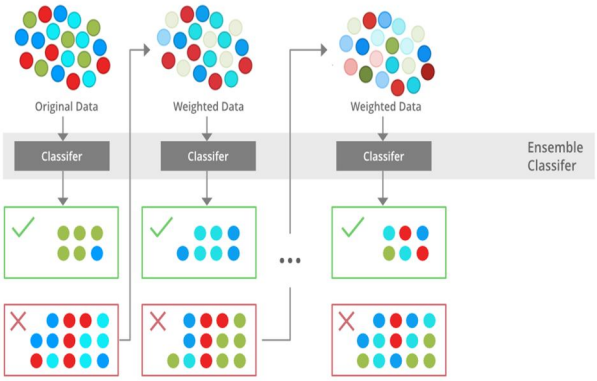


Figure 4. Graphical Representation of XGBoost

Gradient boosting is a well-known machine learning method used for supervised learning tasks like classification and regression. XGBoost is an optimized implementation of this algorithm. A tree ensemble model, used by XGBoost, combines the predictions of numerous trained decision trees to get the final result.

Since big datasets are typical in industry applications, XGBoost is well known for its scalability and performance. Both CPU and GPU processing are supported, and it can deal with missing values. To avoid overfitting and enhance generalization performance, XGBoost also incorporates regularization algorithms.

In numerous machine learning competitions, XGBoost has produced cutting-edge results and has been applied in a number of industries, including banking, healthcare, and natural language processing. Numerous computer languages, including Python, R, and Java, also support it. However, compared to simpler models, tweaking its hyperparameters can be difficult, and its interpretability is constrained.

C. Random Forest

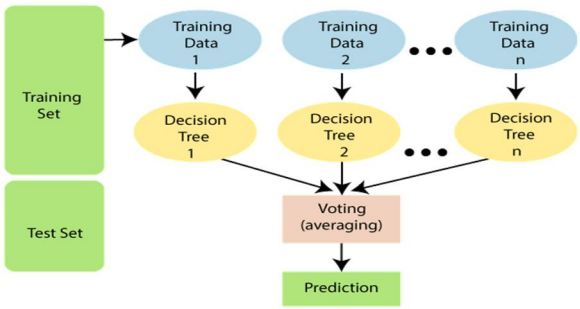


Figure 5. Graphical Representation of Random Forest.

Popular machine learning algorithm Random Forest is a member of the ensemble technique family. In order to increase the model's accuracy and robustness, it integrates several decision trees.

Each decision tree in a random forest is trained using a portion of the data and features that are randomly chosen. This lessens overfitting and enhances the model's generalization capabilities.

Compared to single decision trees, Random Forest has more benefits, such as improved accuracy, lower variance, and better resilience to noise and outliers. It is frequently employed in a variety of applications, including as feature selection, regression, and classification.

However, compared to more straightforward models, Random Forest is more difficult to interpret and can be computationally expensive, especially for large datasets. Additionally, adjusting its hyperparameters, such as the number of trees and the maximum depth of each tree, might be difficult.

When we trained our model based on the three algorithms, we got the following accuracies.

	Model Name	Accuracy
1	Linear Regression	48.52533130856788
2	XGBoost	81.42013612720127
3	Random Forest	76.8494726584375

Figure 6. Accuracies.

As the accuracy of XGBoost is optimum so we selected XGBoost for fitting the data into our model.

VI. . DATASET

A dataset is a group of data that has been compiled, organised, and saved in a structured style for quick access and study. It can be viewed as a systematic collection of knowledge that is employed in the development of machine learning models or in the analysis of specific research problems.

Depending on the sort of data it includes and the use it is intended for, a dataset can take on a variety of distinct shapes. An assortment of pictures, text documents, audio files, or numerical data, for instance, could make up a dataset.

Science, engineering, finance, and the social sciences are just a few of the disciplines that frequently employ datasets. They help to construct models and algorithms, test theories, and provide answers to research issues.

