

**ENHANCING AIR QUALITY INDEX PREDICTION: A COMPARATIVE
ANALYSIS OF TRANSFORMER MODELS, GRAPH NEURAL
NETWORKS, AND TRADITIONAL APPROACHES**

Submitted By – Aditya Sharma

Master of Science in Business Analytics 2023-24

Dublin Business School



Declaration

I, Aditya Sharma, declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this work and this this work is fully compliant with the Dublin Business School's academic honesty policy. Signed: Aditya Sharma

Date: 08/01/2024

Acknowledgement

My profound appreciation goes out to Dublin Business School for giving me the priceless chance to conduct this research. The institute's assistance and wealth of resources made it much easier to carry out my study and ensured a smooth research procedure.

I want to sincerely thank Mr. Obinna Izima for his constant support and ongoing direction during the writing of my thesis. His support and guidance were really helpful in overcoming the challenges of this project, and I sincerely appreciate his commitment.

I also want to express my sincere gratitude to my friends and family for their unwavering support, which allowed me to confidently go through every step of the research process. Their support and comprehension have been crucial in helping me overcome obstacles and accomplish milestones.

I also want to thank everyone in my close vicinity for their support, since their encouragement and help made this research much more successful to complete.

I would want to thank Dublin Business School, Mr. Obinna Izima, and everyone else who helped with this academic endeavor once more. Your assistance has been much appreciated.

Table of Contents

List of Figures	5
Abstract	7
Chapter 1 Introduction	8
1.1 Back Ground scope	8
1.2 Motivation.....	11
1.3 Research Questions	12
1.4 Objective	12
1.5 Research Outline	13
Chapter 2 Literature Review	16
2.1 Introduction.....	16
2.2 Concept	16
2.3 Literature Review.....	17
Chapter 3 Methodology	30
3.1 Dataset.....	34
3.2 Data Analysis using tableau	35
3.3 Data Pre-processing	39
3.4 Data Visualisation.....	45
3.5 Model Training	48

3.5.1 Graph Neural Network (GNN)	49
3.5.2 Hybrid Model	50
3.5.3 Long Short-Term Memory (LSTM)	50
3.5.4 Transformer Model	51
3.5.5 Linear Regression	51
3.5.6 Naive Bayes	52
Chapter 4 Results and Analysis	53
Chapter 5 Conclusion and Future Scope	57
5.1 Future Scope	58

List of Figures

Figure 3.1: CRISP-DM Methodology	30
Figure 3.2: Project Methodology	32
Figure 3.3: Dataset Information	34
Figure 3.4: Dashboard 1	35
Figure 3.5: Dashboard 2	36
Figure 3.6: Dashboard 3	37
Figure 3.7: Dashboard 4	38
Figure 3.8: Dashboard 5	38
Figure 3.9: Importing the libraries	39
Figure 3.10: Load the dataset	40

Figure 3.11: Statistical Description of dataset	40
Figure 3.12: Drop the AQI_Bucket	41
Figure 3.13: Convert date to datetime.....	41
Figure 3.14: Set date as index	41
Figure 3.15: checking the null values	42
Figure 3.16: Drop the missing value.....	42
Figure 3.17: Unique values of columns	42
Figure 3.18: Pearson correlation.....	43
Figure 3.19: Creating a new dataframe.....	43
Figure 3.20: Label Encoder.....	44
Figure 3.21: Unique counts of columns.....	45
Figure 3.22: Line Plot for all columns	47
Figure 3.23: split the dataset	49
Figure 4.1: MSE and RMSE of the models	54
Figure 4.2: Barchart for MSE of the models.....	55
Figure 4.3: RMSE of all models	56

Abstract

This study presents a comprehensive comparative analysis of air quality index (AQI) prediction models, focusing on Transformer Models, Graph Neural Networks (GNN), and traditional approaches such as Linear Regression, Naive Bayes, and Long Short-Term Memory (LSTM). Evaluation metrics include Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Among the models, Graph Neural Networks exhibit superior performance with an MSE of 1601.86 and an RMSE of 40.02. The Hybrid Model follows closely with an MSE of 2057.53 and an RMSE of 45.36. Traditional methods like Linear Regression and Naive Bayes demonstrate moderate accuracy, while the LSTM model exhibits higher errors. Notably, the Transformer Model records the highest errors, suggesting challenges in accurately predicting AQI using this approach. These findings offer valuable insights for selecting optimal models to enhance air quality predictions in environmental research and monitoring.

Chapter 1 Introduction

In response to the escalating global concerns over air quality, this study conducts a comprehensive investigation into advancing the prediction accuracy of the Air Quality Index (AQI). The critical importance of precise AQI forecasting lies in its pivotal role in public health, environmental policy-making, and urban planning. With the increasing availability of diverse machine learning models, this research specifically explores Transformer Models, Graph Neural Networks, and traditional approaches to discern their efficacy in AQI prediction. The examination is rooted in the pressing need for more accurate and reliable predictions to mitigate the adverse effects of air pollution. As urbanization continues to intensify and environmental challenges grow, refining AQI prediction models becomes paramount for informed decision-making and safeguarding public well-being. This study contributes to the ongoing dialogue on environmental sustainability by offering insights into the potential of various models, paving the way for enhanced air quality predictions and proactive interventions.

1.1 Back Ground scope

Urbanization and industrialization have precipitated a concerning rise in air pollution, presenting substantial threats to public health and the environment. The heightened concentrations of primary air pollutants, including particulate matter (PM_{2.5} and PM₁₀), nitrogen dioxide (NO₂), and sulfur dioxide (SO₂), underscore the urgent need for effective air quality management in urban areas globally. To address these challenges, researchers have actively pursued advancements in air quality prediction methodologies, seeking to comprehend, measure, and forecast air pollution dynamics comprehensively.

Liao et al. (2021) undertook a meticulous examination of statistical approaches for forecasting primary air pollutants. These traditional models, grounded in statistical methods, have played a pivotal role in elucidating air quality patterns. By analyzing historical data and identifying trends, these models establish a foundational understanding, offering a basis for comparison with more intricate predictive models.

Limperis et al. (2023) introduced a transformative element by integrating Transformer neural networks into air quality prediction models. This departure from conventional methods reflects the evolution of predictive modeling. Transformers, with their attention mechanisms, prove instrumental in capturing long-range dependencies in data, enhancing the discernment of intricate patterns and relationships within air quality datasets.

Ma et al. (2023) contributed the SpatioTemporal-Informer model, recognizing the interconnected nature of spatial and temporal characteristics in air quality dynamics. This holistic approach considers both dimensions simultaneously, providing a more comprehensive understanding of the factors influencing air quality variations over time. The integration of spatial and temporal information enhances prediction accuracy, particularly in dynamic urban settings.

Mengara et al. (2022) presented an attention-based distributed deep learning model, addressing both the complexities of air quality data and the computational challenges associated with deep learning. By distributing computations across multiple nodes, the model demonstrates scalability, making it apt for handling large and intricate datasets. The incorporation of attention mechanisms refines the model's ability to capture relevant features in the data.

Naz et al. (2023) conducted a meticulous comparative analysis of deep learning and statistical models for predicting air pollutants in urban areas. The study critically evaluates the performance

of diverse approaches, shedding light on their respective strengths and weaknesses. These findings aid researchers and policymakers in making informed decisions regarding model selection for air quality forecasting in specific contexts.

Oliveira Santos et al. (2023) introduced a graph-based deep learning model tailored for forecasting chloride concentration in urban streams. This innovative approach expands the application of deep learning models beyond air quality parameters, showcasing their adaptability to diverse environmental factors. The graph-based model provides insights into water quality challenges in urban settings, highlighting the versatility of deep learning techniques.

Mitreska Jovanovska et al. (2023) conducted an extensive review of methods for urban air pollution measurement and forecasting. This study delves into the challenges and opportunities in the field, proposing innovative solutions. The comprehensive overview aids in understanding the complexities of urban air pollution, guiding the development of more effective and targeted forecasting models.

The intensifying concerns about air quality have spurred research endeavors to advance prediction models. From traditional statistical approaches to cutting-edge deep learning architectures, each model contributes unique insights to the intricate dynamics of air pollution. These studies collectively provide a foundation for understanding, measuring, and predicting air quality, offering valuable tools for researchers, policymakers, and practitioners working towards mitigating the impact of air pollutants on public health and the environment. As the field continues to evolve, the interdisciplinary nature of these advancements promises to shape future strategies for addressing the challenges posed by urban air pollution.

1.2 Motivation

Motivation for the study on "Enhancing Air Quality Index Prediction" stems from a profound commitment to addressing one of the most critical challenges of our time – ensuring the well-being of our environment and the health of communities. The escalating concerns about air quality necessitate innovative and accurate predictive models, prompting a journey into the realms of advanced technologies like Transformer Models, Graph Neural Networks (GNN), and their comparison with traditional approaches.

In the backdrop of escalating environmental issues, the study is propelled by a vision to contribute meaningfully to the field of air quality prediction. The imperative to mitigate the adverse impacts of air pollution on public health and the environment serves as a compelling catalyst for the exploration of cutting-edge models. The transformative potential of technologies like Transformers and GNNs offers a glimpse into a future where predictive accuracy is heightened, and decision-makers are equipped with more reliable tools for environmental management.

Moreover, the motivation extends to bridging the gap between traditional methodologies and state-of-the-art deep learning approaches. Recognizing the importance of both interpretability and predictive power, the study seeks to amalgamate the strengths of these diverse models in a Hybrid approach. This endeavor is driven by a commitment to creating a holistic framework that not only excels in accuracy but also aligns with real-world applicability and understanding.

The urgency to enhance air quality predictions resonates with a global imperative for sustainable living. By delving into the intricacies of these models, the study aspires to contribute valuable insights that can guide policy-making, urban planning, and public awareness initiatives.

Ultimately, the motivation lies in the belief that the fusion of cutting-edge technology and established methodologies can pave the way for a more resilient and informed approach to mitigating the impact of air pollution on our communities and the planet.

