PREDICTIVE MODELLING OF PM2.5

A PROJECT REPORT

Submitted by

NIRAV PARMAR

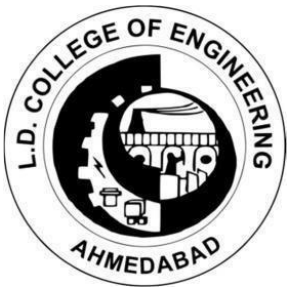
190280116086

In partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

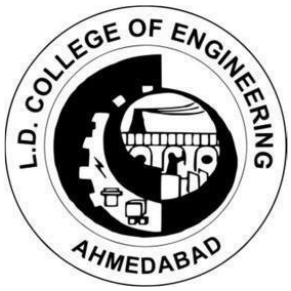
in

Information Technology
L.D College of Engineering, Navrangpura
Ahmedabad-380015



Gujarat Technological University, Ahmedabad

MAY 2023



L.D COLLEGE OF ENGINEERING,

Navrangpura, Ahmedabad, Gujarat 380015

CERTIFICATE

This is to certify that the project report submitted along with the project entitled **Predictive Modelling of PM2.5** has been carried out by **Nirav Parmar (190280116086)** under my guidance in partial fulfillment for the degree of Bachelor of Engineering in Information Technology, **8th Semester** of Gujarat Technological University, Ahmedabad during the academic year **2023-24**.

Prof. Jaimin Chavda
Internal Guide

Dr Hiteishi Diwanji
Head of the Department
Information Technology

Ahmedabad University

Date: 28-04-2023

TO WHOM IT MAY CONCERN

This is to certify that Nirav Parmar, a student of LD college of engineering, Ahmedabad has successfully completed his internship in the field of Atmospheric Physics from 7st February, 2023 to 28th April, 2023 (Total number of Weeks: 12) under my guidance.

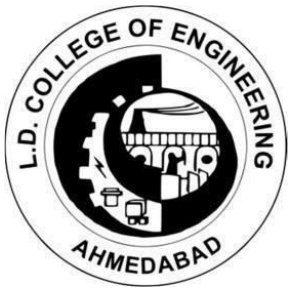
His internship activities include understanding and cleaning satellite data, processing satellite data and reanalysis data, fine-tuning various machine learning algorithms to predict PM2.5 concentrations over the Indian landmass using satellite and reanalysis datasets as inputs.

During the period of his internship program with me, he had been exposed to different processes and was found diligent, hardworking and inquisitive.

I wish him every success in his life and career.

For Ahmedabad University:

Aditya Vaishya PhD
Assistant Professor
Mathematical and Physical Sciences division
School of Arts and Sciences
Ahmedabad University, Ahmedabad 380009, India
+91 7961911 520 https://ahduni.edu.in/sas @ahmedabad



L.D COLLEGE OF ENGINEERING,

Navrangpura, Ahmedabad, Gujarat 380015

DECLARATION

I hereby declare that the Project report submitted along with the Project entitled **Predictive Modelling of PM2.5** submitted in partial fulfillment for the degree of Bachelor of Engineering in Information Technology to Gujarat Technological University, Ahmedabad, is a bonafide record of original project work carried out by me at under the supervision of and that no part of this report has been directly copied from any student's reports or taken from any other source, without providing due reference.

Name of the Student — Sign of Student

NIRAV PARMAR

Acknowledgement

I would like to extend my heartfelt gratitude to all the individuals who have provided me with their unwavering support and guidance throughout the duration of this project.

First and foremost, I would like to express my sincere appreciation to my supervisor, Professor Aditya Vaishya, for his invaluable insights, feedback, and guidance. His expertise and encouragement have been instrumental in shaping this research work and aiding me in overcoming various challenges.

Furthermore, I would like to extend my thanks to my lab mate, Yash Dahima, for his invaluable support and assistance throughout the project. His willingness to share his knowledge, expertise, and resources has been integral in guiding the direction of this research.

Last but not least, I am deeply grateful to my internal guide, Mr. Jaimin Buddhisagar Chavda, for his unwavering support, guidance, and encouragement. His insights and feedback have been crucial in shaping this research work and making it a success.

I extend my sincere thanks to all the individuals who have contributed to the successful completion of this project.

Abstract

Air pollution is a major environmental and public health concern, particularly in large cities, where particulate matter 2.5 (PM2.5) is a significant contributor. PM2.5 is a hazardous air pollutant, as it can penetrate deep into the lungs and cause respiratory and cardiovascular diseases. Therefore, regular monitoring and control of PM2.5 are essential for public health and safety.

This study presents a machine learning model that predicts seasonal levels of PM2.5 by analyzing crucial aerosol parameters, such as Aerosol Optical Depth (AOD) and Boundary Layer Height (BLH), achieving an R2 value of 0.81. The study validates satellite imager data with AERONET and compares different satellite data sources, such as INSAT-3D/3DR AOD and MODIS aqua dark target and deep blue AOD. This study found a linear correlation coefficient (R) of 0.20 ±0.05 and root mean square error (RMSE) value of 0.4 ±0.1 with AERONET and (R) 0.17 ±0.01 and (RMSE) 0.29 ±0.01 with MODIS during 2018-2019. The study's findings highlight the uncertainty in INSAT AOD data due to cloud contamination. Therefore, spatiotemporal method, one-sigma rule filtering, and various re-gridding criteria are used to conduct comparative analysis. This model's reliable predictions make it a useful tool for predicting PM2.5 levels in specific cities like Patna, Delhi, Ahmedabad, Faridabad, Bengaluru, Mumbai, Pune, Chennai, Hyderabad, Agra, Kanpur, and Lucknow, with a high correlation coefficient (R) 0.90 ±0.06 during Winter and Post-monsoon whereas in Pre-monsoon and Monsoon a low correlation coefficient (R) is 0.56 ±0.03 for the year 2019.

However, monitoring PM2.5 concentrations can be challenging in some regions where measuring instruments are not available, but satellite data is available in almost all locations, making it possible to predict PM2.5 concentrations accurately. This enables us to keep a track on air pollution episodes take preventive measures for its mitigation.

List of Figures

List of Tables

List of Abbreviations

PM2.5	**P**articulate **M**ater **2.5**
AOD	**A**erosol **O**ptical **D**epth
BLH	**B**oundary **L**ayer **H**eight
INSAT	**IN**dian **N**ation **SAT**ellite
MODIS	**MOD**erate Resolution **I**maging **S**pectroradiometer
ISRO	**I**ndia **S**pace **R**esearch **O**rganization
CPCB	**C**entral **P**ollution **C**ontrol **B**oard
ERA5	**E**CMWF **R**ean**a**lysis v5
ECMWF	**E**uropean **C**entre for **M**edium range **W**eather **F**orecasts
RMSE	**R**oot **M**ean **S**quare **E**rror
MSE	**M**ean **S**quare **E**rror
R	Correlation Coefficient
IGP	**I**ndo **G**angetic **P**lain
CSV	**C**ommon **S**eparate **V**alue
NetCDF	**Net**work **C**ommon **D**ata **F**orm

Table of Contents

Chapter 1

INTRODUCTION

1.1 AEROSOL PARAMETERS

PM2.5, AOD, and BLH are all key aerosol parameters that are closely related to air pollution. PM2.5 refers to fine particulate matter that is less than 2.5 micrometers in diameter and can penetrate deep into the lungs, causing serious health problems. AOD, or Aerosol Optical Depth, measures the amount of light that is absorbed or scattered by aerosols in the atmosphere, and is an indicator of the concentration of these particles. BLH, or Boundary Layer Height, is the height of the lowest layer of the atmosphere where the air is well-mixed.

There is a strong relationship between PM2.5, AOD, and BLH, as they are all affected by the same factors that influence air pollution levels. As the concentration of PM2.5 particles in the air increases due to factors such as vehicle emissions or industrial activity, it can lead to a higher AOD.

1.2 PURPOSE

Air pollution is a major problem in many cities around the world and is a key pollutant that poses serious health risks. To address this issue, it is important to accurately measure and predict concentration. Traditional ground-based monitoring stations have limited spatial coverage and may not provide a complete picture of the pollution levels in each area.

The purpose of this study is to accurately estimate the concentration of PM2.5, by using satellite remote sensing data. Traditional ground-based monitoring stations have limited spatial coverage, making it difficult to obtain a complete picture of air pollution levels. Satellite remote sensing can provide continuous and spatially extensive observations of key aerosol parameters such as AOD and BLH. Therefore, the goal of this study is to evaluate the accuracy of satellite-based estimates of PM2.5 by comparing them with ground-based measurements.

1.3 OBJECTIVE

The objective of this study is to predict seasonal PM2.5 Concentration using a machine learning model and to investigate the relationship between PM2.5 and key aerosol parameters, including AOD and BLH. Multiple data sources will be used in this study, including ground-based monitoring stations from the Central Pollution Control Board (CPCB), geostationary earth orbit (GEO) satellite INSAT-3D/3DR, moderate resolution imaging spectroradiometer (MODIS) Aqua, and aerosol robotic network (AERONET). Various data preprocessing and filtering techniques will be employed to ensure data quality and reliability.

The study will focus on several cities in India, including Patna, Delhi, Ahmedabad, Faridabad, Bengaluru, Mumbai, Pune, Chennai, Hyderabad, Agra, Kanpur, and Lucknow. The accuracy of the machine learning model will be evaluated by comparing the predicted PM2.5 values with ground-based measurements and calculating key performance metrics such as correlation coefficient (R) and root mean square error (RMSE). The findings of this study will provide important implications for air quality management and public health in India and beyond.

1.4 SCOPE

This project aims to develop a model that can predict seasonal PM2.5 concentration at a specific location using available data sources. While the model will initially be designed to provide accurate and precise predictions at a specific location over India but will not include any measures for controlling or reducing PM2.5 concentration. It is feasible to predict Daily PM2.5 concentration at any location over India with reliable Satellite Imager Data.

Chapter 2

STUDY AREA AND DATA

2.1 SITE LOCATION

The regions that are of interest for study include Patna, Delhi, Ahmedabad, Faridabad, Bengaluru, Mumbai, Pune, Chennai, Hyderabad, Agra, Kanpur, and Lucknow. Locations of all cities are shown in Figure 2.1 and Table 2.1 displays the latitude and longitude coordinates.

Figure 2.1: Study Area location On India Map

The Indo-Gangetic Plain (IGP) region in northern India, Pakistan, and Bangladesh is densely populated and agriculturally productive, but also home to some of the world's most polluted cities. Cities such as Delhi, Faridabad, Agra, Kanpur, Lucknow, and Patna have high levels of air pollution, with particularly high concentrations of PM2.5 nearly above 100 $\mu g/m^3$ and AOD. The boundary layer height (BLH) is also low in these cities during winter months, exacerbating the air pollution problem.

Chennai, Hyderabad, and Bengaluru are in the southern part of India, outside of the IGP. These cities have moderate levels of PM2.5 pollution and concentration are around $60\mu g/m^3$, and its AOD and BLH levels are relatively low compared to the other cities in the IGP.

Mumbai and Pune are located on the western part of India and has moderate levels of

Table 2.1: Latitude and Longitude Coordinates of Sites

SR No.	Sites	Latitude	Longitude
01	Patna	25.59	85.13
02	Delhi	28.65	77.15
03	Ahmedabad	23.00	72.59
04	Faridabad	28.40	77.30
05	Bengaluru	12.95	77.64
06	Mumbai	19.06	72.84
07	Pune	18.50	73.81
08	Chennai	13.08	80.24
09	Hyderabad	17.40	78.44
10	Agra	27.19	78.00
11	Kanpur	26.47	80.32
12	Lucknow	26.86	80.92

PM2.5 pollution and almost below $50\mu g/m^3$ during winter. The city's AOD levels are also relatively low, but the BLH can be high. Ahmedabad is in the western part of the IGP and is known for its high levels of PM2.5 pollution. The city also has high AOD and BLH levels, which contribute to poor air quality.

2.2 CPCB PARTICULATE MATTER 2.5

CPCB stands for Central Pollution Control Board, which is an organization in India that works under the Ministry of Environment, Forest, and Climate Change. CPCB monitors the concentration of PM2.5 in the air in various cities across India to assess air quality and identify areas where pollution levels are high. PM2.5 is a major air pollutant and can cause various health issues, including respiratory and cardiovascular problems. The CPCB regularly publishes reports on air quality in different parts of the country based on their monitoring data.

To address the issue of high PM2.5 levels, the CPCB has implemented various measures,

including regulating industrial emissions, promoting the use of clean energy sources, and enforcing strict vehicular emission norms. The board also works with state pollution control boards and other agencies to monitor and control air pollution levels.

The dataset I have used, consisting of PM2.5 data with a 15-minute temporal resolution from various cities, has been filtered to include only those cities with data available from 2015 to 2021. This dataset provides a valuable resource for studying trends in air quality over time and can be analyzed using statistical and data visualization techniques such as calculating annual average PM2.5 levels and identifying seasonal patterns. Because the data was collected from a ground station, it is likely to have a lower level of uncertainty compared to data obtained from other sources.

Table 2.2: PM2.5 Data Description

Spatial Coverage	Point-location
Temporal Coverage	2015-2021
Temporal resolution	15 minutes
Unit	$\mu g/m^3$
File Format	CSV

2.3 ERA5 BOUNDARY LAYER HEIGHT

ERA5 stand for European Centre for Medium-Range Weather Forecasts (ECMWF) re-analysis version 5 dataset, providing global climate and weather information for the past eight decades. This dataset is obtained through data assimilation, which combines model data with observations from across the world into a complete and consistent dataset. ERA5 provides hourly estimates for a wide range of atmospheric, ocean-wave, and land-surface quantities, including BLH. The uncertainty estimate for ERA5 is sampled by an underlying 10-member ensemble at three-hourly intervals, and ensemble mean, and spread have been pre-computed for convenience. Monthly-mean averages have also been pre-calculated to facilitate many climate applications.

This involves quality control checks to ensure the data is accurate and reliable. The data is also interpolated onto a regular grid, with a spatial resolution of 0.25 degrees latitude by 0.25 degrees longitude, to make it consistent with the other variables in the re-analysis dataset. The preprocessing step is important because it helps to remove any errors or inconsistencies in the raw data and ensures that the final re-analysis product is of high quality and suitable for a wide range of applications.

In summary, the process of estimating the BLH in re-analysis involves combining model output and observations using an algorithm that uses physical relationships to estimate the height of the atmospheric boundary layer. The BLH data undergoes a preprocessing step to ensure accuracy and consistency, which is important for producing a high-quality re-analysis dataset that can be used for a wide range of applications.

Table 2.3: BLH Data Description

Data Type	Gridded
Temporal Coverage	Global
Temporal Resolution	0.25^o x 0.25^o
Temporal Coverage	2015 to 2021
Temporal Resolution	Hourly
Unit	Meter
File Format	NetCDF

2.4 SATELLITE IMAGER DATA

Satellite imagery refers to the pictures or images of the Earth taken by satellites orbiting the planet. These images are captured using different types of sensors, including optical, radar, and infrared sensors. The images obtained from these sensors can be used for various purposes, such as meteorology, agriculture, land-use mapping, urban planning, environmental monitoring, and military intelligence, among others.

2.4.1 INSAT 3D And INSAT 3DR

Dedicated meteorological geostationary satellites, INSAT 3D and INSAT 3DR, have been deployed at 82^o E and 74^o E longitude respectively. These satellites are equipped with multi-spectral 6-channel imagers, 19-channel sounders, data relay transponders, and satellite-aided search and rescue payloads.

The staggered mode operation of the imagers in INSAT-3D and INSAT-3DR allows for an effective temporal resolution of 15 minutes. The sounder payload of INSAT-3DR covers the Indian land region sector data twenty times and Indian Ocean region data at four different times (4, 11, 16, and 23 UTC) on an hourly basis. These details are a subsection of the report chapter on the capabilities and operations of INSAT 3D and

INSAT 3DR. The distribution of AOD data is shown in Figure 2.2.

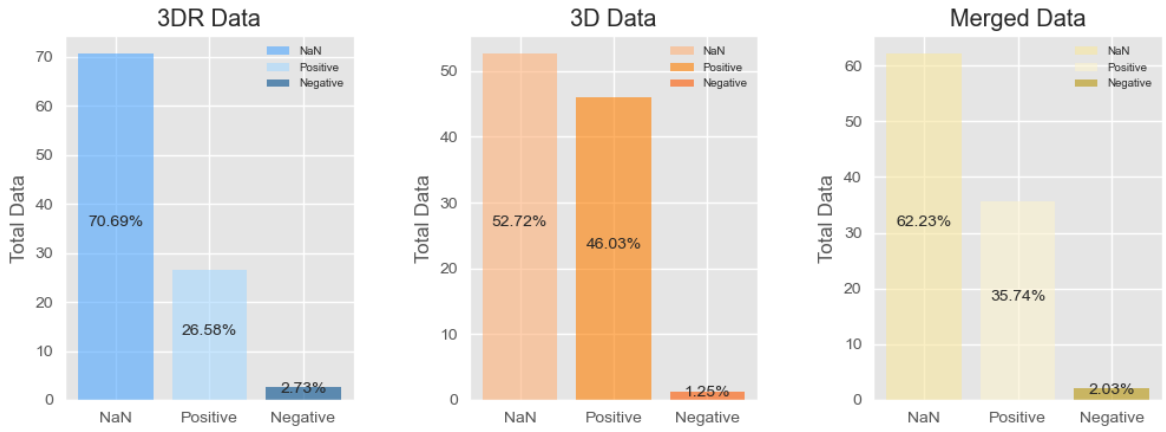


Figure 2.2: INSAT 3D,3DR and 3D and 3DR Combined Data Distribution

Mishra, 2018 exclaimed that the retrieval uncertainty in INSAT-3D AOD may fall in the range of 30% to 45% over land depending on the uncertainty of the aerosol optical properties used in retrieval algorithm, the instrument measurement error, the background AOD, and the gaseous absorption. These factors have resulted in more than 60% missing data over India's land surface from the INSAT 3D and 3DR combined. As a result, caution must be exercised when using the available AOD data to avoid potential inaccuracies.

A comparison was made between the AOD time series of INSAT 3D and 3DR satellites over India, and the results were presented in Figure 2.3. The figure indicates that the average AOD over India does not follow the expected trend, which could be due to various reasons, such as uncertainty in the AOD data and missing values caused by factors such as cloud coverage, high aerosol loading, instrument calibration, and data processing algorithms. The discrepancies between the AOD measurements of INSAT 3D and 3DR could also be attributed to differences in instrument calibration, data processing

algorithms, and other technical factors. Further investigation is required to identify the specific reasons for the deviation from the expected trend and to ensure the accuracy and reliability of the AOD data collected by INSAT 3D and 3DR. These findings highlight the importance of exercising caution while using INSAT 3D and 3DR AOD data, particularly over India's land surface, and call for the continued improvement of satellite instruments and data processing algorithms to enhance the accuracy and reliability of AOD measurements.

The analysis indicates that the AOD levels during the monsoon period should be lower,

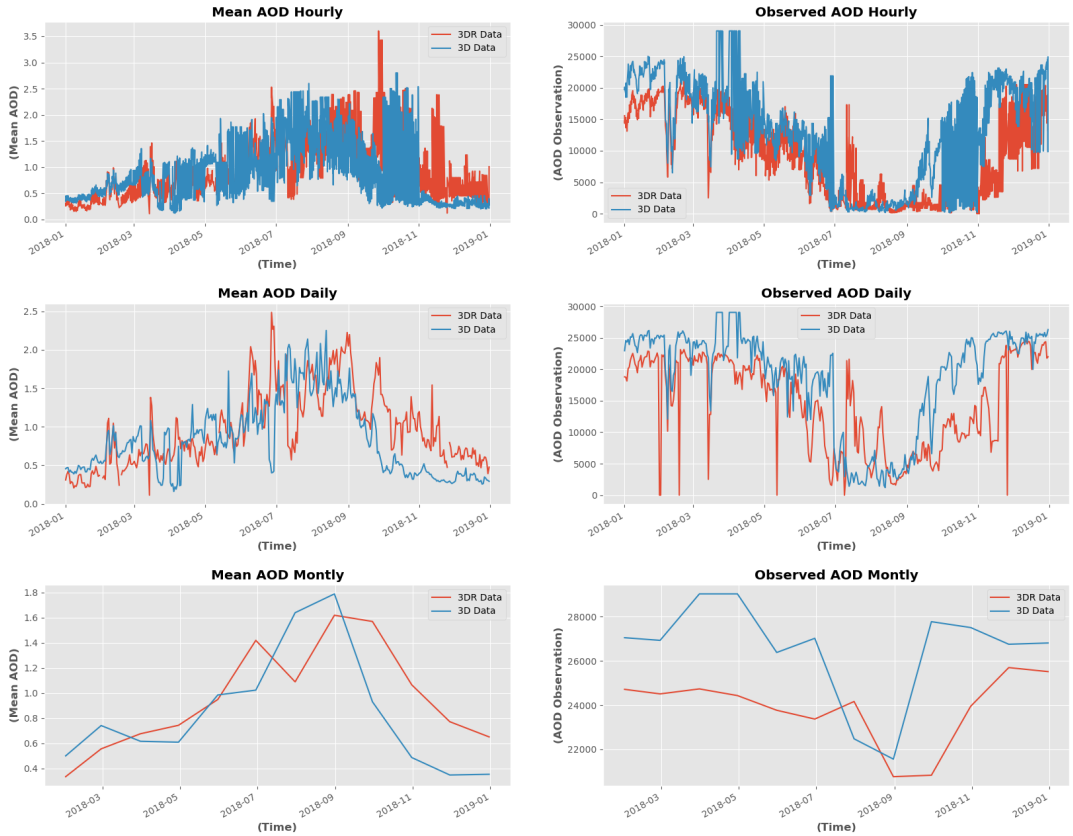


Figure 2.3: Average AOD and Total Valid Pixel Over India

yet the AOD values recorded by INSAT reveal a spike in the months of July-October, which is not consistent with the expected trend. Daily Records Analysis also made

which is shown in Figure 2.4

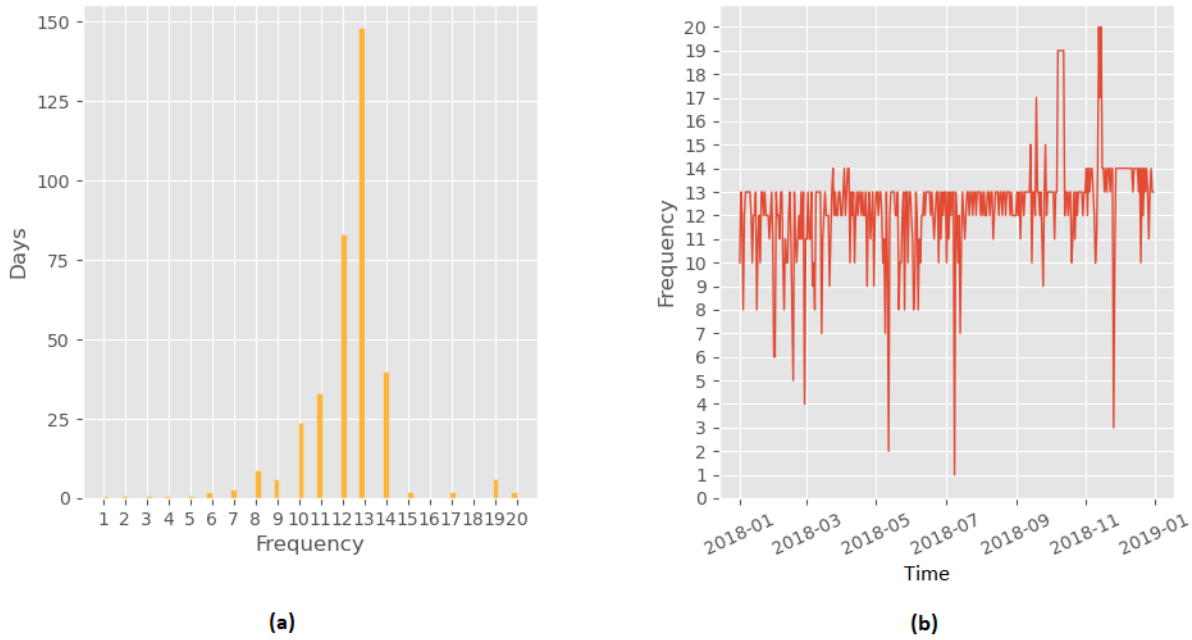


Figure 2.4: Combined INSAT 3D and INSAT 3DR AOD DATA a) Availability and b)Frequency Time Plot

The INSAT satellites, namely 3D and 3DR, have been found to capture up to 20 records per day, with 13 records being the most recorded figure in a single day.

2.4.2 INSAT vs MODIS

To gain further insights into the accuracy of INSAT AOD data, a comparison was made with AOD measurements from MODIS (Moderate Resolution Imaging Spectroradiometer), another satellite-based sensor commonly used to measure AOD levels. MODIS is an instrument on board two NASA Earth Observing System (EOS) satellites: Terra and Aqua. It was launched in 1999 and 2002, respectively. MODIS is designed to collect data about the Earth's surface, atmosphere, and oceans. MODIS collects data in 36

spectral bands, covering wavelengths from 0.4 to 14.4 micrometers. It provides data at a spatial resolution of 250 meters to 1 kilometer, depending on the spectral band. The instrument has a swath width of 2,330 kilometers, allowing it to cover the entire Earth's surface every one to two days. MODIS data is freely available to the public through the NASA Earth Observing System Data and Information System (EOSDIS). A variety of tools and resources are also available to help users visualize and analyze the data.

As shown in Table 2.4, the two satellites differ in their temporal and spatial resolution.

Table 2.4: Satellite Data Specification

SATELLITE AOD DATA	**SPATIAL RESOLUTION**	**TEMPORAL RESOLUTION**	**Wavelength (μm)**
MODIS	1^o X 1^o (100km X 100km)	Daily	0.55
INSAT 3D	0.1^o X 0.1^o (10km X 10km)	Half-Hourly	0.65

Various re-gridding criteria are utilized to match the trend between MODIS AOD and INSAT AOD, aiming to address to resolve the uncertainty occur in INSAT AOD.

Plot 2.5 show re-gridding techniques that involve calculating the spatial average by grouping 10 x 10 pixels around the coordinate of MODIS. A window size of ±5 pixels is created, and only the valid data falling under that window is included in the calculation of the mean. Furthermore, a criterion is set such that if the concentration of valid data is greater than or equal to the threshold, then the mean is computed, otherwise, it is rejected.

Formula,

$\bar{x} = (\Sigma x_i)/n$; if, n/N $\geq$ Threshold

where:

- $\bar{x}$ represents the mean of the dataset

-Σx_i represents the sum of all valid data points in space window

- n represents the number of valid data points in space window

- N represent the Total number of data points in space window

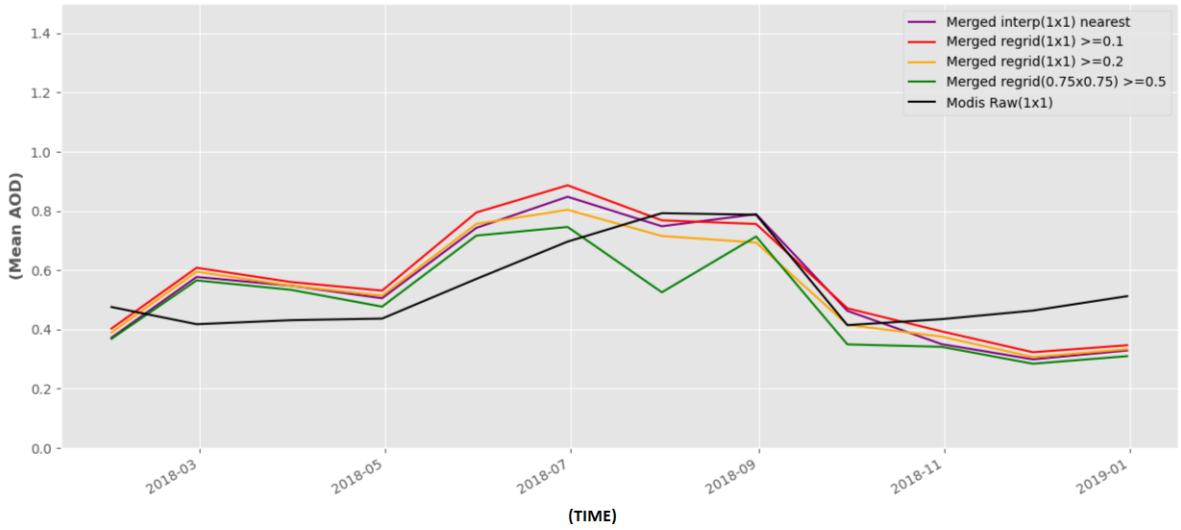


Figure 2.5: Different Regrid Method

Interpolation Method Nearest, 3D and 3DR Merged regrid with threshold value 20% and 10% shows good agreement with MODIS monthly average AOD over INDIA.

The application of a one standard deviation (sigma) rule on INSAT AOD is used to eliminate invalid data records that exhibit excessive daily variation, as a solution to address the abrupt increase in AOD during the monsoon.The one-sigma standard deviation rule is a statistical technique that helps to identify data points that are significantly different from the average value. This technique is commonly used to eliminate outliers in a

dataset and filter valid data records from the pool of data.

The one-sigma standard deviation rule involves calculating the mean (average) value of a dataset, and then determining the amount of variation in the data by calculating the standard deviation. The standard deviation is a measure of how much the data values differ from the mean. According to the one-sigma standard deviation rule, data points that fall outside the range of one standard deviation from the mean are outliers and are removed from the dataset.

Formula,

$$\mu_{loc} - \sigma_{loc} \leq x_i \leq \mu_{loc} + \sigma_{loc}$$

Where,

x_i: valid data at time t for a particular day.

σ_{loc}: Standard deviation over a day for a particular location.

μ_{loc}: Mean over a day for a particular location.

To convert hourly INSAT AOD data into daily data, group the information by day and create a daily vector for each location. Calculate the mean and standard deviation for each location's daily vector, apply a filtering process using the one sigma rule, and ensure that each day has at least 6 records. Finally, resample the data to obtain daily data and create plot 2.6.

It can be observed that INSAT monthly average during Sep 2019 is very high but when filter is applied on INSAT AOD 10x10km spatial resolution, it value reduced significantly but result is massive data loss hence it is trade of between accuracy and data loss.

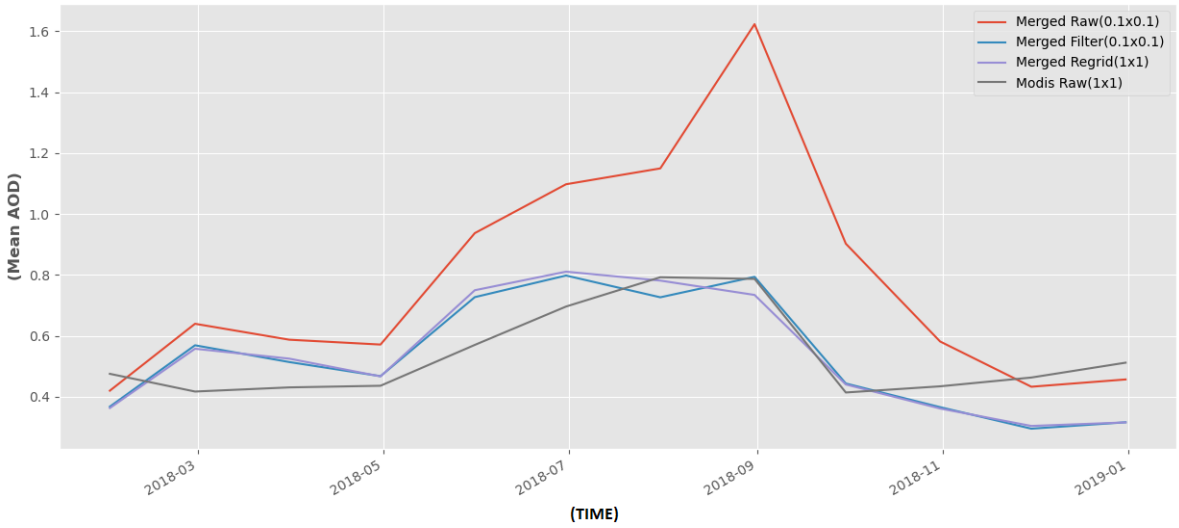


Figure 2.6: Comparison Of Monthly Average of Filter INSAT, Regrid INSAT and MODIS AOD

2.4.3 Validation With Point-Location AERONET

To validate the accuracy and reliability of INSAT 3D's data, it is necessary to compare it with measurements taken from ground-based instruments. One method of doing so is by utilizing a network of sun photometers called AERONET, which can measure atmospheric aerosols and their optical properties. In the case of INSAT 3D, the validation process involved comparing the satellite's data with the measurements obtained from AERONET's sun photometers at three different locations: Gandhi College, Jaipur, and Pune, during the year 2018. This comparison was performed daily.

Since AERONET Data has an Average Temporal Resolution of 15min whereas INSAT 3D have Half-hourly Temporal Resolution we are going to apply Spatiotemporal Method explained by (Ichoku et al., 2002).

- Define a time window of ±30 minutes and a spatial window of 5x5 pixels (50x50 km).

- Select data points that have at least three valid AERONET observations and twelve valid INSAT-3D observations within the defined spatiotemporal window.
- Calculate the temporal average of AERONET AOD measurements within the time window.
- Calculate the spatial average of INSAT-3D AOD measurements within the spatial window.
- Compare the temporal average of AERONET with the spatial average of INSAT-3D AOD measurements.
- Reject any AOD pixels that may have been contaminated by cloud edges or residual subpixel clouds by performing a standard deviation test, with a threshold of less than 0.2, as suggested by Kolmonen et al. (2013).

Although the spatial variation of aerosols is generally smoother, cloud contamination can introduce noise in the retrieved AOD. To address this issue, a standard deviation test can be used to reject AOD pixels that may have been potentially contaminated by cloud edges or residual subpixel clouds, as suggested by Kolmonen et al. (2013).

Table 2.5: AERONET And INSAT Comparison Metric

Location	Root Mean Square Error (RMSE)	Correlation Coefficient (R)
Gandhi College	0.33	0.15
Jaipur	0.34	0.15
Pune	0.57	0.32

The Correlation Coefficient (R) values indicate the strength of the relationship between the AERONET AOD and INSAT AOD, with values closer to 1 indicating a stronger

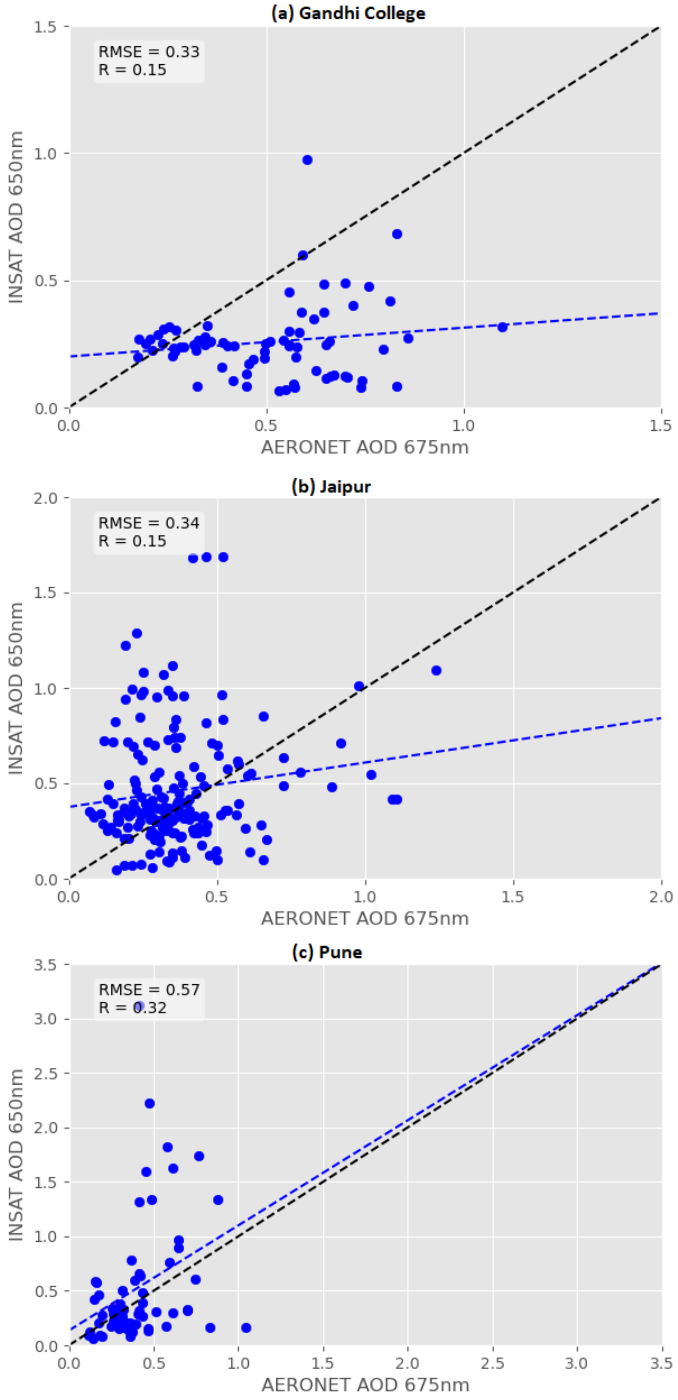


Figure 2.7: Scatter Plot of AERONET (675nm) and INSAT 3D (650nm)

relationship. Based on the table, all three locations have a relatively low correlation coefficient value of around 0.15 to 0.32, indicating a weak relationship between the predicted and actual values.

Overall, the results suggest that the predictive model used to generate these values may not be very accurate, as evidenced by the relatively high RMSE values and low correlation coefficient values. Further investigation and refinement of the predictive model may be needed to improve the accuracy of the predictions.

2.5 SATELLITE OVERPASS TIME

Satellite overpass time refers to the time when a satellite passes over a particular location on Earth. In India, the overpass time of Aqua and Terra varies depending on the latitude and longitude of the location. For example, in northern India, the overpass time for Aqua is typically around 11:00 AM to 12:00 PM local time, while for Terra it is around 10:30 AM to 11:30 AM local time. In southern India, the overpass time for Aqua is typically around 1:00 PM to 2:00 PM local time, while for Terra it is around 12:30 PM to 1:30 PM local time.

When planning to acquire and analyze satellite data, it is crucial to take into account the overpass time of MODIS Aqua and Terra. This timing is significant because acquiring satellite data during overpass time is essential for obtaining accurate daily mean readings of PM2.5 and BLH. Since MODIS only passes over a location once in a day, it is essential to capture data during that precise moment to ensure that the PM2.5, INSAT AOD and BLH parameters are considered.

2.5.1 Seasonal Mean

The methodology for calculating the Seasonal mean of PM2.5, INSAT 3D AOD, and BLH using the satellite overpass time of MODIS Aqua is as follows:

- Retrieve the overpass time of Aqua for each site location separately.
- If the overpass time of Aqua is not available, derive it from the Terra satellite, which is ahead of Aqua by 3:30 GMT.
- If Terra is also not available, discard the entire day.
- Create an Aqua overpass time window of ±30 minutes.
- Calculate the average of only valid data points that fall inside the time window.
- Calculate the daily mean using the data obtained in this manner.
- Select the months for each season: Winter (DJF), Pre-monsoon (MAM), Monsoon (JJAS), and Post-monsoon (ON).
- For the winter season (DJF), the data for December is taken from the previous year, and for the rest of the months, the data is taken from the current year.
- For each season, aggregate the data for the respective months from the current year for PM2.5, MODIS AOD, INSAT AOD, and BLH.

This approach ensures that seasonal data is correctly aggregated and avoids any potential errors or inconsistencies that may arise from using incorrect data.

Chapter 3

MACHINE LEARNING MODEL

3.1 DATA PREPROCESSING AND FEATURE SELECTION

Data preprocessing is a crucial step to ensure that the data collected is of high quality and suitable for analysis. Before merging the data from different site locations and seasons, it underwent several preprocessing steps. Some of the techniques employed for data preprocessing include:

- Data cleaning: This step involved removing any duplicates, missing or irrelevant data points, and correcting any inconsistencies or errors in the data. One approach used was to remove entire rows from the dataset if any column contained NaN values, which could potentially skew the results of the analysis. Additionally, any -999 values were replaced with NaN values to enable easier identification and handling of missing values in the dataset.

- Data transformation: To merge the data from different sources, it was necessary to transform it into a common format that is suitable for merging. This involved converting the PM2.5 values, which were recorded in India Standard Time (IST), to (Universal Time Coordinate) UTC format, so that they could be combined with the AOD and BLH data, which had time values in UTC format. Once the data

was in a consistent format, it was merged into a single dataset.

In addition to this, categorical values were converted into numerical data using a label encoder. The label encoder is a preprocessing technique that assigns a unique numerical value to each category in a categorical variable. This is useful because many machine learning algorithms require numerical data as input. The label encoder allows us to convert categorical data into a format that can be used for modeling and analysis, without losing the information contained in the original categories.

Table 3.1: Categorical Value to Numerical Value Encoding

Column	Name	Unique ID
City	Patna	0
	Delhi	1
	Ahmedabad	2
	Faridabad	3
	Bengaluru	4
	Mumbai	5
	Pune	6
	Chennai	7
	Hyderabad	8
	Agra	9
	Kanpur	10
	Lucknow	12
Season	Winter	3
	Pre-Monsoon	2
	Monsoon	0
	Post-Monsoon	1

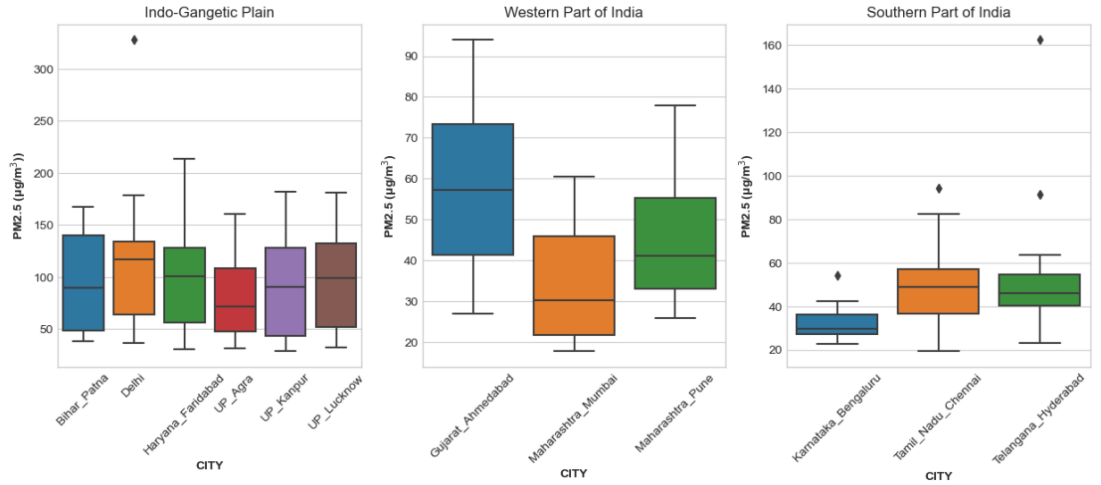
The label encoder assigns a unique numerical value to each category in a categorical variable, but the values are not necessarily assigned in alphabetical order. The label encoder assigns numerical values based on the order in which the categories appear in the dataset, with the first category being assigned a value of 0,

the second category being assigned a value of 1, and so on. Therefore, the numerical values assigned to the city and season names in the dataset will depend on the order in which they appear in the dataset, rather than their alphabetical order. If alphabetical order is desired, the data should be sorted alphabetically prior to applying the label encoder.

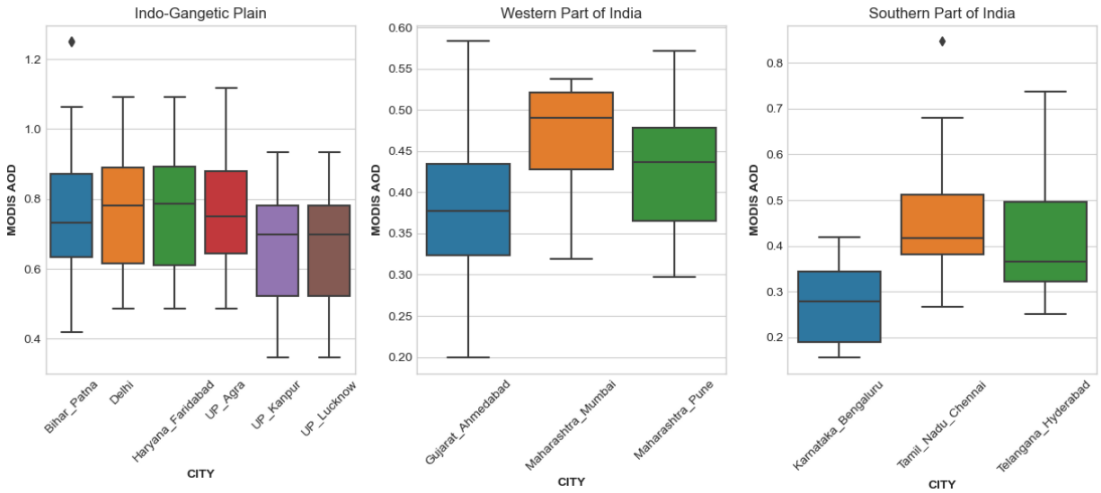
- Data aggregation: The data was aggregated by site location and season to generate the average PM2.5, INSAT AOD, MODIS AOD, and BLH values for each site location and season which is already mentioned is Chapter 2 Section 2.5.1 Seasonal Mean.

- Outlier detection: To avoid obtaining biased results, it is crucial to identify and eliminate outliers from the dataset, as they can have a significant impact on the analysis outcomes. Therefore, using tools such as box and violin plots can help in identifying outliers, which can then be removed from the data frame.
  Fig 3.1 show box plot of different region i.e. Indo Gangetic Plain which is densely populated by aerosol particle , Western Region of India i.e. Mumbai experience lower level of PM2.5 because it is situated at coastal area which helps in reducing the PM2.5 concentration by bringing in fresh air from the sea and The Southern region of India is located on a high plateau, which allows for good ventilation and air circulation.

- Normalization: This technique was applied to ensure that the data from different sources was on the same scale to avoid any undue influence of one variable over another. And, to reduce computation power while training data on different parameters it is required to bring down large number to small scale using Standard



(a)



(b)

Figure 3.1: Box Plot Of (a) PM2.5 Concentration (b) MODIS AOD For Different Region.

Scaler method. The formula for standard scaling (also known as z-score normalization) is:

$$z = (x - \mu)/\sigma$$

Where,

- x is the value of a specific observation or variable

- μ is the mean (average) value of the entire dataset

- σ is the standard deviation of the entire dataset

- z is the standardized value of x, which represents the number of standard deviations x is away from the mean.

The standard scaler rescales the dataset so that it has a mean of 0 and a standard deviation of 1, which makes it easier to compare variables that have different units or scales.

After applying these techniques, the useful feature that related to study are selected, which included seasonal average of PM2.5, INSAT AOD, MODIS AOD, and BLH for each site location and season. This data format enables easy comparison and analysis of the trends and patterns in the air quality data across different locations and seasons.

Table 3.2: Training Dataset Column Description

Columns	Data Type	Total Entries	Unique Value	Range
City	Object	292	12	0-11
Season	Object		04	0-3
Year	Int64		7	2015 - 2021
MODIS AOD	Float64		-	0.155000 - 1.251000
INSAT AOD	Float64		-	0.099633 - 4.658120
BLH	Float64		-	202.345720- 2924.515400
PM2.5			-	17.796308 - 327.505618

3.2 DATA ANALYSIS

Data analysis is a crucial step in building machine learning models, as it helps in gaining insights into the data and identifying patterns that can aid in model selection and feature engineering. In this sub-chapter, we focus on the analysis of three key variables

- PM2.5, AOD, and BLH, and how they are related to each other.

One way to visualize the distribution of these variables is by plotting a histogram. A histogram is a graphical representation of the frequency distribution of a variable. By plotting histograms for PM2.5, AOD, and BLH, we can gain insights into their distribution, such as whether they follow a normal distribution, have a skewed distribution, or have outliers.

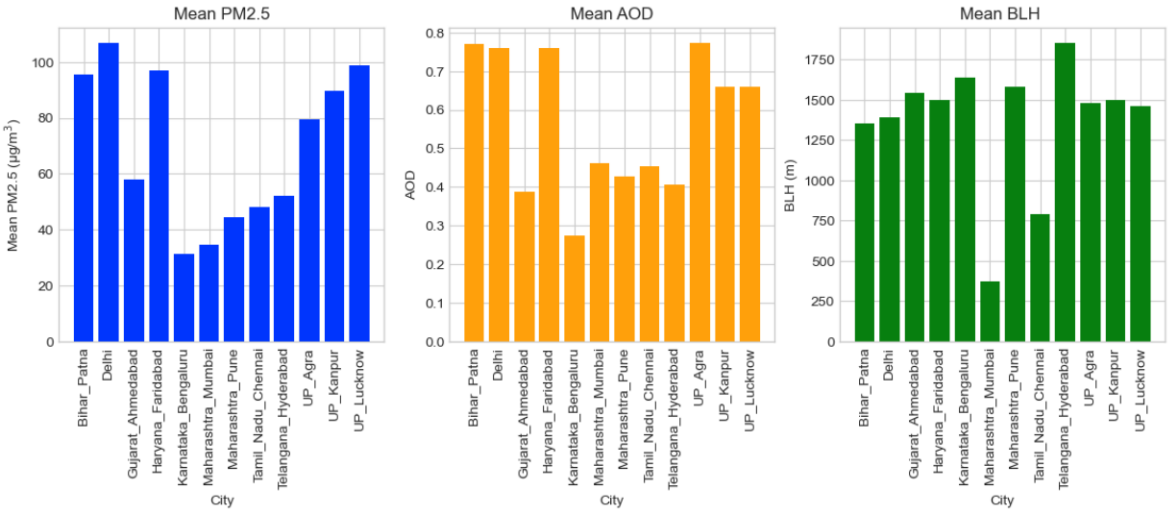


Figure 3.2: Bar plot of Mean PM2.5, MODIS AOD, BLH Level by City

In addition to histograms, we can also create a correlation matrix for the dataset. A correlation matrix is a table that shows the correlation coefficients between different variables in a dataset. The correlation coefficient measures the strength and direction of the linear relationship between two variables. By creating a correlation matrix for PM2.5, AOD, and BLH, we can identify whether there are any strong correlations between these variables. For instance, a high positive correlation of 0.5 between PM2.5 and MODIS AOD is shown in Fig 3.3.

Overall, analyzing the distribution of PM2.5, AOD, and BLH through histograms and

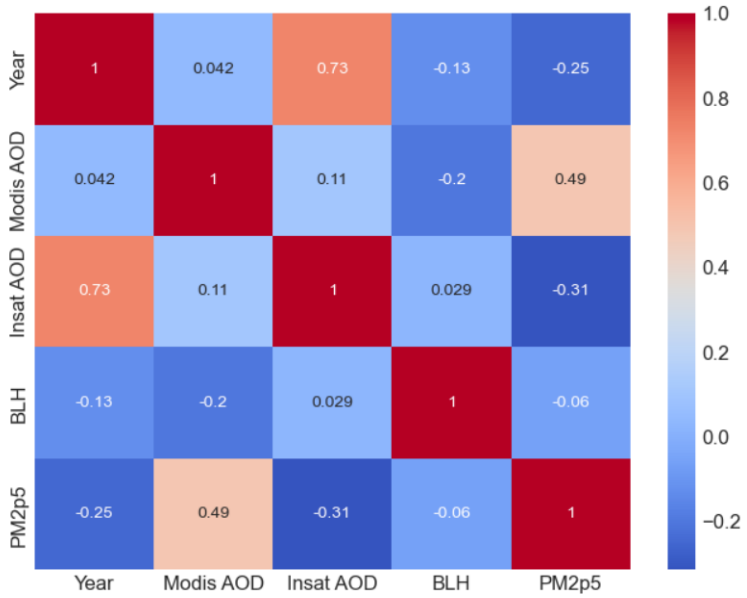


Figure 3.3: Correlation Matrix

creating a correlation matrix can help in gaining insights into the data and identifying important relationships between the variables. These insights can be used to guide model selection and feature engineering, and ultimately improve the accuracy of the machine learning model.

3.3 MODELING

To Train the model, data is split into three parts i.e., Training (65% of Dataset), Testing (20% of Dataset) and Validation (15% of Dataset). For predicting PM2.5 different Supervised Learning Model are trained to determine which model performs better thus the model Extreme Gradient Boosting (XGBoost), Random Forest Estimator and Artificial Neural Network is used.

3.3.1 Random Forest

Random forest is an ensemble method that combines the predictions of multiple decision trees to make a final prediction. In the case of regression, the output of each decision tree is a continuous value, so the final prediction of the random forest is the average of the outputs of all the individual decision trees. The averaging process helps to reduce the variance of the final prediction and improve the generalization performance of the model.

Following are the Random Forest Parameters which are essential to build a good predictive model.

- ‘n_estimators‘: The number of trees in the forest.
- ‘max_depth‘: The maximum depth of each decision tree in the forest.
- ‘max_features‘: The maximum number of features to consider when looking for the best split at each node.
- ‘min_samples_split‘: The minimum number of samples required to split an internal node.
- ‘min_samples_leaf‘: The minimum number of samples required to be at a leaf node.
- ‘bootstrap‘: Whether to use bootstrap samples when building trees. If True, the size of the dataset used for building each tree will be the same as the original dataset. If False, each tree will be built using the entire dataset.

3.3.2 Extreme Gradient Boosting

XGBoost also use ensemble learning algorithm that uses decision trees as base learners to achieve high scalability, efficiency, and accuracy. Here is a simple algorithm for XGBoost:

1. Initialize the ensemble with a simple decision tree.
2. Compute the gradient of the loss function with respect to the predictions of the current ensemble.
3. Fit a new decision tree to the gradient using the training data.
4. Add the new tree to the ensemble.
5. Repeat steps 2-4 until a stopping criterion is met, such as reaching a maximum number of iterations or achieving a minimum level of improvement in the validation error.

Some of the most important hyperparameters include:

- ‘n_estimators‘: The number of trees in the ensemble.
- ‘max_depth‘: The maximum depth of each tree in the ensemble.
- ‘min_child_weight‘: The minimum sum of instance weight (hessian) needed in a child node.
- ‘learning_rate‘: The step size shrinkage used to prevent overfitting.
- ‘colsample_bytree‘: The fraction of columns (features) to be randomly subsampled for each tree.
- ‘subsample‘: The fraction of instances (observations) to be randomly subsampled for each tree.
- ‘lambda‘: L2 regularization term on weights (also known as Ridge regularization).
- ‘alpha‘: L1 regularization term on weights (also known as Lasso regularization).

- ‘gamma‘: Minimum loss reduction required to make a further partition on a leaf node of the tree.

3.3.3 Artificial Neural Network

Artificial neural networks (ANNs) are a branch of artificial intelligence that emulates the structure and functionality of the human brain. They consist of an input layer where input data is fed, followed by randomly generated weights that are used in combination with an activation function to pass data through each subsequent layer of nodes. The output layer of the network provides the result. Following are the algorithm steps.

- Initialize the weights of the neural network using a random number.
- Feed the input data through the network and compute the output.
- Output is computed by summation of all weight multiple by input value and bias is added to keep the node alive.
- Output is then passes through activation function which decided the keep the node activate or make it dead if condition is not satisfied.
- Neural network each layer keeps going forward repeat the same step to calculate the output send it to activation function and reach to output layer where backpropagation start when loss function is calculated.
- Calculate the loss between the predicted output and the actual output.
- Compute the gradients of the loss function with respect to the weights of the network using backpropagation.
- Use an optimizer, such as Adam or SGD, to update the weights of the network in the direction of the negative gradient.

- Apply the activation function to the output of each neuron in the hidden layers of the network to introduce nonlinearity and increase its expressive power.
- Repeat steps 2-6 for a specified number of epochs, which refers to the number of times the entire training dataset has passed through the network.
- After each epoch, evaluate the performance of the network on a validation dataset to monitor for overfitting.
- Repeat steps 2-8 until the validation performance stops improving, or until a maximum number of epochs is reached.
- Once the model is trained, use it to make predictions on new, unseen data.

The following are the parameters of ANN.

- Optimizer: The algorithm used to update the weights of an ANN during training, such as Adam or SGD.
- Loss function: The function used to measure the error between the predicted output and the actual output of an ANN, such as Mean Squared Error (MSE) or Mean Absolute Error (MAE).
- Activation function: The non-linear function applied to the output of each neuron in an ANN to introduce non-linearity, such as ReLU or Sigmoid.
- Epochs: The number of times the entire training dataset is passed through an ANN during training.
- Batch size: The number of training samples used to update the weights of an ANN in each iteration during training of ANN:

3.4 HYPER PARAMETER TUNING

To achieve the best performance and attain state-of-the-art accuracy in predicting PM2.5, all model parameters need to be optimized. This process involves trial and error, but there are Python packages that can reduce the workload by performing an exhaustive search within the given parameter options.

However, it should be noted that there are limitations to the search as it can only be performed within a given set of parameter combinations. As the number of possible combinations increases, the computational requirements also increase, and it can take a considerable amount of time to find the optimal parameter values.

Table 3.3: Optimal Hyper Parameter for each model

Model	Hyper Parameter	Value
Random Forest Regressor	bootstrap	False
	max_depth	10
	max_features	'sqrt'
	min_samples_leaf	2
	min_sample_split	5
	n_estimators	50
XGBoost	alpha	0.02
	colsample_bytree	0.78
	gamma	0.01
	lambda	6.8
	learning rate	0.1
	max_depth	7
	min_child_weight	2
	n_estimators	110
	subsample	0.7
Artificial Neural Network	optimizer	adam
	loss function	rmse(Mean Squared Error)
	epochs	200
	batch_size	10
	activation	Relu(Rectified Linear)

3.5 METRICS

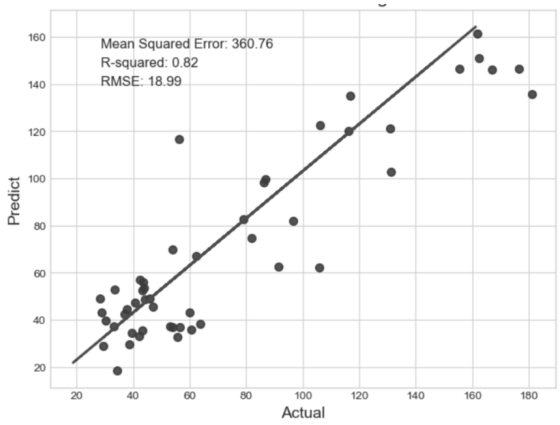
The performance of Trained Models can be compared using certain metrics such as Mean Squared Error, Root Mean Square Error, and R Squared Value.

- Mean Squared Error (MSE): This is a common metric used to evaluate the performance of regression models. It measures the average squared difference between the predicted and actual values. A lower MSE indicates better performance.
- Root Mean Squared Error (RMSE): This is the square root of the MSE and is another popular metric used for regression models. It has the same units as the predicted variable and is more interpretable than the MSE.
- R-squared (R^2) Value: This metric measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, with 1 indicating a perfect fit and 0 indicating no correlation between the independent and dependent variables. A higher R^2 value indicates better model performance. Following Fig 3.4 shows scatter plot for predicted value and actual value.