Datasets are used in machine learning to train models that can categorise or make predictions based on new data. The quality and quantity of the data used to train a machine learning model determines how well it performs, so picking the right dataset is crucial when developing machine learning systems.

Overall, datasets are a crucial tool for machine learning and data analysis because they offer a disciplined method of organising and analysing complex information.

Parameters used in Dataset:

1) T: Temperature, measured in degrees Celsius (°C)
2) TM: Minimum temperature, measured in degrees Celsius (°C)
3) Tm: Maximum temperature, measured in degrees Celsius (°C)
4) SLP: Atmospheric pressure at mean sea level, measured in hectopascals (hPa)
5) H: Relative humidity, expressed as a percentage (%)
6) VV: Visibility, measured in kilometers (km)
7) V: Wind speed, measured in kilometers per hour (km/h)
8) VM: Maximum sustained wind speed, measured in kilometers per hour (km/h)
9) PM 2.5: Concentration of particulate matter with a diameter of 2.5 microns or less, measured in micrograms per cubic meter (µg/m³)

VII. SNAPSHOTS

A. *Web Module*

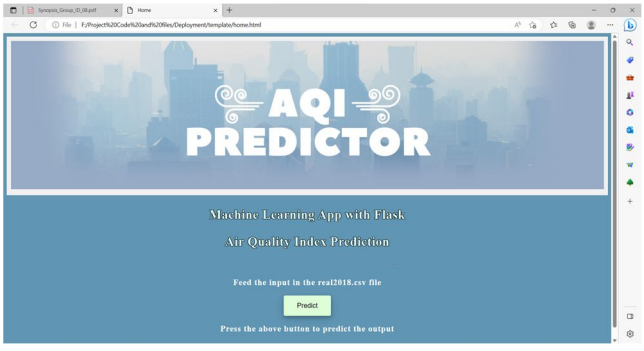


Figure 7. Graphical Representation of Website.

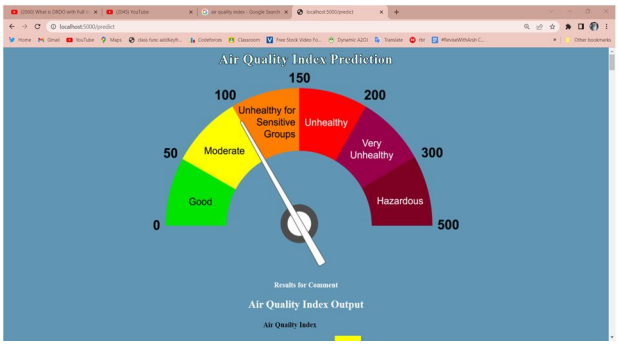


Figure 8. Graphical Representation of Website.

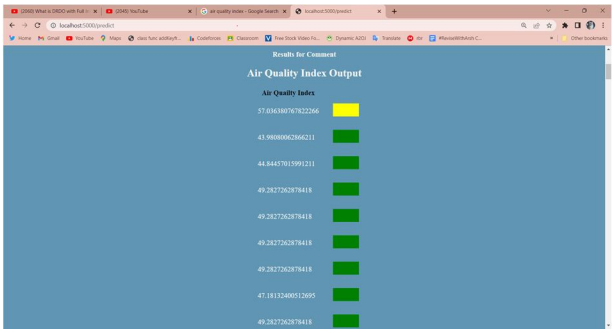


Figure 9. Graphical Representation of Website.