1.3 Research Questions

The main research questions of this study are as follows:

1. How effectively can Transformer Models and Graph Neural Networks be utilized in predicting the Air Quality Index?
2. How do these deep learning approaches compare to traditional methods and other machine learning techniques in terms of prediction accuracy?
3. How can these models be optimized for handling the complexity and high-dimensionality of air quality data?

1.4 Objective

The objectives are as follows:

1. Evaluate the effectiveness of Transformer Models and Graph Neural Networks (GNN) in predicting the Air Quality Index (AQI) through the assessment of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) against established benchmarks.
2. Conduct a comparative analysis of Transformer Models and GNN against traditional methods, such as Long Short-Term Memory (LSTM) and other machine learning techniques, to quantify the improvement in prediction accuracy and identify the most effective approach.

3. Explore various configurations of Transformer Models and GNN to identify the optimal architecture and hyperparameters that yield the highest accuracy in predicting AQI, considering the unique complexities and high-dimensionality inherent in air quality data.
4. Investigate the robustness of Transformer Models and GNN by evaluating their performance across diverse datasets, temporal variations, and geographical locations, aiming to ascertain the reliability of these models under different conditions.
5. Examine the interpretability of features and patterns influencing the AQI predictions made by Transformer Models and GNN, enhancing the understanding of the models' decision-making processes.

1.5 Research Outline

Chapter 1: Introduction

The introduction sets the stage for the research, delving into the critical issue of air quality and the necessity for accurate prediction models. It delineates the existing challenges in air quality prediction, highlighting the gaps in current methodologies. The objectives of the study are clearly defined, centering on the exploration and comparison of Transformer Models, Graph Neural Networks, and traditional approaches. A set of research questions guides the inquiry, aiming to uncover the most effective models for Air Quality Index (AQI) prediction. The significance of the study is underscored by its potential impact on environmental management and public health. The chapter concludes with an overview of the thesis structure, providing a roadmap for the reader.

Chapter 2: Literature Review

The literature review comprehensively surveys existing research on air quality prediction. It outlines the historical evolution of methodologies, scrutinizing traditional approaches and their limitations. Transformer Models and Graph Neural Networks are scrutinized for their applications in environmental predictions, emphasizing their advantages and challenges. The chapter identifies gaps in current literature, setting the groundwork for the present study by addressing these shortcomings and contributing new insights to the field.

Chapter 3: Methodology

This chapter outlines the research design, detailing the comparative analysis framework employed in the study. It elucidates the data collection process, specifying the sources and characteristics of the air quality datasets. The architectures of Transformer Models, Graph Neural Networks, and traditional approaches are expounded upon. Additionally, the chosen evaluation metrics and the steps taken in data preprocessing are meticulously explained. This chapter provides a transparent and systematic approach to the research methodology.

Chapter 4: Results and Evaluation

The results and evaluation chapter presents the outcomes of the comparative analysis, showcasing the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for each model. A detailed analysis of the results is conducted, emphasizing the strengths and weaknesses of Transformer Models, Graph Neural Networks, and traditional approaches. Visual representations, such as graphs and charts, enhance the understanding of the model predictions and their alignment with actual values. This chapter serves as a critical juncture for assessing the effectiveness of each model.

Chapter 5: Conclusion and Future Scope

In the final chapter, the study's findings are succinctly summarized, and conclusions are drawn based on the results and evaluation. The contributions of the research to the field are emphasized, along with acknowledgments of any limitations. Future scope is discussed, presenting opportunities for further research and improvement in air quality prediction models. The chapter concludes with reflections on the overall significance of the study, leaving the reader with a deeper understanding of the complexities and potential advancements in air quality prediction.

Chapter 2 Literature Review

2.1 Introduction

The quality of air significantly impacts public health along with the well-being of the Environment in recent periods. To address this scenario accurately, the prediction of the "Air Quality Index (AQI)" becomes a critical concern. The increasingly urgent need for air pollution-related challenges control requires innovative solutions with surpassing the constraints of traditional forecasting procedures (Liao *et al.*, 2021). Taking environmental concerns as a major consideration, having a strong and precise knowledge of air quality dynamics is a necessity. As the consequences of air pollution become the biggest concern of the whole world, this comparative analysis gives a clear insight into the practical application that enables the advancement of air quality monitoring and prediction frameworks (Huang *et al.*, 2021). The incorporation of advanced technologies and the comparison to the traditional approaches within this research will strengthen the ability to predict and mitigate air pollution impact on a global scale. This paper will explore the strong potentiality of innovative models, specifically "Transformers" and "Graph Neural Networks," in determining and revolutionizing the prediction of AQI with high accuracy. Also, relevant research papers will be reviewed in order to acquire an informative insight into this particular topic.

2.2 Concept

The concept of "Enhancing Air Quality Index Prediction" revolves around the urgent need to enhance the accuracy of prediction and precision of the Air Quality Index (AQI). The "Air Quality Index" is an essential metric that calibrates and captures the concentration levels and volume of different air pollutants (Schulte *et al.*, 2020). An accurate prediction system is vital for prompt intervention and effective decision-making as air pollution becomes a crucial concern,

and it significantly affects the ecosystems, climate, and human health. Enhancing the prediction accuracy of air quality, including "Transformer models" and "Graph Neural Networks" in the conventional approach of AQI prediction techniques, can be a forward-looking initiative (Chen *et al.*, 2023). A Multi-graph Spatial-temporal Attention Network for Air-quality Prediction. Process Safety and Environmental Protection. Transformers deliver a paradigm transformation in managing the complex spatio-temporal data integrated into air quality dynamics. Transformers are widely acknowledged for their proficiency in sequence modeling and contextual understanding. Also, "Graph Neural Networks" is a novel approach for accurate prediction of air quality as it captures interdependencies among air quality parameters in a graph structure.

Well-established, traditional methods are not capable of handling the intricacies of pollution pattern changes with different air pollutant elements, which is a more sophisticated model. Various research analyses considerable implications for environmental management, public health, and policy formulation by enhancing the accuracy of AQI predictions. Precise forecasting allows authorities and communities to create focused measures, reduce health threats, and anticipate reasons for pollution. Moreover, the integration of modern innovative technologies is consistent with the more expansive approach toward employing artificial intelligence to manage intricate environmental issues. It is important to recognize how urgent it is to improve the power for prediction, especially in a scenario of growing urbanization and concerns about climate change.

2.3 Literature Review

The study conducted by Zhang *et al.* (2022) recommends a novel "deep learning" model for "PM2.5" air pollution prediction dubbed "temporal difference-based graph transformer networks (TDGTM)." With the help of its encoder and decoder layers, which include an advanced graph

attention mechanism, TDGTN can extract complicated connections and long-term temporal dependencies from time series “PM2.5 data”. The process considers the temporal differences between adjacent moments and their significance in predicting air quality “PM2.5”, as well as the similarity of distinct time moments. To learn from a graph perspective on air quality “PM2.5 prediction tasks”, the proposed method first constructs graph-structured data from the original time series “PM2.5” data at different moments without precise graph structure. Then, it embeds the developed “graph attention networks” into the encoder and decoder layers.

An accurate air quality prediction is essential for pollution prevention and control as a result of the severe air quality concerns brought on by rapid economic development. To improve air quality index (AQI) prediction, a novel model named Enhanced Autoformer (EnAutoformer) is presented in this study by Feng *et al.* (2023). There are two main parts to the “EnAutoformer”:

- (a) To extract temporal dependencies from AQI time series, "Enhanced Cross-Correlation (ECC)" is suggested;
- (b) To increase computational efficiency, ECC is integrated with cross-stage feature fusion mechanism of "CSPDenseNet."
- (c) To extract smooth features from the original AQI data and improve the long-term prediction ability of the model, time series decomposition and broad-end causal convolution are added to the decoder module.

In demand for management and vulnerable people to take preventive action, reliable air quality predictions are required due to the substantial health and environmental dangers caused by air pollution. Although there are numerous approaches to the prediction of air quality data, it is difficult due to the intricate and nonstationary “spatio-temporal” correlation. To solve this issue, the study by Li *et al.* (2023) presents "GCNInformer," a unique spatiotemporal "neural network" that forecasts air quality data by combining "Informer" with "graph convolution networks (GCNs)." Using "GCN layers," "GCNInformer" documents spatial correlations between

monitoring locations and uses “Informer” layers to obtain both short and long-term temporal data. Low-dimensional representations are known from meteorological and air quality data employing “MLP” layers.

Forecasting air quality accurately is essential for public health, decision-making, and environmental conservation. Low precision in long-term prediction results from existing approaches' inability to accurately model complex and long-term relationships in time series PM2.5 data. According to the study conducted by Zhang *et al.* (2023), the solution to this problem is the "sparse attention-based Transformer networks (STN)" light "deep learning model," which consists of encoder and decoder layers with a multi-head sparse attention mechanism to reduce time complexity. Using “time series PM2.5” data, the suggested approach removes complicated connections and long-term dependencies for air quality forecasting. The suggested “STN” method outperforms “state-of-the-art” techniques, as demonstrated by experiments conducted on two real-world datasets in China.

The issue of predicting water quality, which is essential for farming and everyday human activities like drinking, is the main topic of this review. The Research by Irwan *et al.* (2023) looks at multiple criteria related to water quality that are included in modeling procedures to measure water quality. It also looks at models based on "artificial intelligence" that are frequently used to predict the quality of water in various areas. Based on their significance in modeling and forecasting water quality, taking into account factors, "modelling methods," "time scale scenarios," and "performance assessment indicators," the evaluation assesses 83 studies that were published between the years 2009 and 2023. According to the study, "hybrid-DL" models perform better than "standalone ML," "standalone DL," and "hybrid-ML" models.