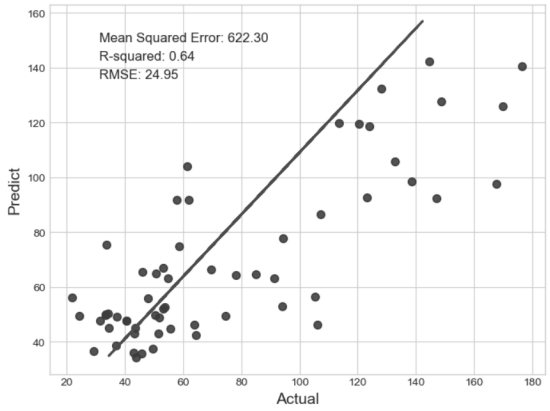
Table 3.4: Different Model Metric Table

MODEL	Mean Square Error	RMSE	R^2 value
XGBoost	360.76	18.9	0.82
ANN	622.3	24.95	0.64
Random Forest	332.32	18.2	0.81

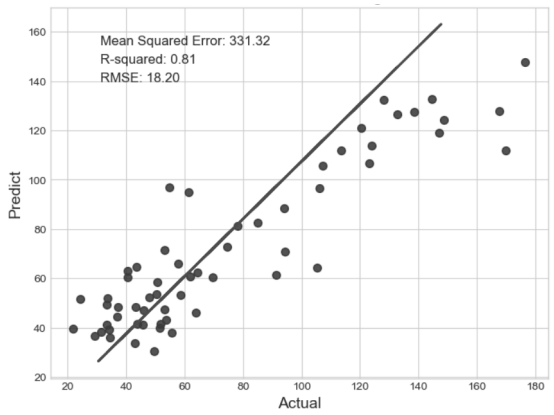
This metric tells us that the performance of XGBoost and Random Forest model are almost same with R^2 value above 81.



(a)



(b)



(c)

Figure 3.4: Scatter Plot for Model (a) XGBoost (b) ANN (c) Random Forest

Chapter 4

VALIDATION

Validation is performed by evaluating the trained model's performance on a validation set, where the model's ability to generalize to new data is assessed. The validation Dataset is collected for all city's season data for one year i.e., 2019. The inclusion of Figure 4.1 is particularly helpful, as it presents the results of the model's performance in a clear and easily understandable manner.

The figure clearly indicates that the seasonal prediction of PM2.5 during the winter and post-monsoon periods is highly accurate, with correlation coefficients (R) of 0.96 and 0.83, respectively. However, the accuracy of the prediction for the pre-monsoon and monsoon periods is significantly lower, with correlation coefficients of only 0.59 and 0.54, respectively.

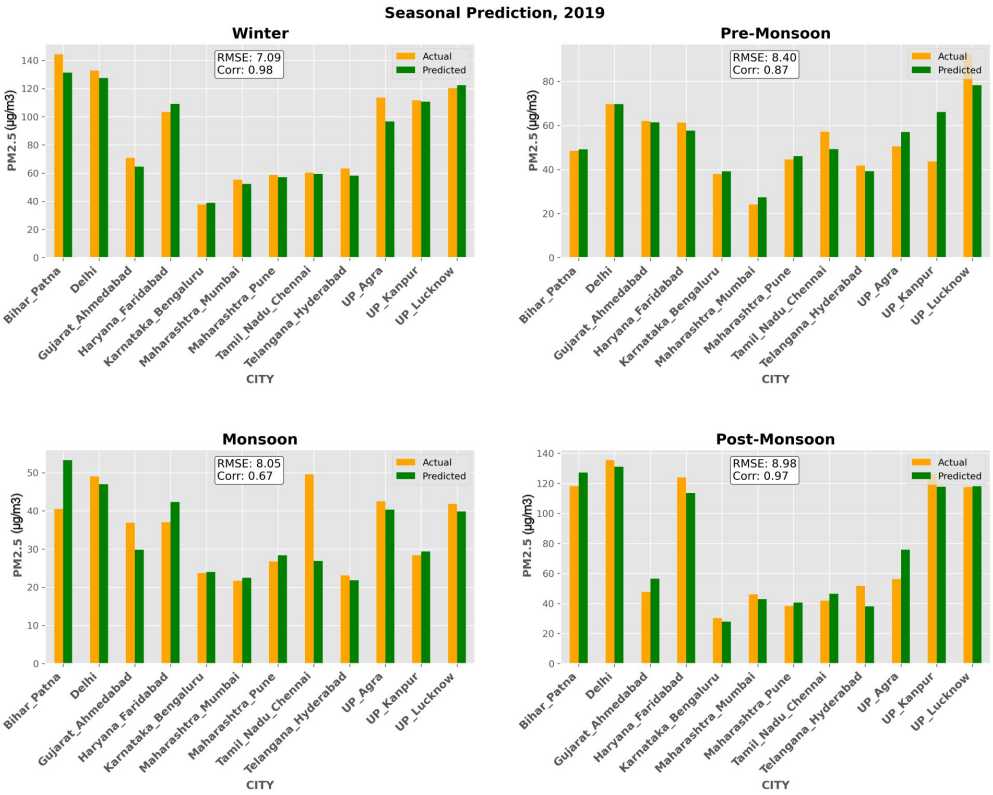


Figure 4.1: Seasonal Prediction of PM2.5 During 2019

Chapter 5

DISCUSSION and CONCLUSION

This study addresses the issue of air pollution by implementing Seasonal Prediction of PM2.5 using a Supervised Machine Learning Model for specific cities such as Patna, Delhi, Ahmedabad, Faridabad, Bengaluru, Mumbai, Pune, Chennai, Hyderabad, Agra, Kanpur, and Lucknow. We performed a comprehensive analysis on INSAT 3D and 3DR Satellite level-2 Imager AOD Data launched by ISRO and compared the results with MODIS Aqua AOD. Unfortunately, the correlation between them was only 0.18, signifying no correlation. Moreover, we validated the results with AERONET Sun Photometer on various locations such as Gandhi College, Jaipur, and Pune, and the correlation coefficients measured were only 0.15, 0.15, and 0.32, respectively, with corresponding RMSE of 0.33, 0.34, and 0.57. This indicates that INSAT 3D and 3DR suffer from a lot of uncertainty and need improvement on its retrieval algorithm and preprocessing method.

However, despite the uncertainty of INSAT 3D and 3DR, we were able to predict PM2.5 levels with good accuracy in different cities using XGBoost and Random Forest models. The XGBoost and Random Forest model outperformed the ANN with R^2 values of 0.82, 0.81, and 0.64 for datasets with a limited number of records and features. These findings suggest that INSAT 3D and 3DR may not be the most reliable sources for PM2.5 prediction due to their uncertain nature and lack of correlation with AERONET data.

However, we were able to overcome this obstacle by using fine-tuned XGBoost and Random Forest models to predict PM2.5 levels in different cities with good accuracy. These models can be useful if it is tuned properly in addressing the problem of air pollution in the future, particularly in areas where air quality data is limited.

Further research is needed to improve the reliability and accuracy of INSAT 3D and 3DR for PM2.5 prediction. Additionally, more diverse data from different locations and representing different environmental conditions should be used and analyzed to improve the performance of the machine learning models. Overall, this study contributes to the growing body of knowledge on air quality prediction and the use of machine learning techniques in addressing environmental problems.

References

Fu, Disong, Christian A. Gueymard, Dazhi Yang, Yu Zheng, Xiangao Xia, and Jianchun Bian. ”Improving aerosol optical depth retrievals from Himawari-8 with ensemble learning enhancement: Validation over Asia.” Atmospheric Research (2023): 106624.

Gupta, Amitesh, Yogesh Kant, Debashis Mitra, and Prakash Chauhan. ”Spatio-temporal distribution of INSAT-3D AOD derived particulate matter concentration over India.” Atmospheric Pollution Research 12, no. 1 (2021): 159-172.

Ichoku, Charles, D. Allen Chu, Shana Mattoo, Yoram J. Kaufman, Lorraine A. Remer, Didier Tanré, Ilya Slutsker, and Brent N. Holben. ”A spatio-temporal approach for global validation and analysis of MODIS aerosol products.” Geophysical Research Letters 29, no. 12 (2002): MOD1-1.

Kolmonen, P., A-M. Sundström, L. Sogacheva, E. Rodriguez, T. Virtanen, and G. de Leeuw. ”Uncertainty characterization of AOD for the AATSR dual and single view retrieval algorithms.” Atmospheric Measurement Techniques Discussions 6, no. 2 (2013): 4039-4075.

Mishra, Manoj K. ”Retrieval of Aerosol Optical Depth From INSAT-3D Imager Over Asian Landmass and Adjoining Ocean: Retrieval Uncertainty and Validation.” Journal of Geophysical Research: Atmospheres 123, no. 10 (2018): 5484-5508.

Singh, Ramesh P., and Akshansha Chauhan. ”Sources of atmospheric pollution in India.” In Asian Atmospheric Pollution, pp. 1-37. Elsevier, 2022.

Xing, Yu-Fei, Yue-Hua Xu, Min-Hua Shi, and Yi-Xin Lian. "The impact of PM2. 5 on the human respiratory system." Journal of thoracic disease 8, no. 1 (2016): E69.

Vadgama, Dhyani, Tejas Turakhia, Akhil S. Nair, Rajesh Iyer, and Abha Chhabra. "Study of Particulate Matter over Ahmedabad and Gandhinagar Cities: A Case Study over Two Years." In 2021 IEEE International India Geoscience and Remote Sensing Symposium (InGARSS), pp. 316-319. IEEE, 2021.