Prediction of Air Quality Index for Mumbai Region: A Machine Learning Approach

MINI PROJECT

SUBMITTED TO

CENTRAL UNIVERSITY OF SOUTH BIHAR

in Partial Fulfilment of The Requirement OF THE DEGREE OF MASTER'S

IN Data science & Applied statistics

Under Supervision of

DR. INDRAJEET KUMAR

Assistant Professor

Department of Statistics

Central University of South Bihar, Gaya

Submitted by:

RAHUL KUMAR MAHATO

CUSB2302222005

Masters in Data Science & Applied Statistics

Department of Statistics

SCHOOL OF MATHEMATICS, STATISTICS AND COMUTER SCIENCE

DEPARTMENT OF STATISTICS
CENTRAL UNIVERSITY OF SOUTH BIHAR
GAYA - 824236

Declaration

I hereby declare that the work presented in this mini-project, titled **"Prediction of Air Quality Index for the Mumbai Region: A Machine Learning Approach,"** is a genuine and original contribution carried out by me as part of the partial fulfillment of the requirements for the award of the degree **Master's in Data Science & Applied Statistics.** This MINI PROJECT was conducted during my III semester under the guidance and supervision of **Dr. Indrajeet Kumar, Assistant Professor, Department of Statistics, Central University of South Bihar.**

I affirm that the content of this dissertation is entirely my own work and has not been submitted, either in part or in full, for the award of any other degree or diploma from this or any other university or institute. This work reflects my academic endeavors and dedication to advancing the field of data science.

Signature

(Rahul Kumar Mahato)

दक्षिण बिहार केन्द्रीय विश्वविद्यालय
Central University of South Bihar
Department of Statistics
School of Mathematics, Statistics, and Computer Science
SH-7, Gaya Panchanpur Road, Village - Karhara, Post. Fatehpur, Gaya-824236
(BIHAR) Phone/Fax :0631-2229530 2229514, Website:www.cush.ac.in

CERTIFICATE

This is to certify that his project entitled "**"Prediction of Air Quality Index for the Mumbai Region: A Machine Learning Approach",** enrolment number CUSB2302222005, in partial fulfilment for award of Master's in Data Science & Applied Statistics of Central University of South Bihar. This work has not been submitted in part or full, to this or any other university or institution, for any degree or diploma.

Date: ……..

Signature
(DR. Indrajeet Kumar)
Supervisor

Date: ……..

Signature
(DR. Sunit Kumar)
Head of the Department

Acknowledgement

I would like to express my heartfelt gratitude to my supervisor, **Dr. Indrajeet Kumar, Assistant Professor, Department of Statistics, Central University of South Bihar,** for his invaluable guidance, unwavering encouragement, and insightful feedback throughout the course of my project work. His mentorship has been instrumental in shaping my approach to research and fostering my interest in the interdisciplinary areas of data science and applied statistics. I am especially grateful for the freedom he granted me to explore my ideas and the essential skills and techniques he imparted during this journey.

I extend my sincere thanks to Dr. Sunit kumar (Head of the Department) , Dr. Kamlesh Kumar and Dr. Sandeep Kumar Maurya for the thought-provoking and enriching courses they taught as part of my Master's program. Their teachings significantly contributed to building my foundational knowledge, which was crucial for pursuing this dissertation. I am also deeply thankful for their constant encouragement and inspiration throughout my academic journey.

I am grateful to all the faculty members of the Department of Statistics, Central University of South Bihar**,** for their support and for creating an intellectually stimulating and collaborative environment. I also appreciate the peaceful and resourceful atmosphere provided by the university, which enabled me to concentrate fully on my project work.

My heartfelt thanks go to the Ph.D. scholars of our department**,** Mrs. Shivani Kumari and Mr. Dhananjay Kumar**,** for their valuable advice and assistance. They have been instrumental in teaching us how to be effective visualizers and uncover meaningful patterns in data through tools like Tableau. Their guidance in data visualization and interpretation has significantly enriched my understanding and enhanced my ability to communicate insights effectively.

Lastly, I bow in gratitude to the Almighty God for granting me the strength, perseverance, and blessings to successfully complete this work.

Rahul Kumar Mahato

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Mumbai, the financial capital of India, stands as a symbol of ambition, resilience, and growth. With an estimated population of 12.5 million, it is not only a hub for commerce but also one of the most densely populated cities in the world. However, rapid urbanization, industrial expansion, and population growth have posed significant environmental challenges, with air pollution emerging as a critical concern.

Vehicular emissions, industrial activities, and construction dust have contributed to the deteriorating air quality, despite numerous efforts to control pollution. Poor air quality has far-reaching consequences, impacting public health, economic productivity, and the overall livability of the city. Moreover, as one of India's busiest cities for tourism and business, Mumbai's air pollution can leave a lasting negative impression on visitors while affecting the quality of life for its residents.

Motivated by these challenges, this project leverages the potential of data science and machine learning to predict the Air Quality Index (AQI). By utilizing advanced algorithms, the study aims to enable early interventions, effective pollution control, and better public health management. This endeavor is rooted in the belief that predictive analytics can play a vital role in addressing Mumbai's air pollution crisis and making the city more livable for everyone.

1.2 Mumbai Air Pollution

Spanning an area of approximately 603.4 km², Mumbai is nestled along the western coast of India, located between the latitudes 18°55' N and 19°4' N, and longitudes 72°45' E and 73° E. As India's most populous city and a bustling urban agglomeration, Mumbai attracts millions to its shores annually for trade, culture, and opportunities. However, this growth comes with its price—severe air pollution that has become a growing public health crisis affecting millions of lives daily.

The city's pollution problem is fueled by a combination of factors, including vehicular emissions, industrial discharges, waste burning, and construction dust. These diverse sources of pollution make it challenging to identify specific causes and implement effective solutions.

Mumbai boasts a network of air quality monitoring stations managed by the Indian Institute of Tropical Meteorology (IITM) and the Maharashtra Pollution Control Board (MPCB). These stations, located in regions such as Bandra Kurla Complex, Deonar, and Colaba, provide valuable data to analyze pollution trends and identify hotspots. Despite these efforts, achieving sustainable improvements in air quality requires a collective and coordinated effort involving the government, industries, and citizens.

1.2.1 What's Polluting Mumbai's Air?

Air pollution in Mumbai is a complex issue driven by multiple sources such as vehicular emissions, industrial discharges, construction dust, and waste burning. Despite its significance, understanding the full extent of the problem remains challenging due to the varied nature of pollution sources and their widespread impact across the city.

1.2.2 The Challenges of Monitoring Air Pollution

To monitor and analyze air quality, Mumbai relies on a network of air quality monitoring stations managed by the Indian Institute of Tropical Meteorology (IITM) and the Maharashtra Pollution Control Board (MPCB). These stations provide valuable data that help identify pollution patterns and hotspots, enabling targeted interventions. Below are some of the key monitoring stations pivotal to this study:

- **Bandra Kurla Complex (IITM) :**
  A bustling commercial hub characterized by high traffic congestion and continuous construction activities.
- **Borivali East (IITM & MPCB) :**
  A residential area surrounded by green spaces but experiencing increasing pollution due to urban sprawl.
- **Chakala, Andheri West (IITM) :**
  A commercial-residential area that sees heavy traffic flow throughout the day.
- **Deonar (IITM) :**
  Known for hosting one of the largest landfill sites, which contributes significantly to particulate pollution from waste burning.
- **Khindipada, Bhandup (IITM) :**
  A suburban locality affected by pollution from nearby industrial activities.

- **Malad West (IITM) :**
  A mixed-use area frequently impacted by vehicular emissions and construction dust.
- **Mazgaon (IITM) :**
  A densely populated zone with significant contributions from maritime and industrial activities.
- **Navinagar, Colaba (IITM) :**
  Situated in South Mumbai, this area faces emissions from shipping activities and urban congestion.
- **Siddharth Nagar, Worli (IITM) :**
  A mixed-use development area experiencing consistent traffic-related pollution.
- **Chhatrapati Shivaji International Airport (IITM) :**
  A major contributor to pollution due to aviation activities and surrounding vehicular traffic.
- **Colaba (MPCB) :**
  Renowned for its heritage sites but heavily affected by traffic and port-related emissions.
- **Kandivali East (MPCB) :**
  A suburban locality facing pollution challenges from rapid urban development.
- **Kurla (MPCB) :**
  A key industrial and transportation hub characterized by significant emissions from vehicles and factories.
- **Mulund West (MPCB) :**
  A blend of residential and industrial areas contributing to variable pollution levels.
- **Powai (MPCB) :**
  A prominent IT hub surrounded by heavy vehicular traffic.
- **Sion (MPCB) :**
  A transit-heavy area plagued by traffic congestion and vehicular emissions.
- **Vasai West (MPCB) :**
  Located on the city's outskirts, it experiences pollution from waste burning and industrial activities.

- **Vile Parle West (MPCB) :**
  A residential neighborhood impacted by airport-related emissions and vehicular congestion.
- **Worli (MPCB) :**
  A high-profile locality that nonetheless suffers from traffic-induced pollution.

These stations provide a detailed understanding of the ways that various parts of Mumbai both contribute to and are affected by air pollution, allowing for targeted study and useful insights.

1.3 Problem Statement

Air pollution is one of the most pressing issues in Mumbai today, driven by the city's rapid urbanization, dense population, and industrial activities. The lack of reliable predictions about air quality often hinders timely interventions and effective policymaking.

The complexity of Mumbai's urban environment demands an intelligent, localized approach to understanding and mitigating air pollution. While existing measures, such as stricter vehicular emission norms and green cover initiatives, provide some relief, they fall short of addressing the scale and dynamics of the problem. A robust, data-driven predictive model is necessary to fill this gap and provide real-time insights into the state of the city's air quality.

This project seeks to address these challenges by creating a machine learning-based AQI prediction tool tailored specifically for Mumbai.

1.4 Objectives

The primary objective of this project is to analyze and predict the AQI for Mumbai using advanced machine learning techniques. By identifying patterns in pollutant concentrations and AQI data, the study aims to forecast air quality levels with high accuracy.

The insights derived from this project will inform policy decisions, enhance pollution control measures, and ultimately contribute to improving public health in Mumbai. The model is designed to not only predict AQI but also provide actionable recommendations to reduce pollution and its impact on residents.

1.5 Scope of the Study (Condensed)

This study aims to develop a machine learning-based AQI prediction model for Mumbai to address escalating air pollution. Key aspects include:

1. **Air Quality Monitoring:**
   Utilizing historical data from key stations to analyze pollution trends and hotspots.
2. **Pollutant Analysis:**
   Assessing critical pollutants (PM2.5, PM10, NO2, SO2, O3, CO) affecting air quality.
3. **Predictive Modeling:**
   Building a robust AQI prediction model and benchmarking it against FB Prophet for accuracy.
4. **Actionable Insights:**
   Providing region-specific pollution severity insights for effective policy and urban planning.
5. **Strategic Planning:**
   Supporting traffic regulation, emission control, and waste management strategies.
6. **Community Awareness:**
   Delivering real-time AQI forecasts to promote public awareness and behavioral change.

This study addresses health, economic, and urban planning challenges, offering data-driven solutions for a sustainable and healthier Mumbai.

Chapter 2

Literature Review

This chapter provides a detailed exploration of the foundational concepts and methodologies pertinent to understanding and addressing air quality challenges. The discussion is structured into three main sections: an overview of AQI and its associated pollutants, the application of machine learning (ML) techniques in AQI prediction, and a review of related studies highlighting their implications for public health, policy-making, and urban planning. By synthesizing insights from the literature, this chapter aims to establish a solid knowledge base for analyzing Mumbai's air quality and developing predictive models.

2.1 An overview of AQI

Air Quality Index (AQI) Awareness of daily levels of air pollution is important to the citizens, especially for those who suffer from illnesses caused by exposure to air pollution. Further, success of a nation to improve air quality depends on the support of its citizens who are well-informed about local and national air pollution problems and about the progress of mitigation efforts. Thus, a simple yet effective communication of air quality is important. The concept of an air quality index (AQI) that transforms weighted values of individual air pollution related parameters (e.g. SO2, CO, visibility, etc.) into a single number or set of numbers is widely used for air quality communication and decision making in many countries. Air quality monitoring procedures and protocols, Indian National Air Quality Standards (INAQS) and dose-response relationships of pollutants, an AQI system is devised. The AQI system is based on maximum operator of a function (i.e. selecting the maximum of subindices of individual pollutants as an overall AQI). The objective of an AQI is to quickly disseminate air quality information (almost in real-time) that entails the system to account for pollutants which have short-term impacts. Eight parameters (PM10, PM25, NO2, SO2, CO, O3, NH3, and Pb) having short-term standards have been considered for near real-time dissemination of AQI. It is recognized that air concentrations of Pb are not known in real-time and cannot contribute to AQI. However, its consideration in AQI calculation of past days will help in scrutinizing the status of this important toxic.

The proposed index has six categories with elegant colour scheme, as shown below:

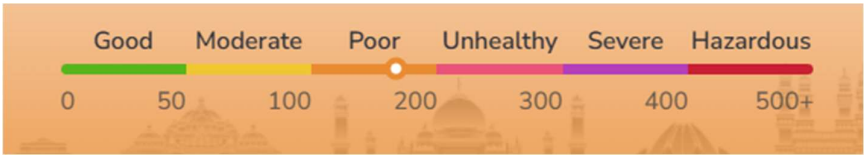

Figure 2.1: Air Quality Index India categories(aqi.in)

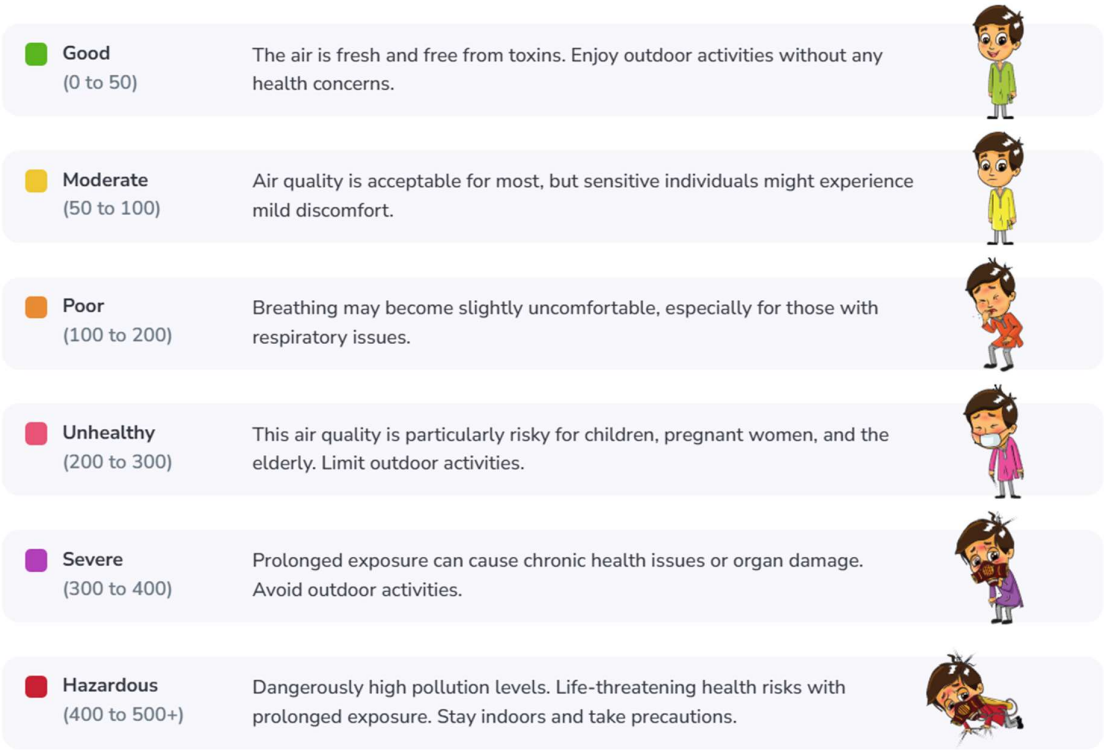

Figure 2.2 : Air Quality index category by Govt of India (aqi.in)

A scientific basis in terms of attainment of air quality standards and dose-response relationships of various pollutant parameters have been derived and used in arriving at breakpoint concentrations for each AQI category.

It is proposed that for continuous air quality stations, AQI is reported in near real-time for as many parameters as possible. For manual stations, the daily AQI is reported with a lag of one week to ensure manual data are scrutinized and available for AQI. AQIs must be identified if these are from continuous or manual station to maintain uniformity and clarity on sources of data. A web-based AQI dissemination system is developed for quick, simple and elegant looking response to an AQI query. The other features of the

website include reporting of pollutant responsible for index, pollutants exceeding the standards and health effects.

2.1.1 Origin and concept of Air Quality Index (AQI)

In addition to land and water, air is the prime resource for sustenance of life. With the technological advancements, a vast amount of data on ambient air quality is generated and used to establish the quality of air in different areas. The large monitoring data result is in encyclopaedic volumes of information that neither gives a clear picture to a decision maker nor to a common man who simply wants to know how good or bad the air is? One way to describe air quality is to report the concentrations of all pollutants with acceptable levels (standards). As the number of sampling stations and pollution pa rameters (and their sampling frequencies) increase, such descriptions of air quality tend to become confusing even for the scientific and technical community. As for the general public, they usually will not be satisfied with raw data, time series plots, statistical analyses, and other complex findings pertaining to air quality. The result is that people tend to lose interest and can neither appreciate the state of air quality nor the pollution mitigation efforts by regulatory agencies. Since awareness of daily levels of urban air pollution is important to those who suffer from illnesses caused by exposure to air pollution, the issue of air quality communication should be addressed in an effective manner. Further, the success of a nation to improve air quality depends on the support of its citizens who are well informed about local and national air pollution problems and about the progress of mitigation efforts. To address the above concerns, the concept of an Air Quality Index (AQI) has been developed and used effectively in many developed countries for over last three decades. An AQI is defined as an overall scheme that transforms weighted values of individual air pollution related parameters (SO2, CO, visibility, etc.) into a single number or set of numbers. There have not been significant efforts to develop and use AQI in India, pri marily due to the fact that a modest air quality monitoring programme was started only in 1984 and public awareness about air pollution was almost non existent. The challenge of communicating with the people in a comprehensible manner has two dimensions:

1.Translate the complex scientific and medical information into simple and precise knowledge and

2. Communicate with the citizens in the historical, current and futuristic sense. Ad dressing these challenges and thus developing an efficient and comprehensible AQI scale is required for citizens and policy makers to make decisions to prevent and minimize air pollution exposure and ailments induced from the exposure.

2.1.2 Applications of Air Quality Index

1. Resource Allocation: To assist administrators in allocating funds and determin ing priorities. Enable evaluation of trade-offs involved in alternative air pollution control strategies.

2. Ranking of Locations: To assist in comparing air quality conditions at different locations/cities. Thus, pointing out areas and frequencies of potential hazards.

3. Enforcement of Standards: To determine extent to which the legislative stan dards and existing criteria are being adhered. Also helps in identifying faulty stan dards and inadequate monitoring programs.

4. Trend Analysis: To determine change in air quality (degradation or improve ment) which have occurred over a specified period. This enables forecasting of air quality (i.e., tracking the behaviour of pollutants in air) and plan pollution control measures.

5. Public Information: To inform the public about environmental conditions (state of environment). It's useful for people who suffer from illness aggravated or caused by air pollution. Thus it enables them to modify their daily activities at times when they are informed of high pollution levels.

6. Scientific Research: As a means for reducing a large set of data to a comprehensible form that gives better insight to the researcher while conducting a study of some environmental phenomena. This enables more objective determination of the contribution of individual pollutants and sources to overall air quality. Such tools become more useful when used in conjunction with other sources such as local emission surveys.

2.1.3 Major pollutants For AQI

There are many pollutants and the critical air pollutants are:

1. Particulate Matter (PM) - one bin with all PM below 10 (μm) and one bin with all PM below 2.5 (μm)

2. Nitrogen oxides (NOx)

3. Ammonia (NH3)

4. Sulphur dioxide (SO2)

5. Carbon monoxide (CO)

6. Ozone (O3**)**

And on the other side, we have the greenhouse gasses (GHGs) like carbon dioxide (CO2) which also has an impact on health, but more decisively linked to climate change. We should never discuss all these criterion pollutants at once, because they are very different in their chemical nature and different in the ways they affect our health.

Only thing common for all the pollutants is that they originate from the same sourcesanything burnt will produce at least one of these pollutants or all of them (an extension to this is that if you implement controls for one pollutant, you are likely to control other pollutants as well a concept reffered to as co-benefits analysis). Sources also contribute differently to these pollutants, meaning a source attribution based on NOx emissions is not the same as a sources attribution based on PM or CO2.

So, of these, if we have to pick one pollutant that is critical for human health, then it is PM. Sometimes, this is also referred as dust, aerosol and soot. The chemical composition of PM2.5 has contributions from all the other gaseous components. For example, SO2 shows up as sulphate aerosols, NOx shows up as nitrate aerosols, volatile organic compounds (VOCs) after undergoing a series of chemical reactions with Ozone, NOx and CO, shows up as secondary organic aerosols (SOA). In simple language, focusing our efforts on PM2.5 to identify urban air pollution sources will be enough to address overall urban air quality scenario in Delhi, without mixing messages and discussing everything under the sun.

PM10 and PM2.5

PM10 is all aerosol under 10µm diameter, PM2.5 all aerosols under 2.5µm diameter. PM2.5 is a subset of PM10 and the ratio varies from city to city and sources to sources. The source apportionment studies conducted for PM2.5 and PM10 samples will result in different contribution charts.

Most of the PM2.5 pollution comes is combustion based.

For example, more than 95% of emission from diesel, petrol and natural gas combustion, open waste burning pollution, biomass burnig pollution, and coal combustion at cookstoves and boilers, falls under PM2.5. Most of the PM10 pollution comes from mechanical processes Like dust, on the roads due to the constant vehicular movement, at the construction sites and the seasonal dust storms. Close to 80% of the dust (that we commonly find on the roads) falls into the size fraction between PM2.5 and PM10. This is the main reason for finding more dust in a PM10 sample compared to a PM2.5 sample. Emission inventory is not pollution source attribution Emission is what comes out of the vehicle tailpipes, chimneys at the huts, industries and power plants trash burning and re-suspension of dust. This is commonly measures and reported as grams of pollutant emitted per kilometer of vehicle travel, grams of pollutant emitted per kilo of fuel burnt or sometimes grams of pollutant emitted per hours (mass over time).

Effect of Particulate Matter (PM)

Inhalation of particulate pollution can have adverse health impacts, and there is understood to be no safe threshold below which no adverse effects would be anticipated. The biggest impact of particulate air pollution on public health is understood to be from long-term exposure to PM2.5, which increases the age-specific mortality risk, particularly from cardiovascular causes. Several plausible mechanisms for this effect on mortality have been proposed, although it is not yet clear which is the most important. Exposure to high concentrations of PM (e.g. during short-term pollution episodes) can also exacerbate lung and heart conditions, significantly affecting quality of life, and increase deaths and hospital admissions. Children, the elderly and those with predisposed respiratory and cardiovascular disease, are known to be more susceptible to the health impacts from air pollution. Potential mechanisms by which air pollution could cause cardiovascular effects are described in the Committee on the Medical Effects of Air Pollution (COMEAP).

Nitrogen Oxides (NOx)

Nitrogen dioxide is an irritant gas, which at high concentrations causes inflammation of the airways. When nitrogen is released during fuel combustion it combines with oxygen atoms to create nitric oxide (NO). This further combines with oxygen to create nitrogen dioxide (NO2). Nitric oxide is not considered to be hazardous to health at typical ambient concentrations, but nitrogen dioxide can be. Nitrogen dioxide and nitric oxide are referred to together as oxides of nitrogen (NOx). NOx gases react to form smog and

acid rain as well as being central to the formation of fine particles (PM) and ground level ozone, both of which are associated with adverse health effects.

Sources of Nitrogen Oxides

Nitrogen oxides are produced in combustion processes, partly from nitrogen compounds in the fuel, but mostly by direct combination of atmospheric oxygen and nitrogen in flames. Nitrogen oxides are produced naturally by lightning, and also, to a small extent, by microbial processes in soils.

Emission Sources and Trends

An emission inventory has many advantages such as providing vital information on the level and assessment of Air pollution also planning for its control measures and tracking its progress. In this work emission inventories were built for NOx for the years 2008 and 2010 for urban part of Delhi. Mainly four sectors were selected; transport sector, power plants, domestic sector and waste burning for preparing the emission inventory. Most of the industrial and construction activity was mainly contributed from Gurgaon, Ghaziabad and Faridabad totalling nearly 8% of the total NOx emissions. Also the cement industries lie in the border area. Therefore we neglected these sectors as they lied outside our study domain and percentage contribution was also insignificant.

Atmospheric chemistry and transport

The primary pollutant, directly emitted, is nitric oxide (NO), together with a small proportion of nitrogen dioxide (NO2). NO is oxidised by ozone in the atmosphere, on a time scale of tens of minutes, to give NO2. In rural air, away from sources of NO, most of the nitrogen oxides in the atmosphere are in the form of NO2. NO and NO2 are collectively known as NOx because they are rapidly inter-converted during the day. NO2 is split up by UV light to give NO and an O atom, which combines with molecular oxygen (O2) to give ozone (O3). Therefore, during the day NO, NO2 and ozone exist in a quasiequilibrium which depends on the amount of sunlight. Eventually, NO2 is oxidised to nitric acid (HNO3, vapour) which is absorbed directly at the ground, is converted into nitrate-containing particles, or dissolves in cloud droplets. At night, different oxidation processes convert NO2 to nitrates. Although nitric acid is rapidly absorbed on contact with surfaces (cloud droplets, soil or vegetation), the other nitrogen

oxides are removed only rather slowly, and may travel many hundreds of km before their eventual conversion to nitric acid or nitrates. Consequently, emissions in one country will be deposited in others. Measured NO2 concentrations show the predominance of traffic and urban sources, with the largest concentrations in the large conurbations and adjacent to the motor way network, with annual mean concentrations in excess of 10 ppb (parts per billion) in these areas.

Ecosystem Impacts

It is likely that the strongest effect of emissions of nitrogen oxides across the Delhi is through their contribution to total nitrogen deposition. However, direct effects of gaseous nitrogen oxides, may also be important, especially in areas close to sources (e.g. roadside verges). The critical level for all vegetation types from the effects of NOx has been set to 30μg/m3. Experimental evidence suggests that moderate concentrations of NOx may produce both positive and negative growth responses, with the potential for synergistic interactions with sulphur dioxide (SO2) being very important. There is substantial evidence to suggest that the effects of NO2 are much more likely to be negative in the presence of equivalent concentrations of SO2. One important effect of NOx may be its influence on insect populations; there is evidence of improved performance of insect pests on plants grown in moderate concentrations of NO2 and SO2.

Ammonia (NH3)

Ammonia (NH3) is a highly reactive and soluble alkaline gas. It originates from both natural and anthropogenic sources, with the main source being agriculture, e.g. manures, slurries and fertiliser application.

Source of Ammonia (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. Other agriculture-related emissions of ammonia include biomass burning or fertiliser manufacture. Ammonia is also emitted from a range of non-agricultural sources, such as catalytic converters in petrol cars, landfill sites, sewage works, composting of organic materials, combustion, industry and wild mammals and birds.

Effect of Ammonia (NH3)

In addition to the impact of NH3 on human health and the environment from aerosol formation, the contribution of anthropogenic NH3 to reactive nitrogen deposition in the form of NH3 and NH_3^+ (and associated NO_3^-) impacts the nitrogen cascade and poses a threat to sensitive ecosystems. Excessive deposition of NH3 causes eutrophication in surface water and soil acidification and can further cause nutrient imbalances in sensitive ecosystems. Levels of reactive nitrogen deposition exceed those deemed critical for the protection of biodiversity in many regions throughout the region, now or in the next few decades, underscoring concerns not only of past but also future impacts of anthropogenic activity on the global nitrogen cycle. Despite the recognized importance of this issue, uncertainties in our ability to model such processes may hinder efforts in the region to develop secondary air quality standards to protect ecosystems from the hazards of this type of pollutant deposition.

Sulphur Dioxides (SO2)

Sulphur dioxide is a colour less gas with a strong, pungent smell. It dissolves very easily in water. This gas can be dissolved in the environment through air and water both which could be converted into sulphuric acid while in the air and sulphuric dioxide while in water.

Sources of Sulphur Dioxides

About 99% of the sulfur dioxide in air comes from human sources. The main source of sulfur dioxide in the air is industrial activity that processes materials that contain sulfur, that is the generation of electricity from coal, oil or gas that contains sulfur. Some mineral ores also contain sulfur, and sulfur dioxide is released when they are processed. In addition, industrial activities that burn fossil fuels containing sulfur can be important sources of sulfur dioxide. Sulfur dioxide is also present in motor vehicle emissions, as the result of fuel combustion. In the past, motor vehicle exhaust was an important, but not the main, source of sulfur dioxide in air. However, this is no longer the case.

Sulphur Dioxide and Health Effects

Exposure to sulphur dioxide through inhaling it is extremely dangerous and life-threatening as it makes the lung functioning changes and causes nasal and throat difficulties and severe airway obstructions. It also irritates the skin and mucous membranes of the eyes, nose, throat, and lungs. High concentrations of SO2 can cause inflammation and irritation of the respiratory system, particularly during heavy physical activity. There are also other several health-related problems that are caused of Sulfur Dioxide:-

1. Corneal Haze, which is the creation of cloudy appearance in the cornea.
2. Breathing difficulty is caused.
3. Airway inflammation causing a lot burning in the airway.
4. Heart Failure.
5. Causing higher mortality rates

Carbon Monoxide (CO)

Carbon monoxide (CO) is a colorless, odorless gas that is a product of the incomplete combustion of carbon-based fuels including gasoline, diesel fuel, crude oil, and wood and other natural and synthetic products. Main contributors of carbon monoxide emissions include vehicle exhaust from cars, trucks, buses, gas-powered furnaces and portable generators.

Effects of Carbon Monoxide (CO)

Breathing air with a high concentration of CO reduces the amount of oxygen that can be transported in the blood stream to critical organs like the heart and brain. At very high levels, which are possible indoors or in other enclosed environments, CO can cause dizziness, confusion, unconsciousness and death. Very high levels of CO are not likely to occur outdoors. However, when CO levels are elevated outdoors, they can be of particular concern for people with some types of heart disease. These people already have a reduced ability for getting oxygenated blood to their hearts in situations where the heart needs more oxygen than usual. They are especially vulnerable to the effects of CO when exercising or under increased stress. In these situations, short-term exposure to elevated CO may result in reduced oxygen to the heart accompanied by chest pain also known as angina.

Ozone (O3)

Ozone (O3) is present throughout the atmosphere although there are concentration peaks at two levels, the stratosphere (15 - 50 km) and troposphere (0-15 km), with the largest fraction and concentrations being in the stratospheric O3 layer . Stratospheric O3 is important as it regulates the transmittance of ultraviolet light to the surface of the earth. Hence reductions in stratospheric O3 in polar regions, particularly the Antarctic "ozone hole", are of concern regarding the health effects of exposure to increased levels of UV-B. In contrast, O3 in the troposphere (ground level) is regionally important as a toxic air pollutant and greenhouse gas. Mixing with stratospheric air provides a natural global average background of around 10-20 parts per billion (ppb), though there is some debate about the concentration. Additional quantities of tropospheric O3 are produced by photochemical reactions from nitrogen oxides (NOx) and volatile organic compounds (VOCs), which include various hydrocarbons.

Source of Ozone (O3)

Ground-level ozone (O3) is not emitted directly from anthropogenic sources. It is a secondary pollutant formed by a complicated series of chemical reactions in the presence of sunlight. Photochemical reactions of NOx and VOCs (originating from largely from combustion processes) govern the concentration of ground-level O3 in the atmosphere. The chemical reactions do not take place instantaneously, but can take hours or days. Ozone levels at a particular location may have arisen from VOC and NOx emissions many hundreds or even thousands of miles away. Maximum concentrations, therefore, generally occur down wind of the source areas of the precursor pollutant emissions. Anthropogenic emissions of the ozone precursors (NOx/VOCs) can also cause large transient increases in ozone concentration, termed episodes or smog. These occur when high concentrations of precursors coincide with weather conditions favorable for ozone production such as when the air is warm and slow moving. These "ozone episodes" provide concentrations of O3 (>40 ppb) which are toxic both to human health and vegetation. Prior to the industrial revolution natural sources of NOx and VOCs would have generated O3 in the troposphere, adding to that transported from the stratosphere. However, the large amounts of O3 and VOCs released by human activities, such as the combustion of fossil fuels, has led to a large increase in the northern hemisphere background concentration.

2.1.4 Indian Air Quality Index system

AQI Category (Range)	PM_{10} 24-hr	$PM_{2.5}$ 24-hr	NO_2 24-hr	O_3 8-hr	CO 8-hr (mg/m^3)	SO_2 24-hr	NH_3 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.6 –1.0
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1- 10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748*	17.1-34	801-1600	1201-1800	3.1-3.5
Severe (401-500)	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+

**One hourly monitoring (for mathematical calculation only)*

Figure 2.3: Annual Standards of Parameters(aqi.in)

Air quality standards are the basic foundation that provides a legal framework for air pollution control. An air quality standard is a description of a level of air quality that is adopted by a regulatory authority as enforceable. The basis of development of standards is to provide a rational for protecting public health from adverse effects of air pollutants, to eliminate or reduce exposure to hazardous air pollutants, and to guide national/ local authorities for pollution control decisions. With these objectives, CPCB notified (http://www.cpcb.nic.in) a new set of Indian National Air Quality Standards (INAQS) for 12 parameters [carbon monoxide (CO) nitrogen dioxide (NO2), sulphur dioxide (SO2), particulate matter (PM) of less than 2.5microns size (PM2.5), PM of less than 10 microns size (PM10), Ozone (O3), Lead (Pb), Ammonia (NH3),Benzo(a)Pyrene (BaP), Benzene (C6H6), Arsenic (As), and Nickel (Ni)] . The first eight parameters have short-term (1/8/24 hrs) and annual standards (except for CO and O3) and rest four parameters have only annual standards.

The objective of an AQI is to quickly disseminate air quality information (almost in real time) that entails the system to account for pollutants which have short-term impacts. It is equally important that most of these pollutants are measured continuously through an

online monitoring network. Consequently, in the proposed AQI system, the following pollutants are considered PM10, PM25, NO2, SO2, CO, O3, NH3, and Pb.

It is recognized that air concentrations of Pb are not known in real-time and cannot contribute to real-time AQI but its consideration in AQI calculation of past days will help in scrutinizing the status of this important toxic.

2.2 Machine Learning Applications in AQI Prediction

why Machine learning is a suitable approach for AQI prediction and highlight the common techniques employed in this field ?

Why Machine Learning for AQI Prediction?

- **Complex Relationships:** Air pollution is influenced by numerous factors (meteorological conditions, traffic, industrial emissions, etc.) with complex, non-linear relationships. ML algorithms are well-suited to capture these intricate patterns that traditional statistical methods might struggle with.
- **Large Datasets:** Air quality monitoring generates vast amounts of data. ML algorithms excel at processing and extracting meaningful information from large datasets.
- **Predictive Power:** ML models can learn from historical data to predict future AQI levels, enabling timely interventions and public health advisories.
- **Adaptability:** ML models can be retrained and updated with new data to improve their accuracy and adapt to changing environmental conditions.

2.3 Related work and Studies

This section consists of a brief overview of previous studies related to the prediction of AQI. The techniques related to predicting AQI have been studied and reviewed. The machine learning approaches were examined to demonstrate why these approaches are capable to perform well in forecasting air quality.

2.3.1 Air Quality Index Predictive Models

An predictive air quality model is a computation tool which predicts the AQI. AQI can give complete description of the air quality. Regression algorithms are the most usual supervised machine learning algorithms that were used in predicting the AQI.

2.3.2 Regression Models

For predicting the AQI, both linear regression and non linear regression models have been used. Linear regression model attempts to establish a relation between the in dependent variables and dependent variable. The common usage of linear regression model is to know about the linear relationship between predictors and a predictand.

P Arulmozhivarman et al. (2017) presented multiple regression models for predicting AQI [4]. They implemented regression models like SVR and multiple linear regression model for forecasting AQI [1]. Among those, SVR exhibited high level of per formance. They considered statistical criteria like MAE, Mean absolute percentange error (MAPE), Correlation coefficient (R), RMSE, and Index of agreement (IA) for evaluating the performance of regression models. In our study we considered some of those statistical metrics for evaluating the performance.

Huixiang Liu et al. (2019) presented a research paper which used regression models for the prediction of air quality [2]. They implemented machine learning models such as SVR and Random forest regression (RFR) for the forecasting of AQI. Between the two models, RFR model performed better than SVR model because the time com plexity of SVR has increased cubically with the increasing number of samples. They used performance metrics like RMSE, correlation coefficient (R), and coefficient of determination (R2).In our research, we have considered some the above mentioned metrics for performance evaluation.

In the research done by Soubhik Mahanta et al. (2019), AQI is predicted using various regression models in the sklearn library [3]. They used models like Linear regression, Neural network regression, LASSO regression, ElasticNet, Decision forest, Extra trees, Boosted decision trees, XGBoost, KNN, and Ridge regression for the prediction of AQI. Out of all the regression models, Extra trees model had the high est accuracy and the

least RMSE. They used statistical criteria like Accuracy and RMSE. In our research, we used RMSE as one our performance metrics for evaluation.

Anikender Kumar and Pramila Goyal (2011) has presented a research paper in which Air quality of Delhi is forecasted [4]. They used regression models like Principal Component Regressor (PCR) and Multiple Linear Regressor (MLR) to predict the AQI. Their study have been made for four seasons namely, Summer, monsoon, post monsoon, and winter. The statistical analysis of the model showed that it performed well in winter. Co-linearity, which has been found in MLR, has been eliminated with the use of Principal Components (PC). It was also found that the performance of PCR was better than MLR. They used Normalised Mean Square Error (NMSE), RMSE, and coefficient of determination (R2) as the performance metrics for evaluation.

Mauro Castelli et al. (2020) presented a research paper in which they used SVR to predict the air quality in California [5]. The study is about the use of SVR to accurately calculate the AQI and the concentrations of the pollutants. The study has produced a highly suitable model for the hourly prediction of AQI, accurate concen trations of pollutants such as o3, co2, and so2, and the hourly atmospheric pollution in the area. They used MAE, Normalised Mean Square Error (NMSE), and RMSE as the performance evaluation metrics.

Bing-chun liu et al. (2017) presented a research paper in which they used multi dimensional collaborative SVR model. In this paper the data used was collabo rative multiple city air quality data [6]. As the data used was from more than one city, the training is complex that leads to increase in training time. The RMSE values for the training and testing data sets were less than 12, and they concluded that SVR is strong and is an efficient model for the prediction of AQI. They used the performance metrics such as RMSE and Mean ab- solute percentange erro (MAPE).

Jasleen Kaur Sethi & Mamta Mittal (2021) presented a research paper in which they used a feature selection method named Correlation based Adaptive LASSO regression

method for air quality index prediction [7]. In this study the model eval uation depicts that the feature subset extracted by proposed model performs better than subset extracted by LASSO with an average classification accuracy of 70 per cent.

Chenchen Li et al. (2021) presented a research paper in which they used ridge regression, k nearest neighbour regression, decision tree regression, random forest regression and gradient boosting regression and other supervised machine learning algorithms for predicting the air quality index [8]. They concluded that random forest regression and gradient boosting regression performed better in predicting the air quality index.

Chapter 3

Methodology

Data Collection and Preprocessing

Accurate air quality prediction relies heavily on the availability of high-quality data. This chapter details the process of acquiring, cleaning, and preparing the air quality and meteorological data used in this study. The raw data, sourced from [Central Pollution Control Board (CPCB)] [9], presented several challenges, including missing values due to sensor malfunctions or data transmission issues. To address these challenges and ensure data integrity, a comprehensive preprocessing strategy was implemented. This involved a multi-stage imputation approach to handle missing data, normalization to standardize feature scales, visualization to understand variable relationships, and multicollinearity analysis to mitigate potential model instability. This chapter outlines the data sources, describes the preprocessing techniques used, and justifies the choices made to ensure data integrity and suitability for machine learning algorithms.

3.1 Overview of Data Collection

Data collection for this project involved gathering historical air quality data specific to the Mumbai region from the Central Pollution Control Board (CPCB) of India. The CPCB is the primary government body responsible for monitoring and regulating air quality across the country. Data was accessed through the CPCB's [9] online portal, specifically the "Central Control Room for Air Quality Management - All India" [10] under the "Data Repository" section. This portal provides access to a comprehensive database of air quality measurements collected from various monitoring stations across India. The data collected encompassed key air pollutants and meteorological parameters relevant to air quality analysis.

3.1.1 Data

We've collected data set from Central pollution control board (CPCB) [9]official web site (https://cpcb.nic.in/) [10]. It is government of India's official portal. The data set

contains air quality data and AQI of various cities in India. The raw data set contains 21,900 rows, 11 columns . The data set contains following variables particulate matter 2.5, particulate matter 10, nitric oxide, nitric dioxide, nitric x-oxide, ammonia, carbon monoxide, sulphur dioxide. The sample of "air quality data in India (2021- 2023) data set is shown below.

To obtain the data for this analysis, the following parameters were selected:

- **Data Type:** Raw data (Raw)
- **Frequency:** 1-day intervals
- **State:** Maharashtra
- **City:** Mumbai
- **Monitoring Stations:** All available monitoring stations within Mumbai
- **Time Period :** 2021-2023

	Timestamp	Station	PM2.5 (µg/m³)	PM10 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	NOx (ppb)	NH3 (µg/m³)	SO2 (µg/m³)	CO (mg/m³)	Ozone (µg/m³)
0	01-01-2021	Bandra_kurla_complex_mumbai	194.9900	228.7000	25.1800	66.3900	54.8700	82.8900	9.8100	0.9100	19.6900
1	01-02-2021	Bandra_kurla_complex_mumbai	144.5400	212.4700	9.3800	32.3500	40.4200	65.5300	2.7400	0.5400	10.0500
2	01-03-2021	Bandra_kurla_complex_mumbai	103.2100	187.7100	12.1400	32.0500	41.8100	65.0500	4.8700	0.7200	17.1700
3	01-04-2021	Bandra_kurla_complex_mumbai	123.1000	232.7400	9.4700	35.6700	45.1100	69.0600	2.5500	0.6000	7.3800
4	01-05-2021	Bandra_kurla_complex_mumbai	99.0100	192.7700	8.0200	38.2800	46.1900	69.1300	1.3900	0.4100	6.7600

Figure 3.1: Sample of air quality data for Mumbai region [10] (2021- 2023)

These selections provided a comprehensive dataset of daily air quality measurements from multiple monitoring stations in Mumbai. The dataset includes key pollutants and meteorological parameters necessary for analyzing and predicting the Air Quality Index (AQI) in the city.

3.2 Raw Data Summary

The dataset contains 21,900 rows and 11 columns.

Timestamp: 2021 – 2023

Station: Unique stations- 20

Air Quality Parameters (all in their respective units – µg/m³):

- PM2.5, PM10, NO, NO2, NOx, NH_3, SO_2, CO, Ozone

3.3 Data preprocessing

Initially the data set contains noisy, inconsistent data and missing values. The data has to be preprocessed to remove the unwanted data and to make the data useful. Data preprocessing helps to transform data into useful format. The following steps were involved in data preprocessing.

1) Data cleaning
2) Data reduction
3) Data transformation

Data cleaning: Data cleaning is the process of removing unwanted data like in correct data, duplicate data, unformatted data from the data set. By cleaning the data, we can improve the accuracy of the result. The following steps were involved while cleaning the data.

• **Remove duplicates:** When the data is collected from different sources, there will be a high possibility for data to have duplicated entries. These duplicates will create confusion in results. It is advisable to remove those duplicates to improve our results.

• **Remove irrelevant data:** Performing analysis on irrelevant data slows down the process as it wasn't useful. For example, if we only need particulate matter concentration for analysis then we've to exclude other components as they're irrelevant to our analysis to save time.

• **Handling missing values:** We can handle missing data by removing the entire tuple of data or else by filling the missing values in it. We can place Chapter 3.3 approximate value in the missing field. If the data is too large then we can remove the tuple data that has missing values or our data is sensitive to loose can replace it by mean , median & 0.

• **Clear formatting:** If the data is formatted heavily, machine learning models can't process the information. If we have different formats in our data it will be confusing.

• **Convert data types:** Sometimes numbers will be inputted as text. Then the data type of those numbers will be string. As they were strings, we can't perform mathematical operations on them. So we have to convert the data types to perform operations on them.

Data transformation: Data is transformed to improve it's structure. It enables better data driven decision making. It is the process of changing the format, structure (or) values of data. It involves normalization, attribute selection, discretization, concept hierarchy generation.

Data reduction: When working with huge amounts of data, analysis becomes more challenging. To handle this we use data reduction technique. Data reduction focuses on enhancing the storage efficiency while minimizing data storage and anal ysis cost. The various steps involved in data reduction were data cube aggregation, attribute subset selection, numerosity reduction, dimensionality reduction.

Data Quality Assessment and Handling Missing Values :

Data Quality and Missing Values: The initial dataset contained missing values across several key pollutants and meteorological parameters, as summarized below:

- **Timestamp:** 0 missing values
- **Station:** 0 missing values
- **PM2.5 (µg/m³):** 2,347 missing values
- **PM10 (µg/m³):** 2,374 missing values
- **NO (µg/m³):** 2,453 missing values
- **NO2 (µg/m³):** 2,342 missing values
- **NOx (ppb):** 2,328 missing values
- **NH3 (µg/m³):** 2,565 missing values
- **SO2 (µg/m³):** 2,916 missing values
- **CO (mg/m³):** 2,346 missing values
- **Ozone (µg/m³):** 2,907 missing values

Addressing these missing values is essential for maintaining data integrity and ensuring the accuracy of the Air Quality Index (AQI) predictions.

	Timestamp	Station	PM2.5 (µg/m³)	PM10 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	NOx (ppb)	NH3 (µg/m³)	SO2 (µg/m³)	CO (mg/m³)	Ozone (µg/m³)
0	0	0	2347	2374	2453	2342	2328	2565	2916	2346	2907

Figure 3.2: Missing values in raw air quality data

3.3.1 Imputation Using Mean for Missing Values

Why mean ?

The mean is a good choice for imputing missing values in AQI data for several reasons:

- **Preserves the overall level of pollution:** The mean is a measure of central tendency that represents the average value of a dataset. By using the mean to fill in missing values, you are essentially preserving the overall level of pollution in the data. This is important because it helps to maintain the integrity of the data and avoid introducing bias.
- **Minimizes bias:** The mean is less susceptible to bias from outliers than other measures of central tendency, such as the median or mode. This is because the mean takes into account all of the values in the dataset, while the median and mode only consider the middle value and the most common value, respectively.
- **Suitable for continuous data:** AQI data, such as PM2.5 and PM10 concentrations, is typically continuous data. The mean is a good choice for imputing missing values in continuous data because it is based on the assumption that the missing values are likely to be similar to the observed values.

Data Cleaning Process: A Hierarchical Imputation Approach

To address missing values in the air quality dataset, a hierarchical imputation strategy was employed. This approach leverages temporal patterns and spatial similarities to fill in gaps in the data while preserving the underlying structure.

Step 1: Quarterly Grouping and Imputation

1. Granular Time Periods: The data was initially grouped by **Station**, **Year**, and Quarter to capture seasonal variations and local trends.
2. Imputation with Similar Stations: Missing values within each quarter were filled using the mean values of similar stations during the same quarter. This approach accounts for both spatial and temporal patterns.

	Timestamp	Station	PM2.5 (µg/m³)	PM10 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	NOx (ppb)	NH3 (µg/m³)	SO2 (µg/m³)	CO (mg/m³)	Ozone (µg/m³)
0	0	0	367	365	184	184	184	184	1187	275	822

Figure 3.3 : Missing values in after Quarterly Grouping and Imputation

Step 2: Semester Imputation

For remaining missing values after the quarterly imputation:

1. Data Grouping: The data was grouped by **Station** and **Year** into six-month periods (semesters).
2. **Imputation:** Missing values within each semester were filled using the mean values of the same station during that semester. This approach leverages the temporal pattern within a year to infer missing values.

	Timestamp	Station	PM2.5 (µg/m³)	PM10 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	NOx (ppb)	NH3 (µg/m³)	SO2 (µg/m³)	CO (mg/m³)	Ozone (µg/m³)
0	0	0	0	181	0	0	0	0	1095	0	546

Figure 3.4: Missing values in after Semester Imputation

Step 3 : 9-Month Period Grouping and Imputation

1. Wider Time Window: To address remaining missing values, a broader 9-month period was considered. This wider window allows for more data points to be used for imputation, especially in cases where quarterly data was insufficient.
2. Hierarchical Grouping: The data was grouped by Station and Year to calculate 9-month averages. This hierarchical approach ensures that both local and temporal patterns are preserved.
3. Imputation: Missing values were filled using the calculated 9-month averages.

	Timestamp	Station	PM2.5 (µg/m³)	PM10 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	NOx (ppb)	NH3 (µg/m³)	SO2 (µg/m³)	CO (mg/m³)	Ozone (µg/m³)
0	0	0	0	0	0	0	0	0	1095	0	0

Figure 3.4.1: Missing values in after 9-Month Period Grouping and Imputation

Results and Limitations:

- The hierarchical imputation process was successful in filling most missing values for various pollutants.
- However, persistent missing values in the SO2 parameter indicate systematic gaps in measurements during the years 2021, 2022, and 2023. This suggests that a

different approach, such as interpolation or model-based imputation, may be necessary for this specific pollutant.

Overall, the hierarchical imputation strategy provides a robust and effective method for handling missing values in air quality datasets. By combining granular and broader time windows, the approach ensures data completeness while preserving essential temporal and spatial patterns.

3.3.2 AQI Calculation

After cleaning the dataset and addressing missing values, an additional column, Calculated AQI, was added to represent the Air Quality Index (AQI) for each day and station. The AQI was computed using the *AQI Calculator*

(version: aqi_calculator_v2-2021.10.03.xlsm), a tool that calculates AQI based on India's AQI breakpoint information provided by the *Central Pollution Control Board (CPCB)*.

The AQI in India is computed based on predefined concentration breakpoints for specific pollutants, including PM2.5, PM10, NO_2, SO_2, CO, and O_3. Each pollutant's concentration level is matched to an AQI category, which reflects its potential health impact. The AQI Calculator automatically assigns an AQI value by comparing pollutant levels to these breakpoints and selecting the highest contributing pollutant's AQI as the final AQI for that day and location.

This **Calculated AQI** column is crucial for the project's analysis, as it provides a standardized measure of air quality that can be used to assess trends, make predictions, and evaluate the impact of various pollutants.

After this AQI calculation process, an **AQI Category** column was created. The AQI Category feature was generated by applying specific conditions based on the AQI values, defined as follows:

- **Good:** AQI values from 0 to 50
- **Moderate:** AQI values from 51 to 100
- **Poor:** AQI values from 101 to 200
- **Unhealthy:** AQI values from 201 to 300
- **Severe:** AQI values from 301 to 400

- **Hazardous:** AQI values from 401 to 500

These conditions were implemented using the **pd.cut()** function, which categorizes the AQI values into the defined bins and assigns the corresponding labels.

The updated dataframe now includes the **AQI Category** column, which reflects these categorizations. Here are the first few rows of the updated dataframe:

3.3.3 Processed Data Summary

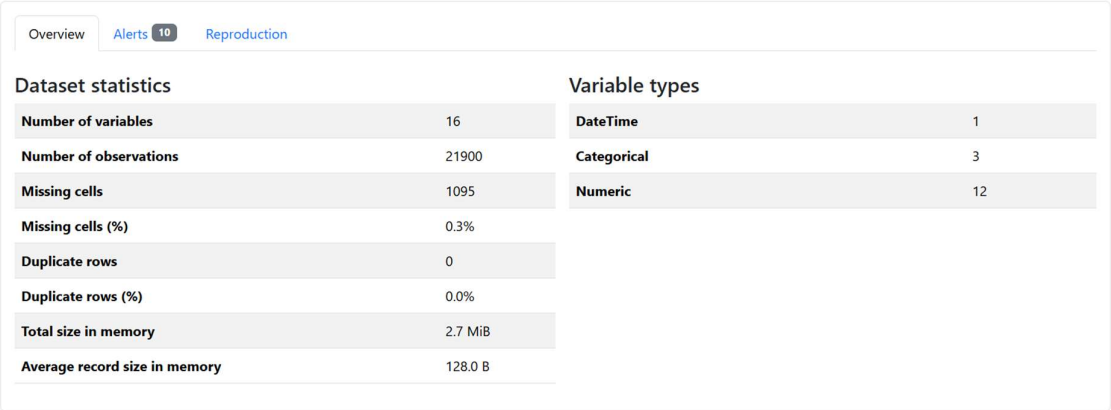

Figure 3.5 : Data Overview Using Pandas Profiling Report

Figure 3.6 : Dataset Quality Summary and Key Correlations

The figure displays a list of alerts generated from data quality analysis, categorized into high correlation, missing values, and uniform distributions. Several fields in the dataset exhibit strong relationships, indicating potential multicollinearity or overlapping information. These are:

- **AQI (Air Quality Index)**:
    - Highly correlated with AQI_Category and 2 other fields.

- **NO (µg/m³)** and **NOx (ppb)**:
    - Strongly correlated, reflecting their interdependence (NOx often includes NO).
- **NO2 (µg/m³)**:
    - Correlates with NOx (ppb) due to their chemical relationships.
- **PM2.5 (µg/m³)** and **PM10 (µg/m³)**:
    - Strongly correlated with AQI, which is expected because particulate matter contributes significantly to air quality.
- These correlations suggest that multiple variables may convey similar patterns and can be considered for dimensionality reduction or further modeling.

2. Missing Values:

- **SO2 (µg/m³)** has **5.0% missing values** (1,095 rows).
    - Missing values can affect the analysis or model performance if not handled appropriately.

3. Uniform Distributions:

- **Year** and **Station** fields are **uniformly distributed**:
    - This means data is evenly spread across these variables, and no particular year or station dominates.
    - Uniform distribution is typically useful for ensuring balanced data representation.

3.3.4 Data Normalization and Scaling

Why Normalize and Scale?

In machine learning, it's often necessary to normalize or scale features to ensure that they have a similar range. This is crucial for several reasons:

1. **Improved Model Performance:** Many machine learning algorithms, especially those that use distance-based calculations (like K-Nearest Neighbors or Support Vector Machines), perform better when features are on a similar scale.

2. **Fair Feature Comparison:** Normalization prevents features with larger ranges from dominating the learning process, allowing for a more equitable comparison of features.
3. **Faster Convergence:** Some optimization algorithms converge faster when features are scaled, leading to quicker training times.

Min-Max Scaling

Min-Max scaling is a technique that transforms features to a specific range, typically between 0 and 1. It's calculated as follows: **Scale** the data so that all features have comparable ranges, improving the performance of machine learning algorithms

To normalize and scale the features **PM2.5(Âµg/mÂ³) , PM10 (Âµg/mÂ³), SO2 (Âµg/mÂ³), NO2 (Âµg/mÂ³), CO (mg/mÂ³), Ozone (Âµg/mÂ³), NO (Âµg/mÂ³), NOx (ppb), NH3 (Âµg/mÂ³)**.

I will use ***Min-Max scaling***, which transforms the data to a range between 0 and 1. This is useful for ensuring that all features contribute equally to any analysis or modeling that follows.

Final Preprocessed Air Quality Index Data :

	Timestamp	Year	Month	Day	Station	AQI	AQI_Category	PM2.5 (µg/m³)	PM10 (µg/m³)	SO2 (µg/m³)	NO2 (µg/m³)	CO (mg/m³)	Ozone (µg/m³)	NO (µg/m³)	NOx (ppb)	NH3 (µg/m³)
0	01-Jan-21	2021	1	1	Bkcm	186	Poor	0.233682	0.231527	0.049385	0.188285	0.116368	0.117661	0.050652	0.110097	0.198734
1	02-Jan-21	2021	1	2	Bkcm	175	Poor	0.173163	0.215036	0.013757	0.091732	0.069054	0.060026	0.018856	0.081103	0.157107
2	03-Jan-21	2021	1	3	Bkcm	158	Poor	0.123583	0.189878	0.024491	0.090881	0.092072	0.102595	0.024410	0.083892	0.155956
3	04-Jan-21	2021	1	4	Bkcm	188	Poor	0.147443	0.235632	0.012800	0.101149	0.076726	0.044063	0.019037	0.090513	0.165572
4	05-Jan-21	2021	1	5	Bkcm	162	Poor	0.118545	0.195019	0.006954	0.108552	0.052430	0.040356	0.016119	0.092680	0.165739
5	06-Jan-21	2021	1	6	Bkcm	234	Unhealthy	0.159007	0.287524	0.031344	0.218068	0.074169	0.029116	0.025819	0.101870	0.179191
6	07-Jan-21	2021	1	7	Bkcm	286	Unhealthy	0.179892	0.340422	0.005997	0.094540	0.069054	0.017637	0.041073	0.107227	0.186841
7	08-Jan-21	2021	1	8	Bkcm	323	Severe	0.227289	0.373689	0.003679	0.084442	0.053708	0.017876	0.039020	0.098620	0.171902
8	09-Jan-21	2021	1	9	Bkcm	304	Severe	0.212605	0.358184	0.008416	0.089576	0.056266	0.018654	0.039765	0.102151	0.173413
9	10-Jan-21	2021	1	10	Bkcm	291	Unhealthy	0.197466	0.345320	0.022475	0.153482	0.057545	0.019730	0.056629	0.102091	0.179959

Figure 3.7 :Final preprocessed AQI Data set

3.4 Exploratory Data analysis

Data Visualization: Data visualization is the graphical representation of data using visual elements like charts, graphs, maps, and plots. It transforms raw data into visual formats that are easy to interpret, enabling better understanding, analysis, and communication of insights. In the context of an AQI prediction project, data visualization involves creating visual representations of air quality data (e.g., pollutant concentrations, meteorological factors, temporal trends) to uncover patterns, relationships in the data.

Correlation Plot:

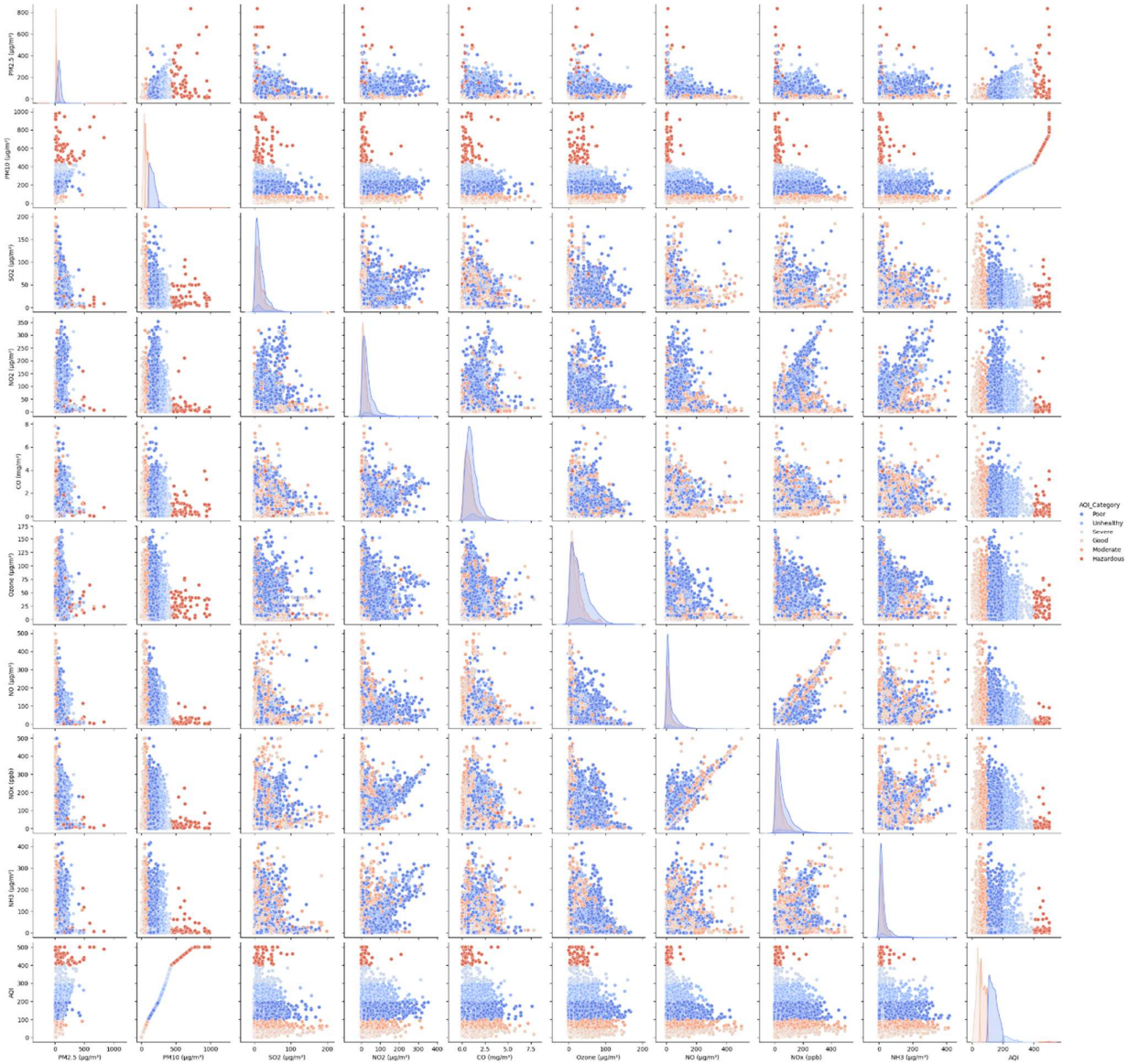

Figure 3.8 Correlation plot Using pairplot

This pairplot now shows the relationships between variables colored by AQI_Category, which adds another dimension to our analysis:

Color Distribution: The points are colored based on AQI categories (Good, Moderate, Poor, etc.) This helps visualize how different pollutant combinations relate to air quality categories

Key Relationships :

PM2.5 vs PM10: Shows strong positive correlation with higher AQI categories (warmer colors) clustering in the upper right.

NO vs NOx: Clear positive correlation with AQI categories showing distinct clustering

PM2.5/PM10 vs AQI: Higher values correspond to worse AQI categories, showing direct influence on air quality

Pollutant Patterns: Higher concentrations of PM2.5, PM10, and NOx tend to correspond with worse AQI categories.

Ozone shows more mixed distribution across AQI categories

CO shows some clustering of higher AQI categories at higher concentrations

Distribution Patterns: The diagonal plots show the distribution of each variable split by AQI category

Better air quality categories (cooler colors) tend to cluster at lower pollutant values

Worse air quality categories (warmer colors) tend to appear at higher pollutant values

This visualization helps identify which combinations of pollutants are most associated with poor air quality conditions and which tend to occur during good air quality periods.

Boxplot :visualization of PM distributions across stations

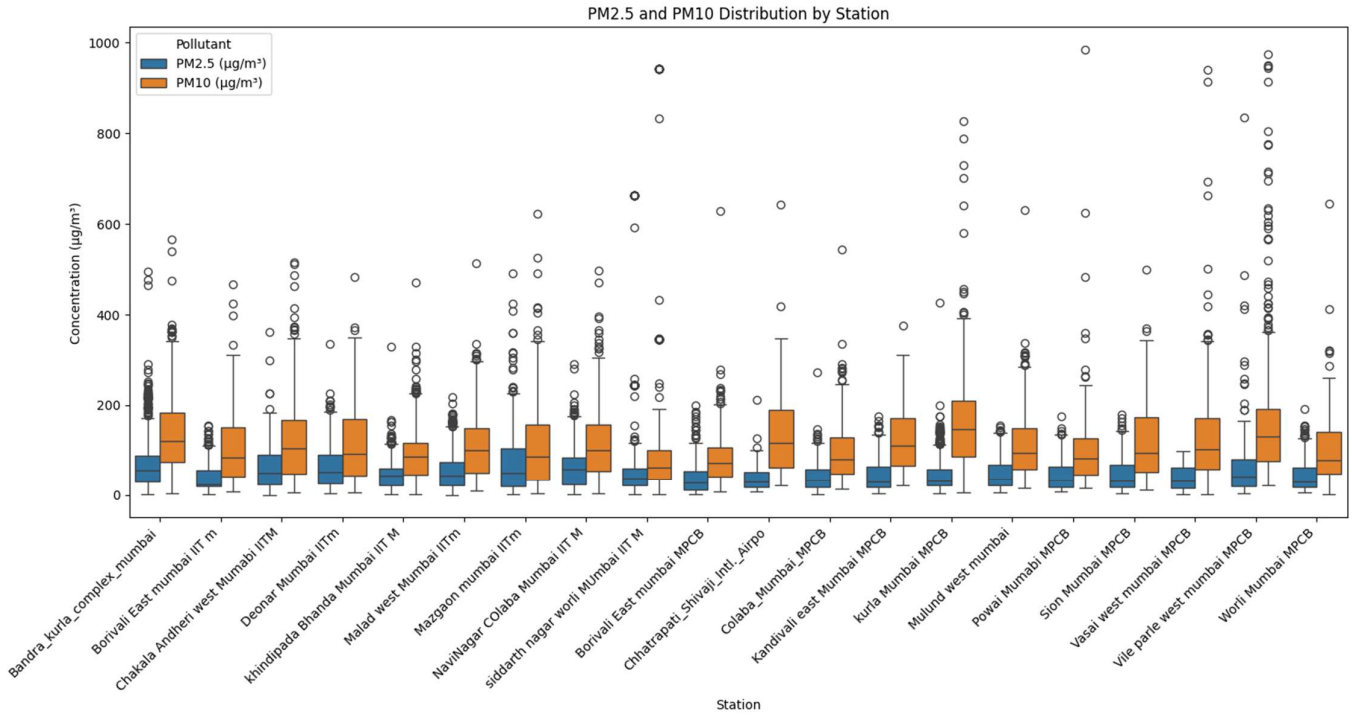

Figure 3.9 Boxplot for visualization of PM distributions across stations

we can observe:

PM10 levels (orange boxes) are consistently higher than PM2.5 levels across all stations.

- Kurla Mumbai MPCB shows the highest median PM10 concentrations
- Mazgaon Mumbai IITm and Bandra Kurla Complex show notably high PM2.5 levels
- Some stations like Vasai West Mumbai MPCB show relatively lower concentrations of both pollutants
- The whiskers indicate significant daily variations in pollution levels at most stations

Pollution parameter trends using linechart : Air pollution parameter for all observed Satations.

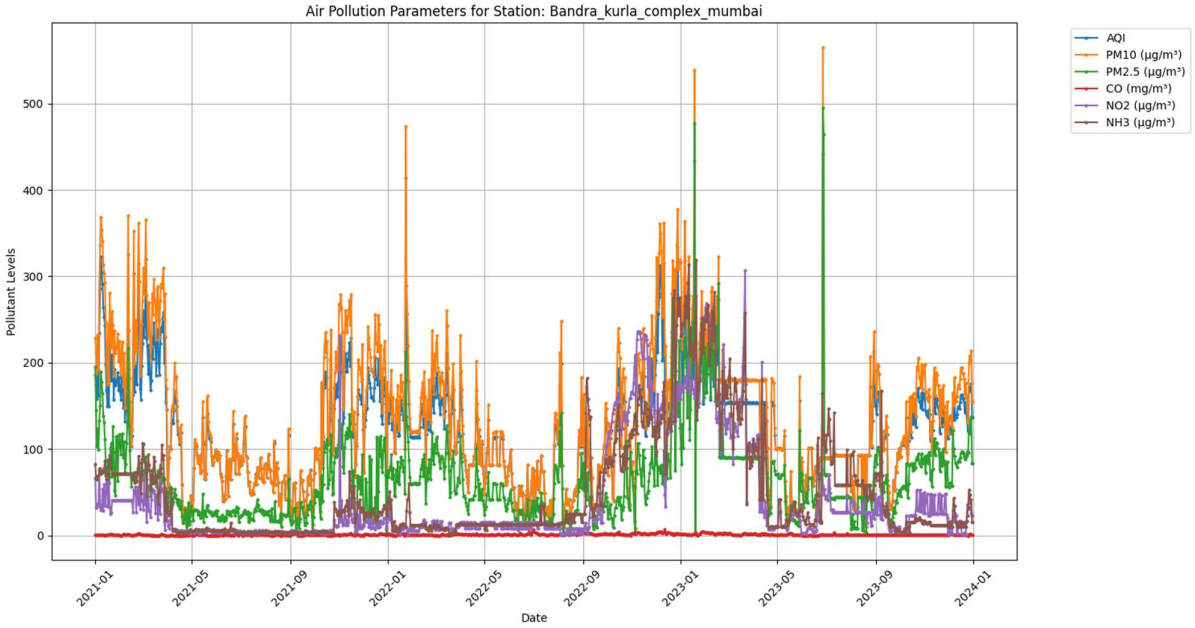

Figure 3.10 Air pollution parameter for satation : Bandra Kurla complex Mumbai

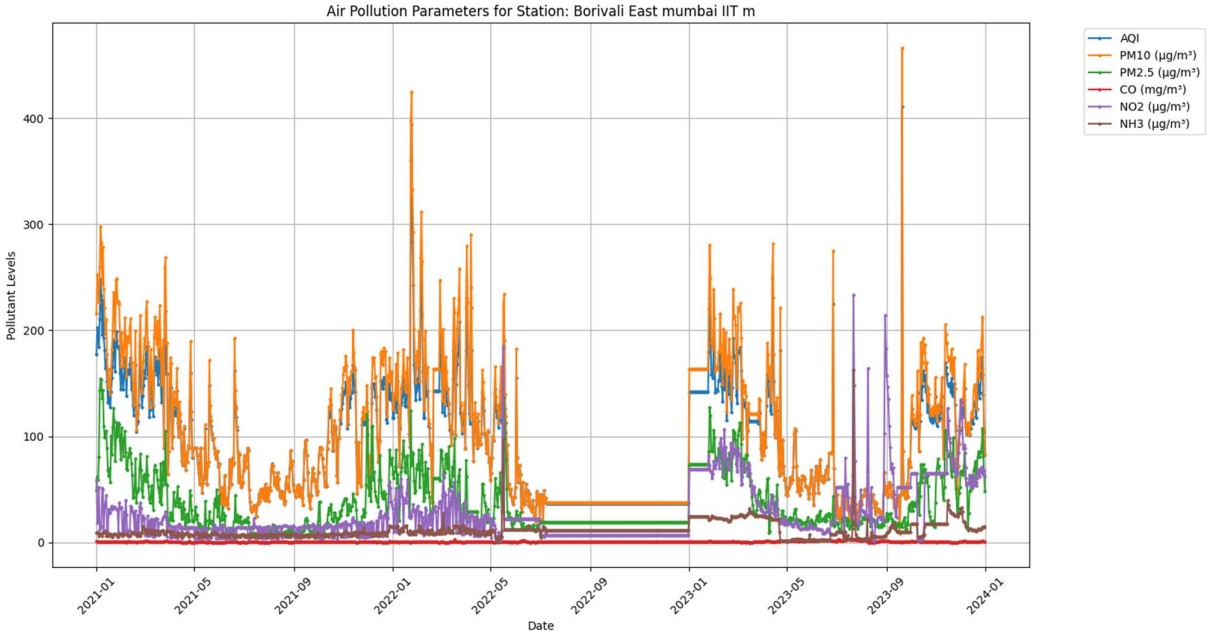

Figure 3.11 Air pollution parameter for satation : Borivali East Mumbai IITM

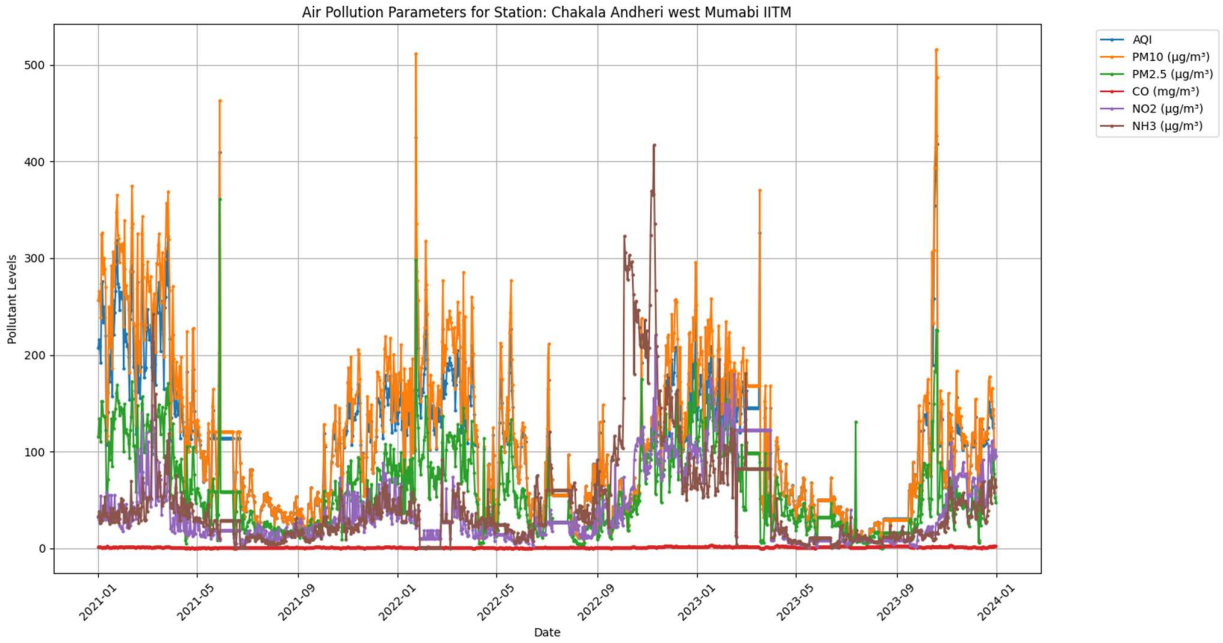

Figure 3.12Air pollution parameter for satation : Chakla Andheri West Mumbai IITM

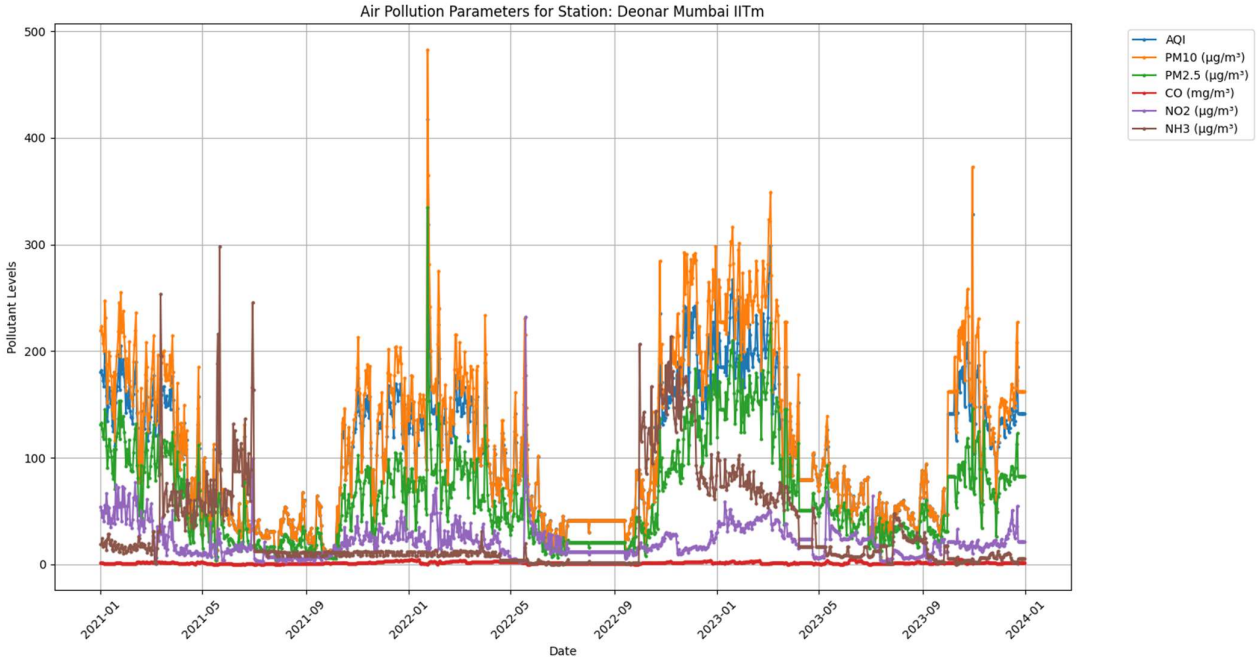

Figure 3.13 Air pollution parameter for satation :Deonar Mumbai IITM

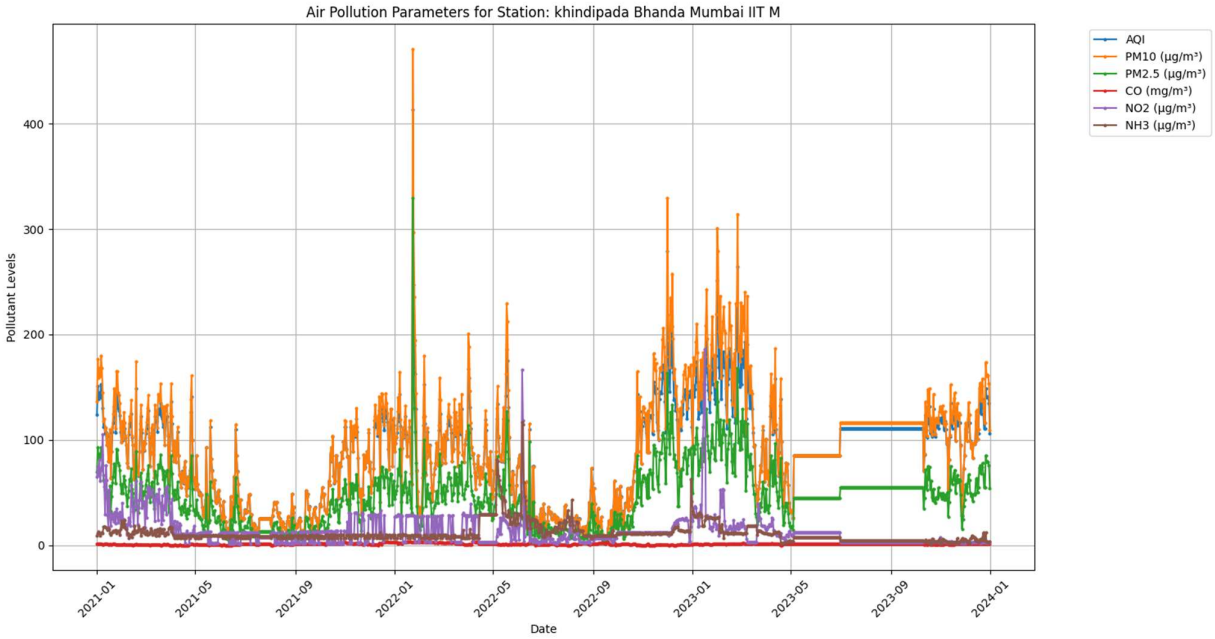

Figure 3.14 Air pollution parameter for satation : Khindipada Bhanda Mumbai IITm

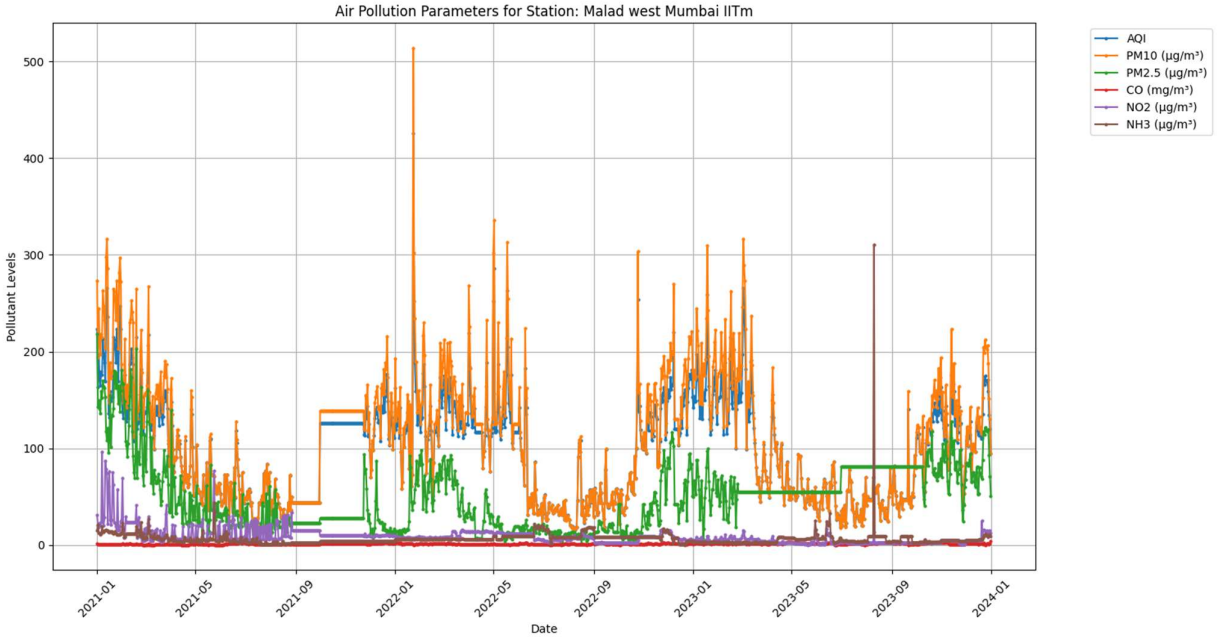

Figure 3.15 Air pollution parameter for satation :Malad West Mumbai IITm

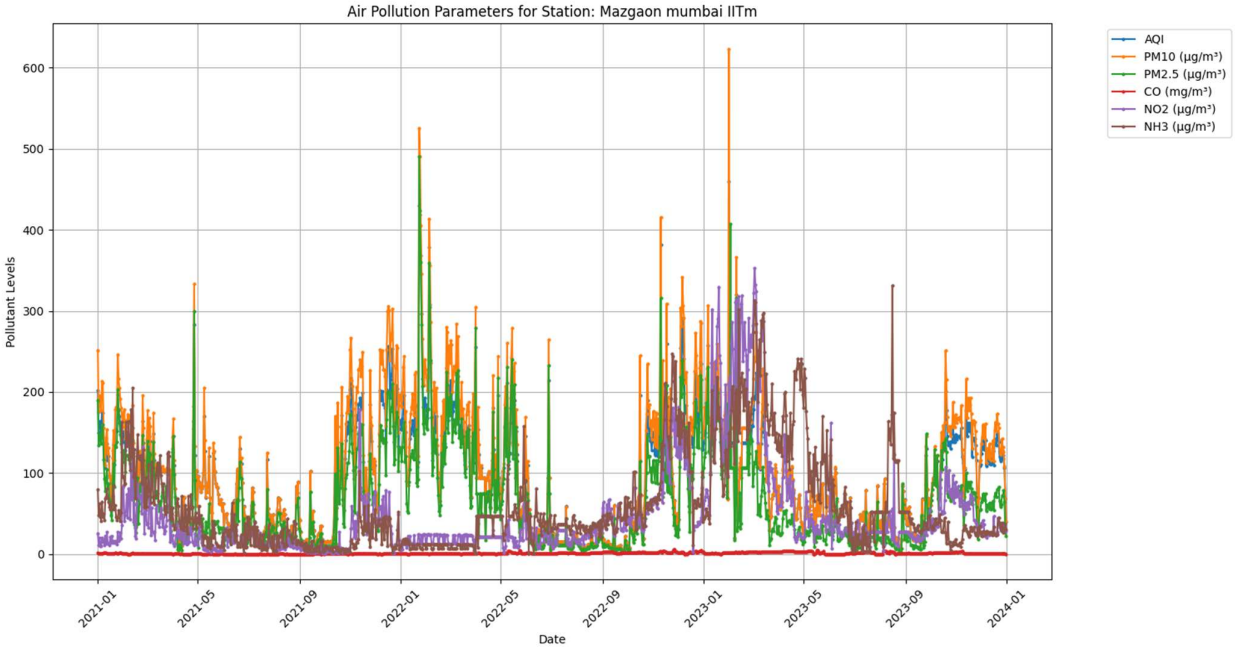

Figure 3.16 Air pollution parameter for satation : Mazgaon Mumbai IITM

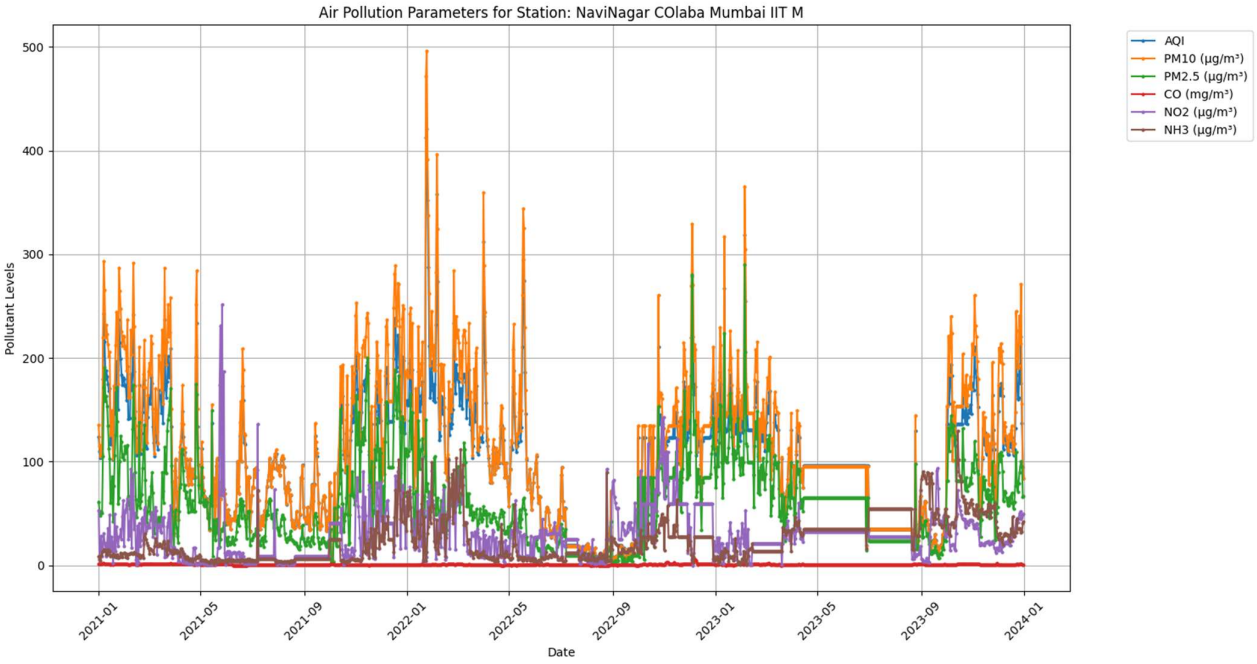

Figure 3.17 Air pollution parameter for satation :Navi nagar colaba Mumbai IITm

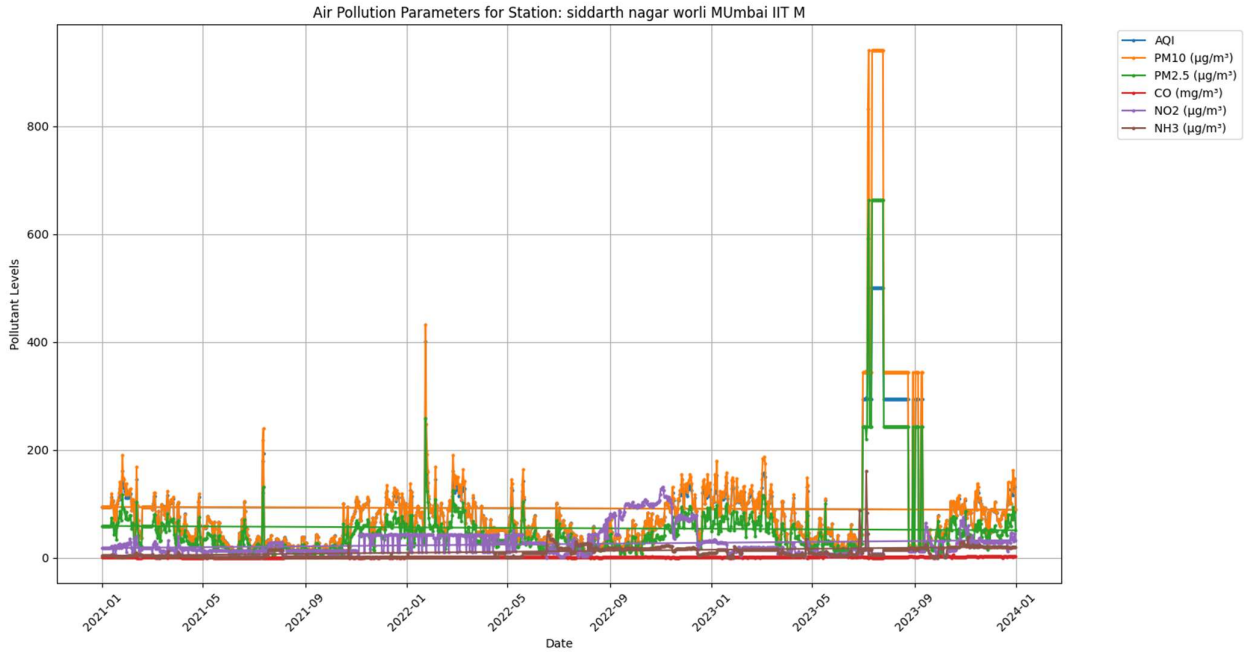

Figure 3.18 Air pollution parameter for satation : siddarth nagar worli MUmbai IIT M

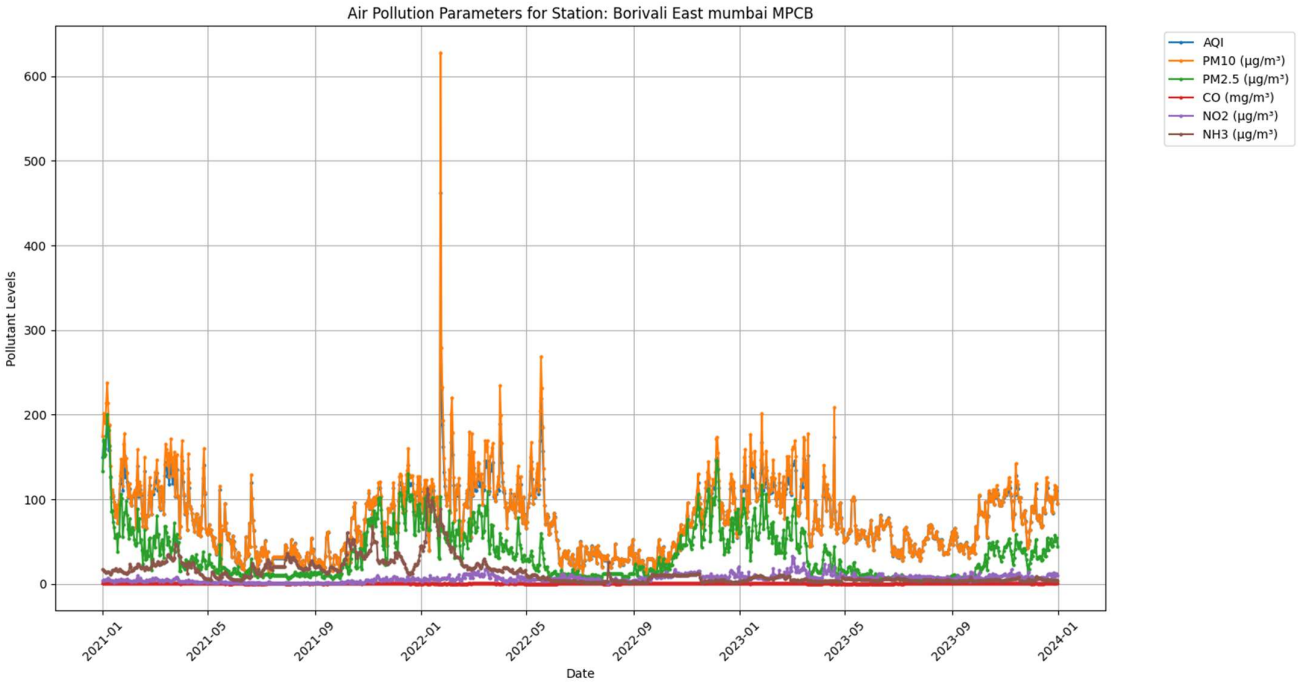

Figure 3.19 Air pollution parameter for satation : Borivali East mumbai MPCB

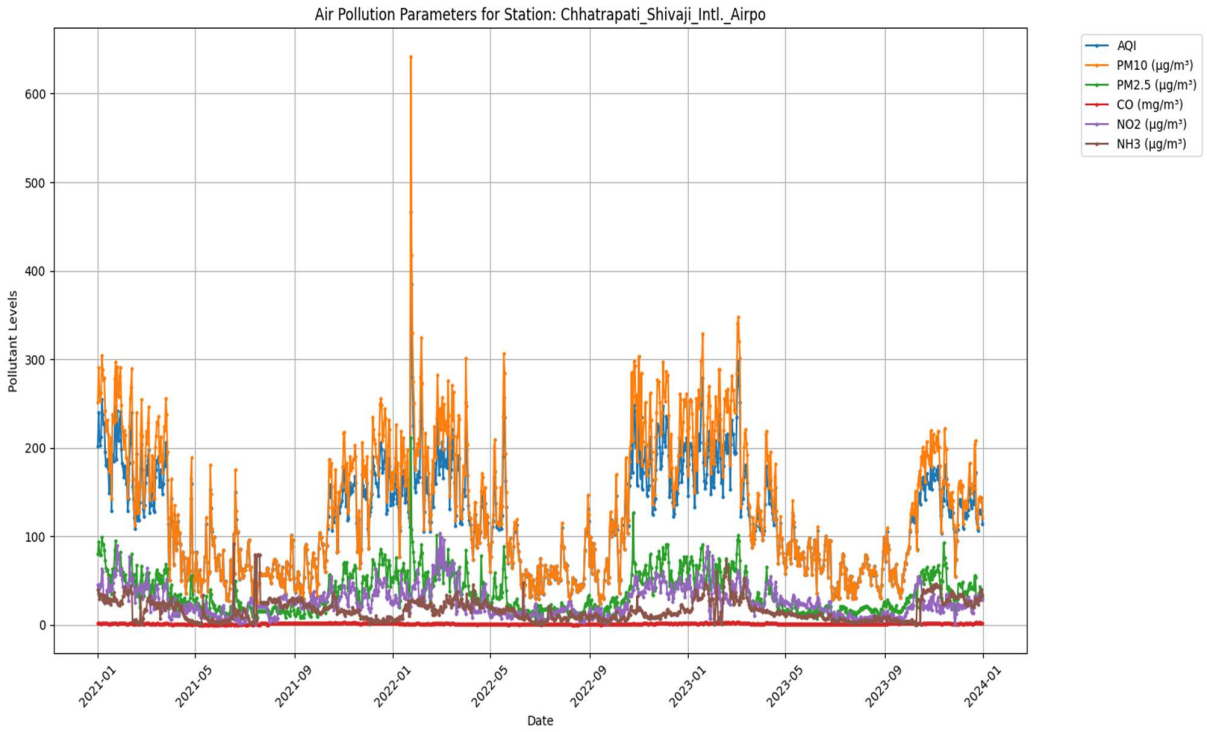

Figure 3.20 Air pollution parameter for satation : Chhatrapati_Shivaji_Intl._Airport

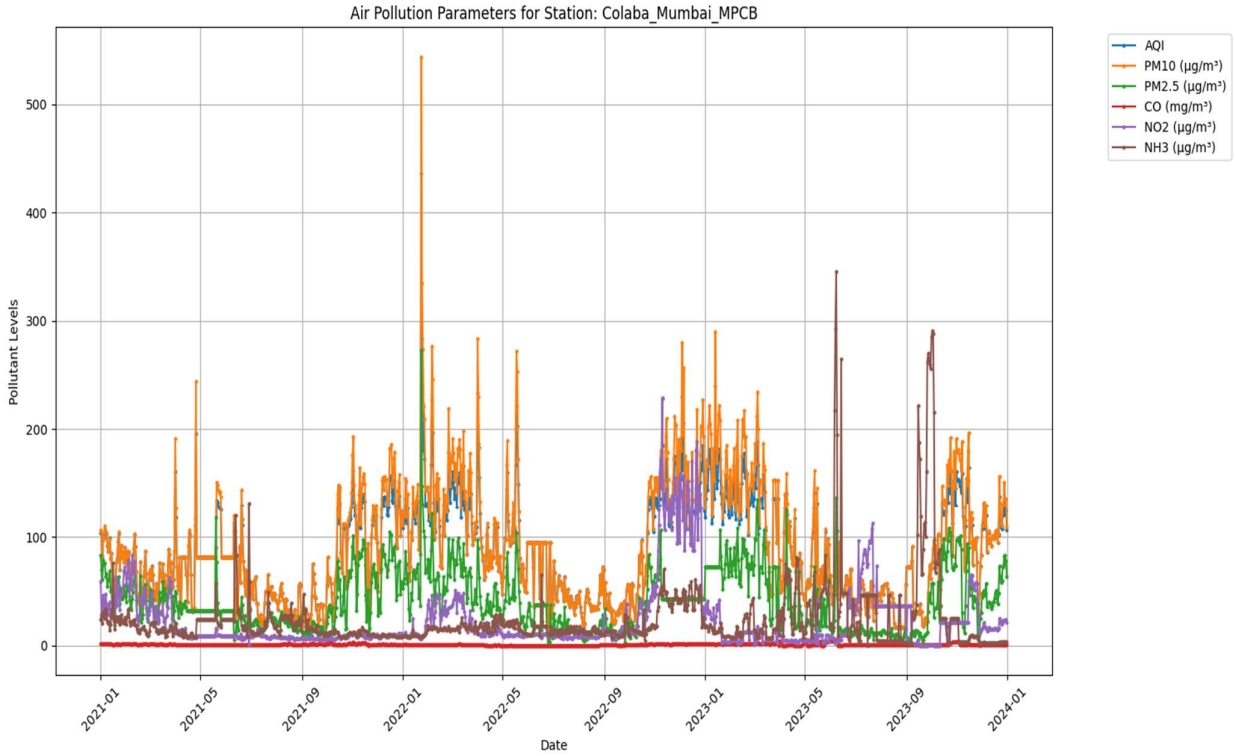

Figure 3.21 Air pollution parameter for satation : Colaba_Mumbai_MPCB

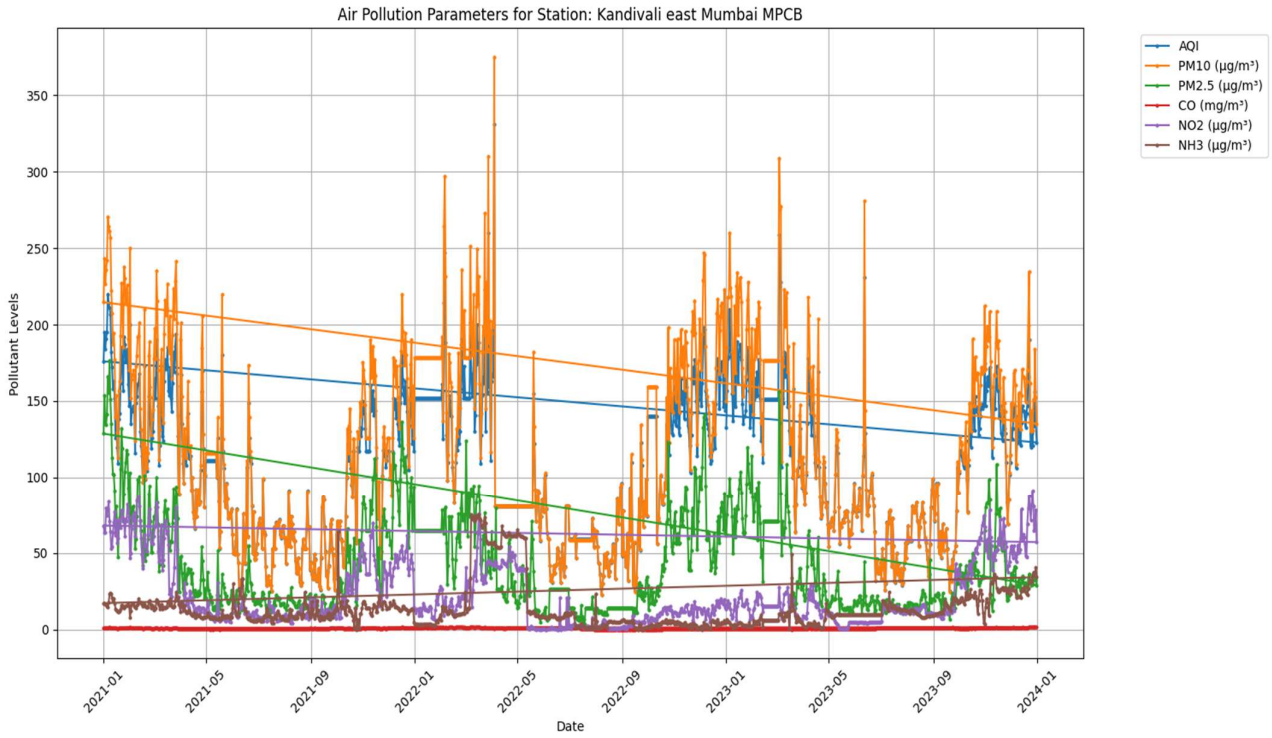

Figure 3.22 Air pollution parameter for satation : Kandivali east Mumbai MPCB

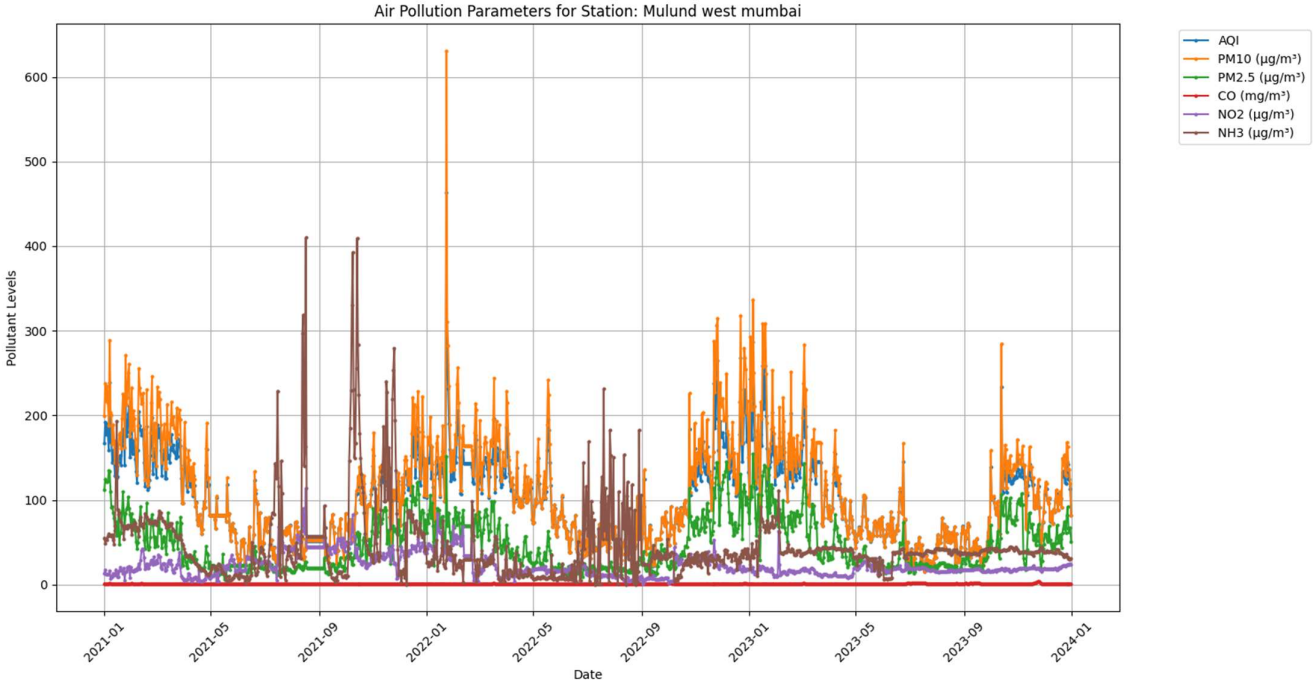

Figure 3.23 Air pollution parameter for satation : Mulund Mumbai MPCB

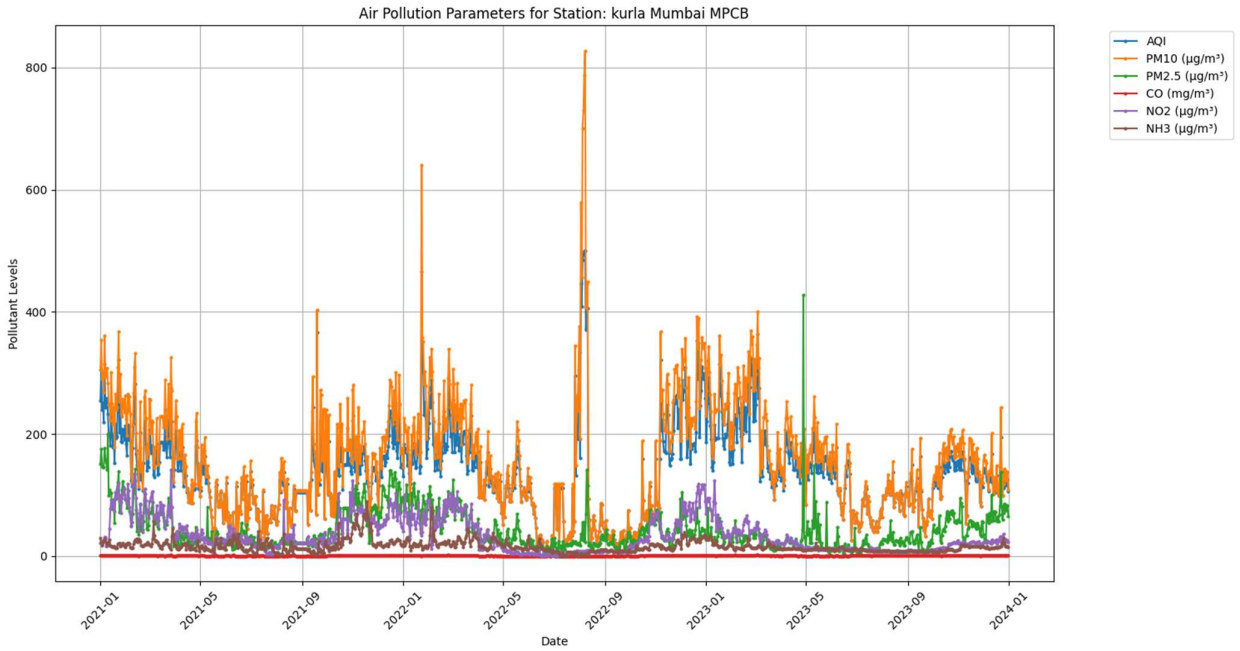

Figure 3.24 Air pollution parameter for satation : Kurla Mumabi MPCB

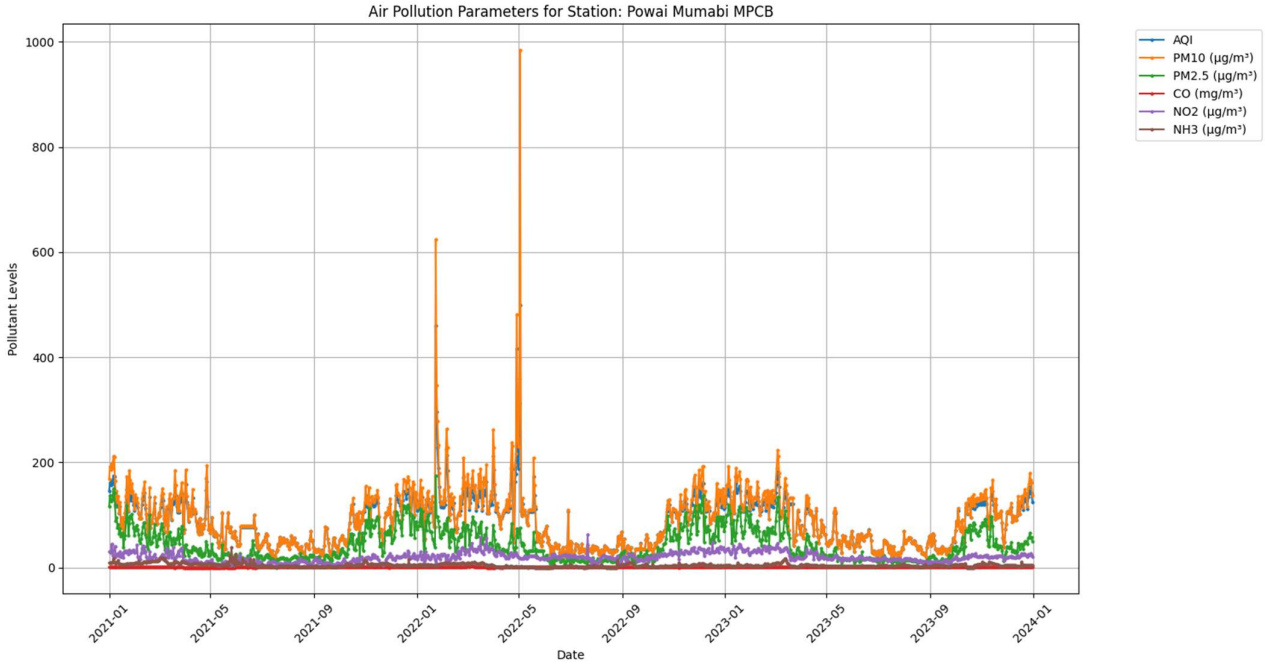

Figure 3.25 Air pollution parameter for satation : Powai Mumbai MPCB

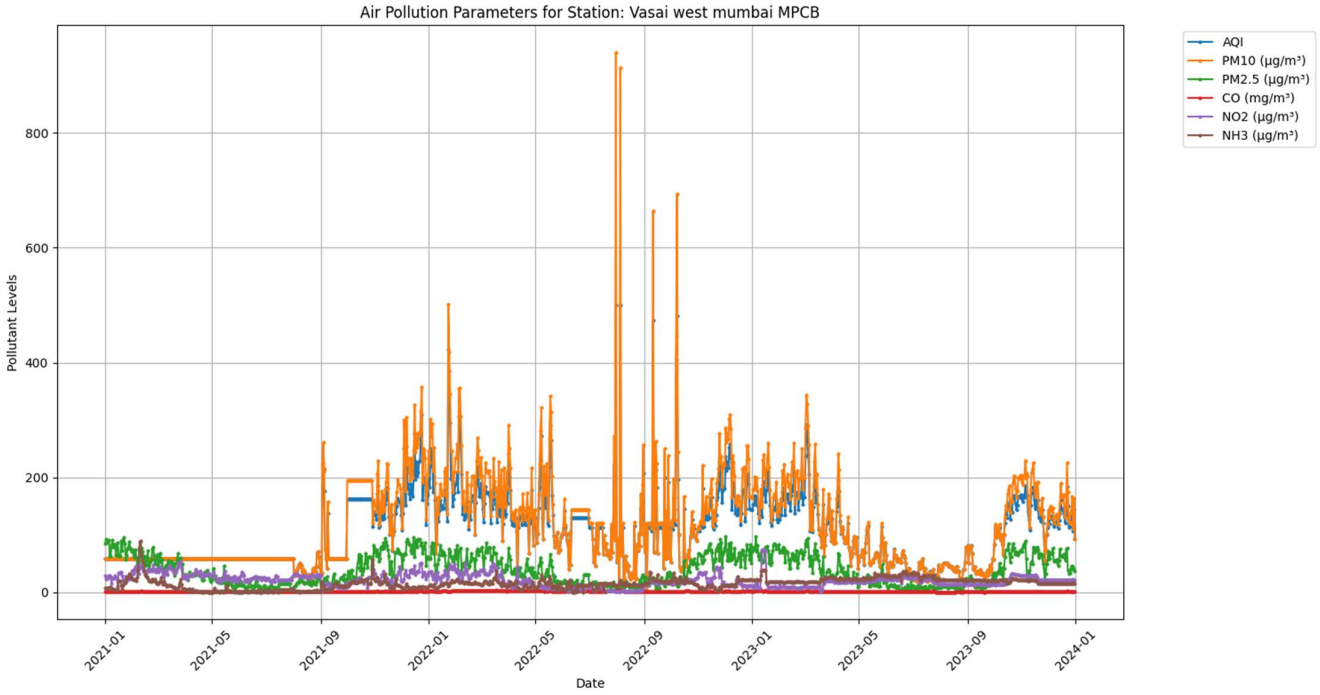

Figure 3.26 Air pollution parameter for satation : Vasai west mumbai MPCB

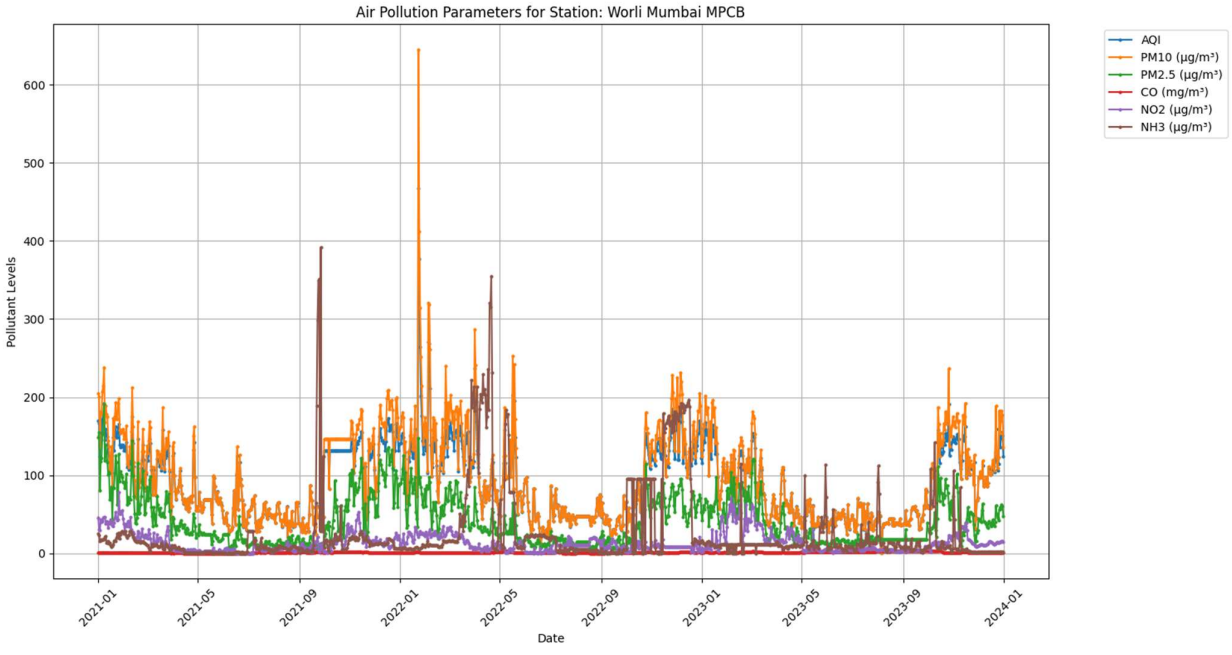

Figure 3.27 Air pollution parameter for satation : Worli Mumbai MPCB

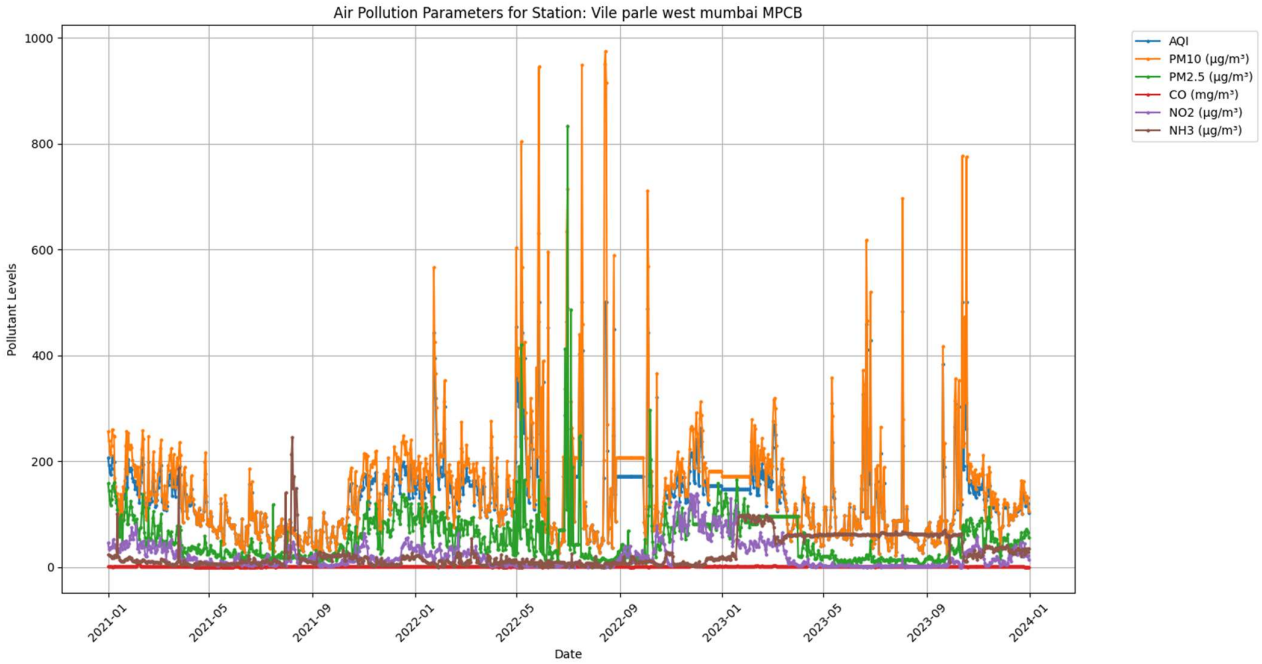

Figure 3.28 Air pollution parameter for satation :Vile parle west mumbai MPCB

The importance of these visualizations and what insights we can derive from them:

1. Purpose of Station-wise Visualization:

- Individual station analysis helps us understand the local air quality patterns
- We can identify which stations consistently show higher pollution levels
- It helps in understanding the spatial distribution of pollution across Mumbai

2. Key Findings from the Data:

- Stations with Highest Average AQI:
    - KM (Kurla Mumbai MPCB): Average AQI of 135.8 Figure
    - VpwM (Vile Parle West Mumbai): Average AQI of 129.5 Figure
    - Bkcm (Bandra Kurla Complex): Average AQI of 120.5 Figure
    - CSIA (Chhatrapati Shivaji International Airport): Average AQI of 116.4 Figure

3. Patterns Observed:

- Seasonal Variations:
    - Most stations show higher pollution levels during winter months
    - Lower levels are generally observed during monsoon season
- Different pollutants show different patterns:
    - PM2.5 and PM10 show more pronounced seasonal variations
    - CO and NO_2 levels often correlate with traffic patterns
    - NH_3 levels show less variation compared to other pollutants

4.Notable Observations:

- Industrial Areas (like Kurla) show consistently higher pollution levels
- Coastal stations (like Colaba) generally show lower average AQI
- Traffic-heavy areas (like Vile Parle, Bandra) show significant variations during peak hours.

summary visualization to better illustrate these patterns:

The reasoning here is to provide a clear visualization of the average AQI levels across stations

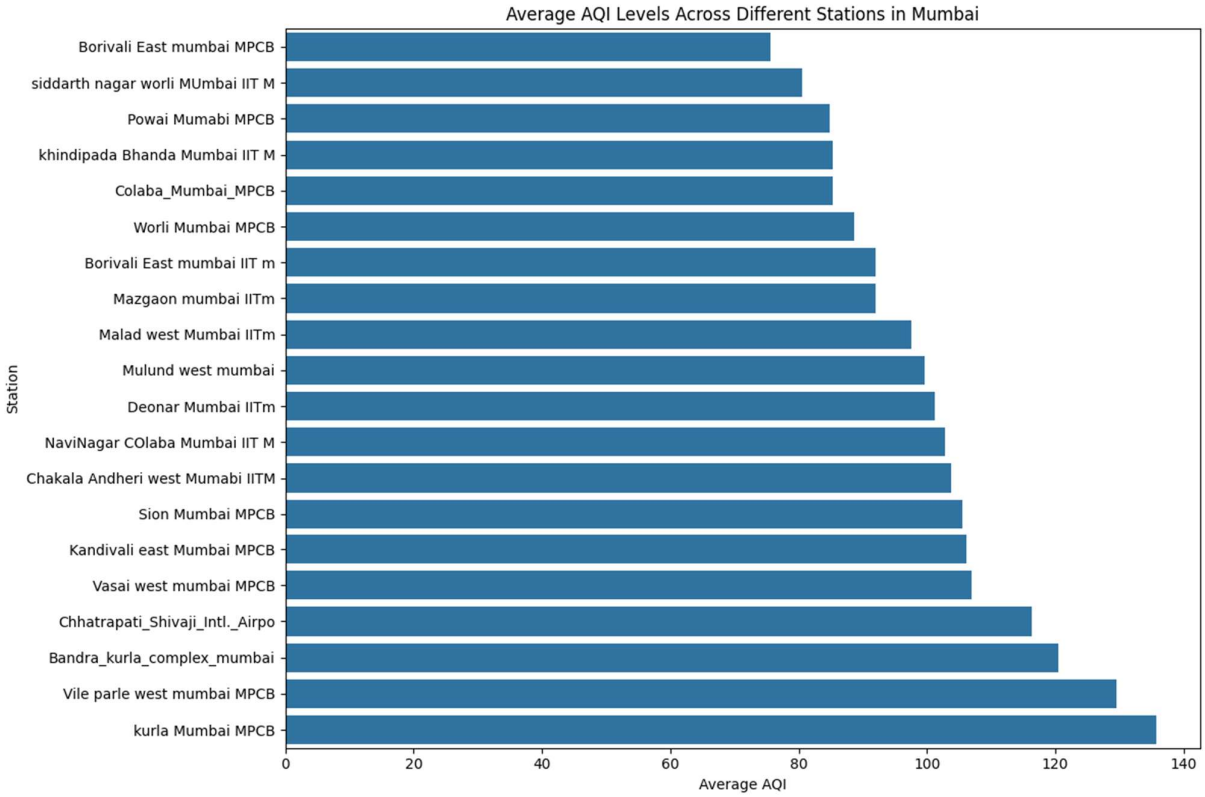

Figure 3.29: Average AQI levels across different stations in Mumbai

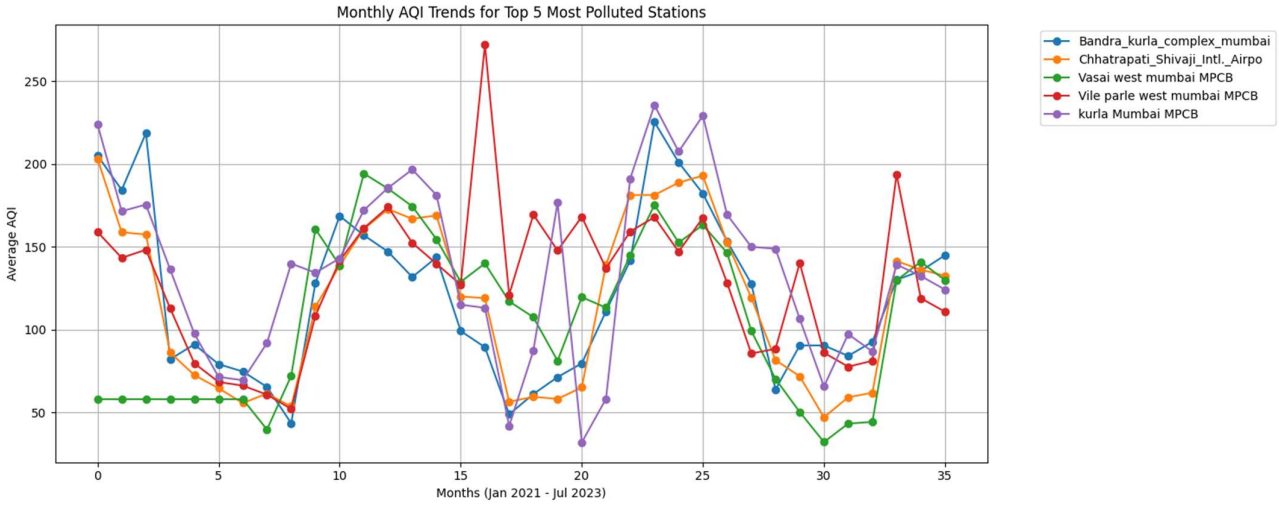

Figure 3.30 : Monthly AQI trends for the top 5 most polluted stations

These Figure help to identify spatial and temporal pollution patterns effectively.

TreeMap for Air pollutant distribution across Mumbai stations

Figure 3.31 TreeMap for Air pollutant distribution across Mumbai stations

The radar chart serves as a visual representation of pollution distribution across 20 stations.

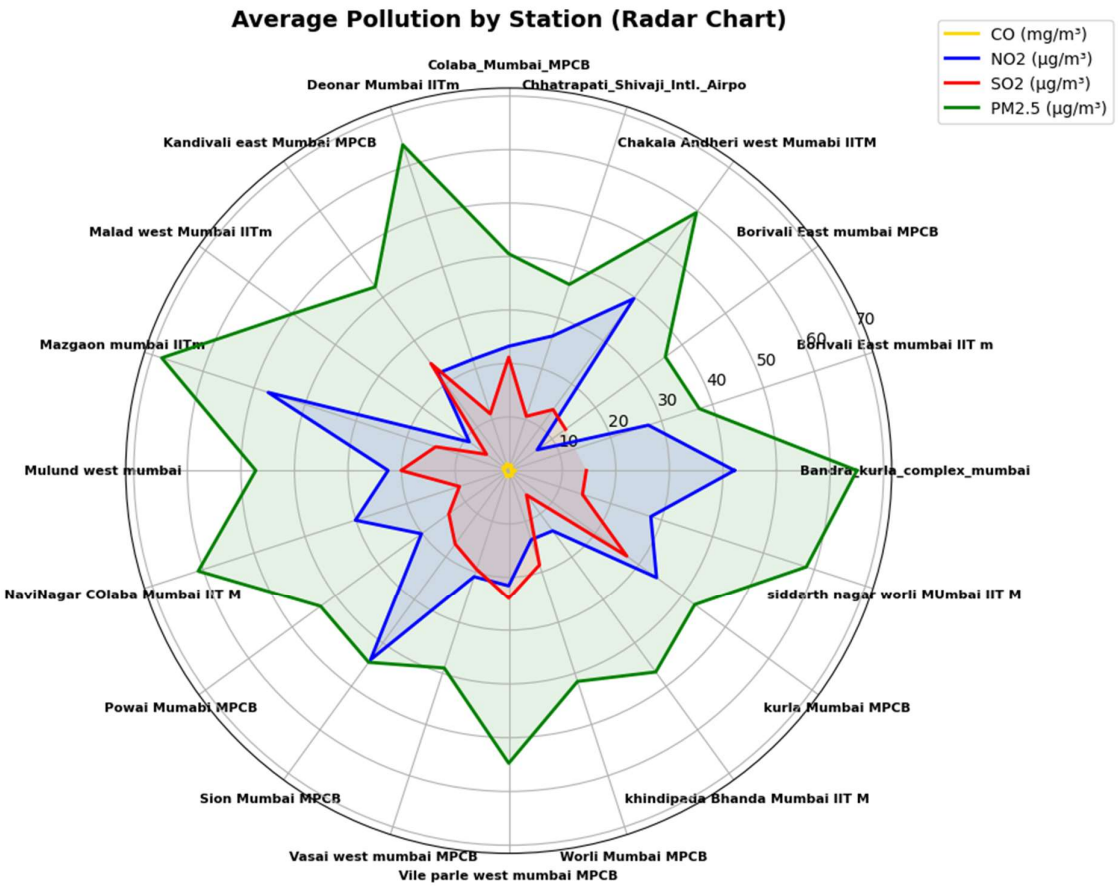

Figure 3.32: Radar chart of Average Pollution Levels Across 20 Stations in Mumbai

The accompanying table provides **quantitative details** for stations and pollutants, making it easy to analyze trends.

Pollution Levels by Station (AQI, PM10, and CO)

Station	AQI	PM10 (µg/m³)	CO (mg/m³)
Borivali East mumbai MPCB	75.59	78.83	0.39
siddarth nagar worli MUmbai IIT M	80.53	90.74	0.91
Powai Mumabi MPCB	84.82	90.74	0.8
khindipada Bhanda Mumbai IIT M	85.25	90.83	1.16
Colaba_Mumbai_MPCB	85.37	91.51	0.86
Worli Mumbai MPCB	88.59	95.79	1.13
Borivali East mumbai IIT m	91.96	100.98	0.57
Mazgaon mumbai IITm	92.04	102.61	1.15
Malad west Mumbai IITm	97.6	107.59	1.18
Mulund west mumbai	99.56	109.23	0.77
Deonar Mumbai IITm	101.29	113.8	1.38
NaviNagar COlaba Mumbai IIT M	102.82	114.22	0.84
Chakala Andheri west Mumabi IITM	103.7	116.28	0.98
Sion Mumbai MPCB	105.47	118.54	0.79
Kandivali east Mumbai MPCB	106.08	117.58	0.72
Vasai west mumbai MPCB	106.97	120.67	1.31
Chhatrapati_Shivaji_Intl._Airpo	116.4	131.92	1.17
Bandra_kurla_complex_mumbai	120.5	136.27	1.14
Vile parle west mumbai MPCB	129.46	149.4	0.88
kurla Mumbai MPCB	135.78	156.27	0.57

Table 3.1 :Pollution levels by stations

Together, these visualizations allow stakeholders to helps toIdentify stations with poor air quality and analyze the contribution of individual pollutants, it alsohelps to make decisions for targeted interventions to improve air quality in specific areas.

3.5 Model development and algorithms

3.5.1 Multicollinearity Analysis

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

- Multicollinearity is a statistical concept where several independent variables in a model are correlated.
- Two variables are considered perfectly collinear if their correlation coefficient is +/- 1.0.
- Multicollinearity among independent variables will result in less reliable statistical inferences.

Effects of Multicollinearity

Although multicollinearity does not affect the regression estimates, it makes them vague, imprecise, and unreliable. Thus, it can be hard to determine how the independent variables influence the dependent variable individually. This inflates the standard errors of some or all of the regression coefficients.

Detecting Multicollinearity

A statistical technique called the variance inflation factor (VIF) can detect and measure the amount of collinearity in a multiple regression model. VIF measures how much the variance of the estimated regression coefficients is inflated as compared to when the predictor variables are not linearly related. A VIF of 1 will mean that the variables are not correlated; a VIF between 1 and 5 shows that variables are moderately correlated, and a VIF between 5 and 10 will mean that variables are highly correlated.

Variance Inflation Factors:	
Variable	VIF
NOx (ppb)	20.345521
NO (µg/m³)	12.483217
PM10 (µg/m³)	6.545564
PM2.5 (µg/m³)	5.725762
NO2 (µg/m³)	5.324356
CO (mg/m³)	2.508781
Ozone (µg/m³)	2.341664
NH3 (µg/m³)	2.123169

Table3.2 : multicollinearity analysis

SO2 has missing values so it we can't calculate VIF . we drop this feature

The Variance Inflation Factor (VIF) values indicate the degree of multicollinearity among the variables. A VIF value above 10 suggests high multicollinearity, which can affect the stability and interpretability of regression coefficients. In this case, **"NOx (ppb)" and "NO (µg/m³)"** have VIF values above 10, indicating significant multicollinearity with other variables. This could lead to issues in regression analysis, and it might be necessary to consider removing or combining these variables to reduce multicollinearity.

3.5.2 Randomization in Data Splitting

Randomization is essential in machine learning, ensuring unbiased training, validation, and testing subsets. Randomly shuffling the dataset before partitioning minimizes the risk of introducing patterns specific to the data order. This prevents models from learning noisy data based on the arrangement. Randomization enhances the generalization ability of models, making them robust across various data distributions. It also protects against potential biases, ensuring that each subset reflects the diversity present in the overall dataset.

3.5.3 Feature Selection and Train-Test Split

Train-Test for regression: Divide the regression data sets into training and testing data sets. Train the model based on training data, and evaluate its performance based on testing data. Then data is split into two parameters which contains air particles composition in one parameter and AQI data in other parameter. We've considered PM 2.5, PM 10 and NO2 variables for predicting the AQI variable. We've spilt the data into 80% training data and 20% testing data. An 80/20 train-test split is a balanced approach that ensures sufficient data for training (80%) while reserving enough data (20%) for testing the model's performance. This ratio is often chosen because it provides a good trade-off between training the model effectively and evaluating its performance reliably on unseen data.

3.5.4 Model selection Algorithms

The machine learning algorithms which were identified from the literature review such as linear regression,Decision tree,Random forest ,KNN and SVR algorithms were used to build models for prediction. The same data which was preprocessed before is used to train the model with all those identified algorithms. Building machine learning model involve analyzing data to identify patterns and make predictions. Then the built models will be tested by calculating the performance metrics. These performance metrics will test how well the model can predict the AQI.

3.5.5 Multiple Linear regression

A forecast can be expressed as a function of a certain number of factors that determine its outcome. Multiple linear regression (MLR) technique [4] includes one dependent variable to be predicted and two or more independent variables. In general, multiple linear regression can be expressed as

$$Y=b1+b_2 X_2+\ldots\ldots\ldots\ldots+b_k X_k+e$$

Where Y is the dependent variable, X_2, $X_{3\ldots\ldots\ldots}$, X_k are the independent variables, b_1, $b_{2\ldots\ldots\ldots}$, b_k are linear regression parameters. In this model, AQI is the dependent variable and, previous day's AQI and meteorological variables, are independent variables, e is an estimated error term which is obtained from independent random sampling from the normal distribution with mean zero and constant variance. The task of regression modeling is to estimate the b_1, $b_{2\ldots\ldots\ldots}$, b_k, which can be done using minimum square error technique.

After using the minimum square error technique, the solution can be obtained as $b = (X' X)^{-1} (X'Y)$. Further, the F–test has been performed to determine whether a relationship exists between the dependent variable and the regressors. The t–test is performed in order to determine the potential value of each of the regressor variables in the regression model. The resulting model can be used to predict future observations.

3.5.6 Decision tree

Decision tree classifiers utilize greedy methodology. It is a supervised learning algorithm where attributes and class labels are represented using a tree . The main purpose of using Decision Tree is to form a training prototype which we can use to foresee class or value of target variables by learning decision rules deduced from previous data (training data). The Decision tree can be described by two distinct types, namely decision nodes and leaves. The leaves are the results or the final end results. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every sub-tree rooted at the new nodes.

- **Root Node***:* The initial node at the beginning of a decision tree, where the entire population or dataset starts dividing based on various features or conditions.
- *Decision Nodes:* Nodes resulting from the splitting of root nodes are known as decision nodes. These nodes represent intermediate decisions or conditions within the tree.
- *Leaf Nodes:* Nodes where further splitting is not possible, often indicating the final classification or outcome. Leaf nodes are also referred to as terminal nodes.
- **Sub-Tree***:* Similar to a subsection of a graph being called a sub-graph, a sub-section of a decision tree is referred to as a sub-tree. It represents a specific portion of the decision tree.
- **Pruning**: The process of removing or cutting down specific nodes in a decision tree to prevent overfitting and simplify the model.
- **Branch / Sub-Tree***:* A subsection of the entire decision tree is referred to as a branch or sub-tree. It represents a specific path of decisions and outcomes within the tree.

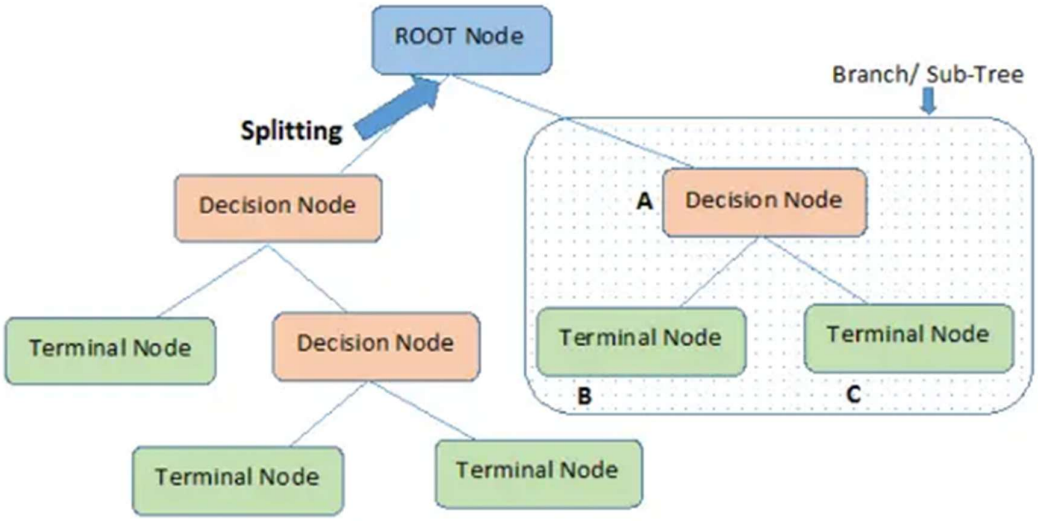

Figure 3.33 Decision tree

Entropy : entropy is a measure of uncertainty or impurity in a dataset. In decision trees, it helps quantify how well a split separates the data into distinct classes.

$$\boldsymbol{Entropy = -\sum_1^n p_i log2\,(p_i)}$$

Gini Index is measures the probability of incorrectly classifying a randomly chosen element if it is labeled according to the distribution of labels in the dataset.

$$\boldsymbol{Gini} = \mathbf{1} - \sum_{\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{p_{i^2}}$$

Information Gain (IG) evaluates the effectiveness of a feature in reducing entropy. It measures how much uncertainty in the dataset is reduced after splitting on a particular feature.

$$\boldsymbol{information\ gain} = \boldsymbol{E(paernt)} - \{\boldsymbol{weighted\ average}\} * \boldsymbol{E(Children)}$$

3.5.7 Random Forest

Random Forest is a ML algorithm. At training situation multitude decision trees are made and the output will be divided based on number of classes i.e., classification, prediction of class i.e., regression. The number of trees is proportional to accuracy in prediction. The dataset includes factors like rainfall, perception, temperature and production. These factors in dataset are used for training. Only two-third of the dataset is considered. Remaining dataset is used for experimental basis. The algorithm random forest has 3 parameters like: n tree which describes the n number of trees which need to grow, m try - mentions how many variables need to be taken at a node split. Node size - In terminal nodes it suggests us the number of observations need to take.

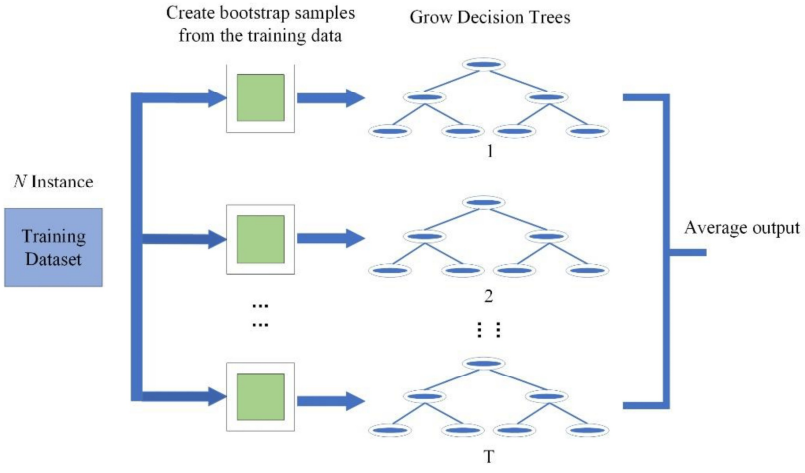

Figure 3.34 Random forest

3.5.8 Support Vector Machines (SVM)

Support vector machines (SVMs) are supervised learning models, with associated learning algorithms that analyze the data used for classification and regression analyses.

A type of SVM for regression, support vector regression (SVRIn SVR, the set of training data includes predictor variables and observed response values. The goal is to find a function *f(x)* that deviates from y_n (sample labels) by a value no greater than ε (bias) for each training point *x*—that is, remain as flat as possible. Therefore, SVR is also known as tube regression. Its schematic diagram is shown below.

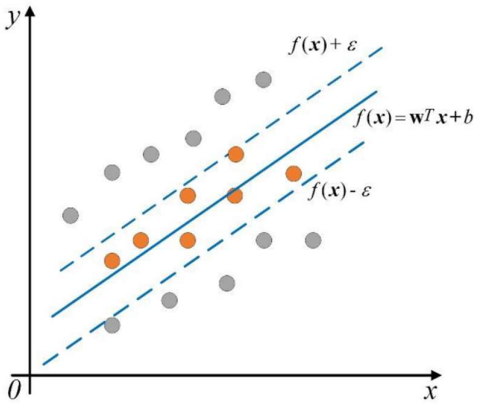

Figure 3.34 : support vector regression

The schematic diagram of support vector regression (SVR).

According to the literature on SVM linear regression , the solution to SVR is as follows:

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} (a_n - a_n^*)\,(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{x}) + b.$$

Where $\boldsymbol{x}$ is the input feature vector, b is the distance parameter, $a$nan and $a*$nan* are the introduced Lagrange multipliers. However, some regression problems cannot be described adequately using a linear model.

In that case, we can obtain a nonlinear SVR model by replacing the dot product $\boldsymbol{x}$T$n\boldsymbol{x}$xnTx with a nonlinear kernel function $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = <\varphi(\boldsymbol{x}_1), \varphi(\boldsymbol{x}_2)>$,

where $\varphi(x)$ is a transformation that maps $\boldsymbol{x}$ to a high-dimensional space. Therefore, the final solution to nonlinear SVR can be obtained as:

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} (a_n - a_n^*)\, K(\boldsymbol{x}, \boldsymbol{x}_n) + b.$$

3.5.9 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple and effective machine learning algorithm used for classification and regression tasks. It works by finding the 'k' closest data points (neighbors) to a given data point and making predictions based on the majority class (for classification) or the average (for regression) of these neighbors.

How KNN Works

Distance Calculation: For a new data point, KNN calculates the distance to all data points in the training set.

Common distance metrics include:

1. Euclidean Distance:

$$d(p, q) = \sqrt{\sum_{1}^{n} (p_i - q_i)^2}$$

2. Manhattan Distance:

$$d(p, q) = \sqrt{\sum_{1}^{n} |p_i - q_i|^2}$$

Selection of Neighbors: The algorithm selects the K nearest neighbors based on the calculated distances. And majority voting for classification ,KNN assigns the class label based on the majority class among the K nearest neighbors .The predicted class of the data point is the one most common among its neighbors.

3.6 Evaluation Metrices

To analyze the performance of a machine learning model we need some metrics. These metrics are statistical criteria that can be used to measure and monitor the performance of a model. As our thesis deals with prediction, we've considered MAE and RMSE as the performance metrics.

Mean absolute error (MAE): MAE is the arithmetic average of the difference between the ground truth and the predicted values. It can also be defined as measure of errors between paired observations expressing same phenomenon. It tells us how far the predictions differed from the actual result. Mathematical representation for MAE is given below.

$$MAE = \frac{1}{n}\sum_{1}^{n}|y_j - y_j|$$

Where,
yj = Prediction
yj = True value
N =Total number of data points

Root mean square error (RMSE): RMSE is the square root of the average of the squared difference between the target value and the value predicted by the model. It is square root of mean square error (MSE). The implementation is very much similar to MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}|y_j - y_j|}$$

Where,
yj = True value
N =Total number of data points

Rsquared (R^2): R square performance metric indicates how well predicted values matches actual values. To compute R squared value, we can use the r2_score function of sklearn.metrics.

$$\mathbf{R_{Square}} = 1 - \sum \frac{(y_J - y_J)^2}{(y_J - y_J)}$$

Adjusted R²: Adjusted R² is a modified version of the R^2 (coefficient of determination) metric used in regression analysis. It accounts for the number of predictors in the model, providing a more accurate measure of how well the model explains the variance in the dependent variable.

$$\textbf{Adjusted } \mathbf{R_{Square}} 1 - \left(\frac{(1 - R^2)(n - 1)}{n - P - 1}\right)$$

The machine learning models are validated by comparing the performance metrics. The lower the MAE, RMSE , higher the r-squared and adjusted R² the machine learning model performs better.

Chapter 4

Result and Conclusion

This chapter shows the analysis of the experiment with the preprocessed data set as well as the identified machine learning algorithms from the literature review. For the collection of air quality data, linear regression,Decision tree,Random forest ,KNN and SVR algorithms are used to build the model individually. For experimentation purpose, we considered PM 2.5, PM 10, and NO2 etc attributes as our input in order to obtain the AQI attribute which is the final output.

4.1 Performance metrics

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R^2 Score	Adjusted R^2 Score
Linear Regression	5.42	90.01	0.97	0.97
Decision tree	0.03	0.1065	1	1
Random forest	0.023	0.020	1	1
SVM	3.156	320.47	0.9147	0.9145
KNN	1.769	14.98	0.9964	0.9964

Table 4.1 : Performance metrics

4.2 Regression Model Performance Analysis

Best Model: Random Forest

1. **Accuracy:**
    - Random Forest has the smallest MAE (0.023) and MSE (0.020), indicating it provides the most accurate predictions.
    - It achieves an R^2 Score and Adjusted R^2 Score of 1, suggesting it perfectly explains the variance in AQI.
2. **Comparison with Decision Tree:**

- Both Decision Tree and Random Forest achieve perfect R2R^2R2 and Adjusted R2R^2R2, but Random Forest has a slightly lower MAE and MSE, showing it makes fewer errors.

3. **Generalization:**
   - Random Forest is less prone to overfitting compared to Decision Trees, as it combines predictions from multiple trees, making it more robust.
4. **SVM and KNN:**
   - SVM has much larger errors (high MAE and MSE) and a lower R_2 square, making it less suitable.
   - KNN performs well but has higher MAE and MSE compared to Random Forest.
5. **Linear Regression:**
   - While linear regression performs decently (with R2=0.97R^2 = 0.97R2=0.97), it has significantly higher errors than Random Forest.

4.3 Comparison of Algorithms

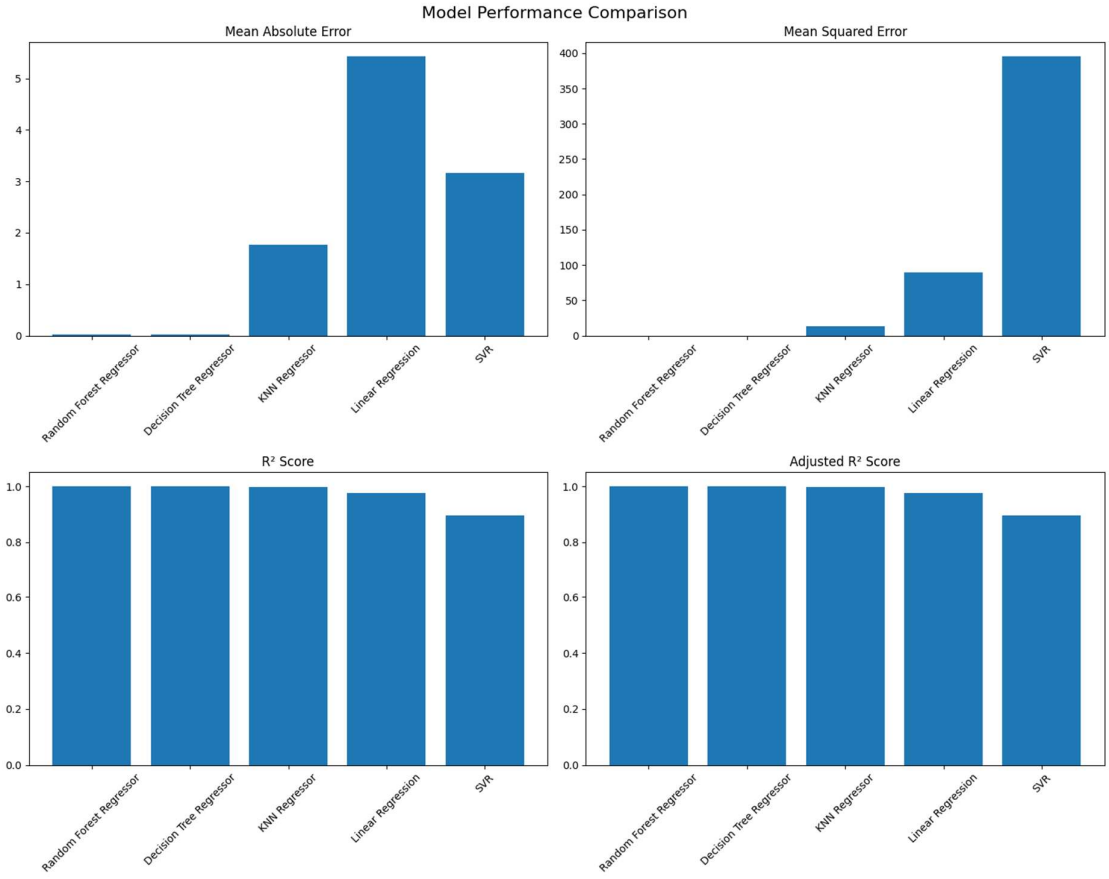

Figure 4.3 : Model Comparision of algorithms

Conclusion:

The **Random Forest** model is the best choice for AQI prediction based on its:

- **Highest accuracy** (lowest MAE and MSE),
- **Perfect R^2 and Adjusted R^2 scores**, and
- **Better generalization compared to Decision Trees**.

The best model based on the evaluation metrics is Random Forest Regressor, as it demonstrates the following:

- Lowest Mean Absolute Error (MAE): 0.0233, indicating the smallest average difference between predicted and actual values.
- Lowest Mean Squared Error (MSE): 0.0202, which minimizes large deviations.
- Highest R^2 Score: 0.98, indicating it good fit to the data.
- Adjusted R^2 Score: 0.98, maintaining its performance even after adjustments for model complexity.

4.5 Conclusion

Work In this study, we've conducted literature review and identified some machine learn ing algorithms to predict the AQI. We identified linear regression, Decision tree,Random forest ,KNN and SVR algorithms from the literature review. We've preprocessed the data and successfully trained these algorithm. Same set of data is used to build every model. In this study we've considered MAE, RMSE and r-square error to evaluate the performance of the models. We can conclude that the models Decision tree,Random forest have shown better performance with lower MAE, root mean squared error and higher r-squared.

Future work

In this study the data used was static that means the data will be fixed and it remains the same after it's collected. However the government updates the data hourly. Further we can use real-time data analysis using cloud to obtain better outcomes for greater performance as the data updates for every particular interval of time. We can further ensemble two or more machine learning algorithms and process large data to get more accurate results.

Chapter 5

References

[1] S. S. G. S. M. S. P. and p. . A. , "Forecasting air quality index using regression models: A case study on Delhi and Houston.," in *International Conference on Trends in Electronics and Informatics (ICEI) (pp. 248–254).*, 2017.

[2] Huixiang.Lui, Q.Li, D. and Y.Gu, ""Air quality index and air pollutant concentration prediction based on machine learning algorithms."," 2019.

[3] S. m. T. R. R. J. and N. , "Urban air quality prediction using regression analysis.," 2019.

[4] A. K. and P. . G. , "Forecasting of air quality in Delhi using principal component regression technique," 2011.

[5] M. C. F. C. A. P. S. S. and L. V. , "A machine learning approach to predict air quality in California.," *Complexity,* vol. 2020, pp. 1-23, 2020.

[6] B. L. A. B. P.-C. .. C. M. k. .. T. and C. -C. .. T. , "Urban air quality forecasting based on multi-dimensional collaborative support vector regression: A case study of Beijing-Tianjin-Shijiazhuang.," *PloS one,* 2017.

[7] J. S. and M. M. , "An efficient correlation-based adaptive LASSO regression method for air quality index prediction.," *Earth Science Informatics,* 2021.

[8] C.Li, Y.Li and Y.Bao, "Research on air quality prediction based on machine learning," in *2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, 2021.

[9] M. o. E. F. a. C. C. and G. o. I. , "Central Pollution Control Board (CPCB)," Air quality monitoring data [Across india]. Central Pollution Control Board., [Online]. Available: https://www.cpcb.nic.in/.

[10] M. o. E. . F. a. C. C. and G. o. I. , "Central Control Room for Air Quality Management - All India," [Online]. Available: https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard-all/caaqm-landing/caaqm-data-repository.

[11] D. P. a. A. J. N. Tomar, "Air quality index forecasting using autoregression models," in *Conference on Electrical,Electronics and Computer Science (SCEECS),*, 2020.