According to the study by Naz *et al.* (2023), air pollution and poor air quality, which have dangerous effects on the Environment and human health, are becoming important global challenges as urbanization and industry advance. Precise prediction of air pollution is essential for stakeholders to make the needed arrangements. Using a publically accessible dataset, this study analyses several single-step forecasting models that are based on "deep learning," including statistical models, "gated recurrent units (GRU)," and "long short-term memory (LSTM)," to predict five air contaminants in Northern Ireland. Three measures are used in the paper to assess the effectiveness of these models: "R-squared (R^2)," "mean absolute error (MAE)," and "root mean square error (RMSE)." The outcomes demonstrate that with an "RMSE" of "0.59", "deep learning" models routinely outperform statistical models.

The goal of the paper by Ghobadi *et al.* (2022) is to improve long-term streamflow prediction accuracy, which is essential for effective water resource management and disaster avoidance. The present data-driven models are not without limitations, especially for multistep projections in inadequately measured basins. The study offers a meta-algorithm to evaluate the power of prediction utilizing booster predictors in direct and direct-recursive hybrid strategies, as well as to analyze complicated "geo-spatio-temporal" environments. For 12-month forward prediction, the study presents four "state-of-the-art" combinations pairs of "TimeDistributed-CNN (TD-CNN)" and "3D-CNN" combined with a "Transformer" network or a "Long- and Short-term Time-series network (LSTNet)."

Deng *et al.*'s (2023) research represents a significant advancement in air quality prediction, recognizing the intricate interplay of weather patterns, exhaust emissions, and air pollution. Unlike previous studies primarily centered on temporal data, Deng *et al.* emphasize the imperative of integrating spatial data to comprehensively address the scattered nature of air and

its influence on nearby regions. The innovative inclusion of spatial dimensions is made possible through the utilization of a "graph neural network," a potent mechanism adept at capturing intricate spatial dependencies. Furthermore, the study incorporates "diffusion convolution" to extrapolate subtle spatial nuances between non-adjacent locations. The forecasting component integrates the "gate recurrent unit," culminating in a comprehensive model that amalgamates both temporal and spatial elements. Deng et al.'s research thus pioneers a holistic approach to air quality prediction, considering the dynamic interplay of factors across both temporal and spatial dimensions, ultimately enhancing the accuracy and robustness of predictive models.

In a parallel domain, Zhao et al. (2022) tackle the intricate task of classifying "Environmental Microorganism (EM)" images, an area witnessing notable advancements attributed to the rise of "deep learning." Despite these strides, the study sheds light on the persisting challenge of dealing with small EM datasets, requiring dedicated efforts to identify models that not only perform well but are also compatible with contemporary tools. Zhao et al. conduct a comprehensive set of experiments involving 21 "deep learning" models, encompassing crucial aspects such as "hyperparameter tuning," imbalanced training, and "direct classification." The insights gleaned from these experiments highlight the complementarity of the proposed model, laying the groundwork for future feature fusion research.

Notably, the research by Zhao et al. delves into the performance of the "VTs" series models—encompassing "ViT," "DeiT," "BotNet," and "T2T-ViT"—in the context of geometric deformation data augmentation. Surprisingly, the findings indicate that the performance of these models is not significantly enhanced by this particular augmentation technique. This revelation

prompts a deeper exploration into the intricacies of the models and the potential nuances in the augmentation process.

In essence, above two studies contribute significantly to their respective domains, pushing the boundaries of research in air quality prediction and environmental microorganism image classification. The methodologies employed, including the integration of spatial data and the meticulous experimentation with diverse deep learning models, showcase the dynamism and innovation prevalent in contemporary research, emphasizing the continuous quest for more robust and accurate models in the face of complex environmental challenges.

Weather, geography, and time all have intricate relationships and effects that make the “Air Quality Prediction (AQP)” project difficult. The strengths of mechanism models and machine learning are combined in a novel AQP technique called “Dynamic Multi-granularity Spatio-temporal Graph Neural Network (DM_STGNN)” to address this research conducted by Liao *et al.* (2023). In order to comprehend the spatiotemporal interactions between pollutants, "DM_STGNN" constructs a dynamic “spatio-temporal graph structure” using the air quality model “HYSPLIT”. The encoder creates a multi-granularity graph structure, assigns node attributes based on time, location, and weather information, creates edges dynamically using “HYSPLIT,” and uses "LSTM" to learn time-series connections of pollutant concentrations. For decoding and "AQP," the decoder employs an "LSTM" based on an attention mechanism.

The work by Limperis *et al.* (2023) presents a novel method for forecasting "PM2.5", a critical air quality indicator: “transformer-based prediction of PM2.5 (TPPM25)”. A “Transformer neural network” and other data embedding techniques are used by “TPPM25” to uncover temporal relationships between various meteorological parameters that affect "PM2.5" levels. On

a standard dataset, "TPPM25" demonstrates better prediction accuracy when compared to a "state-of-the-art" ensemble "deep learning model." Additionally, "TPPM25" performs better at maintaining high prediction accuracy over an extended period than both "Long-Short Term Memory (LSTM)" and "Bidirectional LSTM models."

Air pollution is a significant global issue because it has an impact on sustainable development and public health. Precise forecasts of air pollution are essential for studying air quality because they offer information about expected pollution levels in the future. To describe the quality of the air, one helpful tool is the "Air Pollution Index (API)." Because of air pollution on human health and economic growth worldwide, governments are becoming more and more concerned To order to estimate API for the Malaysian city of Klang. This study by Ragab *et al.* (2020) presents a new forecasting model that uses "One-Dimensional Deep Convolutional Neural Networks (1D-CNNs)" and "Exponential Adaptive Gradients (EAG)" optimization to modify the learning rate adaptively and converge in both convex and non-convex regions, "EAG" exponentially accumulates previous model gradients.

Environmental protection depends on air quality monitoring, but conventional techniques involving sophisticated, large-scale chemical devices are difficult for the general public to use. The objective of the study by Wang *et al.* (2022) is to make it simple for anyone to engage in district-wide air quality monitoring and to acquire a local "Air Quality Index (AQI)" data. To do this, the researchers suggest an enhanced "Transformer-based" technique that utilizes mobile device photos to predict local AQI values with greater accuracy. The strategy performs better in terms of flexibility and frequency than conventional approaches. The "Double Output Vision Transformer (DOViT)" automatically extracts features from images with a greater classification accuracy by using the "multi-head self-attention (MSA)" method.

"Road salt" is frequently used in colder climates to melt snow and ice from roadways throughout the winter. However, freshwater ecosystems may be damaged if this salt washes into neighboring urban streams via storm sewage systems. A specific and reliable model is required to predict future chloride concentrations in urban streams to address this problem. The goal of this study by Oliveira Santos *et al.* (2023) is to predict the chloride concentrations in the Credit River in Ontario, Canada, for a 6-hour forecasting horizon using a "Graph Neural Network"- "Sample and Aggregate" (GNN-SAGE) model. With "RMSE" and "R2" values of "51.16 ppb" and "0.88", respectively, the "GNN-SAGE" beat a "Deep Neural Network-based transformer (DNN-Transformer)" model and a benchmarking persistence model.

According to the study by Chen *et al.* (2023), engine performance calibration has shown encouraging results with recent advances in "Artificial Intelligence (AI)" technologies such as "Transformer" and "Long and Short-Term Memory (LSTM),,,", especially for predicting and optimizing engine emission performance. However, because of the inherent limitations of the "encoder-decoder" structure, "Transformer" cannot fully capture the dynamic features of previous emission information derived from "cycle-by-cycle" engine combustion events, resulting in lower algorithm adaptability and low accuracy. Similarly, "LSTM" suffers from gradient disappearance on long input and output sequences. This research presents an "Embedding-Graph-Neural-Network (EGNN)" model combined with a self-attention mechanism for adaptive graph formation to handle these problems.

By recognizing the correlation between sequences and reducing the number of parameters, this model can predict long-time step sequences with improved accuracy and reduced network topology. Similarly, a technique known as "sensor embedding" is utilized to enable the model to adapt to inconsistent attributes of sensor input, consequently mitigating the influence of

experimental hardware on the precision of the prediction. Based on the experimental results, it can be concluded that the “EGNN” model is a potential tool for future engine calibration operations, as it performs 31.04% better on average in long-time step prediction than five other standard models.

This literary work by Wang *et al.* 2022, constructed two "deep learning" models, "DeepPM" and "APTR," using "PM2.5" and "O3 monitoring" data as well as "WRF-Chem numerical" forecasts in the south-central "Beijing-Tianjin-Hebeito order to improve the prediction accuracy of numerical air quality models. Test datasets and a range of evaluation measures were used to optimize and assess the models. The results demonstrated that for both immediacy forecasts over the next 24 hours and short- to medium-term forecasts over the next 144 hours, the optimized "PM2.5" and "O3" forecast results generated by "DeepPM" and "APTR" significantly outperformed the "WRF-Chem numerical" model. While the "DeepPM" model performed better overall in optimizing short- and medium-term forecasts, the "APTR" model produced the best optimization results in proximity predicting.

This research conducted by Tariq *et al.* (2023) offers a distance adaptive graph convolutional gated network to overcome the shortcomings of the early warning procedures for air quality that are presently in service. Through spatio-temporal sensor fusion, the procedure allows for simultaneous projections of principal air pollutants at different temporal horizons and locations within a mega-city. It manages problems, including estimating pollution levels in new locations and long-term sensor failure. The proposed method dramatically reduces prediction mistakes in comparison to current models, showing improvement of “43.89%” and “52.59%”, respectively, for a 12-hour ahead “PM2.5” predicted over a time-series-Transformer and convolutional long-short-term memory network.

To enhance spatio-temporal predictions of “PM2.5” concentrations in areas exposed to wildfires, this study by Yu *et al.* (2023) delivers the "SpatioTemporal (ST)-Transformer," a multi-head attention-based deep learning architecture. The model uses an attention mechanism to focus on important contextual information that spans variable-wise, temporal, and geographical dimensions. By taking into account critical variables such as road traffic, weather, wildfire parameters, and historical 24-hour temporal indicators, the “ST-Transformer” functions better than present time series forecasting methods, especially when it comes to catching sudden shifts that happen during wildfires. The attention matrix of the model makes it more comfortable to understand complex temporal, spatial, and variable relationships, which makes it possible to differentiate between scenarios, including wildfires, and those that do not.

Mitreska Jovanovska *et al.* (2023) examine the important task of precisely measuring and predicting air pollution in modern metropolitan surroundings, acknowledging the position of "machine learning (ML)" as a key tool for this purpose. The study methodically chooses 30 relevant papers using the "PRISMA methodology" from trustworthy databases like "PubMed," "Springer," "IEEE," "MDPI," and "Elsevier." The study reviews "ML algorithms," models, and statistical approaches used in the forecast of overall urban air contaminants in depth. The outcomes emphasize generally anticipated pollutants, popular machine learning methods, and their significance in assessing urban air quality. The study contains a case study from Skopje, North Macedonia, which emphasizes current work in air pollution prediction and measuring systems.

According to Mengara Mengara *et al.* (2022), Busan, South Korea, is a model city to presents a unique framework for worldwide air quality forecast. The method estimates the pollution intensity of "PM2.5" and "PM10" particles utilizing an attention-based convolutional "BiLSTM

autoencoder model" trained on allocated data parallelism. By using meteorological and transportation data, the model surpasses earlier methods in terms of short- and long-term predictions. The forecasting precision is improved by adding traffic data that is derived from the "YOLO method." According to the outcomes, the suggested model outperforms baseline approaches in terms of accuracy and efficiency, achieving the lowest "MAE," "RMSE," and "SMAPE" for "PM2.5" forecasts. The design, which is hosted on a cloud server, emphasizes its potential for efficient worldwide air quality monitoring and forecast by providing real-time, on-demand data.

This research paper introduces a "spatio-temporal Informer model" to address the problem of precisely predicting air pollution, which is important for safeguarding public health. The model suggested by Feng *et al.* (2023) forecasts multi-positional, multi-temporal Air Quality Index (AQI) values for monitoring stations by altering input data, which includes spatial data and air quality, using novel spatio-temporal embedding and attention procedures. The proposed model performs more useful than others when assessed using data from 34 stations in Beijing, China, with an average "MAPE" of "11.61%" for forecasts made one to twenty hours in advance. Notably, even at extreme locations, it displays remarkable accuracy and stability. These outcomes demonstrate the successful usage of the "spatio-temporal Informer" for precise air quality predictions and validate the significance of the spatio-temporal embedding and attention techniques.

To improve "PM2.5" concentration forecasting for handling hazy weather, this Research by Ma *et al.* (2023) presents the "SpatioTemporal-Informer (ST-Informer) model." The "ST-Informer" combines an independent spatio-temporal embedding layer with parallel computing of long correlations to manage shortcomings in current spatio-temporal prediction models. Exact

predictions rely on the dynamic spatio-temporal correlations captured by this layer. Improving the extraction of contextual information from spatio-temporal data is the "ProbSpare Self-Attention mechanism." The "ST-Informer" functions incredibly satisfactorily at identifying peaks and abrupt deviations in the concentrations of "PM2.5" by using meteorological and air pollution data from several locations.

According to the study conducted by Han *et al.* (2023), the critical relationship between specific weather forecasts and air quality for urban planning and its connection to human welfare is addressed in the research. In contrast to conventional methods, the proposed "MasterGNN+" takes a comprehensive approach, recognizing the interaction between different prediction tasks. "Spatial autocorrelation" is modeled utilizing geographic distance and environmental context utilizing a multi-view graph learning block. Long-range "temporal autocorrelation" is captured by employing a specialized developed recurrent network at every monitoring station and time interval. The research offers a "multi-adversarial graph learning framework" to mitigate the transmission of observation noise.

The study handles the urgent problem of the influence of air pollution on human health by focusing on enhancing air quality predictions for scenarios affecting entire countries. An advanced hierarchical model for sophisticated countrywide city air quality prediction called the "Group-Aware Graph Neural Network (GAGNN)" is suggested by the research conducted by Chen *et al.* (2021). "GAGNN" employs a city group graph and a city graph to determine latent dependencies and spatial dependencies, respectively. Latent dependencies between cities are indicated by the use of a differentiable grouping network, leading to the creation of city groups. A group correlation encoding module efficiently apprehends inter-group dependencies. "GAGNN models" dependencies between cities and city groups through message forwarding.

This Research by Zhang *et al.* (2022) examines the relationship between "indoor air quality (IAQ)" and environmental elements and building components in eleven campus buildings located in Gainesville, Florida. For two weeks, integrated sensor techniques tracked temperature, "relative humidity," "ozone," "nitrogen dioxide," and "airborne particulate matter" in every building. Twenty building details are gathered, which function as the foundation for a Levenberg-Marquardt Backpropagation neural network model with "PCA" assistance. The model accurately indicates the relations of IAQ with controlling factors. The effect of outdoor sources on indoor exposure is significant, as seen by the substantial correlations followed between "O₃" levels and "PM(2.5–10)" levels indoors and outside. The distance from heavy traffic, cracks, outside temperature, and humidity are important affecting factors.

Since urban air pollution is evolving more and more detrimental to human health, accurate forecasting of air quality is essential. This study conducted by Singh *et al.* (2020) proposes a model that uses the "C5.0" and "RPART" algorithms to classify and forecast atmospheric pollution. The model demonstrates its effectiveness by employing supervised machine learning methods on daily data gathered by "CPCB" from the "ITO station" in New Delhi. The research shows that time, place, and unfamiliar factors have a big impact on air quality. Both algorithms indicate effectiveness in indexing and forecasting extensive datasets, providing a practical instrument for reducing the harmful effects of air pollution on urban living and public health.

Chapter 3 Methodology

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely accepted methodology for guiding data mining projects. Organizations use CRISP-DM for its structured approach, ensuring alignment with business objectives, iterative flexibility, and collaborative engagement between business and data professionals. This methodology enables systematic progression through the entire data mining process, from understanding business goals to deploying effective solutions, making it a valuable framework for achieving meaningful insights and successful outcomes in diverse industries.

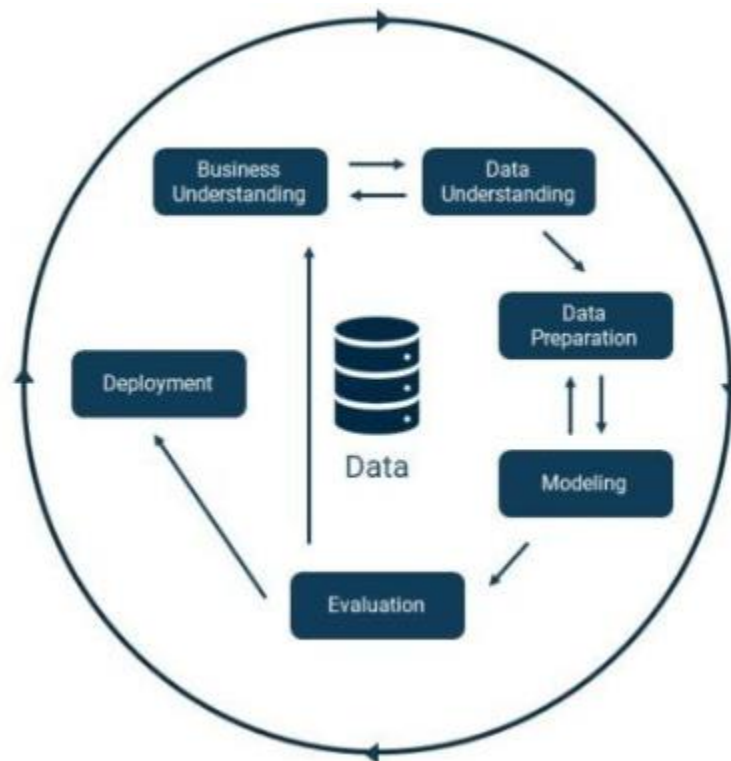


Figure 3.1: CRISP-DM Methodology

The CRISP-DM steps are as follows:

1. Business Understanding:

- Define the business goal: Improve the accuracy of Air Quality Index (AQI) predictions.
 - Specify objectives: Compare and enhance predictive models (Transformer Models, GNN, LSTM) for AQI prediction.
2. Data Understanding:
- Explore available datasets: Analyze historical air quality data.
 - Assess data quality: Examine completeness and reliability of AQI data.
 - Identify relevant variables: Select features impacting AQI.
3. Data Preparation:
- Cleanse and preprocess data: Handle missing values and outliers.
 - Feature engineering: Extract meaningful features for model training.
 - Data transformation: Normalize or scale data as needed for modeling.
4. Modeling:
- Select models: Choose Transformer Models, Graph Neural Networks (GNN), and LSTM.
 - Train models: Utilize historical data to train each model.
 - Optimize hyperparameters: Fine-tune parameters for better performance.
5. Evaluation and Deployment:
- Assess model performance: Use Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics.
 - Compare models: Evaluate the effectiveness of Transformer Models, GNN, and LSTM in AQI prediction.
 - Validate results: Ensure findings align with business objectives.

Applying CRISP-DM ensures a systematic approach, from understanding the business problem to deploying effective models, contributing to the overall success of the study in enhancing air quality index predictions.

The project Methodology diagram are as follows:

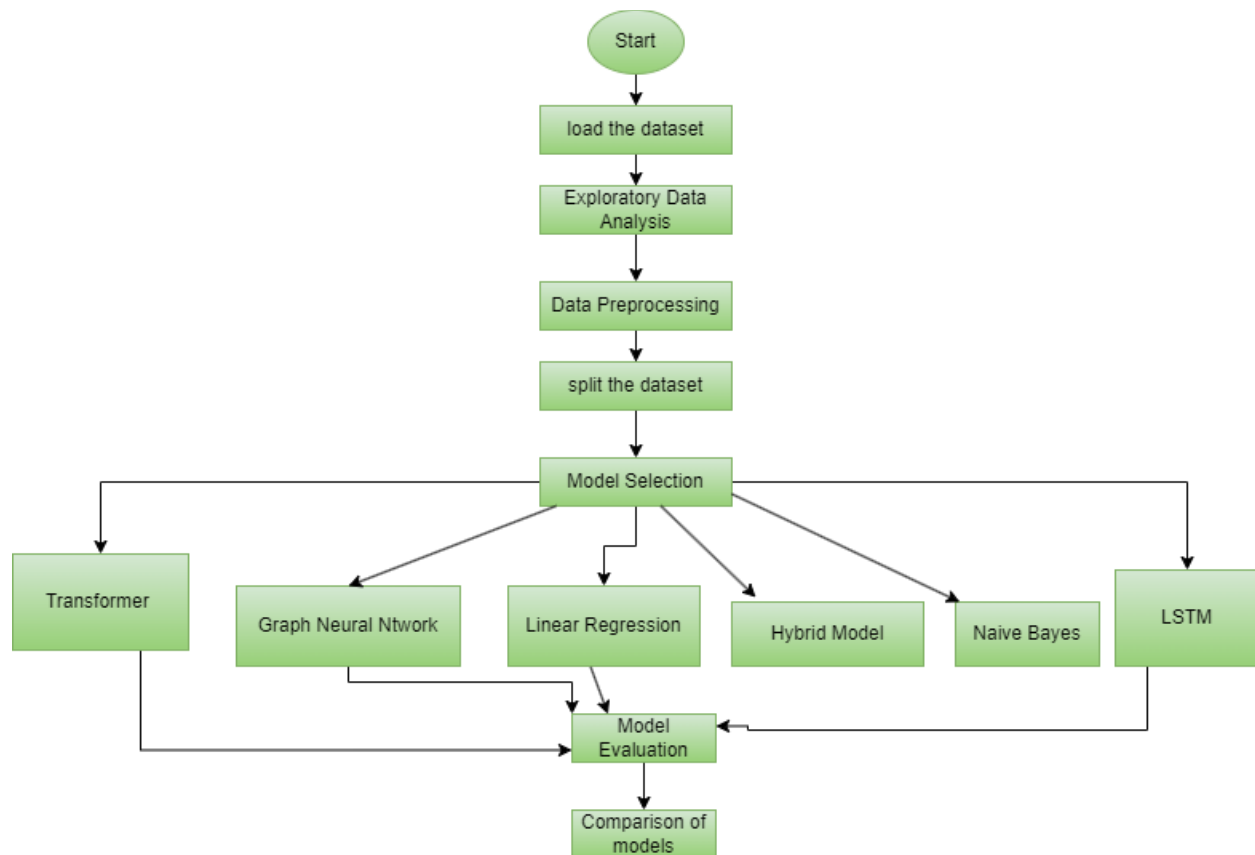


Figure 3.2: Project Methodology

Project Requirements:

In response to project requirements, we opted for Google Colab as our code development and execution platform. Its user-friendly interface, akin to Anaconda Navigator, streamlines the

coding process without the need for installation. This choice not only ensures simplicity and efficiency but also facilitates easy file sharing among collaborators, enhancing overall project collaboration.

1. Python:

Python, a versatile and user-friendly programming language, serves as the backbone for machine learning and data analysis tasks, offering extensive libraries and frameworks.

2. TensorFlow and Keras:

TensorFlow, a powerful open-source machine learning library, collaborates seamlessly with Keras, simplifying the creation and training of deep learning models with high-level abstractions.

3. Scikit-Learn (sklearn):

Scikit-Learn, a robust machine learning library, provides a comprehensive suite of tools for data preprocessing, model training, and evaluation, enhancing the efficiency of predictive modeling.

4. Matplotlib:

Matplotlib, a widely-used plotting library, enables the creation of insightful visualizations, making it indispensable for data exploration and result presentation in Python.

5. Seaborn:

Seaborn, built on top of Matplotlib, offers an aesthetically pleasing interface for statistical data visualization, enhancing the clarity and visual appeal of analytical findings.

3.1 Dataset

The dataset comprises air quality measurements from various stations, featuring parameters such as PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, and AQI_Bucket. Recorded at different datetimes, the data exhibits variations in pollutant concentrations. Missing values are evident in AQI and AQI_Bucket columns. StationId "AP001" is represented, suggesting a focus on a specific location. This comprehensive dataset provides insights into air quality dynamics, enabling in-depth analysis and modeling to enhance understanding and prediction of air quality index variations over time.

```
#   Column      Dtype
---  -
0   StationId  object
1   Datetime   object
2   PM2.5      float64
3   PM10       float64
4   NO         float64
5   NO2        float64
6   NOx        float64
7   NH3        float64
8   CO         float64
9   SO2        float64
10  O3         float64
11  Benzene    float64
12  Toluene    float64
13  Xylene     float64
14  AQI        float64
15  AQI_Bucket object
dtypes: float64(13), object(3)
memory usage: 316.1+ MB
None
```

Figure 3.3: Dataset Information

Within this dataset, records encompass details of air quality measurements conducted at diverse monitoring stations. Each entry is characterized by a distinct identifier for the station, capturing the essence of its spatial location. Correspondingly, temporal information in the form of date and

time accompanies each measurement, allowing for a temporal analysis of air quality trends. The dataset further encapsulates numerical values signifying concentrations of various airborne pollutants, contributing to a comprehensive overview of environmental factors. Additionally, the dataset incorporates the derived Air Quality Index (AQI), a synthesized metric offering a holistic assessment of overall air quality. The AQI is complemented by a qualitative classification in the AQI_Bucket column. Distinct data types, 'object' for categorical variables and 'float64' for numerical variables, reflect the diverse nature of information contained within. The dataset's substantial memory usage, totaling approximately 316.1 MB, underscores the richness and complexity of the air quality data. This foundational information serves as a precursor to in-depth exploratory analysis, fostering a nuanced understanding of air quality dynamics across the monitored stations and time intervals.

3.2 Data Analysis using tableau

Dashboard 1:

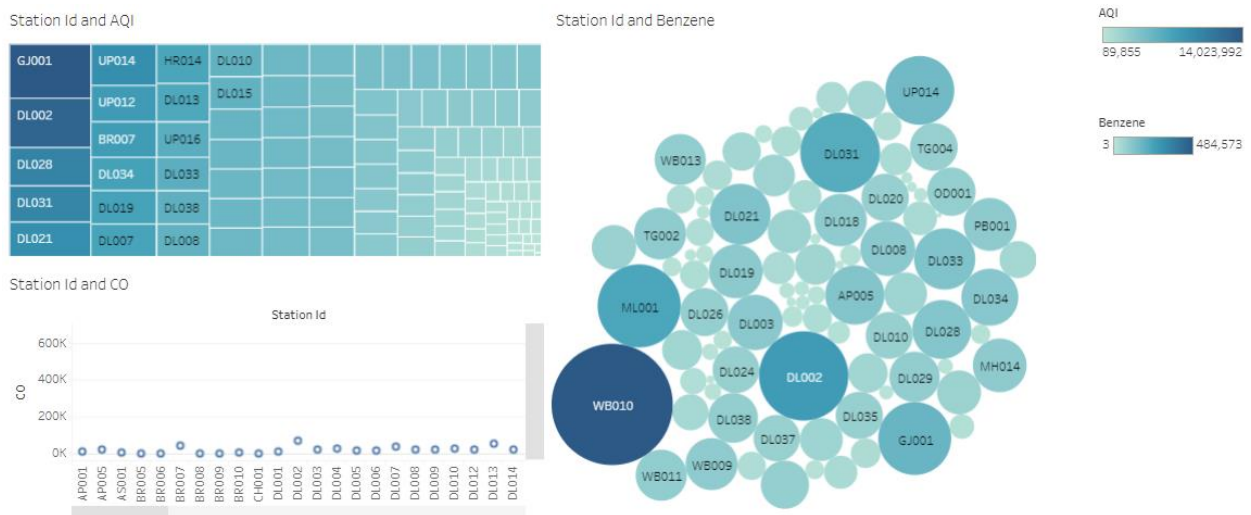


Figure 3.4: Dashboard 1

In the above dashboard here first worksheet show the treemap for Station Id and AQI, here dark region has maximum number of AQI as compare to light region. The second worksheet show the bubble chart for station Id and Benzene, here dark region has maximum count of benzene as compare to light region , here WB010 station Id has highest count of benzene. The third worksheet show the circle view for station Id and CO where station Id is on xx axis and CO is on y axis.

Dashboard 2:

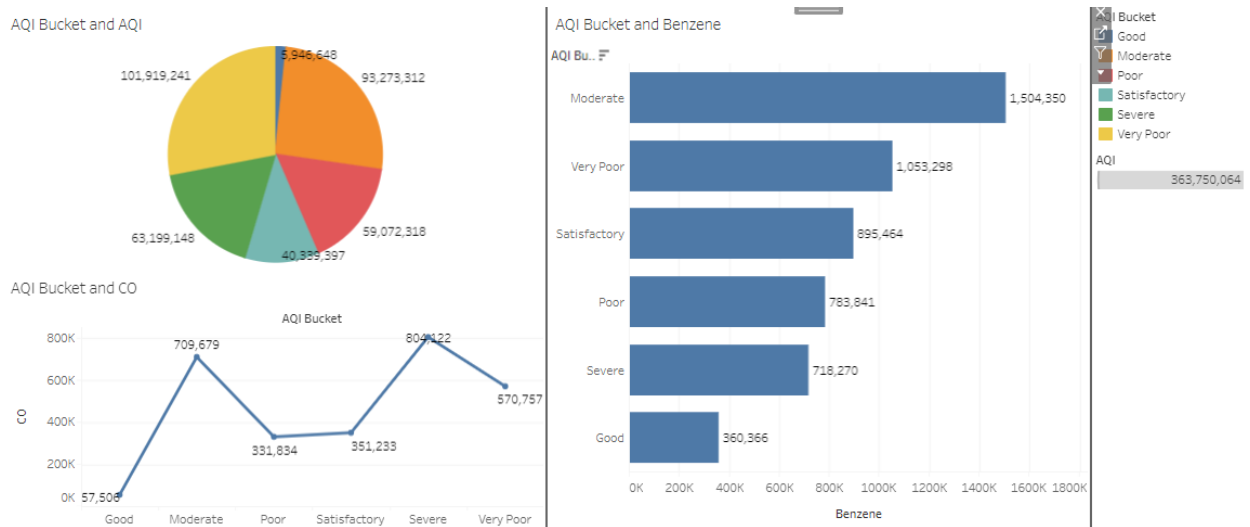


Figure 3.5: Dashboard 2

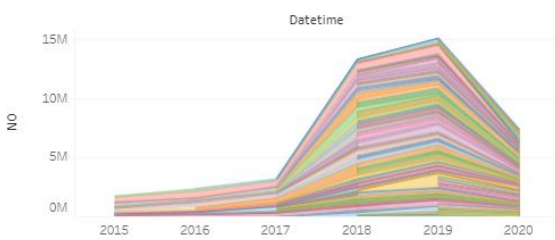
In the above dashboard first worksheet shows the pie chart for AQI Bucket and AQI, here different color indicate the different AQI Bucket. The second worksheet show the horizontal bar chart for AQI Bucket and Benzene, here moderate AQI bucket has highest count as compare to other AQI Bucket. The third worksheet show the line chart for AQI Bucket and CO, here severe AQI Bucket has maximum count of CO as compare to other AQI Bucket.

Dashboard 3:

Station Id and Nox



Datetime, NO and Station Id



Datetime, NH3 and AQI Bucket

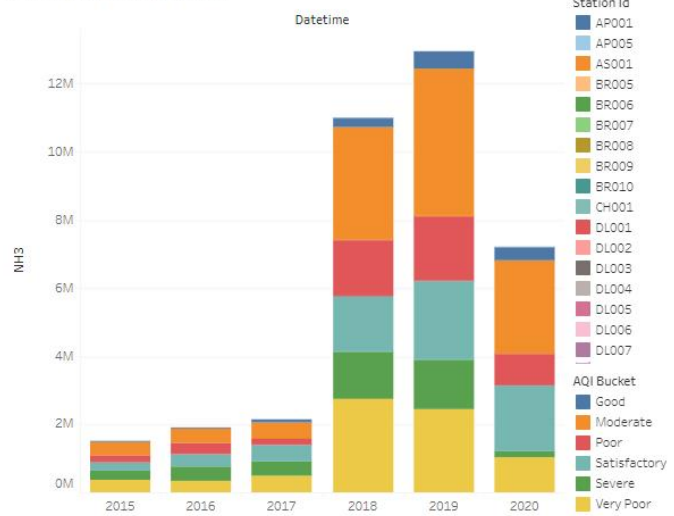


Figure 3.6: Dashboard 3

In the above dashboard here first worksheet show the Station Id and N ox, here different color indicate the different station Id. The second worksheet show the stacked bar plot for Datetime, NH3 and AQI bucket, here different color indicate the different AQI Bucket. The third worksheet show the area plot for Datetime , NO and Station Id, here different color indicate the different station Id.

Dashboard 4:

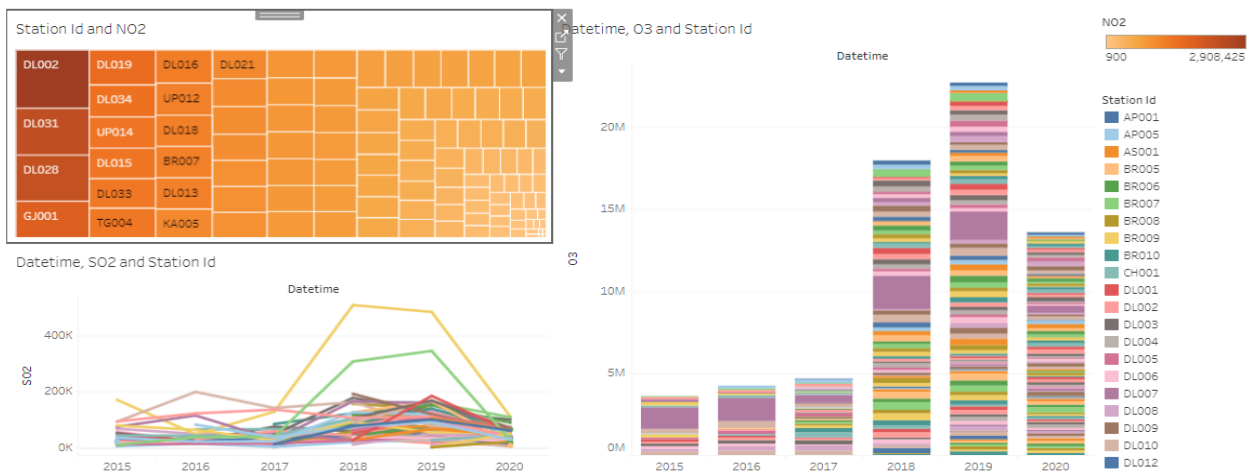


Figure 3.7: Dashboard 4

In the above dashboard first worksheet shows the treemap for Station Id and NO2, here dark region has maximum count of NO2 as compare to light region. The second worksheet show the stacked bar chart for Datetime, O3 and Station Id, here different color indicate the different station Id. The third worksheet show the line chart for Datetime, SO2 and Station Id, here different color indicate the different station Id.

Dashboard 5:

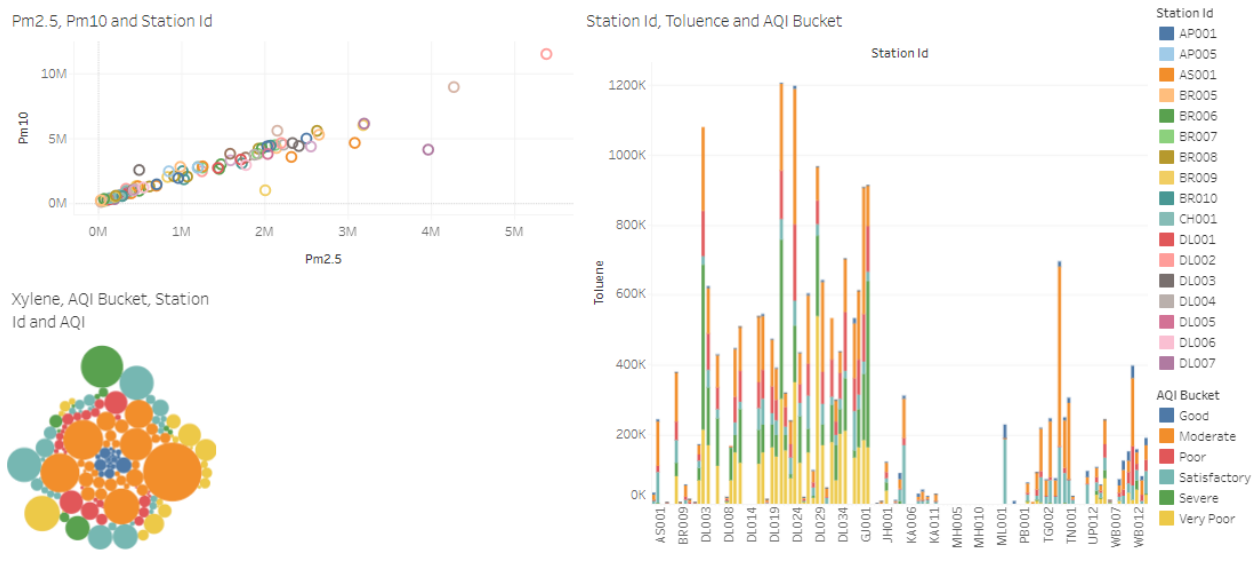


Figure 3.8: Dashboard 5

In the above dashboard first worksheet shows the scatterplot for pm2.5, pm10 and station Id, here different color indicate the different station Id. The second worksheet show the treemap plot for station Id, toluence and AQI Bucket, here different color indicate different AQI Bucket .

The third worksheet show the bubble chart for Xylene, AQI Bucket, Sttation Id and AQI, here different color indicate the different AQI Bucket.

3.3 Data Pre-processing

Data preprocessing is a fundamental step that involves cleaning and organizing raw data to enhance its quality for analysis. This includes tasks such as handling missing values, ensuring data consistency, and structuring the dataset for effective use in subsequent analytical processes.

The goal is to optimize the data for accurate and meaningful insights, ultimately improving the performance of analytical models and facilitating a more robust exploration of the dataset.

```
#installing all libraries
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from seaborn import heatmap
from sklearn.feature_selection import chi2
from sklearn.feature_selection import SelectKBest
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler,MinMaxScaler
import tensorflow as tf
import keras.backend as K
from tensorflow.keras.models import Sequential, Model
from tensorflow.keras.layers import InputLayer, Dense, Flatten, Layer, Input, Dropout, BatchNormalization, LSTM, Bidirectional
from tensorflow.keras.layers import Attention, MultiHeadAttention, LayerNormalization, Add, Multiply
```

Figure 3.9: Importing the libraries

The provided code snippet sets up the environment for data analysis and machine learning. It includes the installation of crucial libraries, such as Pandas for data manipulation, NumPy for numerical operations, Matplotlib and Seaborn for data visualization, and TensorFlow and Keras for machine learning tasks. Additionally, various modules for feature selection, label encoding, and scaling are imported. The code aims to create a streamlined workflow, allowing for efficient

data exploration, feature engineering, and the development of machine learning models. It also incorporates configurations for visualization styles to enhance the readability of generated plots. The inclusion of TensorFlow and Keras suggests a focus on deep learning methodologies, highlighting the potential utilization of neural networks in the subsequent analysis.

```
data = pd.read_csv('archive/station_hour.csv')
```

Figure 3.10: Load the dataset

This code snippet is a common approach to loading tabular data from a CSV file into a Pandas DataFrame, providing a structured and versatile data structure for various data processing tasks in Python.

```
data.describe()
```

Figure 3.11: Statistical Description of dataset

This summary provides key insights into the distribution and central tendencies of the dataset. The 'count' row indicates the number of non-null values, offering an initial assessment of data completeness. The 'mean' represents the average value, while the 'std' (standard deviation) quantifies the spread or dispersion around the mean. The 'min' and 'max' values reveal the range of observed data, and the percentiles (25%, 50%, 75%) offer a nuanced understanding of the data distribution. This concise summary serves as a valuable precursor to exploratory data analysis, aiding in the identification of potential patterns, outliers, and areas for further investigation.

```
data = data.drop(['AQI_Bucket'], axis=1)
```

Figure 3.12: Drop the AQI_Bucket

In essence, this code snippet is used to exclude the 'AQI_Bucket' column from the dataset, potentially because it is not required for the specific analysis or modeling task at hand. Removing unnecessary columns can streamline the dataset and enhance computational efficiency.

```
data.Datetime= pd.to_datetime(data.Datetime , format='%Y-%m-%d')
```

Figure 3.13: Convert date to datetime

The specified 'format' parameter ensures the correct interpretation of the date string, which is in the 'YYYY-MM-DD' format. Converting the 'Datetime' column to a datetime data type allows for seamless temporal analysis, enabling time-based operations and enhancing the dataset's compatibility with time-series analysis or modeling tasks.

```
data.set_index("Datetime",inplace = True)
```

Figure 3.14: Set date as index

This operation modifies the DataFrame in place, meaning it directly alters the existing DataFrame without the need for reassignment. The resulting DataFrame now has a datetime index, facilitating time-based analyses and providing an organized structure for time-series-related operations.

```
data.isnull().sum()
```

Figure 3.15: checking the null values

The **isnull()** method returns a DataFrame of the same shape as 'data' with 'True' where values are missing and 'False' where values are present. By applying **.sum()** on this boolean DataFrame, it calculates the sum along each column, providing a count of missing values in each respective column. This summary is valuable for assessing the extent of missing data and guiding decisions on data imputation or handling strategies.

```
data= data.dropna()
```

Figure 3.16: Drop the missing value

The **dropna()** method is applied to the entire DataFrame, and any row containing at least one missing value is dropped. This operation is useful when dealing with incomplete or inconsistent data, and it helps ensure that the remaining dataset is complete. The modified DataFrame, with missing values removed, is then reassigned to the variable 'data', reflecting the cleaned dataset that can be used for subsequent analysis or modeling.

```
# getting #unique values for each column
uniq= []
for col in data.columns:
    print(col, "=", len(data[col].unique()))
    uniq.append(len(data[col].unique()))
```

Figure 3.17: Unique values of columns

The provided code iterates through each column in the DataFrame 'data' to determine the number of unique values present in each column. This is achieved by utilizing a for loop that prints the column name along with the count of unique values for that specific column. The resulting information is stored in a list named 'uniq,' capturing the uniqueness of values across all

columns. This insightful summary assists in recognizing the distinctiveness of data within each feature, providing a foundational understanding of the dataset's variability. The 'uniq' list becomes a valuable reference for identifying categorical features and guiding subsequent steps in data preprocessing or feature selection with an emphasis on maintaining data quality.

```
# Pearson correlation coefficient
corr = data.corr()["AQI"].sort_values(ascending=False)[1:]

# absolute for positive values
abs_corr = abs(corr)

# random threshold for features to keep
relevant_features = abs_corr[abs_corr>0.1]
relevant_features
```

Figure 3.18: Pearson correlation

The provided code calculates the Pearson correlation coefficient between the 'AQI' (Air Quality Index) column and all other columns in the DataFrame 'data'. The resulting correlation values are then sorted in descending order. To focus on meaningful correlations, the absolute values are taken, and a threshold of 0.1 is set to filter out features with weaker correlations. The 'relevant_features' variable contains the features deemed relevant based on the specified threshold. This process helps identify and retain features that exhibit a significant correlation with the AQI, serving as a valuable step in feature selection for subsequent analysis or modeling tasks.

```
dt= data[relevant_features.index]
dt['AQI']= data['AQI']
data= dt
```

Figure 3.19: Creating a new dataframe

The code creates a new DataFrame, 'dt', incorporating only features deemed relevant based on the previously identified correlation threshold. It includes these selected features along with the 'AQI' column. The original DataFrame 'data' is then updated to reflect this refined dataset, concentrating on the relevant features for subsequent analysis or modeling.

```
le= LabelEncoder()  
data['StationId'] = le.fit_transform(data['StationId'])
```

Figure 3.20: Label Encoder

The provided code snippet employs scikit-learn's **LabelEncoder** to facilitate the transformation of categorical values within the 'StationId' column in the 'data' DataFrame into numerical representations. Beginning with the importation of the **LabelEncoder** class from the scikit-learn library, a LabelEncoder instance is instantiated and assigned to the variable **le**. This instance is then employed to execute the encoding process on the 'StationId' column through the **fit_transform** method. This method not only adapts the encoder to the unique categorical values present in the column but also converts these categorical values into corresponding numerical labels.

In essence, the 'StationId' column is being subjected to a conversion where each distinct categorical value is mapped to a unique numerical label. This encoding is particularly valuable in the realm of machine learning, as it enables the integration of categorical data into algorithms that require numerical input. By executing this transformation, the 'StationId' column becomes a numerical representation, ensuring compatibility with various machine learning models and enhancing the effectiveness of subsequent analyses.

3.4 Data Visualisation

Data visualization is the graphical representation of data to uncover patterns, trends, and insights. It involves creating visual representations such as charts, graphs, and maps to make complex datasets more understandable and accessible. Effective data visualization enhances communication, aiding in the interpretation and communication of data-driven findings, facilitating informed decision-making.

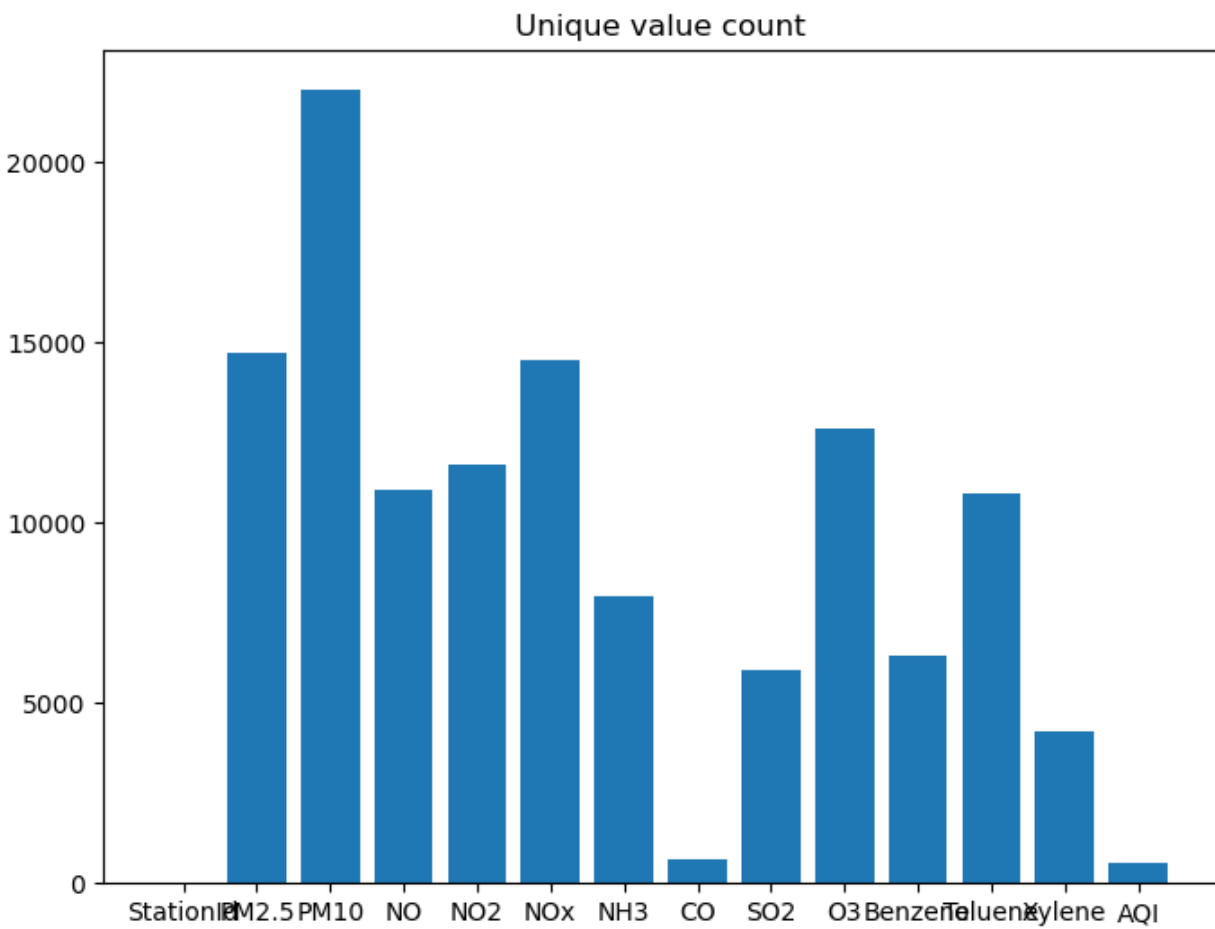
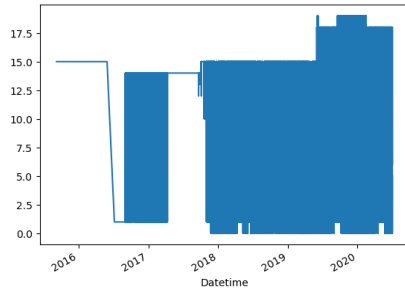


Figure 3.21: Unique counts of columns

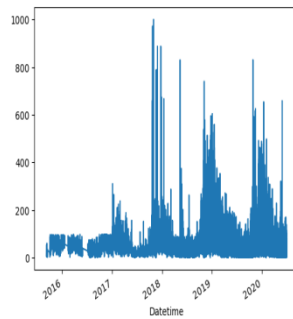
Each bar corresponds to a specific column, and its height signifies the count of unique values for that particular feature. In the resulting chart, the 'PM10' column exhibits the highest bar,

indicating the highest count of unique values among all features. This visualization offers a quick comparison of the diversity in unique values across different columns, highlighting 'PM10' as the feature with the greatest variability or distinctiveness in its data points.