B. ML Model

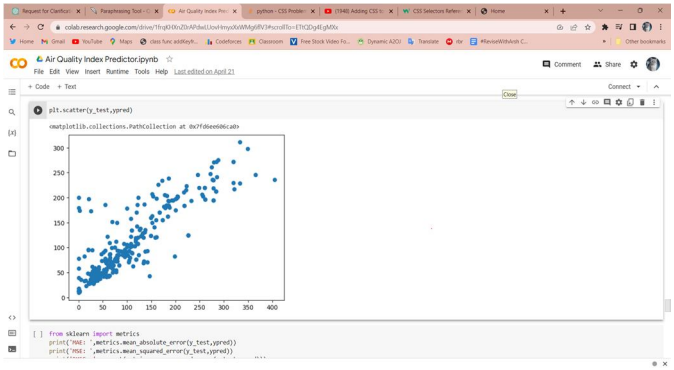


Figure 10. Graphical Representation of ML Model

C. Dataset

	A	B	C	D	E	F	G	H	I
1	T	TM	Tm	SLP	H	VV	V	VM	PM 2.5
2									
3	7.4	9.8	4.8	1017.6	93	0.5	4.3	9.4	219.7208
4									
5	7.8	12.7	4.4	1018.5	87	0.6	4.4	11.1	182.1875
6									
7	6.7	13.4	2.4	1019.4	82	0.6	4.8	11.1	154.0375
8									
9	8.6	15.5	3.3	1018.7	72	0.8	8.1	20.6	223.2083
10									
11	12.4	20.9	4.4	1017.3	61	1.3	8.7	22.2	200.6458
12									
13	16	25.2	10	1013.2	79	0.6	4.8	11.1	285.225
14									
15	13.4	21	9.2	1015.1	87	0.5	1.5	7.6	236.825
16									

Figure 11. Graphical Representation of ML Model

VIII. CONCLUSION

The air quality in a particular area is gauged using a numerical scale called the Air Quality Index (AQI).

1) The AQI gives data on the concentrations of air pollutants such particulate matter, ozone, carbon monoxide, Sulphur dioxide, and nitrogen dioxide in the atmosphere.

2) Governmental organizations and health organizations utilize the AQI to alert the public to poor air quality and to issue advises and warnings when pollution levels are high.

3) High amounts of air pollution can cause a variety of harmful health impacts, including respiratory issues, heart disease, stroke, and cancer.

4) By giving information on the air quality and offering suggestions for lowering exposure, the AQI is a crucial tool for assisting individuals in protecting themselves from the damaging impacts of air pollution.

5) For the purpose of identifying problem regions and taking action to minimize pollution, it is important to constantly evaluate the levels of air quality.

6) In addition to promoting the use of cleaner energy sources and encouraging people to take action to lessen their personal contributions to air pollution, policies and regulations can be put in place to minimize emissions from industry, transportation, and other sources.

7) The AQI is a useful instrument for monitoring air pollution reduction efforts and gauging the success of laws and regulations aimed at enhancing air quality.

8) When deciding on transportation options, planning outdoor activities, and taking precautions for their health, people should take the AQI into account.

9) We can make the environment healthier and more sustainable for present and future generations by adopting steps to reduce air pollution.

REFERENCES

The principles of several cloud computing and machine learning techniques are covered in this section, which can be utilized to develop a new, more secure and reliable air quality forecast system and aid users in understanding their immediate surroundings. It aids in comprehending the numerous concepts advanced by various technical papers written by various writers and how they present a more precise and practical techniques. Below are some concepts, along with their advantages and disadvantages:

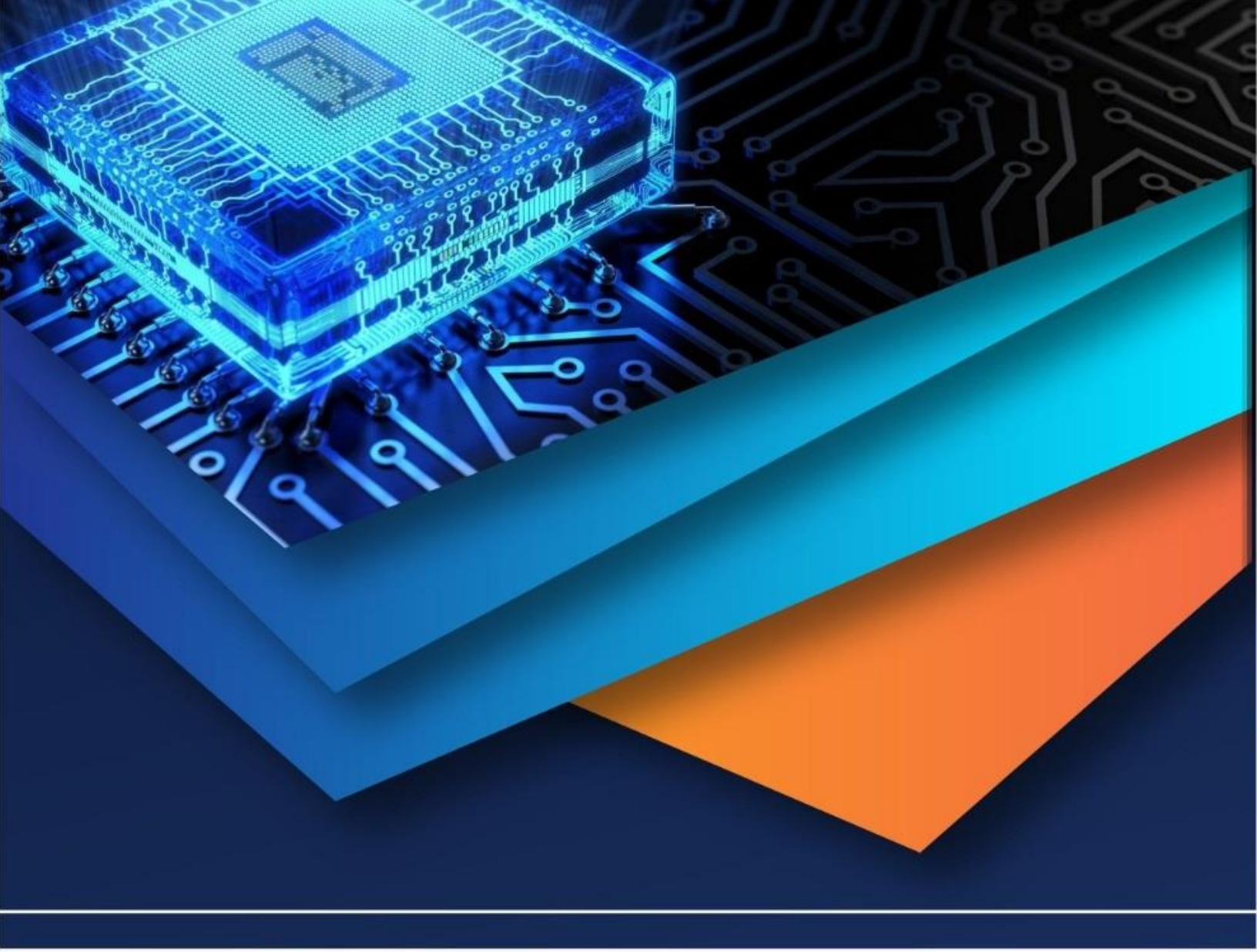
[1] YING ZHANG, YANHAO WANG, MINGHE GAO, QUNFEI MA, JING ZHAO, RONGRONG ZHANG, QINGQING WANG and LINYAN HUANG, ‘‘A Predictive Data Feature Exploration-Based Air Quality Prediction Approach,’’ in IEEE-2019.

[2] Liying Li, Zhi Li, Lara G. Reichmann and Diane Myung-kyung Woodbridge, ‘‘A Scalable and Reliable Model for Real-time Air Quality Prediction,’’ in IEEE-2019.

[3] Nandini K and G Fathima, ‘Urban Air Quality Analysis and Prediction Using Machine Learning,’’ in IEEE-2019.

[4] Anish Singh, Raja Kumar and Nitasha Hasteer, ‘‘Comparative Analysis of Classification Models for Predicting Quality of Air,’’ in IEEE-2020.

[5] Xinyu Zhang, ‘‘ION Channel Prediction Using Lightgbm Model,’’ in IEEE-2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY