**AIR QUALITY PREDICTION**

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

In

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

By

**CH. PREETHI (2103A52047)**

**H. DHARANI (2103A52063)**

**P. BHAVANA (2103A52065)**

Under the guidance of

**Mr. D.Ramesh**

Assistant Professor, Department of CSE.

**

**Department of Computer Science and Artificial Intelligence**

**

**Department of Computer Science and Artificial Intelligence**

**CERTIFICATE**

This is to certify that project entitled **“AIR QUALITY DATA”** is the bonafied work carried out by **P. BHAVANA**  **,CH.PREETHI, H.DHARANI** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

**Mr. D. Ramesh Dr. M. Sheshikala**

Asst. Professor, Assoc. Prof. & HOD

S R University, S R University,

Ananthasagar, Warangal. Ananthasagar, Warangal.

**ABSTRACT**

Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. With this increasing air pollution, We are in need of implementing models which will record information about concentrations of air pollutants(so2,no2,etc).The deposition of this harmful gases in the air is affecting the quality of people’s lives, especially in urban areas. Lately, many researchers began to use Big Data Analytics approach as there are environmental sensing networks and sensor data available. In this paper, machine learning techniques are used to predict the concentration of so2 in the environment. Sulphur dioxide irritates the skin and mucous membranes of the eyes, nose, throat, and lungs. Models in time series are employed to predict the so2 readings in nearing years or months. Contains the responses of a gas multisensor device deployed on the field in an Italian city. Hourly responses averages are recorded along with gas concentrations references from a certified analyzer.

Keywords: Machine Learning, Time Series, Prediction, Air Quality, SO2

**ACKNOWLEDGEMENT**

We express our thanks to Course co-coordinator **Mr.** **D.Ramesh, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude toHead of the department CS&AI, **Dr. M.Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

**Table of Contents**

**Chapter No. Title Page No.**

1. Introduction
   1. Overview 1
   2. Problem Statement 3
   3. Existing system 3
   4. Proposed system 4
   5. Architecture 4
2. Literature survey
   * 1. Document the survey done by you 5
3. Data pre-processing
   1. Dataset description 7
   2. Data cleaning 9
   3. Data augmentation 10
   4. Data Visualization 11
4. Methodology
   1. Procedure to solve the given problem 16
   2. Model architecture 20
   3. Software description 22
5. Results and discussion 23
6. Conclusion and future scope 24 …

**CHAPTER-1**

**INTRODUCTION**

**1.1 Overview**

Air is an invisible substance surrounding the Earth and providing us all with the breathable oxygen and

performs a vital role in supporting life on Earth. But with the passage of time the fresh and pure air is gradually

getting contaminated due to increase in air pollution. Air pollution is the presence of one or more substance at a

concentration above their natural levels, with the potential to produce an adverse effect

Air is an invisible substance surrounding the Earth and providing us all with the breathable oxygen and

performs a vital role in supporting life on Earth. But with the passage of time the fresh and pure air is gradually

getting contaminated due to increase in air pollution. Air pollution is the presence of one or more substance at a

concentration above their natural levels, with the potential to produce an adverse effect

Air is an invisible substance surrounding the Earth and providing us all with the breathable oxygen and

performs a vital role in supporting life on Earth. But with the passage of time the fresh and pure air is gradually

getting contaminated due to increase in air pollution. Air pollution is the presence of one or more substance at a

concentration above their natural levels, with the potential to produce an adverse effect

Air is an invisible substance surrounding the Earth and providing us all with the breathable oxygen and

performs a vital role in supporting life on Earth. But with the passage of time the fresh and pure air is gradually

getting contaminated due to increase in air pollution. Air pollution is the presence of one or more substance at a

concentration above their natural levels, with the potential to produce an adverse effectOur atmosphere is ~21% oxygen and ~78% nitrogen; the remaining 1% is considered “trace gases” and this includes everything else—from carbon dioxide to the noble gases like argon. Scientists and engineers study this 1%, as well as the many types of particles present in the atmosphere. When we think of air quality, we typically think of the air we are breathing and whether or not it is safe. However, air quality can refer to ambient outdoor conditions, indoor conditions, particular sources, good air quality vs. poor air quality, etc. Within the field of air quality, researchers specialized in many other specific topics. Due to the complexity of our atmosphere and the possibility of transport over long distances, we can also think of air quality in terms of scale—that is, pollution may cause local or global problems. For example, pollution in China can make its way over North America and add to existing pollutants there.

Poor air quality can negatively affect human and environmental health. In humans, poor air quality can lead to a multitude of problems that include respiratory and cardiovascular diseases. We tend to think first of asthma and respiratory problems, but some particles are so small that they can enter the blood stream through the lungs and cause inflammation leading to issues beyond our breathing. In plants, poor air quality can also cause disease that can result in crop loss. In addition to human and environmental health, many pollutants that we worry about are greenhouse gases and contribute to climate change. Finally, poor air quality can impact quality of life. Consider visibility issues in National Parks and odors near industrial areas of cities; in addition to potential health dangers, these air quality issues can make daily life unpleasant.

Air is one of the foundational substances the human body needs to survive and thrive. Clean air keeps us healthy and feels good to breathe in. And so, we've put together this comprehensive resource on air quality. You'll learn everything from how air quality is measured to the harmful health and economic effects of poor air quality to some ways to improve the quality of the air you breathe, and so much more. Air is one of foundational substances the human body needs to survive and thrive. Clean air keeps us healthy and feels good to breathe in. And so, we've put together this comprehensive resource on air quality. You'll learn everything from how air quality is measured to the harmful health and economic effects of poor air quality to some ways to improve the quality of the air you breathe, and so much more.

1

As mentioned, air quality measures what kind of pollutants and how much of each is in the air you breathe. Poor air quality can lead to or contribute to all sorts of health problems - especially in groups sensitive to air pollutants - such as:

* Respiratory issues
* Cardiovascular dysfunction
* Reduced cognitive function
* Some cancers

Thus, air quality is a crucial consideration for many in these groups, as well as healthy individuals who want to minimize the chances anything in their body suffers harm.

Additionally, poor air quality can harm the economy in several ways, including but not limited to:

* Medical costs associated with poor air quality
* The human cost - deaths due to poor air quality
* Productivity losses
* Pollution remediation costs
* Negative food production impacts

Indoor air quality is especially critical. Americans spend nearly 90% of their time indores on average, according to the EPA. If you spend much more time inside your home, you’re breathing in much more interior air.

That said, outdoor air quality remains an important factor in choosing a place to live, especially for those with respiratory issues.

There is much to consider when it comes to choosing a place to live. If you or someone in your family is in one of the aforementioned sensitive groups, air quality may be one of the more important matters to consider if you’re moving. It may even be a reason you move from your current location if it’s bad enough.

Various factors can impact outdoor air quality, from industry to weather and more. Consequently, each state has a different overall air quality rating. Even within those states, some cities do much better than others in terms of air pollution. Sometimes, a single city or area of the state may be known for poor air quality, whereas the rest of it isn’t too bad.

Regardless, let’s look at the best five and worst five states for air quality, based on the American Lung Association’s (ALA) State of the Air 2021 report.

2

**1.2 PROBLEM STATEMENT**

We will make a project for prediction of Air Quality. The problem statement is that predicts what are diseases will cause for a person in future.

**1.3 EXISTING SYSTEM**

In the existing systems we only have data but there is no correct prediction of humidity so,our model can predict the above.

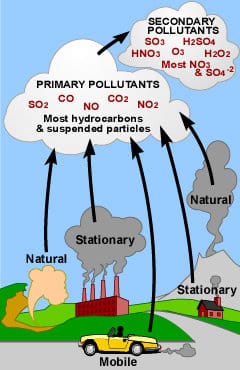
**1.4 PROPOSED SYSTEM**

I get this data from UCI Machine Learning. Here is about descripstion rows and column also another description.

"The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value.

3

**1.6 ARCHITECTURE**



**4**

**CHAPTER-2**

**LITERATURE SURVEY**

**2.1 Document the survey done by you**

With the advancement of IoT infrastructures, big data technologies, and machine learning techniques, real-time air quality monitor and evaluation is desirable for future smart cities. This paper reports our recent literature study, reviews and compares current research work on air quality evaluation based on big data analytics, machine learning models and techniques. Finally, it highlights some observations on future research issues, challenges, and needs.

We have formulated the problem as regularized MTL and employed advanced optimization algorithms for solving different formulations. We have focused on alleviating model complexity by reducing the number of model parameters and on improving the performance by using a structured regularizer. Our results show that the proposed light formulation achieves much better performance than the other two model formulations and that the regularization by enforcing prediction models for two consecutive hours to be close can also boost the performance of predictions. We have also shown that advanced optimization techniques are important for improving the convergence of optimization and that they speed up the training process for big data. For future work, we will further consider the commonalities between nearby meteorology stations and combine them in a MTL framework, which may provide a further boosting for the prediction.

Among the analysed works, 20 of them use open data to perform air quality predictions. These works were carried out from 2014 until now, coinciding with the movement of open data within the cities [70]. Therefore, we can affirm that the open data movement has increased the number of research works in the field of machine learning, especially in the prediction of air quality.

5

Related to the prediction target, the outcome shows that PM2.5 was the main element, applied in 19 papers, 18 of which utilised data of the cities located in China. Most cases, the authors performed a prediction for the next day. Twenty-seven studies used data hourly collected from the sensors.

In order to predict air quality, pm2\_5 is also an important attribute. The values of this must be recorded in future as this particulates are responsible for various health effects including cardiovascular effects such as cardiac arrhythmias

and heart attacks, and respiratoryeffects such as asthma attacks and bronchitis. This model further, also makes us aware of the challenges in future and research needs such as pm2.5,AQI,etc.

Most current studies directly predict air pollutants by inputting data for a fixed period of time in the future. The prediction results performed well in the short-term prediction such as 1h-predictions, but in the long-term prediction (up to or above 24 h-predictions), the time interval between forecasting and training will become very long, which results in lowering the temporal correlations between them. Thus the accuracy of the long-term prediction is generally low.

6

**CHAPTER-3**

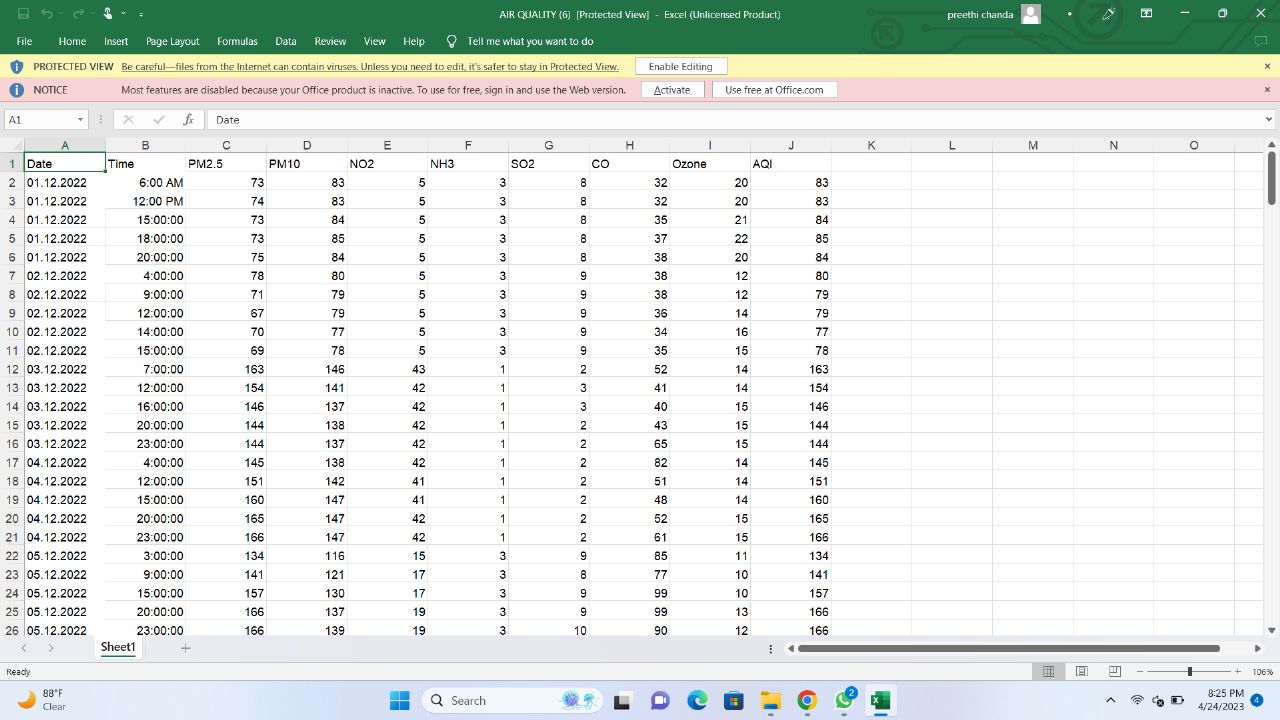
**DATA PRE-PROCESSING**

**2.1 Data Description**

* **PM2.5-** PM2.5 has a significant influence on human health. And with the modern society developing, PM2.5 has been becoming a severe problem for people.
* **PM10-**High levels of PM10 can irritate the eyes and throat.Like PM2.5, increased symptoms can occur when exposed to PM10 if you have asthma or other lung diseases.
* **NO2** -Air quality has been the main concern worldwide and Nitrous oxide (NO2) is one of the pollutants that have a significant effect on human health and environment.
* **NH3**-Ammonia comes from the breakdown and volatilisation of urea. Emissions and deposition vary spatially, with "emission hot-spots" associated with high-density intensive farming practices.
* **S02**-Sulfur dioxide (SO2) is a colorless, reactive air pollutant with a strong odor. This gas can be a threat to human health, animal health, and plant life. The main sources of sulfur dioxide emissions are from fossil fuel combustion and natural volcanic activity.
* **CO**-It is a colorless, odorless gas formed by the incomplete reaction of air with fuel.
* **Ozone**-Air quality regulators are concerned about ozone pollution because of its effects on public health and the environment.

7

**DATASET**



8

**2.2 Data Cleaning**

We have converted all our dataset strings into numerical values.

Contains the responses of a gas multisensory device deployed on the field in an italian city.

Hourly responses averages are recorded along with gas concentrations references from a

certified analyzer.

Air pollution is considered to occur whenever harmful or excessive quantities of defined

substances such as gases , particulates , and biological molecules are introduce in to

atmosphere.

With the economic and technological development of cities , environmental pollution

problems are arising , such as water ,noise ,and ,air pollution . In particular, air pollution has

a direct impact on human health through the exposure air of pollutants and particulates

,which has increased interest in air pollution.

The purpose of air quality can deliver substantial health benefits ; reducing air pollution

levels means reducing premature deaths and diseases from stroke , heart diseases , lung

cancer ,and both chronic and acute respiratory diseases including asthama.

The substances(NO2,CO,SO2) are in compositions of a common index is called air called

indes(AQI), indicating how clean or polluted the air is currently or forecasted to become in

areas.

AQI is an index for reporting daily air quality .It tells you how clean or unhealthy.

what associated health effects might be a concern . The AQI focuses on health effects you

may experience within a few hours or days after breathing unhealthy air.

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal

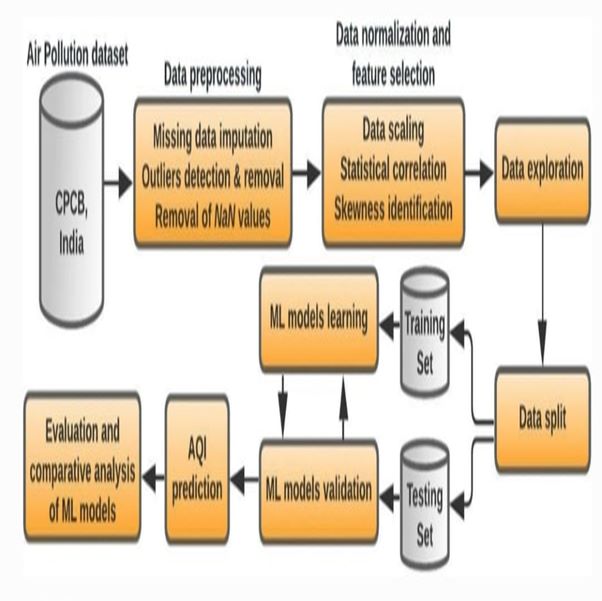
oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The

device was located on the field in a significantly polluted area, at road level,within

an Italian city.

9

**Data cleaning**

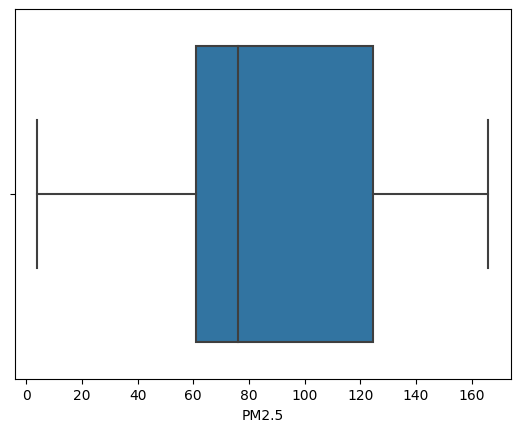


10

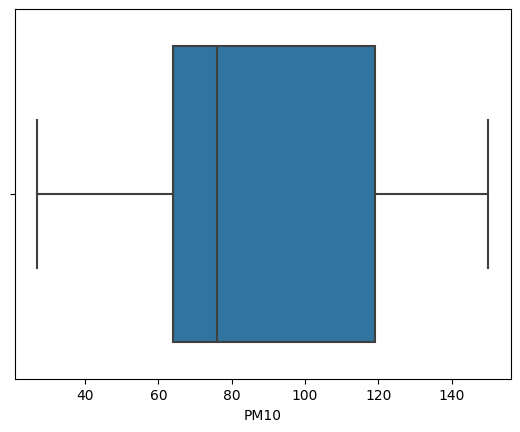
**2.3-Data visualisation**

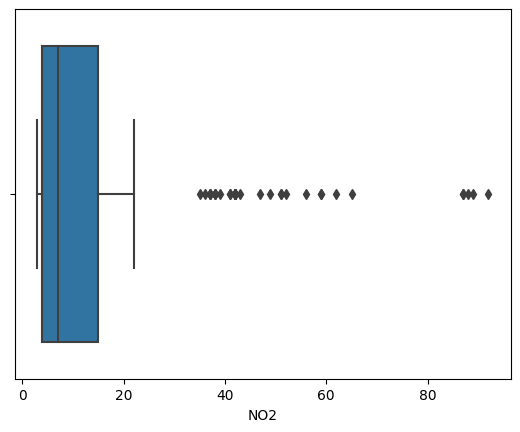
**BOX PLOT**

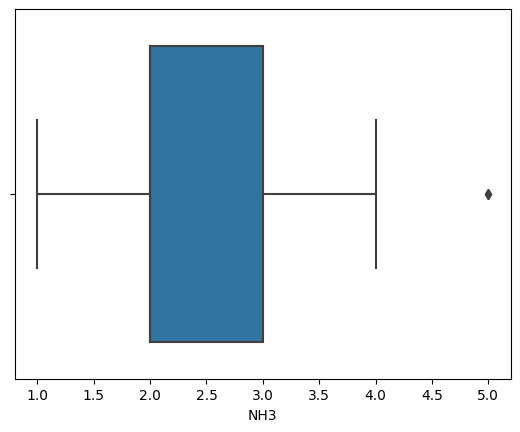
* A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.
* In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.
* Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups. They are built to provide highlevel information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers.
* A boxplot is a way to show a [five number summary](https://www.statisticshowto.com/statistics-basics/how-to-find-a-five-number-summary-in-statistics/) in a chart. The main part of the chart (the “box”) shows where the middle portion of the data is: the interquartile range. At the ends of the box, you” find the first [quartile](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-are-quartiles/)(the 25% mark) and the third quartile (the 75% mark). The far left of the chart (at the end of the left “whisker”) is the minimum (the smallest number in the set) and the far right is the maximum (the largest number in the set). Finally, the median is represented by a vertical bar in the center of the box.



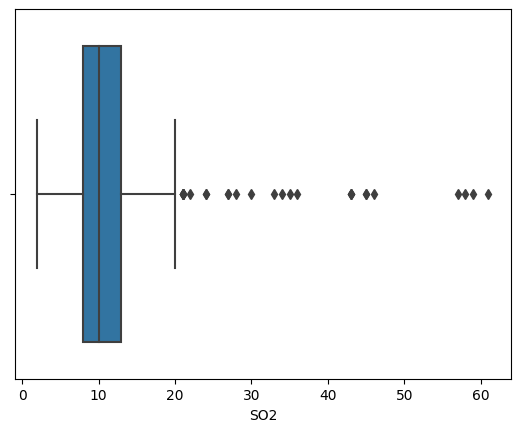
11

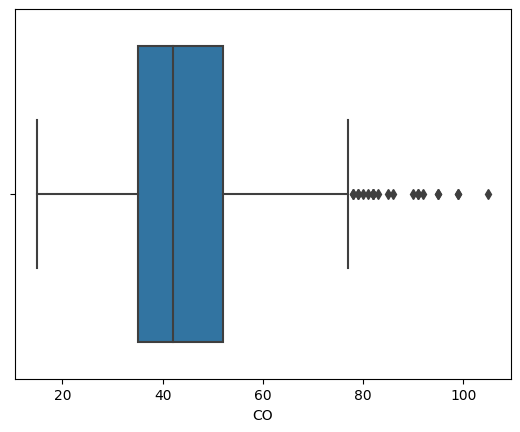


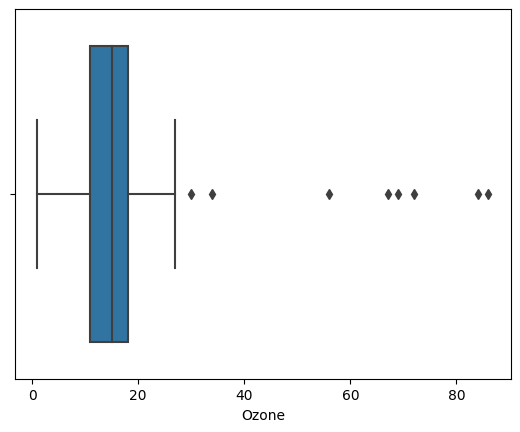




12

****

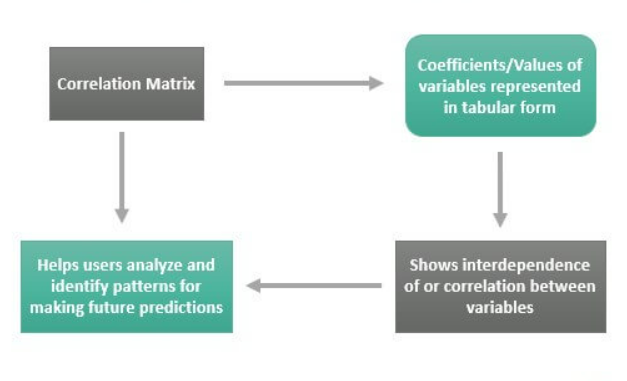
****

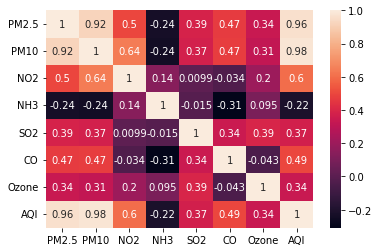
****

13

**CORRELATION MATRIX**

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect. A correlation matrix is simply a table showing the correlation coefficients between variables.





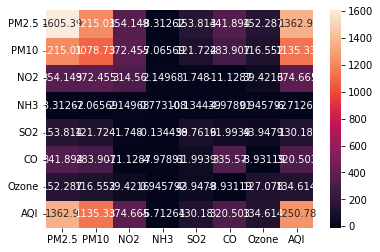
14

**COVARIANCE MATRIX**

Covariance matrix is a square matrix (has the same number of rows and columns) that gives the covariance between each pair of elements available in the data. Covariance measures the extent to which to variables move in the same direction.

Covariance matrix is a type of matrix that is used to represent the covariance values between pairs of elements given in a random vector. The covariance matrix can also be referred to as the variance covariance matrix. This is because the variance of each element is represented along the main diagonal of the matrix.

A covariance matrix is always a square matrix. Furthermore, it is positive semi-definite, and symmetric. This matrix is very useful in stochastic modeling and principle component analysis. In this article, we will learn about the variance covariance matrix, its formula, examples, and various important properties associated with it.



15

**CHAPTER-4**

**METHADOLOGY**

**1.1Procedure to solve a given problem**

* We collected the real time values of the gases which we used in our dataset.
* We choosed some specific area and collected the data of the gases involved in air quality prediction. The main gases which are responsible in damaging the air are CO,NO2 and SO2.
* We calculated and plotted the corelation and covariance matrix.
* Ours is a regression model, in our dataset we have 10 attributes.
* AQI is the y value and all others are x values. We are predicting the y value based on the x values.
* Contains the responses of a gas multisensory device deployed on the field in an italian city. Hourly responses averages are recorded along with gas concentrations references from a certified analyzer.
* Air pollution is considered to occur whenever harmful or excessive quantities of defined substances such as gases , particulates , and biological molecules are introduce in to atmosphere.
* The substances(NO2,CO,SO2) are in compositions of a common index is called air called indeces(AQI), indicating how clean or polluted the air is currently or forecasted to become in areas.
* The purpose of air quality can deliver substantial health benefits ; reducing air pollution levels means reducing premature deaths and diseases from stroke , heart diseases , lung cancer ,and both chronic and acute respiratory diseases including asthama.
* With the economic and technological development of cities , environmental pollution problems are arising , such as water ,noise ,and ,air pollution . In particular, air pollution has a direct impact on human health through the exposure air of pollutants and particulates ,which has increased interest in air pollution.

16

**K-NEAREST NEIGHBOUR**

* K-nearest neighbor (KNN) is a non-parametric algorithm that can be used for classification and regression tasks. It works by finding the k closest data points in the training set to a given test point and then using the labels or values of those data points to predict the label or value of the test point.
* To use KNN for air quality prediction, you would first need to collect data on the air quality at various locations and times. This data could include measurements of pollutants such as particulate matter (PM), ozone (O3), nitrogen dioxide (NO2), and sulfur dioxide (SO2), as well as temperature, humidity, wind speed, and other relevant variables.
* It is important to note that KNN assumes that similar data points have similar values, which may not always be the case for air quality data. Additionally, KNN can be computationally expensive, especially for large datasets, so it may not be the best choice for real-time air quality prediction. Other machine learning algorithms such as decision trees, random forests, and neural networks may also be useful for air quality prediction.

## **Advantages of KNN Algorithm:**

* It is simple to implement.
* It is robust to the noisy training data
* It can be more effective if the training data is large.

## **Disadvantages of KNN Algorithm:**

* Always needs to determine the value of K which may be complex some time.
* The computation cost is high because of calculating the distance between the data points for all the training samples.

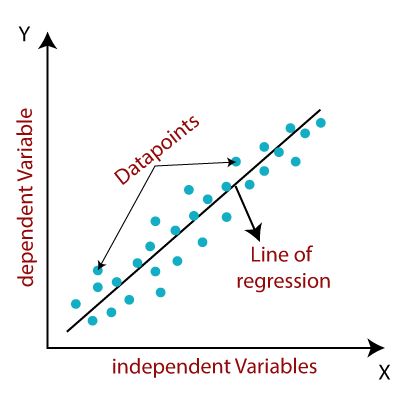
**Calculations:**

* Mean Square Error:25.623478260869575
* Mean Absolute Error:3.2913043478260877
* Root Mean Square Error:5.061963873919842

17

**LINEAR REGRESSION**

* Linear regression is a statistical technique used to analyze the relationship between a dependent variable (Y) and one or more independent variables (X). In the context of air quality prediction, linear regression can be used to model the relationship between air quality measurements (such as particulate matter concentration or ozone levels) and various environmental factors, such as temperature, humidity, and wind speed.
* To use linear regression for air quality prediction, you would first need to collect data on the air quality measurements and environmental factors of interest. This data could be collected using sensors installed at various locations throughout a city or region, or by collecting data from existing monitoring stations.
* It is important to note that linear regression is just one of many statistical techniques that can be used for air quality prediction.



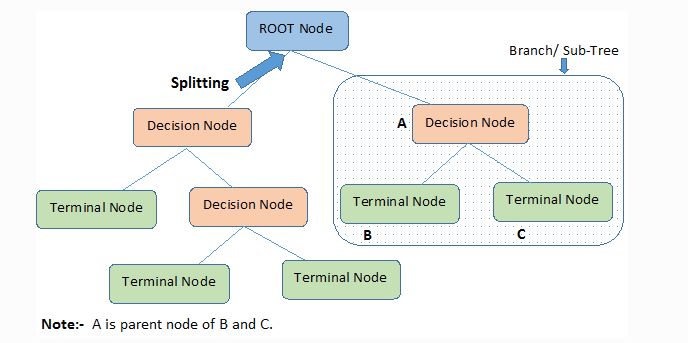
**Calculations:**

* Mean Square Error: 20.186750203222648
* Mean Absolute Error: 3.6098028860602294
* Root Mean Absolute Error: 4.492966748510682

18

**DECISION TREE**

* Decision tree algorithm is a computational method used to build decision trees from data. The goal of the algorithm is to identify the most important variables and their values that are predictive of a particular outcome or class.
* Once the decision tree is built, it can be used to predict the outcome or class of new data based on the values of the variables. This is done by traversing the tree from the root node to a leaf node, where each internal node corresponds to a test on a variable and each leaf node corresponds to a prediction of the outcome or class.



**Calculations:**

* Mean Square Error:12.021739130434783
* Mean Absolute Error:2.108695652173913
* Root Mean Square Error:3.467237968532703

19

**SUPPORT VECTOR MACHINE**

* Support Vector Machine (SVM) is a popular machine learning algorithm that can be used for air quality prediction. SVMs are particularly effective in handling non-linear relationships between variables and have been successfully applied to various environmental prediction problems, including air quality.
* To use SVM for air quality prediction, you would first need to collect data on the air quality measurements and environmental factors of interest. This data could be collected using sensors installed at various locations throughout a city or region or by collecting data from existing monitoring stations.

**Calculations:**

* **Mean Square Error: 24.142612927396755**
* **Mean Absolute Error:3.6699528174496265**
* **Random Mean square Error:4.9135132977734735**

20

**RANDOM FOREST**

* Random Forest is a popular machine learning algorithm used for predictive modeling in a wide range of applications, including air quality prediction. Random Forest is an ensemble learning method that combines multiple decision trees to improve the accuracy and generalization of the model.
* In the context of air quality prediction, Random Forest can be used to model the relationship between various air pollutants, meteorological variables, and other environmental factors, and their impact on air quality. The algorithm works by building multiple decision trees on random subsets of the data and combining their predictions to produce a final prediction.
* Overall, Random Forest is a powerful algorithm for air quality prediction that can take into account multiple variables and their complex interactions.

**Calculations:**

* MeanSquareError:6.253336956521742
* Mean Absolute Error:1.6089130434782613
* Root Mean Suqare Error2.500667302245891

21

**Software Requirements:**

**OS : Windows or higher version.**

**Platform : Google collab**

**Program Language :Python**

**GOOGLE COLLAB:** Google Colab is a cloud-based service provided by Google that allows users to write, run, and share code in various programming languages, including Python, R, and Julia. Colab provides a Jupyter notebook environment that allows users to create and edit code cells, execute them, and view the output. The service is free to use and provides access to a high-performance virtual machine with GPU and TPU support. Users can save their notebooks in Google Drive and share them with others for collaboration. Colab also allows users to install and use third-party libraries and frameworks, such as TensorFlow, PyTorch, and scikit-learn, making it a popular choice for machine learning and data analysis projects.

PYTHON:

Python is a popular high-level programming language that is widely used for web development, scientific computing, data analysis, artificial intelligence, and many other applications. Python is known for its simplicity and readability, which makes it easy to learn and use. It has a large standard library and a wide range of third-party packages available, which makes it easy to find and use modules for many different purposes.

22

**RESULTS:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Linear regression** | **Decision Tree** | **SVM** | **KNN** | **Random Forest** |
| **MEAN SQUARE ERROR** | 20.186750203222648 | 12.021739130434783 | 24.142612927396755 | 25.623478260869575 | 6.253336956521742 |
| **MEAN ABSOULTE ERROR** | 3.6098028860602294 | 2.108695652173913 | 3.6699528174496265 | 3.2913043478260877 | 1.6089130434782613 |
| **ROOT MEAN SQUARE ERROR** | 4.492966748510682 | 3.467237968532703 | 4.9135132977734735 | 5.061963873919842 | 2.50066730224589 |

**By the above table we say that mean square error is smaller for random forest so, random forest is the best model for air prediction.**

23

**CONCLUSION AND FUTURE SCOPE**

Overall, air quality prediction is a critical tool for managing and reducing the impact of air pollution on human health and the environment. By using accurate air quality predictions, individuals and organizations can take proactive steps to reduce their exposure to harmful pollutants and contribute to a cleaner, healthier environment.

Air quality provides standards and objective for key air pollutants ,which are designed to protect human health and the environment.

Air pollution can cause both short term and long term effects on health and many people are concerned about pollution in the air that they breathe.

REFERENCES:

https://www.kaggle.com/code/kerneler/starter-uci-air-quality-b68c5049-e

24