A dive into India's smog crisis

BSDCH ZC229T: Design Project

by

Aryan Bhardwaj

202117B3728

Design Project work carried out at

HCLTech, Noida

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)

June 2024

A dive into India's smog crisis

BSDCH ZC229T: Design Project

by

Aryan Bhardwaj

202117B3728

Design Project work carried out at

HCLTech, Noida

Submitted in partial fulfillment of B.Sc. (Design and Computing) degree programme

Under the Supervision of
Subhadip Chakrabarti,
HCLTech, Kolkata

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)

June 2024

CERTIFICATE

This is to certify that the Design Project entitled **A dive into India's smog crisis** and submitted by **Aryan Bhardwaj** having BITS ID No. **202117B3728** for the partial fulfillment of the requirements of B.Sc. (Design and Computing) degree program of BITS, embodies the Bonafede work done by him/her under my supervision.

Subhadip

Signature of the Mentor

Place : Kolkata

Date : 18th June 2024

Subhadip Chakrabarti,Technical Architect,HCL,Kolkata

Name, Designation & Organization &Location

Abstract

This project aims to analyze the air quality index (AQI) of various cities and states in India using python. The dataset used contains historical data of AQI readings from monitoring stations across India. We conducted a comprehensive exploratory analysis of historical AQI data from Indian states and cities. Key visual insights included trends of AQI throughout the years, AQI trends for different states and cities in India, correlation of toxic gases with AQI (top 3 gases) and AQI visualizations on India maps. The dataset was thoroughly cleaned before usage to avoid bias and errors. Our dataset includes over 100,000 data points, significantly more than previous research papers. We aimed to extract meaningful insights. Unlike other studies that used limited data, we leveraged a larger dataset and created interactive, visually appealing maps to showcase AQI trends in India. The quality of air can be measured. AQI is a measure that determines the quality of the air. AQI is a random scale ranging from 0 to 500. Generally, the more the AQI, the more the level of air pollution and more the health concern. For example, an AQI of 50 or less is said to be good air quality and AQI of 300 or above resembles hazardous quality. AQI is basically classified into six categories of health concerns. Each category is given a color for easy interpretation. The six categories are Green (0-50), Yellow (51-100), Orange (101-150), Red (151-200), Purple (201 300), and Maroon (300 and higher). Calculating AQI is essential because it provides critical information about the state of the air. We can stop people getting effected due to pollution by warning about the air quality whenever AQI exceeds the maximum value

Acknowledgements

I would like to express my deepest gratitude to everyone who contributed to the completion of this project.

First and foremost, I am immensely grateful to my project mentor, **Subhadip Chakrabarti**, for their invaluable guidance, support, and encouragement throughout this project. Their insights and feedback were crucial in shaping the direction and outcome of this work.

I would also like to thank **HCLTech** for providing the resources and facilities necessary for this project. Special thanks to **Samuel Santha Sekaran** for their assistance and cooperation.

Lastly, I am forever grateful to my family and friends for their unwavering support and encouragement throughout this journey. Your belief in me has been a constant source of motivation.

Thank you all for your contributions and support.

Table of Contents

1. Chapter 1 Introduction
2. Chapter 2 Data collection and cleaning
3. Chapter 3 Exploratory Data Analysis
4. Chapter 4 Descriptive Statistics
5. Chapter 5 State wise analysis
6. Chapter 6 Harmful gas analysis
7. Chapter 7 Machine learning model application

List of Figures

1. Graph on India’s AQI for 5 years
2. Graph on India’s AQI for major cities
3. Heatmap for affect of gas on AQI
4. Line graph of major pollutants in Delhi
5. Line graph of major pollutants in Gurugram

Chapter 1

The Air Quality Index (AQI) is a standardized tool used to communicate the quality of air in a specific area. In India, the AQI system is designed to make air quality data easily understandable by converting complex air pollution data of various pollutants into a single number, nomenclature, and color. This system plays a critical role in public health by informing citizens about the quality of the air they breathe and potential health impacts.

Key Components of India's AQI

1. **Pollutants Monitored**: The AQI in India includes eight pollutants: Particulate Matter (PM10 and PM2.5), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), Carbon Monoxide (CO), Ozone (O3), Ammonia (NH3), and Lead (Pb). These pollutants are chosen due to their prevalence and potential health impacts.
2. **Categories and Color Coding**: The AQI is divided into six categories:
   - **Good (0-50)**: Minimal impact on health.
   - **Satisfactory (51-100)**: Minor breathing discomfort to sensitive people.
   - **Moderately Polluted (101-200)**: Breathing discomfort to people with lung diseases, children, and older adults.
   - **Poor (201-300)**: Breathing discomfort to most people on prolonged exposure.
   - **Very Poor (301-400)**: Respiratory illness on prolonged exposure.
   - **Severe (401-500)**: Affects healthy people and seriously impacts those with existing diseases.
3. **Data Collection and Dissemination**: The Central Pollution Control Board (CPCB), under the Ministry of Environment, Forest, and Climate Change (MoEFCC), is responsible for monitoring and disseminating AQI data. Data is collected from various monitoring stations across the country and is made available to the public through websites, mobile applications, and public displays.

Delhi Pollution

Air pollution in Delhi has become a critical public health and environmental issue, garnering national and international attention due to its severe impacts. The capital city of India frequently records some of the highest levels of air pollution in the world, particularly during the winter months when weather conditions exacerbate the problem.

Sources of Air Pollution in Delhi

1. **Vehicular Emissions**: With millions of vehicles on the road, emissions from cars, trucks, and two-wheelers contribute significantly to Delhi's air pollution. Traffic congestion and the prevalence of older, more polluting vehicles worsen the situation.
2. **Industrial Emissions**: Factories and industrial units, especially those using coal and other fossil fuels, emit large amounts of pollutants, including sulfur dioxide (SO2), nitrogen oxides (NOx), and particulate matter (PM).
3. **Construction Activities**: Rapid urbanization and construction activities generate substantial dust and particulate matter, contributing to the city's poor air quality.
4. **Biomass Burning**: The burning of agricultural residues in neighboring states, particularly Punjab and Haryana, contributes heavily to seasonal air pollution in Delhi. During harvest season, smoke from these fires drifts into Delhi, severely affecting air quality.
5. **Domestic Sources**: The use of wood, coal, and other solid fuels for cooking and heating in poorer households adds to indoor and outdoor air pollution.
6. **Meteorological Factors**: During winter, Delhi experiences temperature inversions where cooler air gets trapped near the ground under a layer of warmer air, preventing pollutants from dispersing. This leads to higher concentrations of pollutants in the air.

