

AIR QUALITY INDEX ANALYSIS USING MACHINE LEARNING

*A Project report submitted in the partial fulfilment of the requirements for the award of
the degree of*

BACHELOR OF TECHNOLOGY In COMPUTER SCIENCE AND ENGINEERING

Submitted by

M.V.Jyoshna (19471A0534)

Ch.Chandana (19471A0507)

V.Sri Mallika (19471A0562)

Under the esteemed guidance of

M. Venkata Rao M.Tech, Assist Prof.



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**NARASARAOPETA ENGINEERING COLLEGE: NARASARAOPET
(AUTONOMOUS)**

**Accredited by NAAC with A+ Grade and NBA under Cycle -1
NIRF rank in the band of 251-320 and an ISO 9001:2015 Certified
Approved by AICTE, New Delhi, Permanently Affiliated to JNTUK, Kakinada
KOTAPPAKONDA ROAD, YALAMANDA VILLAGE, NARASARAOPET-522601
2022-2023**

**NARASARAOPETA ENGINEERING COLLEGE: NARASARAOPET
(AUTONOMOUS)**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE



This is to certify that the main project entitled “Air Quality Index Analysis Using Machine Learning” is a bonafide Work done by “M.V.Joshna (19471A0534),Ch.Chandana (19471A0507),V.Sri Mallika (19471A0562)in partial fulfilment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY** in the Department of **COMPUTER SCIENCE AND ENGINEERING** during the academic year 2022- 2023.

PROJECT GUIDE

M.Venkata Rao M.Tech., Assist Prof.

PROJECT CO-ORDINATOR

Dr. M. Sireesha M.Tech., Ph.D.

HEAD OF THE DEPARTMENT

Dr. S. N. TirumalaRao M.Tech.,Ph.D.

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We wish to express our thanks to carious personalities who are responsible for the completion of the project. We are extremely thankful to our beloved chairperson sir **M. V. Koteswara Rao**, B.sc who took keen interest on us in every effort throughout this course. We owe out gratitude to our principal **Dr.M. Sreenivasa Kumar**, M.Tech., Ph.D(UK), MISTE, FIE(1) for his kind attention and valuable guidance throughout the course.

We express our deep felt gratitude to **Dr. S. N. Tirumala Rao**, M.Tech., Ph.D. head of the department (HOD),computer science and engineering(CSE) department and our guide **M.Venkata Rao** AssistProf ,M.tech of CSE department whose valuable guidance and unstinting encouragement enable us to accomplish our project successfully in time.

We extend our sincere thanks to **Dr. M. Sireesha** M.Tech., Ph.D. Coordinator of the project for extending her encouragement. Their profound knowledge and willingness have been a constant source of inspiration for us throughout this project work.

We extend our sincere thanks to all other teaching and non-teaching staff of department for their cooperation and encouragement during our B. Tech degree. we have no words to acknowledge the warm affection, constant inspiration and encouragement that we receive from our parents.

We affectionately acknowledge the encouragement received from our friends and those who involved in giving valuable suggestions and clarifying out doubts, which had really helped us in successfully completing our project.

By

M.V.Joyshna (19471A0534)

Ch. Chandana (19471A0507)

V.Sri Mallika (19471A0562)



INSTITUTE VISION AND MISSION

INSTITUTION VISION

To emerge as a Centre of excellence in technical education with a blend of effective student centric teaching learning practices as well as research for the transformation of lives and community,

INSTITUTION MISSION

M1: Provide the best class infra-structure to explore the field of engineering and research

M2: Build a passionate and a determined team of faculty with student centric teaching, imbibing experiential, innovative skills

M3: Imbibe lifelong learning skills, entrepreneurial skills and ethical values in students for addressing societal problems



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VISION OF THE DEPARTMENT

To become a centre of excellence in nurturing the quality Computer Science & Engineering professionals embedded with software knowledge, aptitude for research and ethical values to cater to the needs of industry and society.

MISSION OF THE DEPARTMENT

The department of Computer Science and Engineering is committed to

M1: Mould the students to become Software Professionals, Researchers and Entrepreneurs by providing advanced laboratories.

M2: Impart high quality professional training to get expertise in modern software tools and technologies to cater to the real time requirements of the industry.

M3: Inculcate team work and lifelong learning among students with a sense of societal and ethical responsibilities.



Program Specific Outcomes (PSO's)

PSO1: Apply mathematical and scientific skills in numerous areas of Computer Science and Engineering to design and develop software-based systems.

PSO2: Acquaint module knowledge on emerging trends of the modern era in ComputerScience and Engineering

PSO3: Promote novel applications that meet the needs of entrepreneur, environmental and social issues.



Program Educational Objectives (PEO's)

The graduates of the programme are able to:

PEO1: Apply the knowledge of Mathematics, Science and Engineering fundamentals to identify and solve Computer Science and Engineering problems.

PEO2: Use various software tools and technologies to solve problems related to academia, industry and society.

PEO3: Work with ethical and moral values in the multi-disciplinary teams and can communicate effectively among team members with continuous learning.

PEO4: Pursue higher studies and develop their career in software industry.



Program Outcomes

- 1. Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
- 2. Problem analysis:** Identify, formulate, research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
- 3. Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
- 4. Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
- 5. Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Project Course Outcomes (CO'S):

CO425.1: Analyse the System of Examinations and identify the problem.

CO425.2: Identify and classify the requirements.

CO425.3: Review the Related Literature

CO425.4: Design and Modularize the project

CO425.5: Construct, Integrate, Test and Implement the Project.

CO425.6: Prepare the project Documentation and present the Report using appropriate method.

Course Outcomes – Program Outcomes mapping

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1		✓											✓		
C425.2	✓		✓		✓								✓		
C425.3				✓		✓	✓	✓					✓		
C425.4			✓			✓	✓	✓					✓	✓	
C425.5					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
C425.6									✓	✓	✓		✓	✓	

Course Outcomes – Program Outcome correlation

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1	2	3											2		
C425.2			2		3								2		
C425.3				2		2	3	3					2		
C425.4			2			1	1	2					3	2	
C425.5					3	3	3	2	3	2	2	1	3	2	1
C425.6									3	2	1		2	3	

Note: The values in the above table represent the level of correlation between CO's and PO's:

1. Low level

2. Medium level

3. High level

Project mapping with various courses of Curriculum with Attained PO's:

Name of the course from which principles are applied in this project	Description of the device	Attained PO
C3.2.4, C3.2.5	Gathering the requirements and defining the problem, plan to develop a mechanism to identify the air quality .	PO1, PO3
CC4.2.5	Each and every requirement is critically analyzed, the process model is identified and divided into four modules	PO2, PO3
CC4.2.5	Logical design is done by using the unified modelling language which involves individual team work	PO3, PO5, PO9
CC4.2.5	Each and every module is tested, integrated, and evaluated in our project	PO1, PO5
CC4.2.5	Documentation is done by all our three members in the form of a group	PO10
CC4.2.5	Each and every phase of the work in group is presented periodically	PO10, PO11
CC4.2.5	Implementation is done and the project will be handled by the college management and in future updates in our project can be done based on air quality value occurring in environment	PO4, PO7
CC4.2.8 CC4.2.	The physical design includes hardware components like intel core software and windows 10.	PO5, PO6

ABSTRACT

Most of the industry applications create pollution in the air and the vehicle emissions are also dangerous to the health of the people. In the developing countries, air pollution is severe in most of the areas. Most of the air quality measuring systems uses air quality index to tell the people about the air quality of their location. The primary objective of the system is to analyze and visualize air quality from the real time sensor data. The proposed system analyses four critical air pollutants which are Nitrogen dioxide (NO₂) , Sulphur dioxide (SO₂) , SPM and RSPM are the most widespread health threats.

INDEX

S.NO	CONTENTS	PAGE NO
I.	LIST OF FIGURES	IV
1.	INTRODUCTION	1
	1.1 Introduction	1
	1.2 Existing System	2
	1.3 Proposed System	2
	1.4 System Requirements	3
	1.4.1 Hardware Requirements	3
	1.4.2 Software Requirements	3
2.	LITERATURE SURVEY	4
	2.1 Machine Learning	4
	2.2 Some Machine Learning Methods	5
	2.3 Applications of Machine Learning	5
3.	SYSTEM ANALYSIS	6
	3.1 pairplots of gases	6
	3.2 Types of gases in air quality index	6
	3.3 Importance of Machine learning	9
	3.4 Implementation of Machine Learning using python	9
	3.5 Scope of the project	11

4.	METHODOLOGY	12
	4.1 Data Set	12
	4.2 Data Preprocessing	12
	4.2.1 Missing Values	13
	4.2.2 Correlation coefficient method	14
	4.3 Machine learning algorithms for Classification	15
5.	IMPLEMENTATION CODE	16
	5.1 Backend	16
	5.2 Frontend	18
	5.3 Connection	21
	5.4 Result Analysis	23
6.	OUTPUT SCREENS	24
7.	CONCLUSION AND FUTURE SCOPE	26
8.	BIBLIOGRAPHY	27

LIST OF FIGURES

S.NO	LIST OF FIGURES	PAGE NO
1.	Fig.3.1 pairplots of gases	6
2.	Fig.3.2.1 Sulfur Dioxide	7
3.	Fig.3.2.2 Nitrogen Dioxide	7
4.	Fig.3.2.3 Suspended Particulate Matter	8
5.	Fig.3.2.4 Respirable Suspended Particulate Matter	8
6.	Fig 4.1 Dataset	12
7.	Fig.4.2.1.1 Before missing data visualization	13
8.	Fig.4.2.1.2 After missing data visualization	14
9.	Fig.4.2.2 Correlation	14
10.	Fig.5.4 Comparison of models	23
11.	Fig.6.1 web page design	24
12.	Fig.6.2 data entered	24
13.	Fig.6.3 Output is Good	25
14.	Fig.4.5 Output is Moderate	25

1. INTRODUCTION

1.1 Introduction

One of the world's unembellished environment issues is air pollution. World Health Organization (WHO) said in its report that, in developing countries there are more than 2 million people died annually due to air pollution in India, China and Pakistan. Also, air pollution causes health issues around the world and 7 million people died annually. Most of the industries generate polluted air because of material processing and other industry related works. Urban areas are in danger to live due to air pollution. Air pollution is due to the lack of knowledge and carelessness of the people. Also, air may pollute due to natural calamity such as flooding, rocks collapse and gas leakage which causes severe health problems. The diseases such as heart disease, respiratory disease, stroke, chronic disease and some types of cancers are due to air pollution.

The daily quality of air is measured with the help of Air Quality Index (AQI). This measure indicates the cleanliness of the air around us. AQI mainly focuses on the air quality in quantitative measures and also tells the types of health issues due to different pollution level. The air quality standard was developed by EPA and the clean air act indicates that there are four major air pollutants such as , SPM, SO₂, RSPM and NO₂.

1.2 Existing System

Nowadays, the system function shows that the how much gases are releasing to the air can only be calculated , but it doesnot show it will cause our health or not.

Disadvantages:

1. Doesn't generate accurate and efficient results.
2. Computation time is very high.
3. Lacking of accuracy may result in lack of efficient further treatment.

1.3 Proposed System

By using this system we can observe that if the releasing of gases is in limited from at then it will show that good environment or moderate etc , and can be calculated by hourly based.

Advantages:

1. Generates accurate and efficient results.
2. Computation time is greatly reduced.
3. Reduces manual work.
4. Efficient further treatment.

1.4. System Requirements

1.4.1 Hardware Requirements:

- System type : intel®core™i7-7500UCPU@2.70gh
- Cache memory : 4 MB
- RAM : 12 GB
- Hard Disc : 8 GB

1.4.2 Software Requirements:

- Operating system : windows 10, 64 bit OS
- Coding language : Python
- Python distribution : Anaconda, Flask

2. LITERATURE SURVEY

2.1 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

A popular saying goes that we are living in an “information age”. Terabytes of data are produced every day. Data mining is the process that transforms information into a set of data. The healthcare system produces massive quantities of data every day. Much of it isn't used successfully though. Animesh Hazra has suggested that any of the latest work on Air quality index prediction using data mining techniques analyzes the various combinations of mining algorithms used and conclude which techniques are successful and efficient.

Air pollution has become the subject of many current environmental studies , and the quality of air is directly related to the quality of life and health of human beings. In this paper , the Random forest model is used to calculate the Air quality. Four pollutants so₂,no₂,spm,rspm are used as the evaluation factors of the model, and AQI value is the output of the model, and then the model is established. Finally the model is used to predict the Air quality and compare with the actual value. The results show that the accuracy of air quality prediction is over 80%, and the predicted value is close to the actual value.

2.2 Some machine learning methods

Machine learning algorithms are often categorized as supervised and unsupervised.

- **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
- **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
- **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best. This is known as the reinforcement signal.

2.3 Applications of machine learning

1. Virtual Personal Assistants
2. Predictions while Commuting
3. Videos Surveillance
4. Social Media Services
5. Email Spam and Malware Filtering

3. SYSTEM ANALYSIS

3.1 sns.pairplot(data=df)

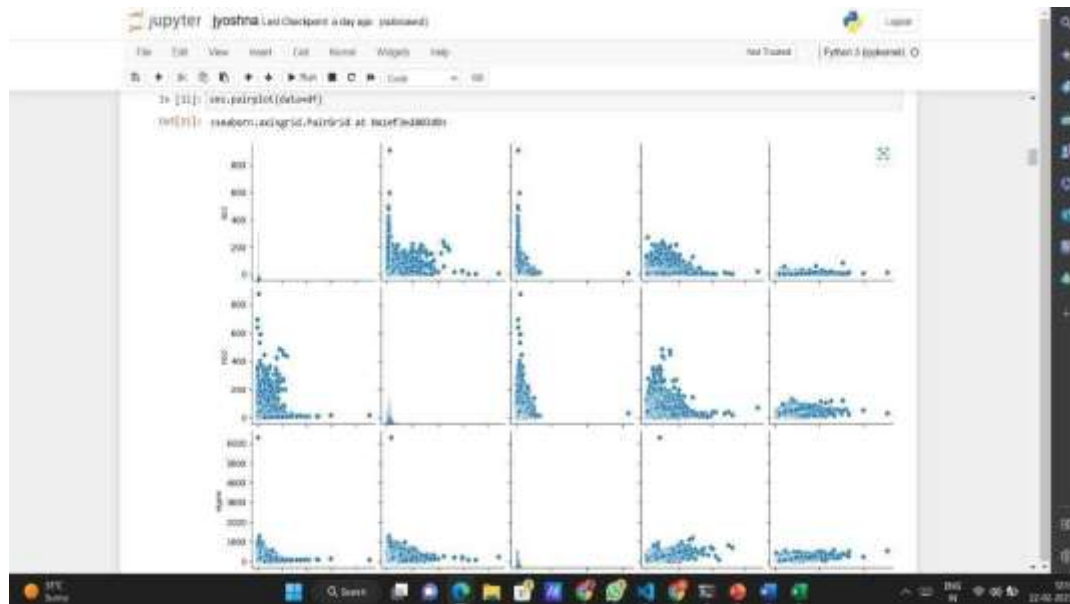


Fig:3.1 pairplots of gases

3.2 Types of gases in Air Quality Index:

1. Sulfur Dioxide (SO_2)
2. Nitrogen Dioxide (NO_2)
3. Suspended particulate Matter (SPM)
4. Respirable Suspended Particulate Matter (RSPM)

3.2.1. Sulfur Dioxide (So2):

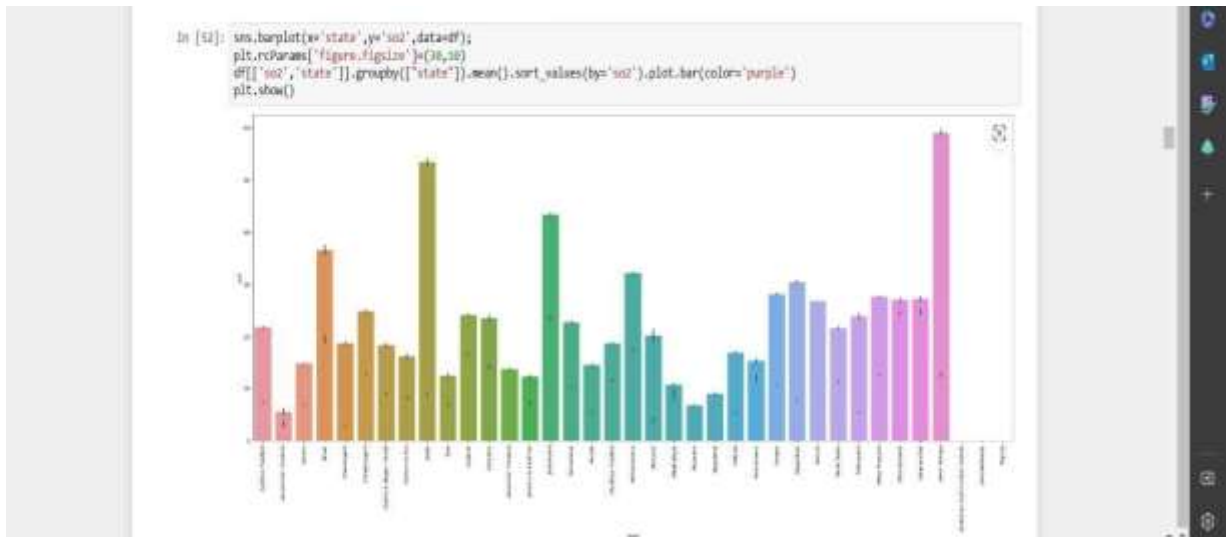


Fig:3.2.1 Sulfur Dioxide

3.2.2. Nitrogen Dioxide (No2):

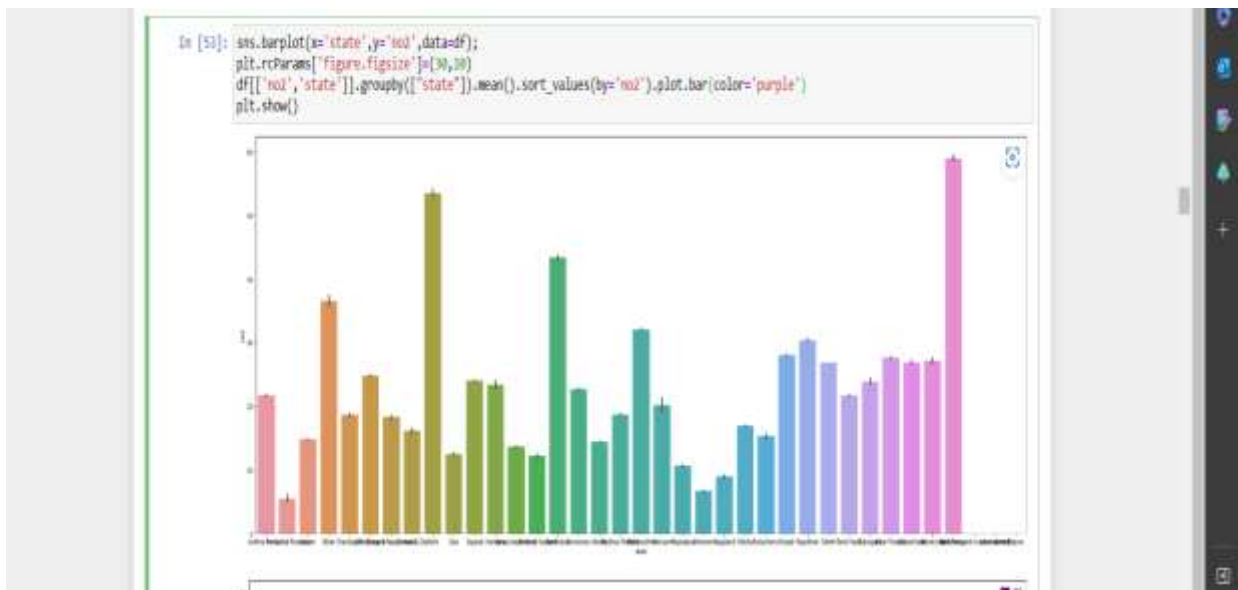


Fig:3.2.2 Nitrogen Dioxide

3.2.3. Suspended Particulate Matter (Spm):

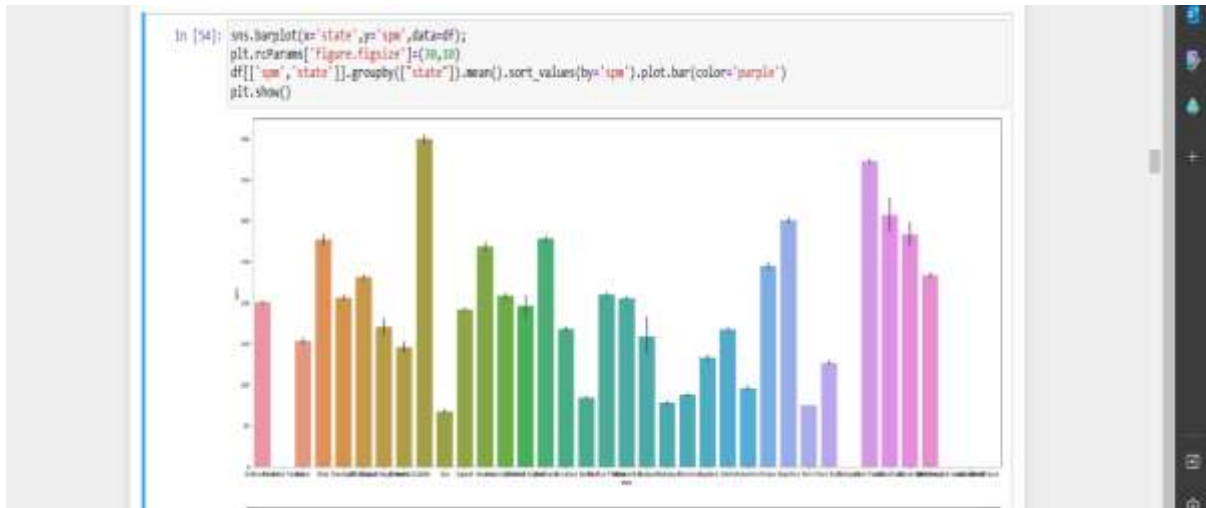


Fig:3.2.3 Suspended Particulate Matter

3.2.4. Respirable Suspended Particulate Matter (Rspm):

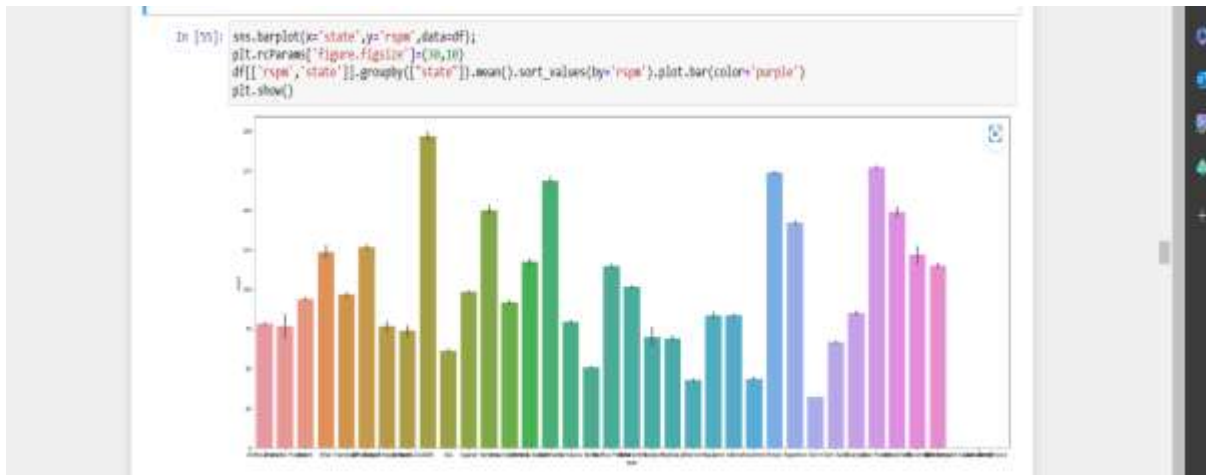


Fig:3.2.4 Respirable Suspended Particulate Matter

3.3 Importance of machine learning in Air Quality Index

The importance of machine learning in Air Quality Index is increasing because of its ability to process huge datasets efficiently beyond the range of human capability, and then dependably convert analysis of that data into clinical insights that assist people in planning and providing care, which ultimately leads to better outcomes and know about the environment .Using these types of advanced analytics, we can provide better information to society at the point of people care.

3.4 Implementation of machine learning using Python

Python is a popular programming language. It was created in 1991 by Guido van Rossum.

It is used for:

- 1.web development (server-side),
- 2.software development,
- 3.mathematics,
- 4.system scripting.

The most recent major version of Python is Python 3. However, Python 2, although not being updated with anything other than securityupdates, is still quite popular.

It is possible to write Python in an Integrated Development Environment, such as Thonny, Pycharm, Netbeans or Eclipse, Anaconda which are particularly useful when managing larger collections of Python files.

Python was designed for its readability. Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.

Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

In the older days, people used to perform Machine Learning tasks manually by coding all the algorithms and mathematical and statistical formula. This made the process time consuming, tedious and inefficient. But in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks, and modules. Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reason is its vast collection of libraries. Python libraries that used in Machine Learning are:

- 1.Numpy
- 2.Scipy
- 3.Scikit-learn
- 4.Pandas
- 5.Matplotlib

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is veryuseful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.

SciPy is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics. There is a difference between the SciPy library and the SciPy stack. The SciPy is one of the core packages that make up the SciPy stack. SciPy is also veryuseful for image manipulation.

Skikit-learn is one of the most popular Machine Learning libraries for classical Machine Learning algorithms. It is built on top of two basic Python libraries, NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with Machine Learning.

Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.

Matplotlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, histogram, error charts, bar chats, etc.

3.5 Scope of the project

The scope of this system is to maintain gases details in datasets, train the model using the large quantity of data present in datasets and predict whether air is good or etc on new data during testing.

4.METHODOLOGY

4.1 DataSet

stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_name	pm2_5
150	February	Andhra Pradesh	Hyderabad	NA	Residential	4.8	17.4	NA	NA	NA	NA
151	February	Andhra Pradesh	Hyderabad	NA	Industrial	3.1	7	NA	NA	NA	NA
152	February	Andhra Pradesh	Hyderabad	NA	Residential	6.2	28.5	NA	NA	NA	NA
150	March	Andhra Pradesh	Hyderabad	NA	Residential	6.3	14.7	NA	NA	NA	NA
151	March	Andhra Pradesh	Hyderabad	NA	Industrial	4.7	7.5	NA	NA	NA	NA
152	March	Andhra Pradesh	Hyderabad	NA	Residential	6.4	25.7	NA	NA	NA	NA
150	April	Andhra Pradesh	Hyderabad	NA	Residential	5.4	17.1	NA	NA	NA	NA
151	April	Andhra Pradesh	Hyderabad	NA	Industrial	4.7	8.7	NA	NA	NA	NA
152	April	Andhra Pradesh	Hyderabad	NA	Residential	4.2	23	NA	NA	NA	NA
151	May	Andhra Pradesh	Hyderabad	NA	Industrial	4	8.9	NA	NA	NA	NA
152	May	Andhra Pradesh	Hyderabad	NA	Residential	3.6	18.6	NA	NA	NA	NA
150	June	Andhra Pradesh	Hyderabad	NA	Residential	3.9	14.1	NA	133	NA	NA
151	June	Andhra Pradesh	Hyderabad	NA	Industrial	5.6	11.8	NA	82	NA	NA
152	June	Andhra Pradesh	Hyderabad	NA	Residential	3.3	19.3	NA	111	NA	NA
150	July	Andhra Pradesh	Hyderabad	NA	Residential	3.9	8.2	NA	118	NA	NA
152	July	Andhra Pradesh	Hyderabad	NA	Residential	3.5	12.1	NA	135	NA	NA
151	July	Andhra Pradesh	Hyderabad	NA	Industrial	7.9	10.2	NA	80	NA	NA
150	August	Andhra Pradesh	Hyderabad	NA	Residential	4	9.9	NA	179	NA	NA
151	August	Andhra Pradesh	Hyderabad	NA	Industrial	12.4	11.5	NA	58	NA	NA
152	August	Andhra Pradesh	Hyderabad	NA	Residential	4	12.3	NA	99	NA	NA
150	September	Andhra Pradesh	Hyderabad	NA	Residential	6.3	11.5	NA	270	NA	NA
151	September	Andhra Pradesh	Hyderabad	NA	Industrial	44.8	13.7	NA	97	NA	NA
152	September	Andhra Pradesh	Hyderabad	NA	Residential	8.1	17.8	NA	167	NA	NA
150	October	Andhra Pradesh	Hyderabad	NA	Residential	7.7	11.3	NA	145	NA	NA
151	October	Andhra Pradesh	Hyderabad	NA	Industrial	20.6	13.6	NA	75	NA	NA
152	October	Andhra Pradesh	Hyderabad	NA	Residential	20.4	27.5	NA	212	NA	NA
150	November	Andhra Pradesh	Hyderabad	NA	Residential	13.9	7.2	NA	93	NA	NA

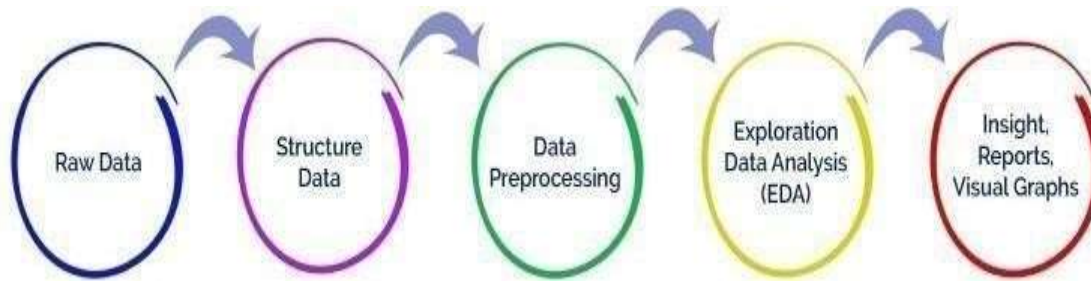
Fig 4.1 Dataset

4.2 Data Pre-processing

Before feeding data to an algorithm we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data.

Data Pre-processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.



Need of Data Preprocessing: For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format. For example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.

4.2.1 Missing values

Filling missing values is one of the pre-processing techniques. The missing values in the dataset is represented as ‘?’ but it a non-standard missing value and it has to be converted into a standard missing value NaN. So that pandas can detect the missing values. The Fig:

```
In [39]: df
Out[39]:
```

	em_code	sampling_date	state	location	agency	type	so2	no2	spen	open	location_monitoring_station	pm2_5	date
0	160-0	February - 0021990	Andhra Pradesh	Hyderabad	NaH	Residential, Rural and other Areas	4.8	17.4	NaH	NaH	NaH	NaH	01-02-1990
1	161-0	February - 0021990	Andhra Pradesh	Hyderabad	NaH	Industrial Area	5.1	7.0	NaH	NaH	NaH	NaH	01-02-1990
2	162-0	February - 0021990	Andhra Pradesh	Hyderabad	NaH	Residential, Rural and other Areas	6.3	20.5	NaH	NaH	NaH	NaH	01-02-1990
3	160-0	March - 0031990	Andhra Pradesh	Hyderabad	NaH	Residential, Rural and other Areas	6.5	14.7	NaH	NaH	NaH	NaH	01-03-1990
4	151-0	March - 0031990	Andhra Pradesh	Hyderabad	NaH	Industrial Area	4.7	7.5	NaH	NaH	NaH	NaH	01-03-1990
...													
435737	5AMH	24-12-2018	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RPHJC	22.0	50.0	143.0	NaH	NaH	NaH	24-12-2018
435738	6AMH	25-12-2018	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RPHJC	20.0	40.0	175.0	NaH	NaH	NaH	25-12-2018
435739	NaH	NaH	Andaman and Nicobar Islands	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH
435740	NaH	NaH	Lakshadweep	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH
435741	NaH	NaH	Tripura	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH	NaH

Fig:4.2.1.1 Before Missing data visualization

air_quality	sm_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
0	100.0	February - MO21990	Andhra Pradesh	Hyderabad	0	Residential, Rural and other Areas	4.8	17.4	0.0	0.0	0	0.0	01-02-1990
1	101.0	February - MO21990	Andhra Pradesh	Hyderabad	0	Industrial Area	3.1	7.0	0.0	0.0	0	0.0	01-02-1990
2	102.0	February - MO21990	Andhra Pradesh	Hyderabad	0	Residential, Rural and other Areas	0.3	30.8	0.0	0.0	0	0.0	01-02-1990
3	100.0	March - MO31990	Andhra Pradesh	Hyderabad	0	Residential, Rural and other Areas	0.3	14.7	0.0	0.0	0	0.0	01-03-1990
4	101.0	March - MO31990	Andhra Pradesh	Hyderabad	0	Industrial Area	4.7	7.8	0.0	0.0	0	0.0	01-03-1990
436737	SAAMP	24-12-2015	West Bengal	ULJBEHUA	West Bengal State Pollution Control Board	RAWES	20.0	50.0	143.0	0.0	Hazrat Rampal Industrial, ULJBEHUA	0.0	24-12-2015
436738	SAAMP	29-12-2015	West Bengal	ULJBEHUA	West Bengal State Pollution Control Board	RAWES	20.0	40.0	171.0	0.0	Hazrat Rampal Industrial, ULJBEHUA	0.0	29-12-2015
436739	0	0	andaman and nicobar islands	0	0	0	0.0	0.0	0.0	0.0	0	0.0	0
436740	0	0	Lakshadweep	0	0	0	0.0	0.0	0.0	0.0	0	0.0	0
436741	0	0	Tripura	0	0	0	0.0	0.0	0.0	0.0	0	0.0	0

Fig:4.2.1.2. After filling Missing data visualization

4.2.2 Correlation coefficient method

We can find dependency between two attributes p and q using Correlation coefficient method using the formula. $r_{p,q} = \frac{\sum (p_i - \bar{p})(q_i - \bar{q})}{n\sigma_p\sigma_q} = \frac{\sum (p_i q_i) - n\bar{p}\bar{q}}{n\sigma_p\sigma_q}$ n is the total number of patterns, p_i and q_i are respective values of p and q attributes in patterns i, \bar{p} and \bar{q} are respective mean values of p and q attributes, σ_p , σ_q are respective standard deviations values of p and q attributes. Generally, $-1 \leq r_{p,q} \leq +1$. If $r_{p,q} < 0$, then p and q are negatively correlated. If $r_{p,q} = 0$, then p and q are independent attributes and there is no correlation between them. If $r_{p,q} > 0$, then p and q are positively correlated. We can drop the attributes that are having correlation coefficient value as 0 as it indicates that the variables are independent with respect to the prediction attribute.

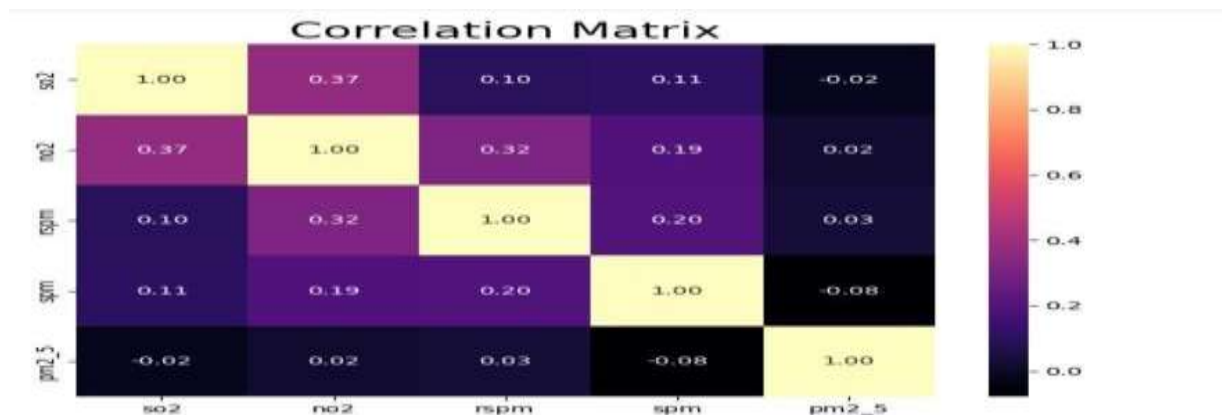


Fig:4.2.2 Correlation

4.3 Machine learning algorithms for classification

Data was gathered from Kaggle, one of the most providers of data sources for the purpose of learning, and hence the data is collected from the Kaggle, which had two sets of details, one of which was for the preparation and the supplementary tests. The dataset for training is the model in which datasets are further divided into datasets was used to train the model train and the minor dataset. For the measuring of the value of attrition, many regression models are applied during this study. The dataset is split into 2 sections. One half for model training and also the other part for model analysis or testing. During this study.

1. Decision Tree:

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.

2. RandomForest:

The random forest algorithm improves the flexibility and decision-making capacity of individual trees. It is another machine learning algorithm incorporating the ensemble learning theorem as its foundation, combining results from various decision trees to optimize training. In some use cases of loan and credit risk prediction, some features are more important than the rest or, more specifically, some features whose removal would improve the overall performance. Since we know the fundamentals of decision trees and how they choose features based on information gain, random forests would incorporate these benefits to give superior performance.

3.Linear Regression:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

5. IMPLEMENTATION CODE

5.1 BACK END

air.py

```
import pandas as pd
df=pd.read_csv('air.csv.csv',encoding=unicode_escape')
print(df)
df.head()
df.info()
df.columns
df.shape
df.describe
df.isnull()
df.isnull().sum()
df.fillna(0,inplace=True)
df.isnull().sum()
k = sns.heatmap(df.isnull(), cbar=False)
sns.heatmap(df.corr())
%matplotlib inline
f,ax=plt.subplots()
f.set_size_inches(8,6)
sns.heatmap(data.corr(),annot=True,fmt=".2f",cmap="magma")
ax.set_title("Correlation Matrix", fontsize=20)
plt.show()
fromsklearn.model_selection import train_test_split

X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2,random_state=70)
print(X_train.shape,X_test.shape,Y_train.shape,Y_test.shape)
```



```

import numpy as np
import seaborn as sns
import pandas as pd
import pickle
import matplotlib.pyplot as plt
#Using DecisionTreeClassifier of tree class to use Decision Tree Algorithm

from sklearn.tree import DecisionTreeClassifier
DT=DecisionTreeRegressor()
DT.fit(X_train,Y_train)
train_preds=DT.predict(X_train)
test_preds=DT.predict(X_test)
RMSE_train=(np.sqrt(metrics.mean_squared_error(Y_train,train_preds)))
RMSE_test=(np.sqrt(metrics.mean_squared_error(Y_test,test_preds)))

#Using Linear Regression class to use Linear Regression Algorithm

from sklearn.linear_model import LinearRegression
model=LinearRegression()
model.fit(X_train,Y_train)
train_pred=model.predict(X_train)
test_pred=model.predict(X_test)
RMSE_train=(np.sqrt(metrics.mean_squared_error(Y_train,train_pred)))
RMSE_test=(np.sqrt(metrics.mean_squared_error(Y_test,test_pred)))

```

```

#Using RandomForest class to use Random Forest Algorithm
from sklearn.ensemble import RandomForestRegressor
RF=RandomForestRegressor().fit(X_train, Y_train)
train_preds1=RF.predict(X_train)
test_preds1=RF.predict(X_test)
RMSE_train=(np.sqrt(metrics.mean_squared_error(Y_train,train_preds1)))
RMSE_test=(np.sqrt(metrics.mean_squared_error(Y_test,test_preds1)))
#Fitting model with trainig data
regressor.fit(X_train,y_train)
# Saving model to disk
import pickle
filename='savedmodel.sav'
pickle.dump(RF,open(filename,'wb'))
# Loading model to compare the results
load_model=pickle.load(open(filename,'rb'))

```

5.2 FRONTEND

index.html

```

<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <!-- <link rel="stylesheet" href="result.css">-->
  <style>
    body{ background:url('https://thumbs.dreamstime.com/z/air-quality-index-concept-
knob-button-air-quality-index-concept-knob-button-changing-speed-g-to-g-reverse-
217530469.jpg') no-repeat center center/cover;
  }

```

```

.form{
  display:flex;
  flex-direction: column;
  height: 700px;
  width: 500px;
  border: 1px solid black;
  align-items: center;
  margin: auto;
  margin-top: 50px;
  background-color: rgba(0,0,0,0.5);
  box-shadow: inset -5px -5px rgba(0,0,0,0.5);
  border-radius: 25px;
}
.form h1{
  color: white;
  font-size: 2rem;
  border-bottom: 4px solid
  rgba(255,255,255,0.5);
  margin: 50px;
}
.form p{
  color:white;
  font-size:2rem;
}
.form br{
  color:white;
  font-size:2rem;
}

```

```

.box{
  padding: 12px;
  margin: 20px;
  width: 65%;
  border: none;
  outline: none;
  border-radius: 20px;
  background-color: rgba(0,0,0,0.5);
  box-shadow: inset -3px -3px
  rgba(0,0,0,0.5);
  color: white;
  font-size: 1rem;
}
#submit{
  padding: 10px 20px;
  margin-top: 50px;
  width: 50%;
  background-color: green;
  box-shadow: inset -3px -3px green;
  color: white;
  border: none;
  outline: none;
  border-radius: 20px;
  font-size: 1rem;
}
#submit:hover{
  cursor:pointer;
  background-color:aqua;
  color:aqua;
}

```

```

</style>
</head>
<body>
    <form action='/predict' method="post"
class="form">
        <h1> Air Quality Index Prediction</h1>
        <input type="number" name="so2" class="box" step="0.1" placeholder="so2">
        <input type="number" name="no2" class="box" step="0.1" placeholder="no2">
        <input type="number" name="rspm" class="box" step="0.1" placeholder="rspm">
        <input type="number" name="spm" class="box" step="0.1" placeholder="spm">
        <button type="submit" id="submit">Predict</button>
        <p align="center",color='green'>{{ a }}</p>
    </form>
</body>
</html>

```

5.3 CONNECTION

App.py

```

import pandas as pd
from flask import Flask,render_template
from flask import request
import pickle
import numpy as np
filename='savedmodel.sav'
classifier=pickle.load(open(filename,'rb'))
# df=pd.read_csv("final_csv.csv")
app=Flask(__name__)

```

```

@app.route('/')
def home():
    return render_template('index.html')
@app.route('/predict',methods=['POST'])
def predict():
    if request.method=='POST':
        so2=request.form['so2']
        no2=request.form['no2']
        spm=request.form['spm']
        rspm=request.form['rspm']
        data=[[so2,no2,spm,rspm]]
        print(data)
        x= classifier.predict(data)[0]
        if x <= 50:
            x= "Good"
        elif x > 50 and x <= 100:
            x= "Moderate"
        elif x > 100 and x <= 200:
            x= "Poor"
        elif x > 200 and x <= 300:
            x= "Unhealthy"
        elif x > 300 and x <= 400:
            x= "Very unhealthy"
        elif x > 400:
            x= "Hazardous"
        print('Data Belongs to Cluster', x)

        return render_template('index.html',a=x)
if __name__=='_main_':
    app.run(debug=True)

```

5.4 Result Analysis

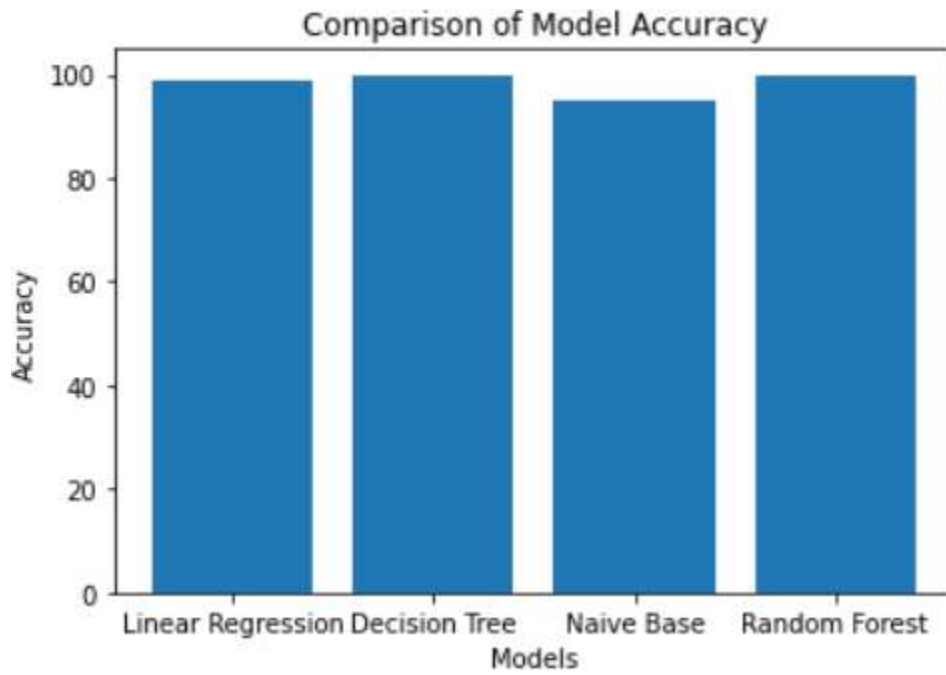


Fig :5.4 Comparison of models

Algorithms	Accuracy
Linear Regression	98.30
Decision Tree Classifier	99.95
Naïve base Classifier	95.06
Random Forest	99.96

6. OUTPUT SCREENS



Fig.6.1 web page design



Fig .6.2 data entered



Fig .6.3 Output is Good



Fig .6.4 Output is Moderate

7. CONCLUSION AND FUTURE SCOPE

7.1 CONCLUSION

We have used 4 algorithms like Linear Regression, Decision Trees, Naive bayes, Random Forest in- order to predict the range of air quality index. The accuracy varies for different algorithms. The accuracy for Linear Regression algorithm is 98.30. The accuracy of Decision Tree algorithm is 99.95 . The accuracy of Naive Bayes algorithm is 95.06. The accuracy of Random Forest algorithm is 99.96. The highest accuracy obtained is Random Forest so the predicting the range by using this algorithm so that we can get the approximate correct result, and also it is easy to know about the environment situation and condition and can be protected.

7.2 FUTURE SCOPE

To develop more accuracy using machine learning algorithms and advanced techniques . The work can be extended and improved that the measuring the Air Quality Index can decrease the pollution in the air so that the health issues can be reduced.

8. Bibliography

- [1] Anikender Kumar, Pramila Goyal, “Forecasting of air quality in Delhi using principal component regression technique”, Atmospheric Pollution Research, 2 (2011) 436-444.
- [2]. <https://www.aqi.in/blog/aqi/>
- [3]. https://www.researchgate.net/profile/Shovan_Sahu/publication/315725810/figure/tbl1/AS:668795018440728@1536464566616/Breakpoints-of-different-pollutants-in-IND-AQI-CPCB-2014.png
- [4]. https://app.cpcbcr.com/ccr_docs/FINALREPORT_AQI.pdf
- [5]. Huixiang Liu, Qing Li, Dongbing Yu, Yu Gu, “Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms”, Applied Sciences, ISSN 2076-3417; CODEN: ASPCC7, 2019, 9, 4069; doi:10.3390/app9194069.
- [6]. Pooja Bhalgat, Sejal Pitale, Sachin Bhoite, “Air Quality Prediction using Machine Learning Algorithms”, International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367- 370, 2019, ISSN:-2319–8656.
- [7]. <https://www.lung.org/clean-air/outdoors/air-quality-index>
- [8]. Ziyue Guan and Richard O. Sinnott, “Prediction of Air Pollution through Machine Learning on the cloud”, IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), 978-1-5386-5502-3/18/\$31.00 ©2018 IEEE DOI 10.1109/BDCAT.2018.00015.
- [9]. Heidar Malek, Armin Sorooshian, Gholamreza Goudarzi, Zeynab Baboli, Yaser Tahmasebi Birgani, Mojtaba Rahmati, “Air pollution prediction by using an artificial neural network model”, Clean Technologies and Environmental Policy, (2019) 21:1341–1352.
- [10]. Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, “Detection and Prediction of Air Pollution using Machine Learning Models”, International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018
- [11]. <https://www.iqair.com/us/india>

Air Quality Index Analysis using Machine Learning

Maddula Venkata Jyoshna
Student

*Department of Computer Science and
Engineering
Narasaraopet, India
jyoshnamaddula7@gmail.com*

Chandana Chala
Student

*Department of Computer Science and
Engineering
Narasaraopet, India
chandanaachandu2043@gmail.com*

Vankayala Sri Mallika
Student

*Department of Computer Science and
Engineering
Narasaraopet, India
vsrimallika2000@gmail.com*

M.Venkata Rao

Asst.Professor,M.Tech

*Department of Computer Science and Engineering
Narasaraopet, India
venkatarao61@gmail.com*

Abstract - Air pollution is a major issue in today's world, caused by the release of hazardous gases into the atmosphere from industries, vehicles, and other sources. To maintain good air quality, this mechanism measures various air toxins in different areas. However, the pollution level in all cities has exceeded the air quality index value set by the government, which significantly affects human health. Thankfully, machine learning (ML) research has advanced to the point where it is now able to forecast contaminants using historical data. This study describes a gadget that can measure current pollution levels and run an ML-based algorithm for estimating future pollution levels using historical pollution data.

Keywords—Machine Learning, Random Forest, Gaussain Naïve Bayes, Decision tree, Linear Regression.

I. INTRODUCTION

Air pollution monitoring is crucial in today's world, as it has a significant impact on both human health and the environment. Harmful emissions not only affect the environment, but also impact the productivity and efficiency of individuals. Thus, continuous monitoring is necessary to effectively control and mitigate air pollution. It is impossible to overlook the effects of climate change on human health., as it has been linked to numerous adverse health effects. Furthermore, air pollution can also negatively impact the environment and its delicate balance. Thus, effective monitoring of air pollution levels is essential.

Given the dangers of air pollution, it is important to monitor and control its levels to minimize its harmful effects. This can be achieved through continuous monitoring and effective measures to mitigate pollution levels. By implementing these measures, we can work towards maintaining a healthier environment and promoting human well-being. It's critical to pinpoint the cause, extent, and origin of air pollution in order to control it. The state government's environmental department typically observes pollution levels by tracking the concentration of toxic gases in various regions. The World Health Organization (WHO) also provides data on pollution levels in the country, which highlights the urgent need for air monitoring.

Monitoring air pollution has become increasingly critical due to the rising levels of pollution. Air tracking has grown to be a significant task for measuring continuous levels of air contaminants in the environment. It is essential to monitor air pollution levels regularly to take appropriate measures and control its impact on human health and the environment.

II. LITERATURE SURVEY

The public is informed of the degree of polluted air in such a specific location using the index of air quality (AQI). The AQI has been widely used by governments and organizations worldwide as a tool for air quality management and public health protection. In recent years, many researchers have

focused on developing and improving the AQI system to make it more accurate and useful for the general public.

One study by Kaur and Bhangra (2021) reviewed the literature on AQI and its application in India. The authors found that AQI is an important tool for air quality management in India, which is facing severe air pollution problems. The study also highlighted the need for public awareness and education regarding the AQI system to improve its effectiveness.

Another study by Zhang et al. (2020) proposed a new AQI system based on machine learning algorithms. The authors used a combination of multiple linear regression and random forest algorithms to predict AQI values based on meteorological and air pollutant data.

The proposed AQI system was found to be effective in communicating air pollution levels to the public and in guiding air quality management efforts. Overall, the literature suggests that AQI is an important tool for air quality management and public health protection. Improvements in AQI systems, such as those proposed in recent studies, can help to make the AQI more accurate and effective in communicating air pollution levels to the public. It has been discovered that linear regression models are suitable for estimating or predicting pollution. While SVM-based and neural network approaches are chosen for forecasting pollution levels.

III. EXISTING SYSTEM

A gap in the literature was identified as previous papers only focused on predicting PM2.5 levels using ML algorithms. However, this project aims to predict levels of all pollutants including CO, O3, NO2, SO2, PM2.5, and PM10 by utilizing meteorological data for improved prediction accuracy. These systems typically assign an AQI value on a scale from 0 to 500, where lower air quality is indicated by greater readings.

IV. METHODOLOGY

In extremely polluted locations, the central environmental control board has built a number of pollution monitoring stations. This monitoring station data are used for research and further evaluation.

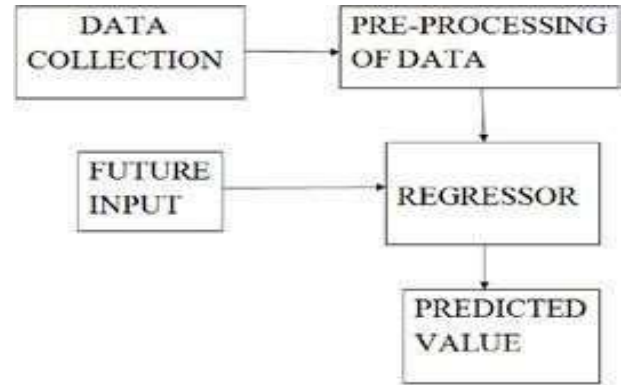


Fig. 1: Proposed system work flow

Data Collection

The Data set is collected from the Kaggle website: <http://www.kaggle.com/code/sharmamanali/air-quality>

The dataset contains data related to the air gases values to predict the air quality. The dataset size is 4,80,534 rows and 11 columns namely so2,no2,spm,rspm.

city	month	location	agency	type	so2	no2	spm	rspm	location	city
550	February	Amritsar	Prakash	Residential	4.8	11.0	NA	NA	Amritsar	NA
551	February	Amritsar	Prakash	Industrial	3.1	7.0	NA	NA	Amritsar	NA
552	February	Amritsar	Prakash	Industrial	6.3	28.3	NA	NA	Amritsar	NA
553	March	Amritsar	Prakash	Industrial	6.3	14.7	NA	NA	Amritsar	NA
554	March	Amritsar	Prakash	Industrial	4.7	7.5	NA	NA	Amritsar	NA
555	March	Amritsar	Prakash	Industrial	6.4	23.7	NA	NA	Amritsar	NA
556	April	Amritsar	Prakash	Industrial	5.4	17.1	NA	NA	Amritsar	NA
557	April	Amritsar	Prakash	Industrial	6.7	8.7	NA	NA	Amritsar	NA
558	April	Amritsar	Prakash	Industrial	4.2	8.0	NA	NA	Amritsar	NA
559	May	Amritsar	Prakash	Industrial	4	8.0	NA	NA	Amritsar	NA
560	May	Amritsar	Prakash	Industrial	3.6	10.0	NA	NA	Amritsar	NA
561	June	Amritsar	Prakash	Industrial	3.9	14.1	NA	NA	Amritsar	NA
562	June	Amritsar	Prakash	Industrial	10.3	22.0	NA	NA	Amritsar	NA
563	June	Amritsar	Prakash	Industrial	8.4	10.3	NA	NA	Amritsar	NA
564	July	Amritsar	Prakash	Industrial	5.9	8.2	NA	NA	Amritsar	NA
565	July	Amritsar	Prakash	Industrial	5.5	12.1	NA	NA	Amritsar	NA
566	July	Amritsar	Prakash	Industrial	7.9	10.3	NA	NA	Amritsar	NA
567	August	Amritsar	Prakash	Industrial	4	8.0	NA	NA	Amritsar	NA
568	August	Amritsar	Prakash	Industrial	12.4	12.2	NA	NA	Amritsar	NA
569	August	Amritsar	Prakash	Industrial	4	12.2	NA	NA	Amritsar	NA
570	September	Amritsar	Prakash	Industrial	6.3	13.3	NA	NA	Amritsar	NA
571	September	Amritsar	Prakash	Industrial	44.8	13.7	NA	NA	Amritsar	NA
572	September	Amritsar	Prakash	Industrial	6.1	17.0	NA	NA	Amritsar	NA
573	October	Amritsar	Prakash	Industrial	7.3	11.3	NA	NA	Amritsar	NA
574	October	Amritsar	Prakash	Industrial	20.6	13.0	NA	NA	Amritsar	NA
575	October	Amritsar	Prakash	Industrial	20.4	17.0	NA	NA	Amritsar	NA
576	November	Amritsar	Prakash	Industrial	10.0	7.0	NA	NA	Amritsar	NA

Fig. 2: Dataset diagram

A. Data Cleaning and Feature Extraction

Before giving the data to any machine learning algorithms the data must be cleaned. In data cleaning process null values and outliers are removed. In the collected dataset no null values are present. After data cleaning ,features are extracted from the data set. In dataset we have 6 different features which affects our model output and we have to check the correlation among the features.

	str_code	so2	no2	rpm	spm	pm2_5	soi	nei	hpi	spm6	aqi
str_code	1.000000	-0.074017	-0.250173	-0.065792	-0.070391	NaN	-0.073918	-0.252804	NaN	-0.095725	-0.678123
so2	-0.074017	1.000000	0.429088	-0.007817	0.170888	NaN	0.992479	0.430772	NaN	0.165794	0.192352
no2	-0.250173	0.429088	1.000000	0.187989	0.391681	NaN	0.441215	-0.060327	NaN	0.349513	0.412102
rpm	-0.065792	-0.007817	0.187989	1.000000	0.235294	NaN	-0.006148	0.189495	NaN	0.214176	0.250063
spm	-0.070391	0.170888	0.391681	0.235294	1.000000	NaN	0.189506	0.350006	NaN	0.902247	0.568764
pm2_5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
soi	-0.073918	0.992479	0.441215	-0.006148	0.189506	NaN	1.000000	0.443056	NaN	0.170238	0.190361
nei	-0.252804	0.430772	-0.060327	0.189495	0.355006	NaN	0.443056	1.000000	NaN	0.352811	0.415356
hpi	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
spm6	-0.095725	0.165794	0.349513	0.214176	0.902247	NaN	0.170238	0.352811	NaN	1.000000	0.963589
aqi	-0.678123	0.192352	0.412102	0.250063	0.568764	NaN	0.190361	0.415356	NaN	0.963589	1.000000

Fig. 3:Correlation

Correlation helpful to remove the features which shows negligible affect on the model output. Fig 3 shows the correlation between the features of the dataset.

B. Model Architecture

In this step dataset is splitted in 75% and 25% for training and testing. During training, machine learning algorithms finds the relation between the input and output features. By using this relation the model able to predict the outputs to the new input values. In this we used the following four machine learning models:

1. Random Forest: A machine learning approach called random forest is employed for both classification and regression applications. It is a member of the family of evolutionary algorithms, which combine a number of weak learners to produce a powerful learner. In a random forest, multiple decision trees are built on randomly selected subsets of the data and features. Each tree is expanded to its maximum depth throughout the training process, and the algorithm chooses the optimum split from a random group of features at each node.

