Air Quality Analysis Using Machine learning

A PROJECT REPORT

Submitted by

M RAGHAVA VARMA [RegNo:RA2011003010887]

K SAI MANIKANTA PITCHAIAH [Reg No:RA2011003010893]

Under the Guidance of

DR. S. PADMINI

Associate Professor, Department of Computing Technologies

In partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTING TECHNOLOGIES

COLLEGE OF ENGINEERING AND

TECHNOLOGY SRM INSTITUTE OF SCIENCE

AND TECHNOLOGY KATTANKULATHUR –

603203

AUG 2023

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that this B.Tech project report titled "**Air Quality Analysis Using Machine learning**" is the bonafide work of M Raghava Varma[Reg. No.: RA2011003010887] and K SAI MANIKANTA [Reg. No.RA2011003010893] who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

DR. S.PADMINI
Associate Professor
Department of Computing Technologies

DR. M.PUSHPALATHA
HEAD OF THE DEPARTMENT
Department of Computing Technologies

SIGNATURE OF
INTERNAL EXAMINER

SIGNATURE OF
EXTERNAL EXAMINER

Department of Computing

Technologies **SRM Institute of**

Science and Technology Own Work

Declaration Form

Degree/ Course **:** B.Tech in Computer Science and Engineering

Student Names **:** Raghava , Manikanta

Registration Number: RA2011003010887, RA2011003010893

Title of Work **:** Air Quality Analysis Using Machine learning

We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism, as listed in the University Website, Regulations, and the Education Committee guidelines.

We confirm that all the work contained in this assessment is our own except where indicated, and that we have met the following conditions:

- Clearly references / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc.)
- Given the sources of all pictures, data etc. that are not my own

- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:
I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.
If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

ACKNOWLEDGEMENT

We express our humble gratitude to **Dr. C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, **Dr. T.V.Gopal**, for his invaluable support.

We wish to thank **Dr. Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We are incredibly grateful to our Head of the Department, **Dr. M. Pushpalatha,** Professor, Department of Computing Technologies,SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our Project Coordinator, , Assistant Professor, Panel Head, **Dr. S.Padmini**, Associate Professor and members,**,** Assistant Professor, **Dr. T. Karthick,** Assistant Professor and**,** Assistant Professor, Department of computing technologies, SRM Institute of Science and Technology, for their inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor, **Dr.P.Murali**, Assistant Professor, Department of Computing Technologies, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to our guide, **Dr. S. Padmini**, Associate Professor, Department of Computing Technologies, SRM Institute of Science and Technology, for providing us with an opportunity to pursue our project under his mentorship. She provided us with the freedom and support to explore the research topics of our interest. Her passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Computing Technologies staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

M Raghava Varma[RA2011003010887]

K SAI MANIKANTA[RA2011003010893]

TABLE OF CONTENTS

ABSTRACT

- Air pollution alludes to the issue of toxins into the air that are harmful to human wellbeing and the entire planet. It can be described as one of the most dangerous threats that the humanity ever faced.
- It causes damage to animals, crops, forests etc. To prevent this problem in transport sectors have to predict air quality from pollutants using machine learning techniques. Subsequently, air quality assessment and prediction has turned into a significant research zone.

- Generally, Air pollution alludes to the issue of toxins into the air that are harmful to human well being and the entire planet. It can be described as one of the most dangerous threats that the humanity ever faced.
- It causes damage to animals, crops, forests etc. To prevent this problem in transport sectors have to predict air quality from pollutants using machine learning techniques. Subsequently, air quality assessment and prediction has turned into a significant research zone.
- The aim is to investigate machine learning based techniques for air quality prediction. The air quality dataset is preprocessed with respect to univariate analysis, bi-variate and multi-variate analysis, missing value treatments, data validation, data cleaning/preparing.
- Then, air quality is predicted using Auto Regression Model. This application can help the meteorological Department in predicting air quality. In future, this work can be optimized by applying Artificial Intelligence techniques.

  - `The survival of mankind cannot be imagined without air. Consistent developments in almost all realms of modern human society affected the health of the air adversely. Daily industrial, transport, and domestic activities are stirring hazardous pollutants in our environment.

- Monitoring and predicting air quality have become essentially important in this era, especially in developing countries like India.
- In contrast to the traditional methods, the prediction technologies based on machine learning techniques are proved to be the most efficient tools to study such modern hazards. The present work investigates six years of air pollution data from 23 Indian cities for air quality analysis and prediction. The dataset is well preprocessed and key features are selected through the correlation analysis.
- An exploratory data analysis is exercised to develop insights into various hidden patterns in the dataset and pollutants directly affecting the air quality index are identified. A significant fall in almost all pollutants is observed in the pandemic year, 2020.
- The data imbalance problem is solved with a resampling technique and five machine learning models are employed to predict air quality. The results of these models are compared with the standard metrics.

LIST OF SYMBOLS AND ABBREVIATIONS

1. AI - Artificial Intelligence
2. SVM - Support Vector Machine
3. RF – Random Forest
4. ML – Machine Learning

CHAPTER 1

INTRODUCTION

- Air quality is a critical environmental factor that profoundly impacts human health, ecosystems, and the overall quality of life. The rise of industrialization, urbanization, and transportation has led to increased air pollution, making it a pressing global concern.
- Poor air quality is associated with various health issues, including respiratory diseases, cardiovascular problems, and even premature mortality. To address these challenges, accurate and timely air quality prediction has become essential.

- Air quality prediction involves forecasting the concentration and distribution of air pollutants over specific geographic areas and time periods.

- This predictive capability is vital for governments, public health agencies, and urban planners to develop effective strategies for pollution control, implement timely interventions, and inform the public about potential health risks.

- In recent years, machine learning has emerged as a powerful tool for improving the accuracy and reliability of air quality prediction.

- By harnessing large volumes of data from diverse sources, machine learning models can capture complex relationships between air pollutants, meteorological variables, geographical factors, and temporal patterns. This multidimensional analysis enables more precise predictions, aiding in pollution reduction efforts and ensuring public safety.

- This comprehensive review explores the multifaceted field of air quality prediction using machine learning techniques. It delves into the methodologies, data sources, algorithms, model evaluation, real-time applications, challenges, and future directions in this critical domain.

- By enhancing our understanding of air quality prediction through machine learning, we aim to promote cleaner air, healthier communities, and a sustainable environment for future generations.

1.1Purpose:

- The purpose of conducting air quality prediction using machine learning is multifaceted and encompasses several important objectives and goals:

- Protecting Public Health: The primary purpose of air quality prediction is to safeguard public health. By accurately forecasting air pollutant levels, especially hazardous pollutants like particulate matter (PM2.5 and PM10) and ozone (O3), authorities can issue timely health advisories and warnings to vulnerable populations, allowing individuals to take precautions and reduce exposure.

- Environmental Conservation: Air quality prediction plays a crucial role in environmental protection and conservation efforts.

- Predictive models help in assessing the impact of pollution on ecosystems, wildlife, and vegetation, enabling the implementation of measures to mitigate harm and preserve biodiversity.

- Urban Planning and Policy Formulation: Air quality forecasts are essential for urban planners and policymakers.

- They provide critical information for designing sustainable cities, optimizing transportation systems, and developing effective pollution control measures. Predictive models support evidence-based decision-making to reduce emissions and improve air quality in urban areas.

1.2Scope:

- The scope of air quality prediction using machine learning is broad and encompasses various dimensions, applications, and areas of impact. Below are the key aspects that define the scope of this field:

- Spatial Coverage: Air quality prediction models can be applied at various spatial scales, from local and urban levels to regional, national, and even global scales.

- The scope of prediction can range from specific neighborhoods and cities to entire regions or countries, depending on the objectives and data availability.

- Pollutants: Air quality prediction models can target a wide range of air pollutants, including but not limited to particulate matter (PM2.5 and PM10), ground-level ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), volatile organic compounds (VOCs), and various hazardous air pollutants (HAPs).

- The choice of pollutants depends on local air quality concerns and regulatory requirements.

- Time Horizons: Air quality prediction can involve short-term forecasting (e.g., hours to days ahead) for real-time decision-making, medium-term prediction (e.g., days to weeks ahead) for urban planning and public health advisories, and long-term projections (e.g., years to decades) for assessing the impact of policy measures and climate change on air quality.

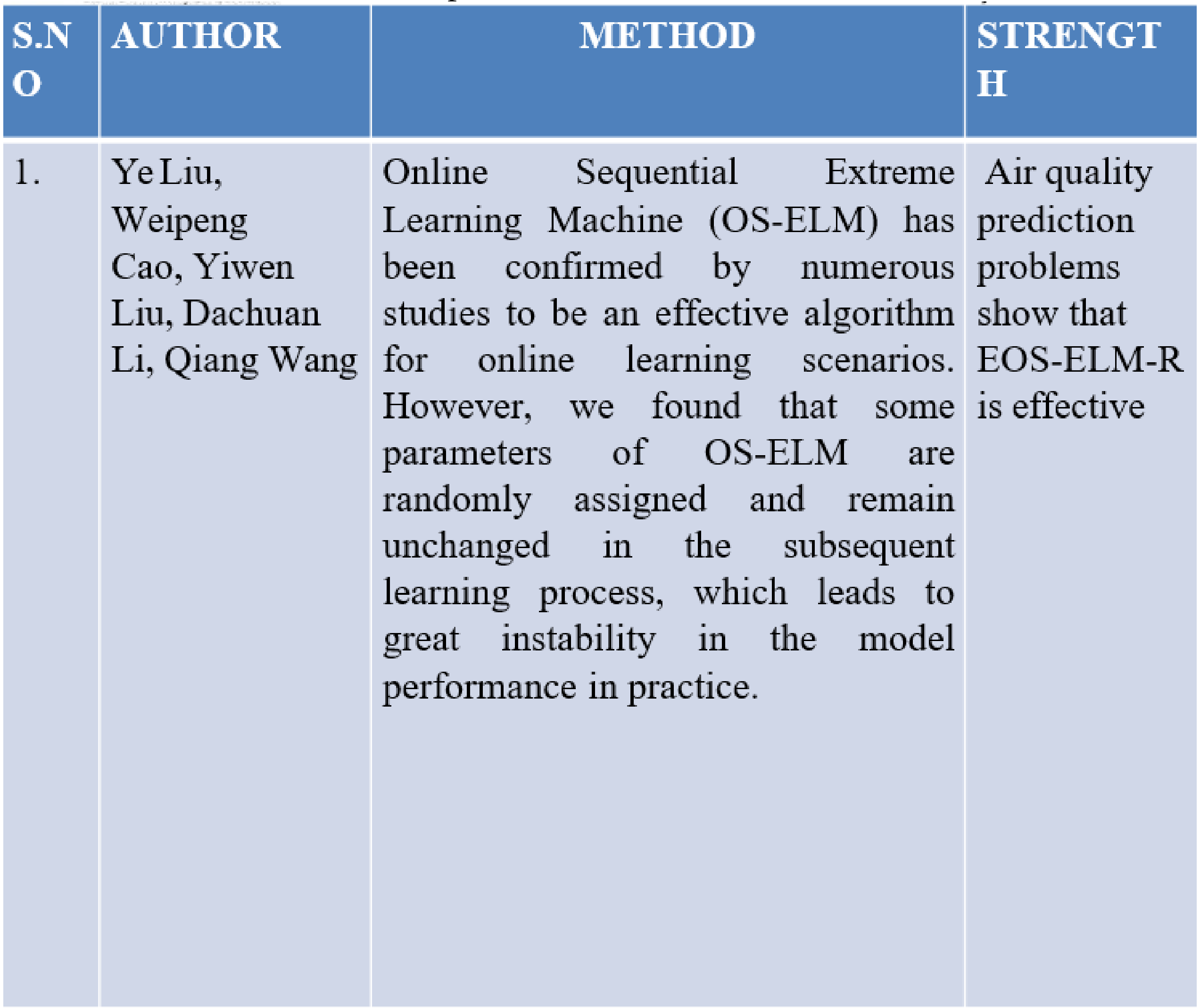
- Data Sources: The scope of data sources includes historical air quality data, real-time monitoring data from ground-based stations, satellite imagery, weather data, geographical information, traffic data, emissions inventories, and data from IoT devices and sensors.

- Integration and analysis of these diverse data sources are critical for accurate predictions.

CHAPTER 2

LITERATURE REVIEW

In this section, we discuss the previous research that has been conducted in this domain.



S.NO	AUTHOR	METHOD	STRENGTH
1.	Ye Liu, Weipeng Cao, Yiwen Liu, Dachuan Li, Qiang Wang	Online Sequential Extreme Learning Machine (OS-ELM) has been confirmed by numerous studies to be an effective algorithm for online learning scenarios. However, we found that some parameters of OS-ELM are randomly assigned and remain unchanged in the subsequent learning process, which leads to great instability in the model performance in practice.	Air quality prediction problems show that EOS-ELM-R is effective

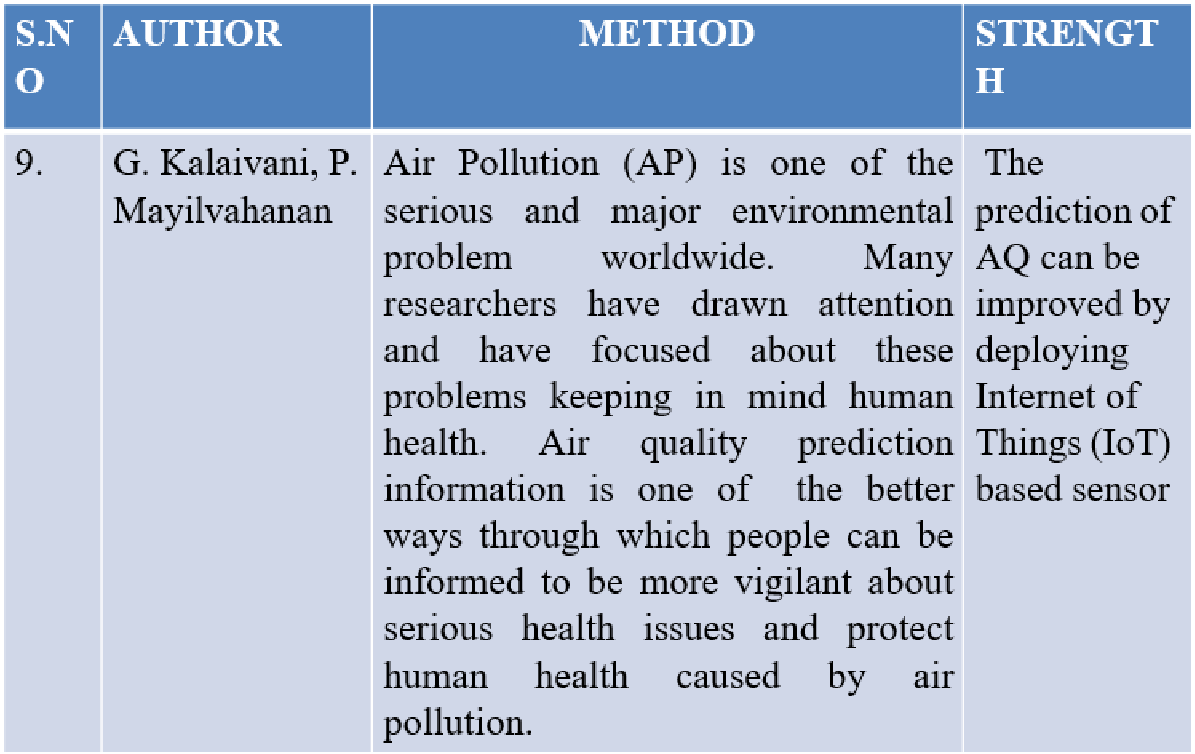
S.NO	AUTHOR	METHOD	STRENGTH
2.	M Sitha Ram, Chintamreddy Reshmasri, Shaik Shahila, Juluru Venkata Pavan Saketh	Identification of fresh air by predicting air quality Index is very important for providing better healthy environment to the society. Air pollution causes a severe health issues for the humans as well as threat to the environment.	XGboost helps to predict the air quality with high accuracy rate.
3.	Chenchen Li, Yan Li, Yubin Bao	The prevention and control of environmental pollution attracted much attention, and the haze weather directly affects people's travel health. In order to effectively prevent and control air pollution, optimize the air quality evaluation system. In this paper, PM 2.5 , PM 10 , SO 2 , NO 2 , CO and O 3 _8h are used as characteristic factors, and air quality index is used as a decision factor.	The Gradient Boosting Regression algorithm can effectively predict the Air Quality Index (AQI) and the air quality level.
4.	Hairong Qu, Runnan Zhang	After entering the new era, people's living standard has been significantly improved, the concept of environmental protection has been deeply rooted. People pursue a greener and healthier lifestyle, and the concern for air quality has become more and more intense.	All the prediction components are superimposed to obtain the final results.

S.NO	AUTHOR	METHOD	STRENGTH
5.	Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth,	The air quality monitoring system measures various air pollutants in various locations to maintain good air quality. It is the burning issue in the present scenario. Air is contaminated by the arrival of dangerous gases into the climate from the industries, vehicular emissions, etc..	Algorithm based on the machine learning to predict the future data of pollutants.

S.NO	AUTHOR	METHOD	STRENGTH
6.	Senlin Li, Xiaowu Deng, Bo Tang	Air quality is a closely relation to people' life and agricultural operations, due to the steep mountain and slippery road in a rainy day in the Wuling mountain area. In this study, machine learning methods is adopted to predict air quality of the Wuling mountain area, in order to give a better model for the air quality prediction of Wuling mountain area.	Random forest obtained the better performance than decision tree and deep back propagation neural network.

S.NO	AUTHOR	METHOD	STRENGTH
7.	N S Aruna Kumari, K S Ananda Kumar, S Hitesh Vardhan Raju	Air quality monitoring and prediction in many industrial and urban areas, it has become one of the most important activities. Owing to different types of pollution, air quality is heavily affected. With increasing air pollution, efficient air quality monitoring models is to be implemented; these models gather data on the concentration of air pollutants.	To solve three problems- prediction, interpolation and feature analysis, previously these problems were solved using three different models

S.NO	AUTHOR	METHOD	STRENGTH
8.	Timothy M. Amado Jennifer C. Dela Cruz	One of the biggest environmental problems right now is air pollution. Air quality is needed to be consistently monitored and assessed to ensure better living conditions. The U.S. Environmental Protection Agency (EPA) uses the air quality index (AQI) to standardize the air quality.	Having the neural network to be the best performing model.



S.NO	AUTHOR	METHOD	STRENGTH
9.	G. Kalaivani, P. Mayilvahanan	Air Pollution (AP) is one of the serious and major environmental problem worldwide. Many researchers have drawn attention and have focused about these problems keeping in mind human health. Air quality prediction information is one of the better ways through which people can be informed to be more vigilant about serious health issues and protect human health caused by air pollution.	The prediction of AQ can be improved by deploying Internet of Things (IoT) based sensor

CHAPTER 3

PROPOSED METHODOLOGY

There are four different kinds of algorithms which is being used in our system for diagnosis purposes. Those are :

1. Decision tree algorithm
2. Random Forest tree algorithm
3. Support Vector Machine
4. Linear Regression

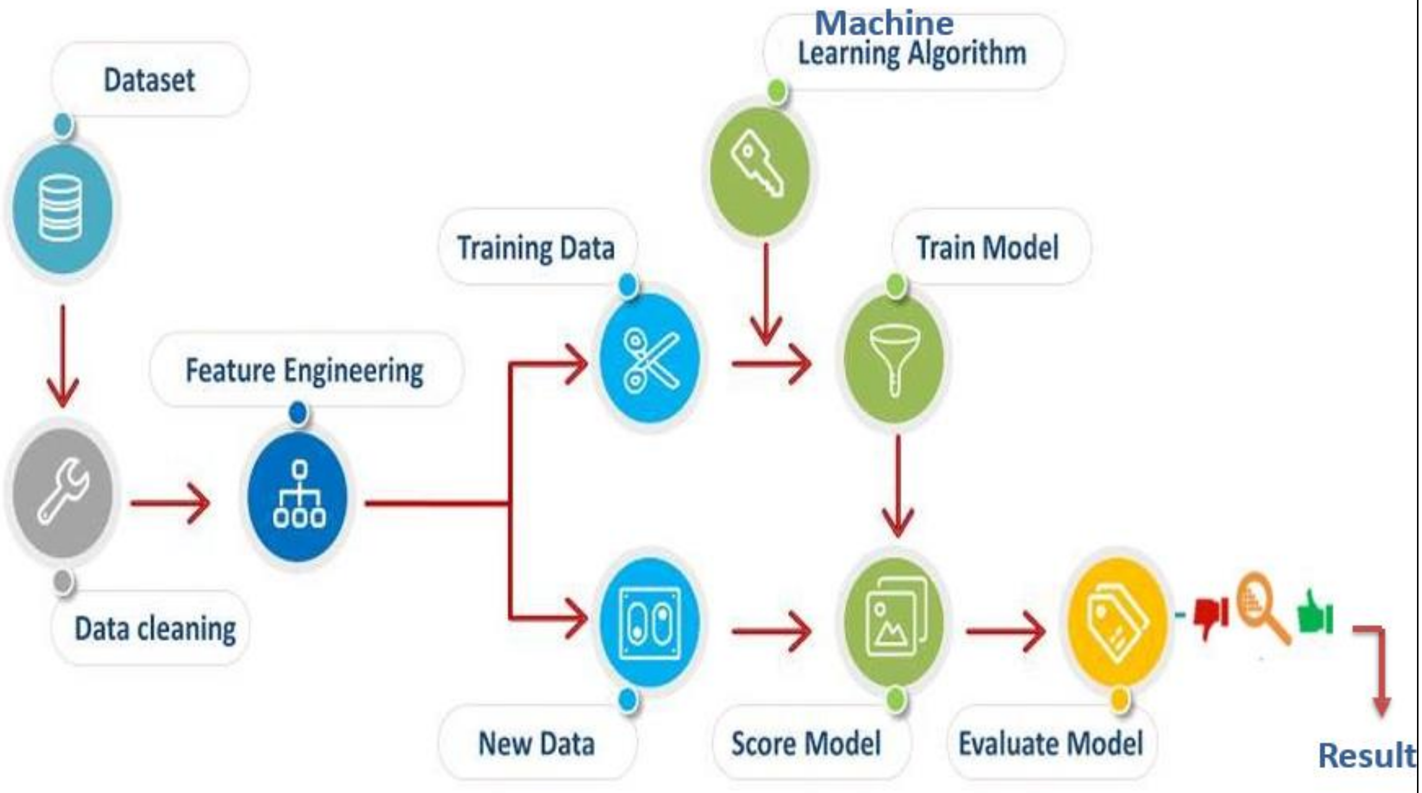
Certainly, you've mentioned several machine learning algorithms: Support Vector Machines (SVM), Random Forest (RF), Decision Trees, and Linear Regression. Each of these algorithms has its own strengths and applications. Here's a brief overview of each:

- Support Vector Machines (SVM):
- Type: Supervised learning algorithm.
- Application in Air Quality Prediction: SVM can be used for regression and classification tasks in air quality prediction. It's effective in handling high-dimensional data and capturing complex relationships between input features and air pollutant levels.
- Advantages: Works well in both linear and nonlinear scenarios. It aims to find the best separation (hyperplane) between classes or regression points.
- Considerations: SVMs can be computationally intensive, especially with large datasets. Proper kernel selection and hyperparameter tuning are critical for optimal performance.

- Random Forest (RF):
- Type: Ensemble learning algorithm (bagging).
- Application in Air Quality Prediction: RF is well-suited for air quality prediction due to its ability to handle complex relationships between multiple input features (e.g., weather variables, geographical data) and pollutant levels. It's known for its robustness and ability to handle noisy data.
- Advantages: High accuracy, ability to handle large datasets, and resistance to overfitting. Provides feature importance rankings.
- Considerations: May be less interpretable compared to single decision trees. Hyperparameter tuning is still important for optimal results.
- Decision Trees:
- Type: Supervised learning algorithm.
- Application in Air Quality Prediction: Decision trees can be used for both classification and regression tasks in air quality prediction. They are straightforward to understand and can capture complex decision boundaries.
- Advantages: Intuitive and interpretable. Can handle both categorical and numerical data. Good for capturing nonlinear relationships.
- Considerations: Prone to overfitting, especially with deep trees. Ensemble methods like Random Forest can mitigate this issue.

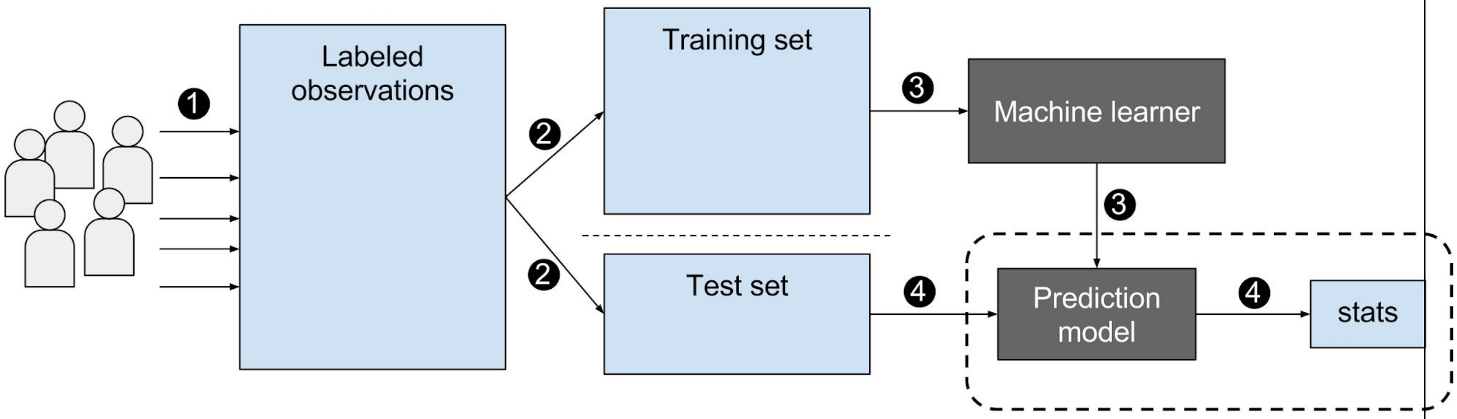
- Linear Regression:
- Type: Supervised learning algorithm.
- Application in Air Quality Prediction: Linear regression can be used for predicting air quality levels when the relationship between input features (e.g., weather variables) and pollutant levels is approximately linear. It provides simple and interpretable models.
- Advantages: Simplicity and interpretability. Fast training and prediction.
- Considerations: Assumes a linear relationship between predictors and the target variable. May not capture complex nonlinear patterns well. Model performance heavily relies on the linearity assumption.

ARCHITECTURE:



Data preprocessing:

- Quality of data is the first and most important prerequisite for effective visualization and creation of efficient ML models. The preprocessing steps help in reducing the noise present in the data which eventually increases the processing speed and generalization capability of ML algorithms. Outliers and missing data are the two most common errors in data extraction and monitoring applications.
- The data preprocessing step performs various operations on data such as filling out not-a-number (NAN) data, removing or changing outlier data, etc. Figure 2 shown below presents a view of the missing values in each feature of the dataset. Observe that among all other features, Xylene has the most missing values and CO has the least missing values. A large number of missing values may be existing due to a variety of factors, such as a station that can sense data but does not possess a device to record it.



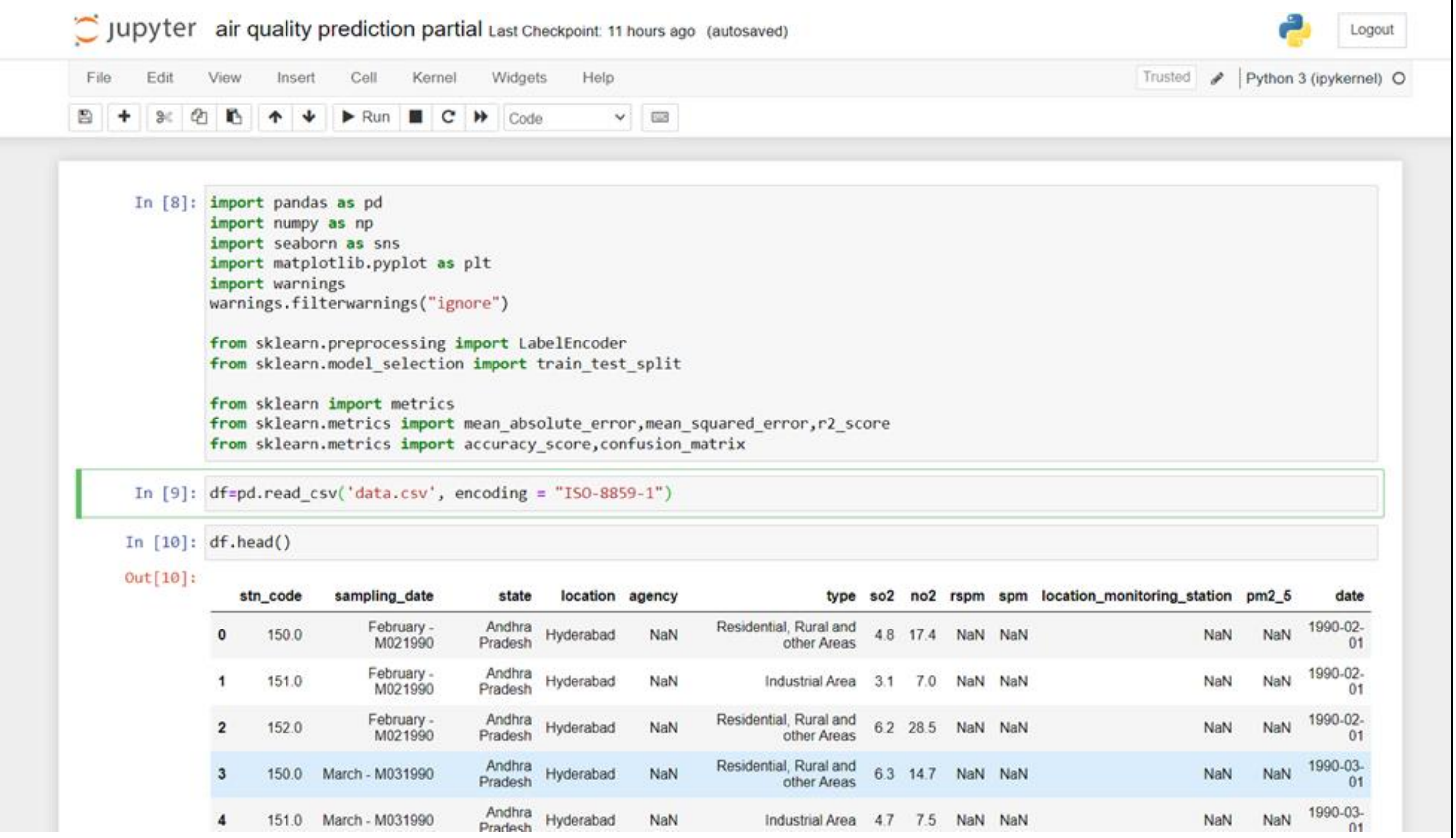
- In this project, standard libraries for database analysis and model creation are used. The following are the libraries used for the initial implementation :

1. **tkinter :** It’s a standard GUI library of python which provides various widgets such as

   Buttons, Label, Entry, Check Button, Drop box, List box and etc.
2. **Numpy :** is a core library of scientific computing in python. It provides powerful tools to deal with various multidimensional arrays in python.
3. **Pandas :** is the most popular python library used for data analysis. Data in python can be analyzed with 2 ways : Series and Dataframes.
4. **sklearn :** is an open source python library which implements a huge range of machine learning, pre-processing, cross-validation and visualization algorithms

CHAPTER 4 RESULTS

- This section discusses the results of the proposed models and methodologies :


jupyter air quality prediction partial Last Checkpoint: 11 hours ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

In [8]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split

from sklearn import metrics
from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score
from sklearn.metrics import accuracy_score,confusion_matrix
```

In [9]:

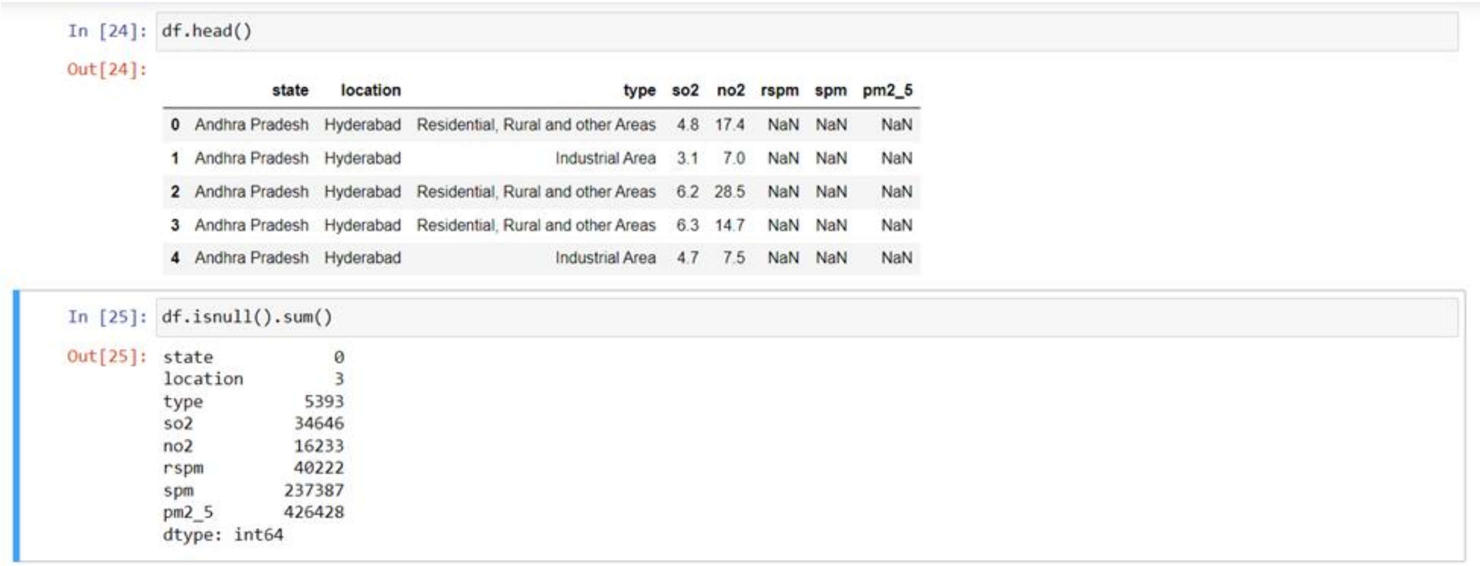
```
df=pd.read_csv('data.csv', encoding = "ISO-8859-1")
```

In [10]:

```
df.head()
```

Out[10]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	NaN	1990-02-01
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	NaN	1990-02-01
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	NaN	1990-02-01
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	NaN	1990-03-01
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	NaN	1990-03-01

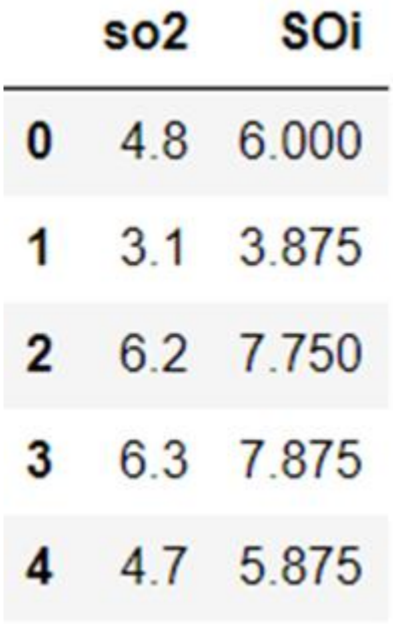


```
In [24]: df.head()
```

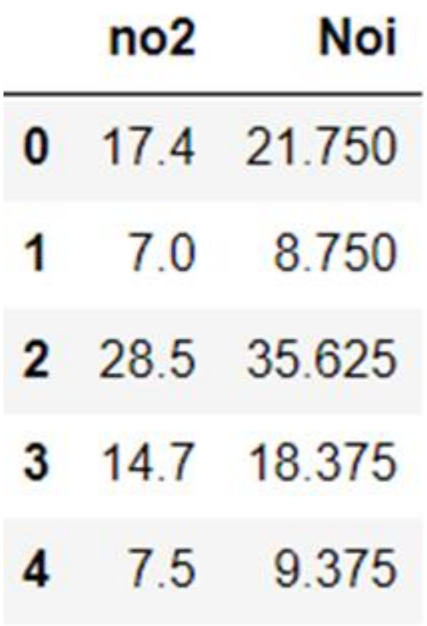
Out[24]:

	state	location	type	so2	no2	rspm	spm	pm2_5
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	NaN	NaN	NaN
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	NaN	NaN	NaN

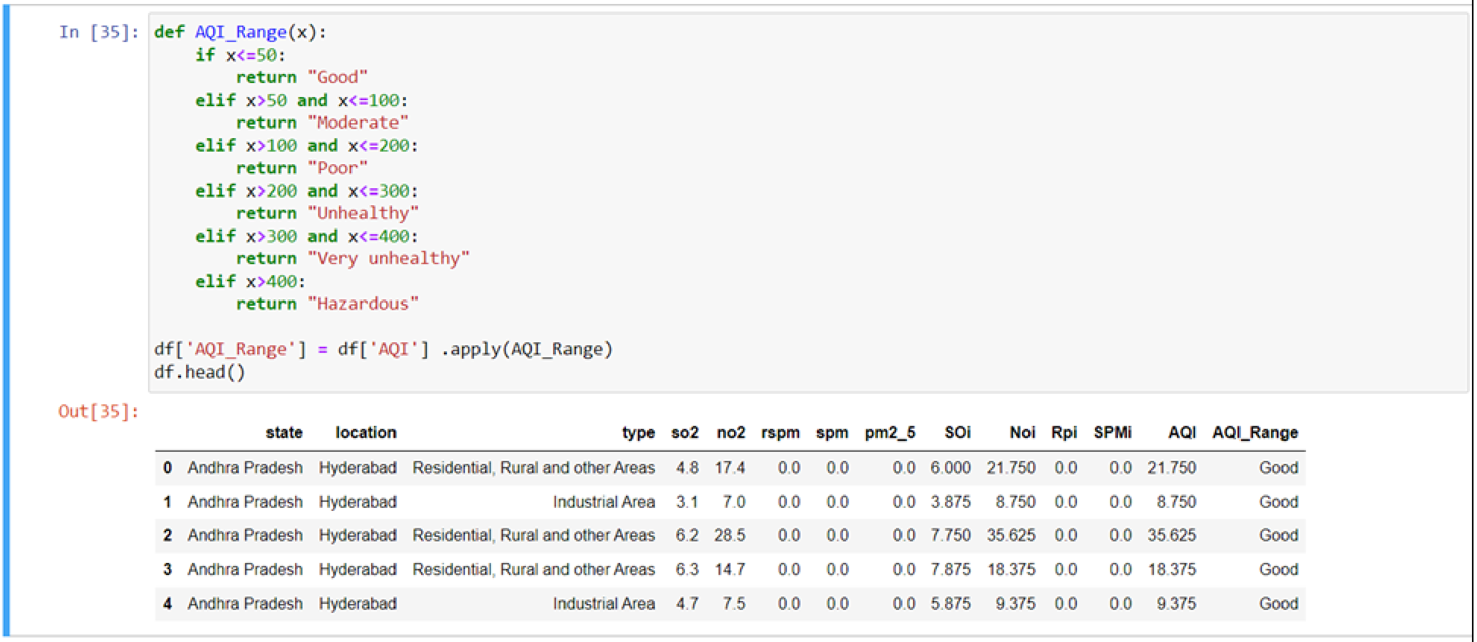
```
In [25]: df.isnull().sum()

Out[25]: state            0
         location         3
         type          5393
         so2          34646
         no2          16233
         rspm         40222
         spm         237387
         pm2_5       426428
         dtype: int64
```



	so2	SOi
0	4.8	6.000
1	3.1	3.875
2	6.2	7.750
3	6.3	7.875
4	4.7	5.875



	no2	Noi
0	17.4	21.750
1	7.0	8.750
2	28.5	35.625
3	14.7	18.375
4	7.5	9.375

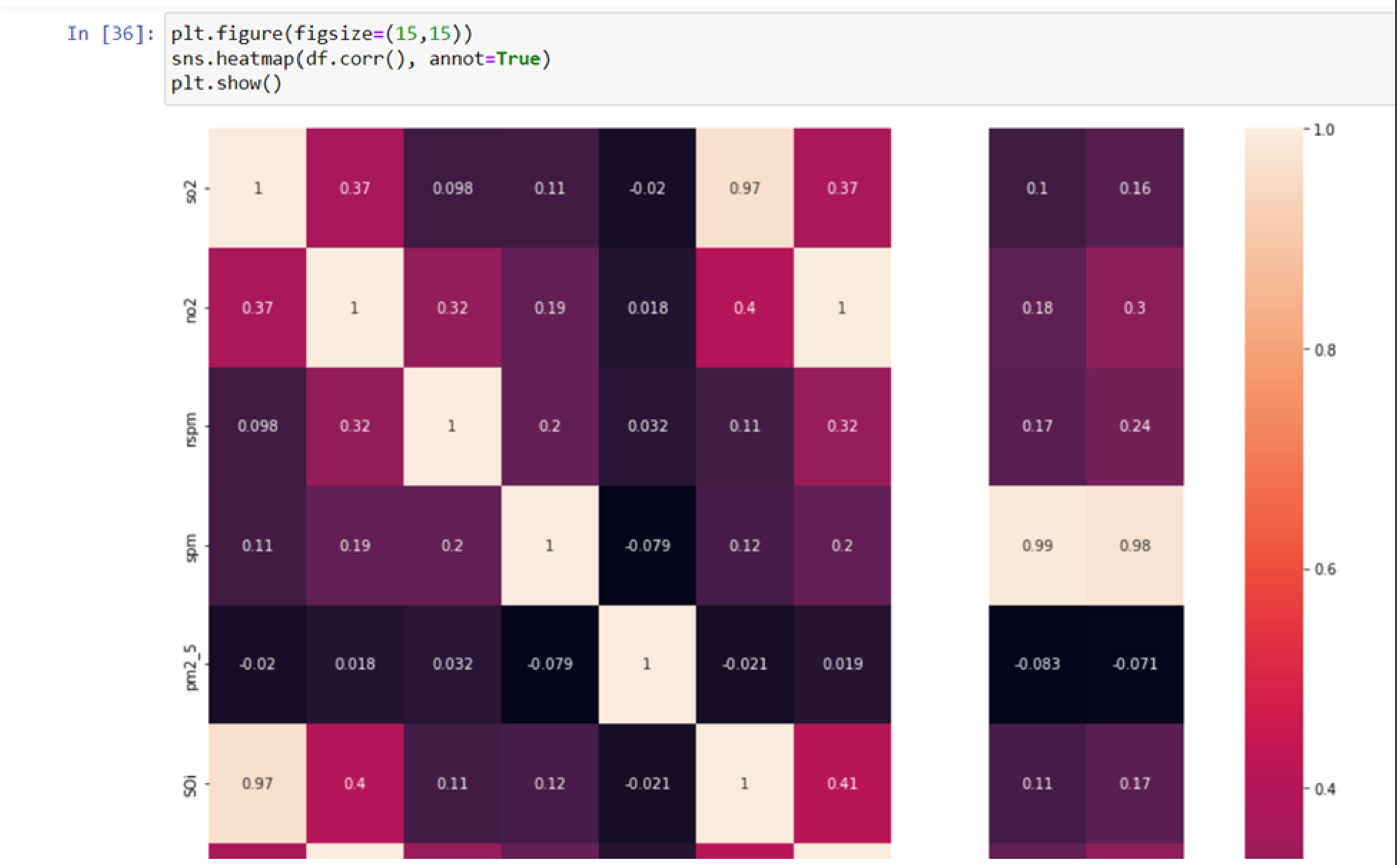


In [35]:

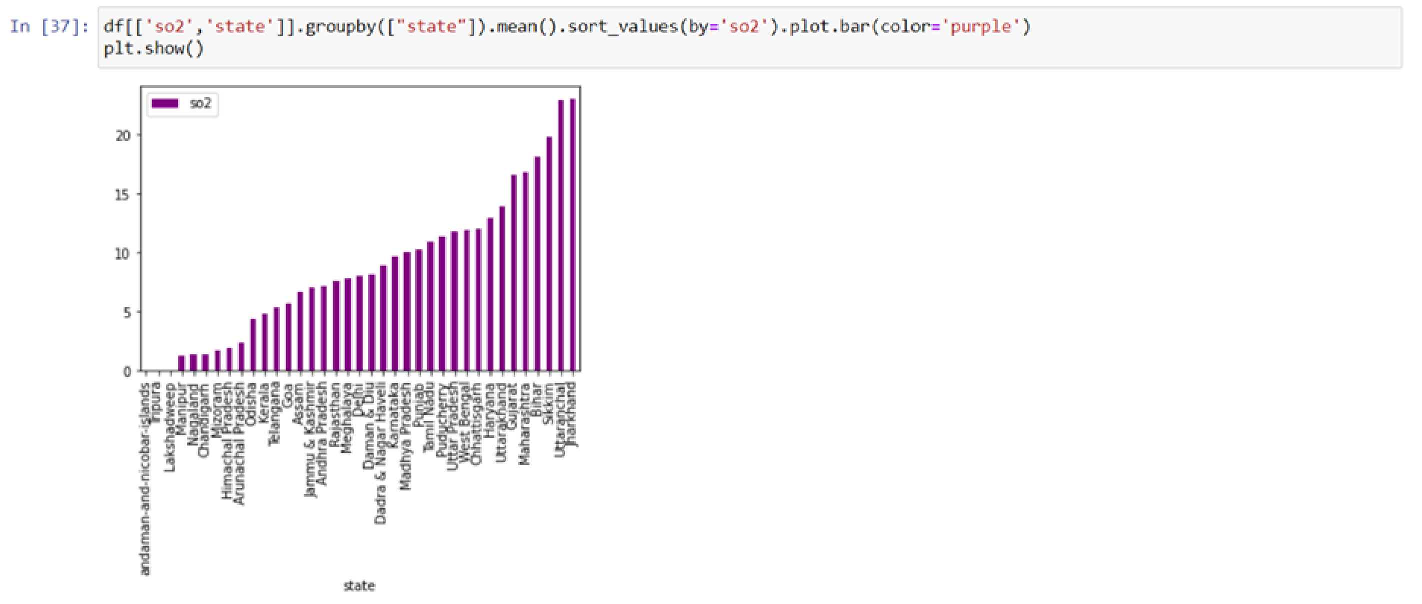
```
def AQI_Range(x):
    if x<=50:
        return "Good"
    elif x>50 and x<=100:
        return "Moderate"
    elif x>100 and x<=200:
        return "Poor"
    elif x>200 and x<=300:
        return "Unhealthy"
    elif x>300 and x<=400:
        return "Very unhealthy"
    elif x>400:
        return "Hazardous"

df['AQI_Range'] = df['AQI'] .apply(AQI_Range)
df.head()
```

Out[35]:

	state	location	type	so2	no2	rspm	spm	pm2_5	SOi	Noi	Rpi	SPMi	AQI	AQI_Range
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	0.0	0.0	0.0	6.000	21.750	0.0	0.0	21.750	Good
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	0.0	0.0	0.0	3.875	8.750	0.0	0.0	8.750	Good
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	0.0	0.0	0.0	7.750	35.625	0.0	0.0	35.625	Good
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	0.0	0.0	0.0	7.875	18.375	0.0	0.0	18.375	Good
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	0.0	0.0	0.0	5.875	9.375	0.0	0.0	9.375	Good



```
In [36]: plt.figure(figsize=(15,15))
         sns.heatmap(df.corr(), annot=True)
         plt.show()
```



```
In [37]: df[['so2','state']].groupby(["state"]).mean().sort_values(by='so2').plot.bar(color='purple')
         plt.show()
```

CHAPTER 5

EXSITING SYSTEM & PROPOSED SYSTEM:

- Urban air pollutant attention forecast is coping with a surge of large ecological monitoring data and intricate alterations in air pollution.
- This necessitates effective estimating methods to strengthen prediction accuracy and avoid grave contamination episodes, thereby improving ecological administration resolution-making capacity.
- A brand new contaminant concentration estimation process is established on sizeable amounts of ecological knowledge and deep learning approaches. This integrates colossal data using two forms of deep networks.
- Disadvantages of existing system:
- They provide limited accuracy as they are unable to predict the extreme points i.e. the pollution maximum and minimum
    - Cut-offs cannot be determined using such approach
- They use inefficient approach for better output prediction.

- **PROPOSED SYSTEM :**

- One versions of auto regression model prediction system will be implemented.
- We will test and evaluate both the systems with same test data to find their prediction accuracy.

- •The proposed system is developed to predict air quality using air database and then predict the air quality of the region.

CHAPTER 6

CONCLUSION

- Prediction of air quality is a challenging task because of the dynamic environment, unpredictability, and variability in space and time of pollutants.

- The grave consequences of air pollution on humans, animals, plants, monuments, climate, and environment call for consistent air quality monitoring and analysis, especially in developing countries.

- However, lesser attention for researchers has been observed for AQI prediction for India. In the present work, air pollution data of 23 Indian cities for a tenure of six years are investigated. The dataset is cleaned and preprocessed first by filling NAN values, addressing outliers, and normalising data values.

- Then correlation-based feature selection technique is exercised to filter AQI affecting pollutants for further study and logarithmic transformations are applied to the skewed features. The exploratory data analysis methods are exercised to find various hidden patterns present in the dataset.

- It was found that almost all pollutants exhibited a significant fall in 2020. The data imbalance problem is addressed by the SMOTE analysis.

- The dataset is split into train-test subsets by the ratio of 75–25% respectively. ML-based AQI prediction is carried out with and without SMOTE resampling technique and a comparative analysis is presented. The results of

ML models for both the train-test subsets are presented in terms of standard metrics like accuracy, precision, recall, and F1-Score. For both the train-test sets, the XGBoost model attained the highest accuracy and the SVM model exhibited the lowest accuracy.

- The classical statistical error metrics, namely MAE, RMSE, RMSLE, and R2 are then evaluated to assess and compare the performances of ML models. The XGBoost model comes out to be the overall best performer by attaining the optimum values in both training and testing phases. For the training phase, the RF model performed relatively good when exercised with SMOTE. On the other hand, almost all ML models exhibited improvements in the testing phase.
- In this phase, the GNB model attained the best results for R2 in target predictions. The present research endeavors to contribute to the literature by addressing air quality analysis and prediction for India which might have not been properly studied. This work can be extended by employing deep learning techniques for AQI prediction.

CHAPTER 7

REFERENCES

- [1] (2016). PhishMe Q1 2016 Malware Review. [Online]. Available: https://phishme.com/project/phishme-q1-2016-malware-review/
- [2] A. Belabed, E. Aimeur, and A. Chikh, ‘‘A personalized whitelist approach for phishing webpage detection,’’ in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012, pp. 249–254.
- [3] Y. Cao, W. Han, and Y. Le, ‘‘Anti-phishing based on automated individual white-list,’’ in Proc.4th ACM Workshop Digit. Identity Manage., 2008, pp. 51–60.
- [4] T.-C. Chen, S. Dick, and J. Miller, ‘‘Detecting visually similar Web pages: Application to phishing detection,’’ ACM Trans. Internet Technol., vol. 10, no. 2, pp. 1–38, May 2010.
- [5] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell, ‘‘Clientside defense against Web-based identity theft,’’ in Proc. 11th Annu. Netw. Distrib. Syst. Security Symp. (NDSS), 2004, pp. 1–16
- [6] C. Inc. (Aug. 2016). Couldmark Toolbar. [Online]. Available:
- http://www.cloudmark.com/desktop/ie-toolbar
  - [7] J. Corbetta, L. Invernizzi, C. Kruegel, and G. Vigna, ‘‘Eyes of a human, eyes of a program: Leveraging different views of the Web for analysis and detection,’’ in Proceedings of Research in Attacks, Intrusions and Defenses (RAID). Gothenburg, Sweden: Springer, 2014.
  - [8] X. Deng, G. Huang, and A. Y. Fu, ‘‘An antiphishing strategy based on visual similarity assessment,’’ Internet Comput., vol. 10, no. 2, pp. 58–65, 2006.
- [9] Z. Dong, K. Kane, and L. J. Camp, ‘‘Phishing in smooth waters: The state of banking certificates in the US,’’ in Proc. Res. Conf. Commun., Inf. Internet Policy (TPRC), 2014, p. 16.