Health Impacts

The high levels of air pollution in Delhi have serious health consequences. Short-term exposure can cause respiratory issues, eye irritation, and reduced lung function. Long-term exposure has been linked to chronic respiratory diseases, cardiovascular diseases, lung cancer, and premature death. Vulnerable groups such as children, the elderly, and those with pre-existing health conditions are particularly at risk.

Chapter 2

Data Collection

The dataset encompasses comprehensive air quality data and Air Quality Index (AQI) measurements, meticulously recorded at both hourly and daily intervals across various monitoring stations situated in multiple cities throughout India. This detailed dataset provides a robust framework for analyzing temporal variations in air quality, capturing the intricate patterns of pollution levels within different urban environments. By offering granular insights at an hourly level, alongside broader daily trends, the dataset facilitates a thorough examination of the factors influencing air pollution. Researchers, policymakers, and public health officials can leverage this extensive data to identify pollution hotspots, evaluate the effectiveness of pollution control measures, and devise strategies for mitigating air pollution's adverse impacts on public health and the environment.

The dataset encompasses comprehensive air quality data for major cities across India, covering the period from 2015 to 2020. This extensive dataset provides a detailed overview of air quality trends over these six years, with AQI values ranging from a minimum of 13, indicating very clean air, to a maximum of 2049, reflecting extremely hazardous pollution levels. The average AQI for the entire country during this period is 166, which falls into the 'Moderate' category but indicates the potential for adverse health effects for sensitive groups. This broad spectrum of data allows for an in-depth analysis of air quality fluctuations and trends in different urban areas, highlighting cities with persistently poor air quality and those maintaining relatively better conditions. Such insights are invaluable for developing targeted air quality management strategies and improving public health outcomes.

Data Cleaning

The data cleaning process involved a meticulous and structured approach to enhance the dataset's quality and reliability. Initially, the dataset was scanned for duplicate entries, which can occur due to various reasons such as repeated data collection or recording errors. These duplicates were systematically identified and removed to prevent data redundancy, which can skew analysis results and lead to inaccurate conclusions.

Following the removal of duplicates, the dataset contained gaps where these entries had been. To address this, the missing values were filled using the mean values of the respective parameters. This imputation technique is effective because it preserves the overall distribution and central tendency of the data, ensuring that the imputed values do not disproportionately affect the dataset's variability or trends.

Moreover, a specific focus was placed on the Air Quality Index (AQI) values. Rows with null AQI values were completely removed from the dataset. This step was crucial because the AQI is a composite measure of air quality, and any missing values in this key metric could significantly impair the accuracy of any analysis or modelling performed on the data. By ensuring that all retained rows had complete AQI information, the data cleaning process reinforced the dataset's robustness, making it more reliable for subsequent analyses.

Overall, these cleaning steps—removing duplicates, imputing missing values with means, and dropping rows with null AQI values—collectively ensured that the dataset was both comprehensive and accurate.

To enhance the robustness and scope of the analysis, the data was integrated to create a comprehensive dataset comprising over 100,000 data points. This integration involved aggregating data from multiple sources, including various air quality monitoring stations across different cities in India. Each station's hourly and daily air quality measurements were meticulously compiled, ensuring a wide temporal and spatial coverage. The integration process also included standardizing the data formats and aligning the measurement units to ensure consistency across the dataset. This extensive compilation of data points significantly enriches the dataset, providing a vast pool of information that captures a detailed and nuanced picture of air quality trends and patterns. With over 100,000 data points, the dataset now offers a robust foundation for in-depth statistical analysis, machine learning applications, and the development of predictive models, ultimately supporting more effective air quality management and policy-making efforts.

Chapter 3

To gain a comprehensive understanding of air quality dynamics across different regions, an in-depth analysis of AQI readings was conducted over several years to identify discernible trends, seasonal variations, and long-term changes. This analysis utilized the extensive dataset, which now includes over 100,000 data points, enabling a detailed examination of air quality patterns.

Identifying Trends

The AQI data was plotted over time to visualize trends in air quality across different regions. By creating time-series graphs and applying statistical methods such as moving averages and trend lines, we could detect whether air quality was improving, deteriorating, or remaining stable. This long-term perspective helped highlight significant shifts, such as the impact of policy changes, technological advancements, or socio-economic developments on air quality.

Seasonal Variations

Seasonal variations were analyzed by segregating the data according to different times of the year. This analysis revealed recurring patterns, such as the increase in air pollution during the winter months. Factors contributing to this seasonal spike include temperature inversions, which trap pollutants close to the ground, and the widespread burning of agricultural residues in neighboring states, which significantly raises particulate matter levels. Conversely, the monsoon season typically showed improved air quality due to the cleansing effect of rainfall, which helps disperse pollutants.

Regional Analysis

Regional analysis involved comparing AQI trends across various cities and states. This was achieved by aggregating data from different monitoring stations and conducting a spatial analysis. The regional differences highlighted how localized sources of pollution, such as industrial zones, high traffic density, and proximity to agricultural areas, contribute to varying levels of air quality. For instance, industrial cities might exhibit higher levels of sulphur dioxide and nitrogen oxides, while urban areas with heavy traffic might show elevated levels of carbon monoxide and particulate matter.

Long-Term Changes

To understand long-term changes, we looked at data spanning multiple years, examining how air quality has evolved in response to regulatory measures, economic growth, and environmental policies. Statistical tools such as regression analysis and hypothesis testing were used to determine the significance of observed changes. This analysis identified periods of significant improvement, potentially linked to the implementation of stringent emission standards or the adoption of cleaner technologies. Conversely, periods of degradation could be correlated with urbanization, increased vehicular traffic, or lax enforcement of environmental regulations.

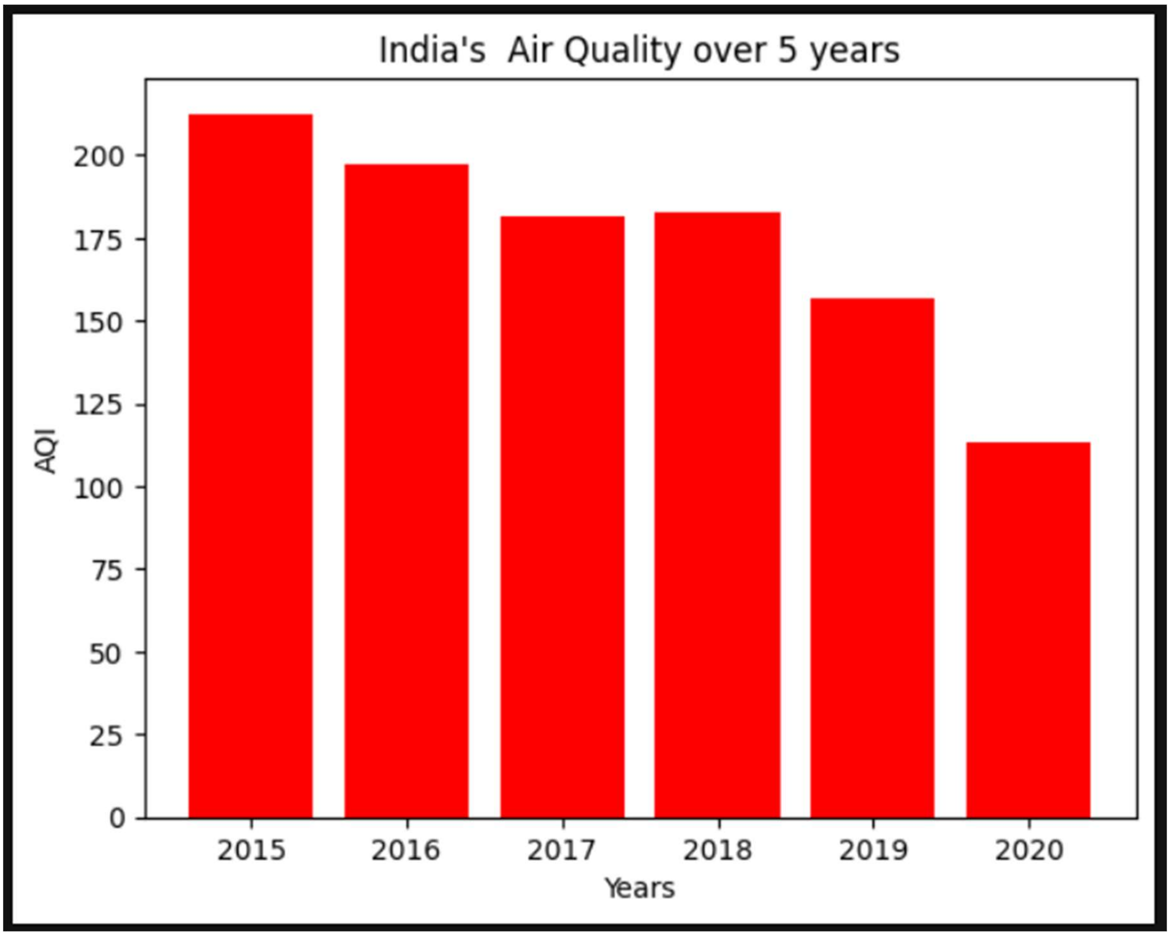
Implications and Recommendations

The insights gained from this comprehensive analysis are crucial for informing policy and public health strategies. By understanding the temporal and spatial dynamics of air pollution, policymakers can design targeted interventions to mitigate air pollution more effectively. For instance, regions with severe winter pollution may benefit from measures such as controlling biomass burning, enhancing public transportation, and promoting cleaner heating methods. The analysis also underscores the importance of continuous monitoring and data collection to adapt strategies in response to emerging trends and new challenges.

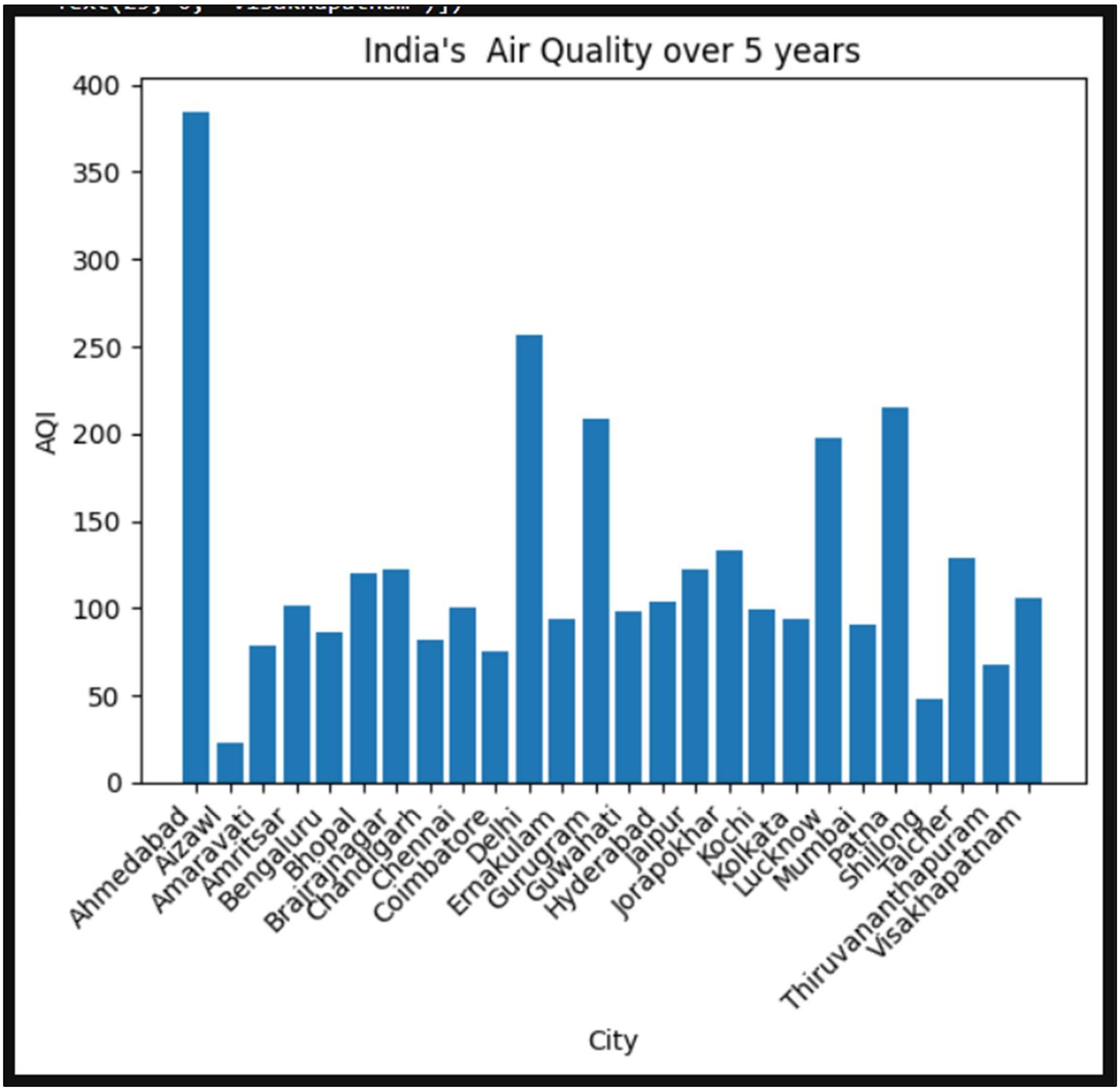
Overall, this detailed examination of AQI trends provides a robust foundation for improving air quality management, protecting public health, and ensuring sustainable development across different regions.

Chapter 4

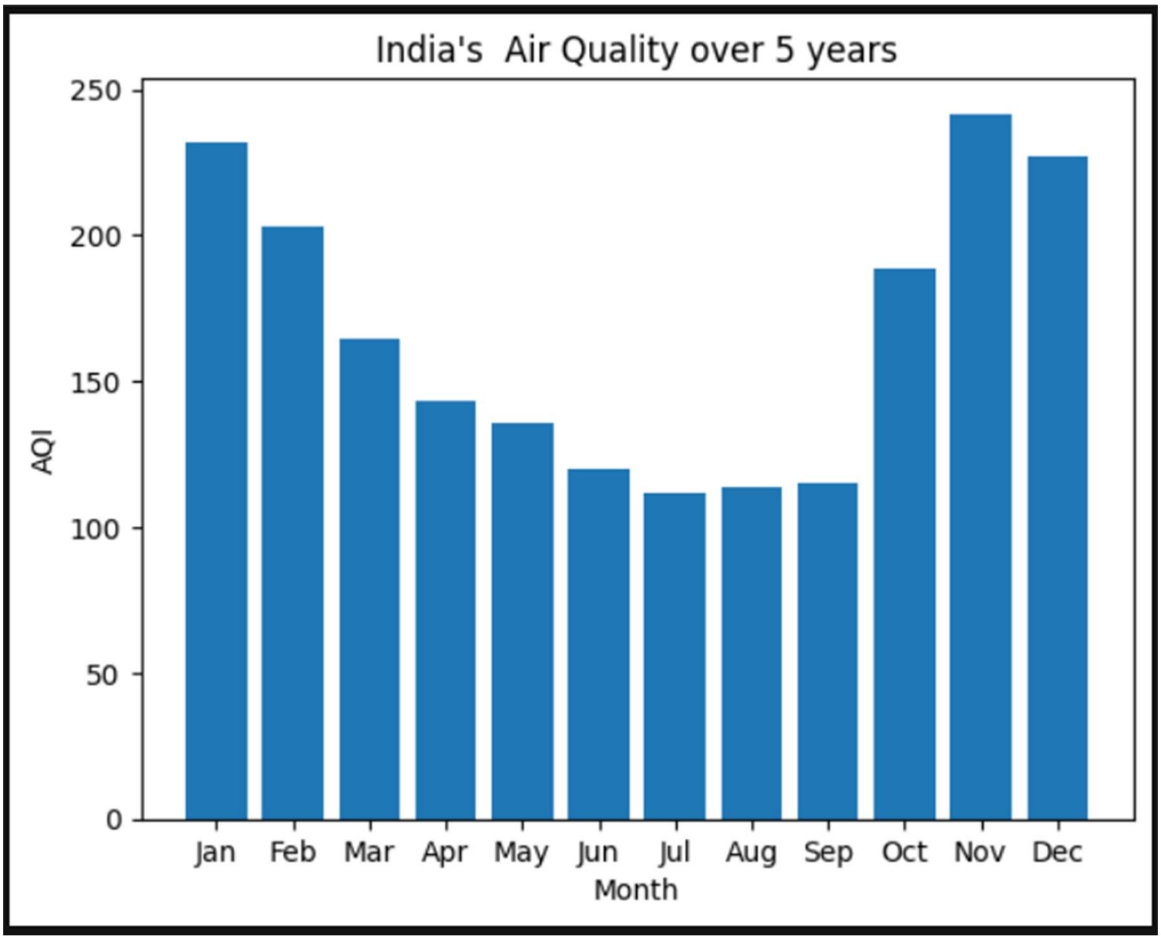
The examination of AQI trends across different states and cities in India revealed significant variations in air quality, highlighting regions with consistently high or low pollution levels. By analyzing the extensive dataset, patterns emerged showing that certain areas suffer from chronic poor air quality, while others maintain relatively better conditions. Metropolitan cities like Delhi, Mumbai, and Kolkata frequently recorded high AQI values, indicating severe pollution levels primarily due to vehicular emissions, industrial activities, and seasonal factors such as crop burning in neighboring states. Conversely, cities in states like Kerala and Himachal Pradesh generally exhibited lower AQI readings, reflecting cleaner air influenced by less industrial activity, higher green cover, and favorable meteorological conditions. The analysis also identified seasonal spikes in pollution in cities such as Lucknow and Kanpur, correlating with wintertime temperature inversions and increased particulate matter from local sources. These insights into regional air quality trends are crucial for targeted policy interventions, enabling authorities to prioritize areas with severe pollution for stricter regulatory measures and promote best practices in regions with better air quality to maintain their status.



The analysis of the dataset reveals significant variations in air quality over the years, with 2015 emerging as the year with the worst AQI levels, averaging 212. This indicates that, on average, the air quality in 2015 fell into the 'Poor' category, posing considerable health risks to the general population, especially those with pre-existing health conditions. Various factors, such as high vehicular emissions, industrial activities, and adverse meteorological conditions, likely contributed to this deterioration. In contrast, 2020 witnessed a marked improvement in air quality, primarily due to the widespread lockdowns imposed to control the Covid-19 pandemic. These lockdowns led to a significant reduction in industrial activities, vehicular traffic, and other pollution sources, resulting in cleaner air and the lowest average AQI recorded during the period analyzed. This drastic improvement underscores the impact of human activities on air quality and highlights the potential benefits of sustained environmental and regulatory measures to achieve long-term air quality improvement.



The analysis of the air quality data indicates that Ahmedabad stands out as the most polluted city in India, consistently recording the highest AQI values over the period from 2015 to 2020. Following closely are Delhi and Patna, both of which also exhibit alarmingly high levels of air pollution. Ahmedabad's position as the most polluted city can be attributed to its significant industrial activities, high vehicular emissions, and rapid urbanization, all of which contribute to its severe air quality issues. Delhi, the capital city, grapples with similar challenges, compounded by its dense population and frequent temperature inversions that trap pollutants. Patna, another major city, faces high pollution levels due to a mix of vehicular emissions, industrial discharge, and biomass burning. These findings highlight the urgent need for targeted air quality management and intervention strategies in these cities to mitigate the adverse health impacts associated with poor air quality and to improve the living conditions for their residents.



The analysis of air quality data over the years highlights a clear seasonal pattern in pollution levels across India. November emerges as the worst month for air quality, with AQI values peaking due to various contributing factors. In contrast, July records the lowest AQI levels, benefiting from the cleansing effects of the monsoon rains, which help to disperse pollutants. The winter months, particularly, see a significant increase in pollution levels. This seasonal spike is primarily due to a meteorological phenomenon where cold air near the ground is trapped under a layer of warmer air, a condition known as temperature inversion. This prevents pollutants from dispersing, leading to higher concentrations of harmful particles in the air.

Additionally, many cities in northern India, such as Delhi and Patna, are situated in geographical settings like valleys or basins that further exacerbate this issue by trapping pollutants. The situation is compounded by increased industrial emissions and vehicular pollution, which are prevalent in urban areas. Another critical factor is the burning of crop residues in agricultural states like Punjab and Haryana, which significantly contributes to the pollution levels in northern cities. This seasonal agricultural practice, combined with other pollution sources, leads to a severe degradation of air quality during the winter months. Understanding these patterns underscores the need for targeted measures during specific times of the year to mitigate the adverse health effects and improve air quality.

Chapter 5

Over the past five years, the data reveals that November consistently had the worst AQI in Ahmedabad, with February also showing significantly high pollution levels. These months stand out due to their exceptionally poor air quality, aligning with broader pollution trends observed during colder periods. The phenomenon of "cold air settles down," known as temperature inversion, plays a crucial role in this seasonal increase in AQI. During the winter months, the cooler air near the ground becomes trapped under a layer of warmer air above, preventing pollutants from dispersing into the upper atmosphere. This results in higher concentrations of pollutants at ground level, exacerbating air quality issues. In Ahmedabad, this effect is particularly pronounced, with industrial emissions, vehicular exhaust, and other local pollution sources contributing to the elevated AQI levels. Consequently, the colder months see a significant deterioration in air quality, highlighting the need for targeted interventions and policies to address the seasonal spikes in pollution and protect public health.

In 2018, the Air Quality Index (AQI) in Ahmedabad reached unprecedented levels, soaring to an alarming 600, indicating extremely hazardous air quality. This marked the highest AQI level ever recorded, reflecting severe pollution and its detrimental impact on public health and the environment. However, in 2020, there was a significant and unexpected improvement in air quality, with AQI levels dropping below 300. This dramatic decline was largely attributed to the COVID-19 lockdowns implemented worldwide. As industries shut down, vehicular traffic decreased, and human activity slowed, pollution levels plummeted, offering a temporary respite for the planet. The lockdown inadvertently became a blessing for Earth, providing a glimpse into the potential for cleaner air and a healthier environment with reduced human interference.

Delhi, often referred to as the pollution capital of India, consistently shows alarmingly high AQI levels during the cold months of November, December, and January. This period is marked by a significant increase in air pollution, driven by a combination of meteorological and anthropogenic factors. The cold weather leads to temperature inversions, where a layer of warm air traps the colder air at the surface, along with pollutants, preventing their dispersion. Additionally, Delhi's dense population and high vehicular traffic contribute substantially to the city's pollution levels. The winter months also coincide with agricultural burning in neighboring states like Punjab and Haryana, where farmers burn crop residues, further adding to the particulate matter in Delhi's air. Industrial emissions and construction activities within the city exacerbate the situation, leading to hazardous air quality levels that pose serious health risks to the residents. This persistent pollution problem during the colder months underscores the urgent need for comprehensive and sustained measures to mitigate air pollution in Delhi.

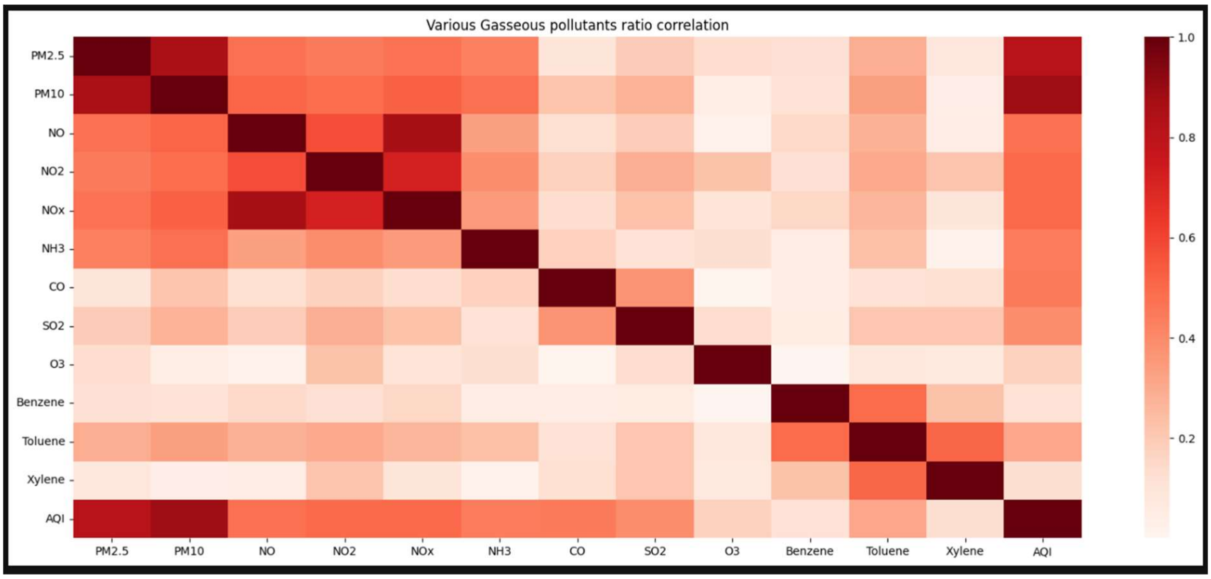
Delhi's Air Quality Index (AQI) peaked in 2016, reaching the 300 mark, a level that signifies hazardous air quality and poses serious health risks to residents. This peak highlighted the severe pollution challenges faced by the city, driven by factors such as vehicular emissions, industrial activities, and construction dust. However, in 2020, Delhi's AQI showed a remarkable improvement, mirroring the trend observed in Ahmedabad. During this year, the AQI in both cities dipped below 200, a substantial reduction that can be attributed to the widespread lockdowns and restrictions imposed due to the COVID-19 pandemic. These measures led to a significant decrease in industrial activities, transportation, and overall human movement, resulting in cleaner air and improved environmental conditions. This period underscored the potential for achieving better air quality through concerted efforts and the reduction of pollution sources.

In 2018, Gurugram experienced a significant peak in its Air Quality Index (AQI), with levels soaring above 250. This peak indicated a severe air quality issue, posing considerable health risks to the city's inhabitants due to the high concentration of pollutants. Factors contributing to this spike included rapid urbanization, vehicular emissions, and industrial activities. However, by 2020, a notable improvement in air quality was observed in Gurugram, with AQI levels dropping to 150. This substantial decrease can be attributed to the impact of the COVID-19 lockdowns, which led to reduced industrial operations, decreased vehicular traffic, and a general slowdown in economic activities. The lockdown inadvertently resulted in cleaner air and better environmental conditions, highlighting the significant influence human activities have on air quality and the potential for achieving healthier living conditions through targeted measures.

In 2016, Lucknow's Air Quality Index (AQI) levels peaked at 250, reflecting a severe pollution crisis in the city. This high AQI level indicated a substantial presence of pollutants, posing serious health risks to residents and highlighting the environmental challenges faced by the region. The primary contributors to this spike were likely urbanization, vehicular emissions, industrial activities, and other pollution sources prevalent in rapidly growing cities. However, by 2020, Lucknow experienced a significant improvement in air quality, with AQI levels dipping to 150. This improvement coincided with the global COVID-19 lockdowns, which led to a marked reduction in industrial operations, vehicular traffic, and overall human activity. The lockdowns inadvertently resulted in cleaner air, offering a glimpse of the positive impact that reduced pollution sources can have on the environment. This period underscored the potential for achieving better air quality and healthier living conditions through concerted efforts to manage and mitigate pollution.

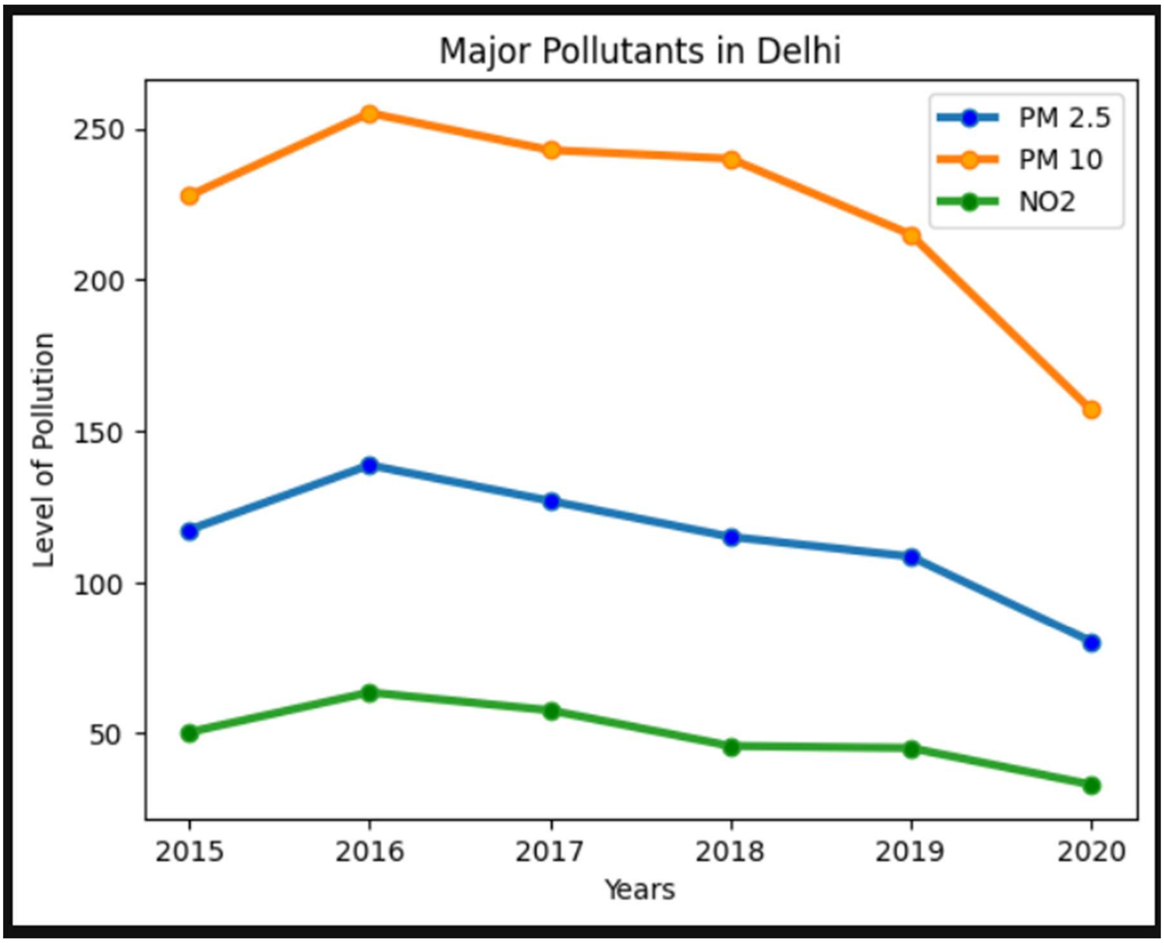
In 2015, Patna experienced a significant peak in its Air Quality Index (AQI), with levels reaching an alarming 350. This high AQI level indicated extremely poor air quality, posing severe health risks to the city's residents and underscoring the environmental challenges faced by the region. The primary contributors to this spike included vehicular emissions, industrial activities, construction dust, and other pollution sources common in rapidly urbanizing areas. However, by 2020, Patna saw a remarkable improvement in air quality, with AQI levels dipping to 150. This improvement was in line with trends observed in other states, largely due to the global COVID-19 lockdowns. These lockdowns led to a significant reduction in industrial operations, vehicular traffic, and overall human activity, resulting in cleaner air and improved environmental conditions. The lockdown period highlighted the potential for substantial air quality improvements through targeted pollution control measures and a reduction in human-induced pollution.

Chapter 6

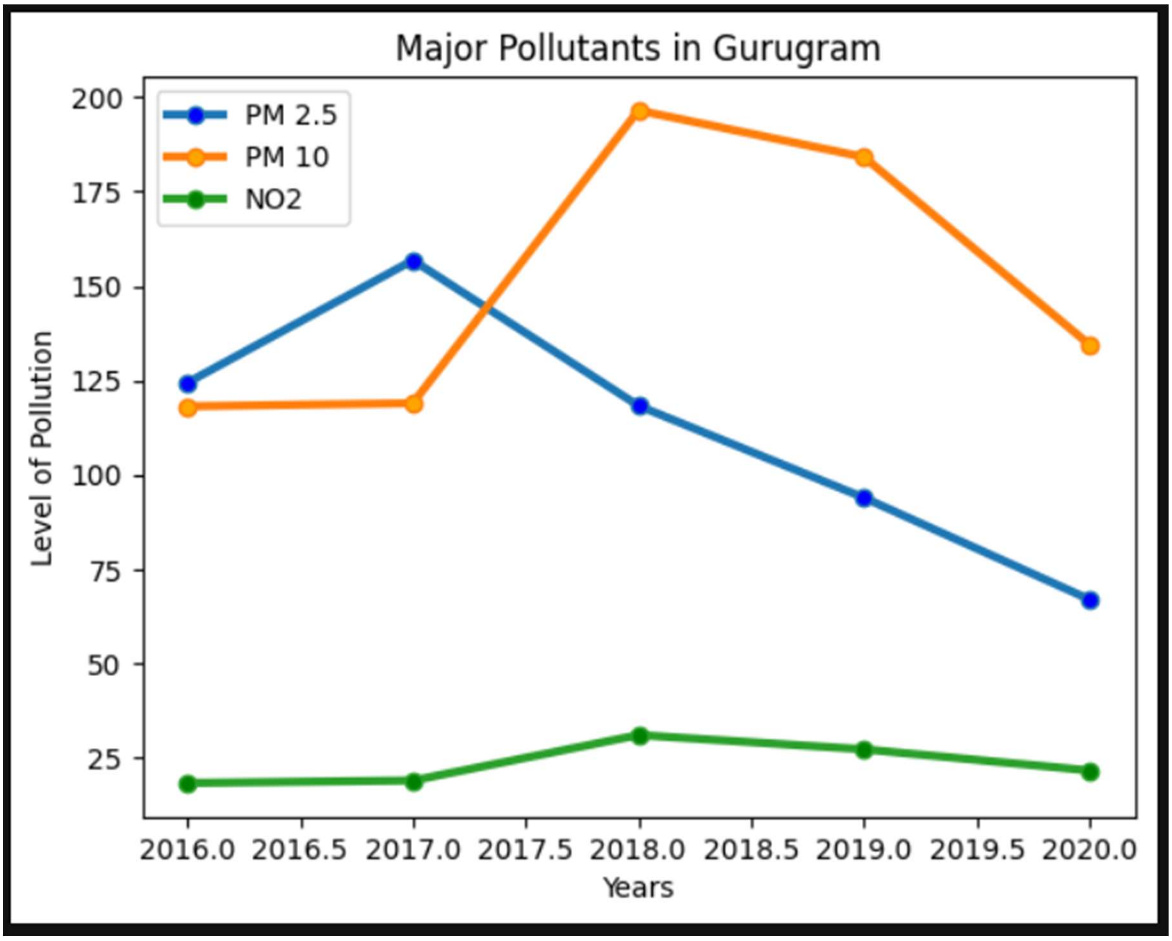


In Indian cities and states, the three primary pollutants of concern are PM 2.5, PM 10, and NO2. These pollutants have varying impacts on air quality and public health. Over the years, the levels of nitrogen dioxide (NO2) have remained relatively constant, indicating that the sources of this pollutant, such as vehicular emissions and industrial activities, have not seen significant changes in their emission rates. In contrast, particulate matter pollutants, specifically PM 2.5 and PM 10, have shown a declining trend. This decrease suggests that measures taken to reduce emissions from sources like construction, road dust, and industrial processes are having a positive effect. Additionally, the temporary reduction in human activity during the COVID-19 lockdowns contributed to this improvement. The overall decline in PM 2.5 and PM 10 levels is a promising sign for air quality management efforts and public health in the region.

In Ahmedabad, NO2 levels have shown dramatic fluctuations over several years. In 2016, there was a slight decrease in NO2 levels, offering a brief respite from the pollutant. However, this was short-lived as 2017 saw a sharp increase, with levels spiking to 80, marking a significant rise. From 2017 to 2019, NO2 levels continued their upward trajectory, reflecting increased industrial activity and vehicular emissions. The trend took a turn in 2020 when NO2 levels saw a significant decline, primarily attributed to the COVID-19 lockdown. The lockdown led to reduced industrial activity and fewer vehicles on the road, which in turn brought down NO2 emissions. Similarly, PM 2.5 levels, which had peaked in 2017, also experienced a decline by 2020, further highlighting the impact of reduced human activity during the pandemic. Meanwhile, PM 10 levels remained steady at 120 for several years but declined to 100 in 2020, indicating an overall improvement in air quality. These trends underscore the significant influence of human activities on air pollution and the potential benefits of targeted measures to control emissions.



The graph illustrates the trends in major pollutants in Delhi from 2016 to 2020. Notably, the levels of nitrogen dioxide (NO2) remained consistent over these years but experienced a decline in 2020. In 2016, there was a slight increase in the concentrations of NO2, particulate matter 10 micrometers or less in diameter (PM10), and particulate matter 2.5 micrometers or less in diameter (PM2.5). However, this was followed by a slight decrease in 2017. By 2020, there was a marked decline in the levels of all three pollutants. This significant reduction in 2020 can be attributed to the COVID-19 lockdown measures, which led to a substantial decrease in industrial activity and vehicular emissions, thereby improving air quality in the city.



The graph illustrates the trends in major pollutants in Gurugram from 2016 to 2020. During this period, particulate matter with a diameter of 10 micrometers or less (PM10) and particulate matter with a diameter of 2.5 micrometers or less (PM2.5) exhibited significant increases, with PM10 levels spiking sharply in 2017 and PM2.5 levels rising notably in 2018. However, by 2020, both pollutants had experienced a significant decline. Meanwhile, nitrogen dioxide (NO2) levels remained relatively stable throughout these years, though there was a slight increase in 2018. This trend was reversed in 2020, with NO2 levels declining, likely as a result of reduced industrial activities and vehicular emissions brought about by the COVID-19 lockdown measures. These reductions highlight the impact of decreased human activity on air quality in Gurugram.

Chapter 7

An analysis of the average Air Quality Index (AQI) across 87,000 readings from cities all over India reveals notable patterns. The data indicates two distinct spikes in AQI values. The first significant increase is observed in the range of 100-200, suggesting a considerable number of instances where air quality was classified as moderate to unhealthy for sensitive groups. The second spike occurs in the range of 300-400, indicating severe pollution levels that are hazardous to health. These spikes highlight periods when air quality significantly deteriorated, posing serious health risks to the population. The presence of these spikes underscores the persistent and widespread air pollution challenges faced by urban areas in India.

An analysis of air quality data from 87,000 readings taken from stations across India reveals distinct patterns in pollutant levels. Benzene, xylene, carbon monoxide (CO), and toluene typically show negligible or zero spikes, indicating these pollutants generally remain at low levels. In contrast, particulate matter (PM10 and PM2.5), nitrogen dioxide (NO2), nitrogen oxides (NOx), and ammonia (NH3) exhibit significantly higher levels, frequently falling within the 100-200 range. This suggests that these pollutants are prevalent and contribute substantially to air quality issues across the country, reflecting a persistent challenge in managing and mitigating air pollution.

Based on the analysis of the performance metrics across different machine learning models, we have derived several key insights regarding their efficacy in predicting air quality indices (AQI).

XGBRegressor

The XGBRegressor, an implementation of the eXtreme Gradient Boosting algorithm, emerged as the fastest and best-performing model in our study. This model leverages gradient boosting principles, which involve sequentially adding predictors to an ensemble, each one correcting its predecessor's errors. It achieved a remarkable cross-validation score of 0.84 and an even higher testing score of 0.91. These high scores reflect the model's robustness and its ability to generalize well to new, unseen data. The XGBRegressor's strength lies in its efficiency and accuracy, making it particularly suitable for complex datasets with non-linear relationships. Additionally, it achieved the best mean absolute error (MAE) score of 21.22. Given the AQI range of 0-400, this level of error is acceptable, although there remains scope for further improvement through hyperparameter tuning and feature engineering.

Random Forest Regressor

The Random Forest Regressor is another powerful ensemble learning method, utilizing a multitude of decision trees to enhance predictive accuracy and control overfitting. Each tree in the forest is built on a random subset of the data, and the final prediction is made by averaging the predictions of all trees. This model showed commendable performance, with strong cross-validation and testing scores, though slightly lower than those of the XGBRegressor. Its robustness to overfitting and ability to handle large datasets with high dimensionality make it a reliable choice. However, it typically requires more computational resources compared to gradient boosting methods.

Support Vector Regressor (SVR)

The SVR model, based on the principles of Support Vector Machines, aims to find a hyperplane in an N-dimensional space that optimally fits the data points. Despite its theoretical strengths, the SVR model underperformed in our study, achieving a cross-validation score of 0.70 and a testing score of 0.67. These lower scores indicate that the SVR struggled to capture the complexities and non-linear patterns present in the AQI data as effectively as the other models. This performance gap highlights the limitations of SVR in this specific context, suggesting that it may not be the best fit for datasets with intricate relationships and noise.

Conclusion

In conclusion, the XGBRegressor stands out as the most effective model for predicting AQI, due to its superior speed and accuracy. While the Random Forest Regressor also performed well, it did not match the efficiency of the XGBRegressor. The SVR model, on the other hand, lagged behind significantly, indicating a need for alternative approaches or further tuning when dealing with similar datasets. These findings guide us towards prioritizing the use of XGBRegressor for future predictive modeling efforts while exploring avenues for further optimization to enhance its performance even more.

Bibliography

DATA SOURCE

Air Quality Data in India (2015 - 2020) (kaggle.com)

DESIGN PROJECT AND THESES

Aryan Rathore Air_Quality_Analysis_and_Prediction_for_Indian_Cities_and_States [Design Project]. cse core student at vit bhopal

Avan Chowdary Gogineni Vamsi Sri Naga Manikanta Murukonda Prediction of Air Quality Index [Bachelor’s in Computer Science Engineering thesis]. Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

Checklist of items for the Final Design Project Report

1.	**Is the final report neatly formatted with all the elements required for a technical Report?**	[✓ Yes] / No
2.	Is the Cover page in proper format as given in Annexure A?	[✓ Yes] / No
3.	Is the Title page (Inner cover page) in proper format?	[✓ Yes] / No
4.	(a) Is the Certificate from the Mentor in proper format? (b) Has it been signed by the Mentor?	[✓ Yes] / No [✓ Yes] / No
5.	Is the Abstract included in the report properly written within one page? Have the technical keywords been specified properly?	[✓ Yes] / No [✓ Yes] / No
6.	Is the title of your report appropriate?	[✓ Yes] / No
7.	Have you included the List of abbreviations / Acronyms, if any?	[✓ Yes] / No
8.	Does the Report contain a summary of the literature survey, if any?	[✓ Yes] / No
9.	**Does the Table of Contents include page numbers?** (i). Are the Pages numbered properly? (Ch. 1 should start on Page # 1) (ii). Are the Figures numbered properly? (Figure Numbers and Figure Titles should be at the bottom of the figures) (iii). Are the Tables numbered properly? (Table Numbers and Table Titles should be at the top of the tables) (iv). Are the Captions for the Figures and Tables proper? (v). Are the Appendices numbered properly? Are their titles appropriate	[✓ Yes] / No [✓ Yes] / No [✓ Yes] / No [✓ Yes] / No [✓ Yes] / No [✓ Yes] / No
10.	Is the conclusion of the Report based on discussion of the work?	[✓ Yes] / No
11.	Are References or Bibliography given at the end of the Report? Have the References been cited properly inside the text of the Report? Are all the references cited in the body of the report	[✓ Yes] / No [✓ Yes] / No [✓ Yes] / No
12.	Is the report format and content according to the guidelines? The report should not be a mere printout of a Power Point Presentation, or a user manual. Source code of software need not be included in the report.	[✓ Yes] / No

Declaration by Student:
I certify that I have properly verified all the items in this checklist and ensure that the report is in proper format as specified in the course handout.

Aryan

Place: Delhi **Signature of the Student**

Date: 16.06.2024 **Name: Aryan Bhardwaj**

ID No.: 202117B3728