2. Decision Tree: Decision trees are highly interpretable, easy to understand, and can handle both categorical and numerical data. They can also handle high-dimensional data and are highly scalable. The risk of overfitting with decision trees increases with tree depth and noisy data, respectively. Techniques such as pruning, regularization, and ensemble methods such as

random forests can help to overcome these limitations.

3. Linear Regression: A statistical technique used to evaluate the connection between two parameters is linear regression, where one variable is dependent on the other. It involves finding a linear equation that best describes the relationship between the variables. The premise underlying the linear regression technique is that there exists a linear connection among the variables.

V. RESULT AND ANALYSIS

After created all the four models by using the training data , we test the models by using the test data and calculates the accuracies of the models. The following table shows the accuracies of different models:

Type of Models	Accuracy
Linear Regression	98.91
Decision Tree	99.95
Naïve Base	95.00
Random Forest	99.96

Fig.4: Accuracy Table

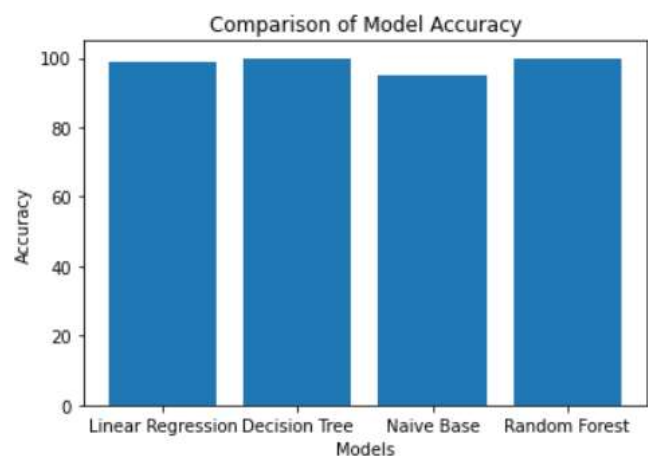


Fig.5: Accuracy Bar Chat

VI. DEPLOYMENT

The final model is deployed into an application by using Flask module of python by creating the user interface which can be accessible easily by the people. The application simply takes the input and provides output.

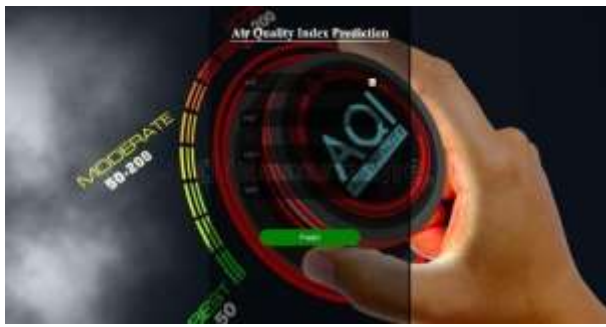


Fig.6 Input Screen



Fig.7 Output Screen

VII. CONCLUSION

By using unique formulas like necessarily imply, Decision Tree, and Random Forest, we forecast the air quality list. We concluded from the results that its Random Forest approach provides a superior expectation of the list of air quality.

This project has highlighted the benefits of Machine Learning in knowing its value of air quality index value, and how it can help improve its efficiency and effectiveness. By automating the recommendation process, farmers can save time and effort, while also reducing the risk of making costly mistakes.

VIII. REFERENCES

- [1] https://en.wikipedia.org/wiki/Air_quality_index
- [2]. Kennedy Okokpuije, 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.
- [3]. 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.
- [4]. 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.
- [5]. A. Masih, "Machine learning algorithms in air quality modeling," Global Journal of Environmental Science and Management, vol. 5, no. 4, pp. 515–534, 2019.
- [6]. <https://archive.ics.uci.edu/ml/datasets/Air+quality>
- [7]. Rokach, Lior, Maimon, O. (2008). Data mining with decision trees : theory and applications. World Scientific Pub Co Inc .ISBN 978-9812771711.
- [8]. Breiman L (2001). "Random Forests". Machine Learning. 45(1):32. doi:10.1023/A:1010933404324

Ritik Sharma, Gaurav Shilimkar, Shivam Pisal, " Air Quality Prediction by Machine Learning", International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN:2395-6011, Online ISSN:2395-602X, Volume 8, Issue 3, pp.486-492, May-June-2021. Available at doi:<https://doi.org/10.32628/IJSRST218396>
Journal URL:<https://ijsrst.com/IJSRST218396>

ORIGINALITY REPORT

6%

SIMILARITY INDEX

3%

INTERNET SOURCES

3%

PUBLICATIONS

1%

STUDENT PAPERS

PRIMARY SOURCES

1	Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth, Hari Kiran Reddy. "Air Quality Prediction Of Data Log By Machine Learning", 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020 Publication	2%
2	"Classification of Student Performance Dataset using Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering, 2019 Publication	1%
3	Submitted to Liverpool John Moores University Student Paper ijrst.com	1%
4	Internet Source	1%
5	ijireeice.com Internet Source	1%
6	neuroquantology.com Internet Source	1%

CERTIFICATE

**NARASARAOPETA**
ENGINEERING COLLEGE
(AUTONOMOUS)

**NBA**
NATIONAL BOARD
ACCREDITATION

**A+**
NAAC
(CYCLE - 2)

**nirf**
NATIONAL INSTITUTE
RANKING FRAMEWORK

**RANKED**
4
PRIVATE
ENGINEERING
COLLEGE
IN A.P.

**25**
YEARS

**COMPUTER SOCIETY OF INDIA**
ESTD 1983

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website: www.nrtec.in

PAPER ID
NECICAIEA2K23061

International Conference on
Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023

Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that **Maddula Venkata Jyoshna** **Narasaraopeta Engineering College**
has presented the paper title **Air Quality Index Analysis using Machine learning** in the
International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-
ICAIEA-2K23], Organized by Department of **Computer Science and Engineering in**
Association with CSI on 17th and 18th March 2023 at **Narasaraopeta Engineering**
College, Narasaraopet, A.P., India.


Convenor
Dr. S.V.N. Srinivasu


Chief-Convenor
Dr. S.N. Tirumala Rao


Principal, Patron
Dr. M. Sreenivasa Kumar





NARASARAOPETA
ENGINEERING COLLEGE
(AUTONOMOUS)



Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website: www.nrtec.in

International Conference on

PAPER ID

NECICAIEA2K23061

Artificial Intelligence and Its Emerging Areas

NEC-ICAIEA-2K23

17th & 18th March, 2023

Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that **Chandana Chala**, **Narasaraopeta Engineering College** has presented the paper title **Air Quality Index Analysis using Machine learning** in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of **Computer Science and Engineering in Association with CSI** on 17th and 18th March 2023 at **Narasaraopeta Engineering College, Narasaraopet, A.P., India.**

Convenor
Dr. S. V. N. Srinivasu

Chief-Convenor
Dr. S. N. Tirumala Rao

Principal, Patron
Dr. M. Sreenivasa Kumar





NARASARAOPETA
ENGINEERING COLLEGE
(AUTONOMOUS)



Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website: www.nrtec.in

International Conference on

PAPER ID
NECICAIEA2K23081

Artificial Intelligence and Its Emerging Areas

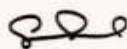
NEC-ICAIEA-2K23

17th & 18th March, 2023

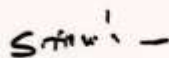
Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that **Vankayala Sri Mallika**, **Narasaraopeta Engineering College** has presented the paper title **Air Quality Index Analysis using Machine learning** in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of **Computer Science and Engineering in Association with CSI** on 17th and 18th March 2023 at **Narasaraopeta Engineering College, Narasaraopet, A.P., India.**



Convenor
Dr. S. V. N. Srinivasu



Chief-Convenor
Dr. S. N. Tirumala Rao



Principal, Patron
Dr. M. Sreenivasa Kumar





NARASARAOPETA
ENGINEERING COLLEGE
(AUTONOMOUS)



Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website: www.nrtec.in

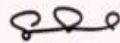
PAPER ID
NECICAIEA2K23061

Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023

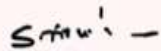
Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that M.Venkata Rao, **Narasaraopeta Engineering College** has presented the paper title **Air Quality Index Analysis using Machine learning** in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of **Computer Science and Engineering** in Association with CSI on 17th and 18th March 2023 at **Narasaraopeta Engineering College, Narasaraopet, A.P., India.**



Convenor
Dr.S.V.N.Srinivasu



Chief-Convenor
Dr.S.N.Tirumala Rao



Principal, Patron
Dr.M.Sreenivasa Kumar