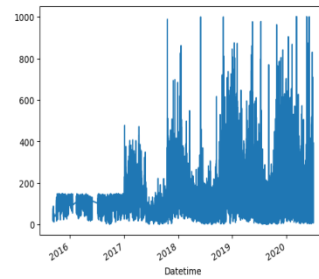
StationId



PM2.5

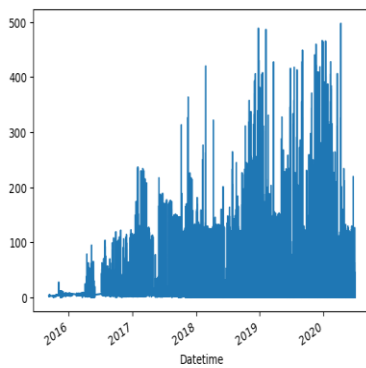


PM10

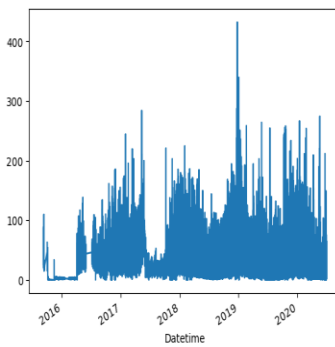


NO2

NO



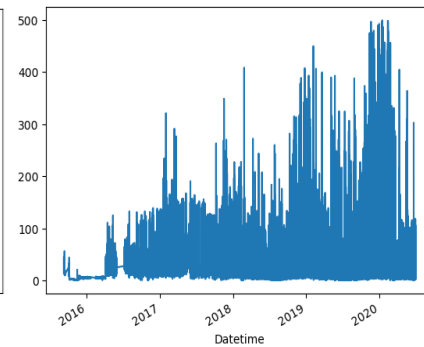
NOx

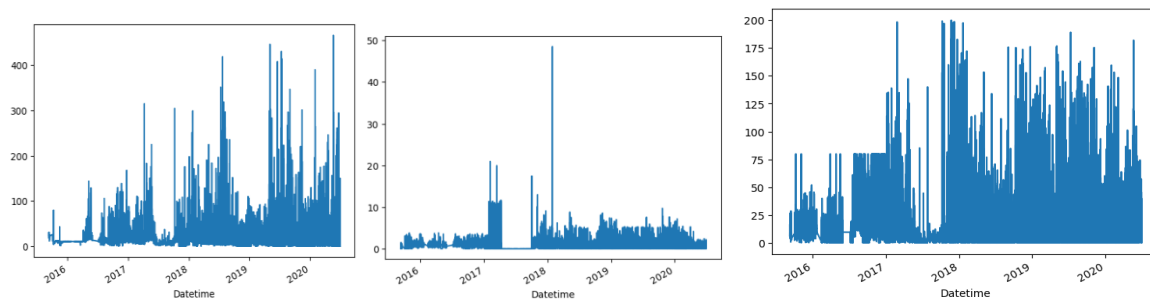


NH3

CO

SO2

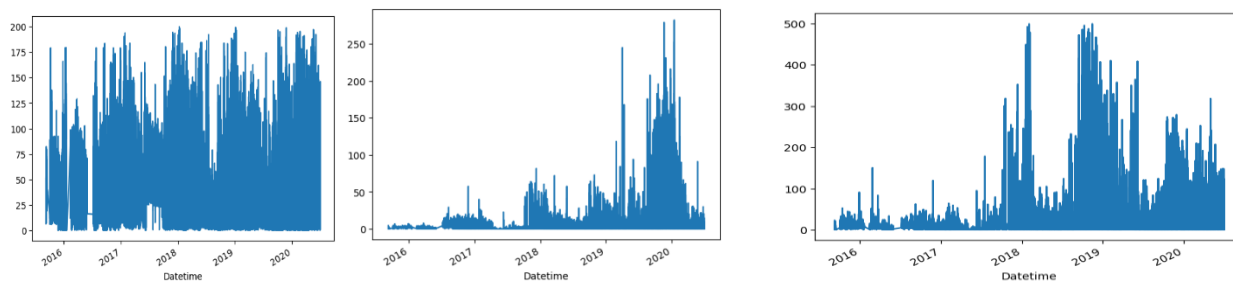




O3

Benzene

Toluene



Xylene

AQI

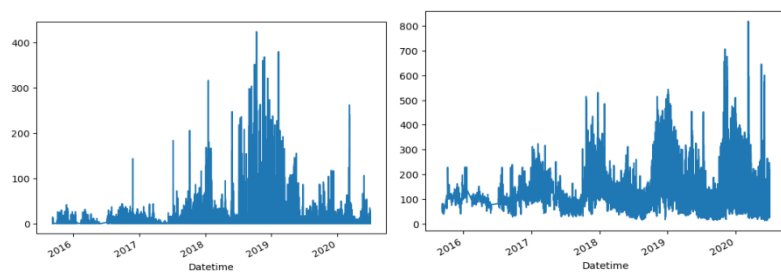


Figure 3.22: Line Plot for all columns

This visualization strategy enables a detailed inspection of the temporal or sequential patterns inherent in the data. By generating separate line plots for every column, the code provides a unique visual perspective on how each variable evolves or changes over the specified range.

Following the creation of each line plot, the corresponding column name is printed, adding a textual reference to the visual representation. This approach not only facilitates the identification of distinct trends within each feature but also assists in pinpointing any noteworthy fluctuations or anomalies that may exist in the dataset.

The interpretation of these line plots is guided by the conventional understanding that the horizontal axis typically represents the index or time, while the vertical axis depicts the respective feature's values. Consequently, the visual inspection of these sequential patterns aids in uncovering valuable insights about the dataset's temporal dynamics and can be crucial for making informed decisions during data analysis or modeling processes. The repetitive nature of this process, creating a series of line plots, contributes to a comprehensive exploration of the dataset's temporal nuances and enhances the overall understanding of its evolving characteristics.

3.5 Model Training

Model training is a crucial step in machine learning where an algorithm learns patterns and relationships within a labeled dataset. The dataset comprises known input features along with their corresponding output labels. The model iteratively adjusts its internal parameters based on the provided examples, aiming to minimize the discrepancy between its predictions and the actual labels. The process involves exposing the model to various instances from the dataset, allowing it to learn and generalize patterns. The ultimate objective is to enhance the model's

ability to make accurate predictions on new, unseen data by capturing inherent patterns and relationships present in the training dataset.

```
Xtrain, Xtest, y_train, y_test = train_test_split(X, y, test_size= 0.2)
```

Figure 3.23: split the dataset

The code snippet utilizes the **train_test_split** function from the scikit-learn library, a crucial step in machine learning model development. This function systematically divides the dataset into training and testing sets, with X representing the feature matrix and y denoting the target variable. By specifying **test_size=0.2**, 20% of the data is reserved for testing, leaving 80% for model training. This strategic split enables a comprehensive evaluation of the model's generalization capabilities on new, unseen data.

The function returns four distinct sets: Xtrain and y_train for model training, and Xtest and y_test for assessing the model's performance on previously unseen data. This partitioning is pivotal for ensuring the model's effectiveness in real-world scenarios. By maintaining the integrity of the data and adhering to best practices, this approach facilitates a robust evaluation of the model's predictive accuracy and its ability to extrapolate learned patterns to novel instances.

3.5.1 Graph Neural Network (GNN)

Graph Neural Networks (GNNs) have emerged as a potent tool for the analysis of graph-structured data, making them particularly relevant for air quality prediction where spatial dependencies and interrelationships between monitoring stations play a crucial role. In the air quality context, each monitoring station can be viewed as a node in a graph, and GNNs excel at capturing the complex, non-linear relationships between these nodes. By leveraging the graph

structure, GNNs enable the modeling of spatial correlations, offering a nuanced understanding of how pollution levels propagate across different geographic locations. The ability to aggregate information from neighboring nodes makes GNNs well-suited for capturing intricate patterns in air quality data, providing a holistic perspective on pollution dynamics.

3.5.2 Hybrid Model

A Hybrid Model in the realm of air quality prediction signifies a fusion of diverse modeling approaches to harness the combined strengths of different methodologies. This fusion often involves integrating traditional statistical techniques, which offer interpretability and a solid theoretical foundation, with modern deep learning methods, known for their ability to capture complex patterns. By amalgamating these approaches, a Hybrid Model aims to achieve a comprehensive understanding of air quality dynamics. For instance, the model might use statistical methods for feature engineering and interpretability, while leveraging the representation learning capabilities of deep learning to capture intricate relationships. This integration enhances the model's adaptability to varying complexities within air quality datasets, contributing to improved predictive accuracy and robustness.

3.5.3 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), are well-suited for handling sequential and time-series data, making them invaluable for air quality prediction tasks. In the context of air quality, where pollutant concentrations exhibit temporal dependencies, LSTMs excel at capturing patterns over time. The architecture of LSTMs, equipped with memory cells and gating mechanisms, enables them to selectively store and retrieve information across different time steps. This unique ability is particularly beneficial when modeling the dynamic and evolving nature of air quality data. LSTMs provide a powerful

framework for understanding and predicting pollution trends, contributing to more accurate and informed decision-making in environmental management.

3.5.4 Transformer Model

Transformer models, initially designed for natural language processing, have proven to be highly adaptable and effective in various domains, including time-series analysis such as air quality prediction. Their architecture, characterized by self-attention mechanisms, allows them to capture long-range dependencies and relationships within sequential data. In the context of air quality, where intricate patterns may emerge over different time scales, Transformers excel at dynamically weighing the significance of different time steps. This adaptability is crucial for accurately capturing the complex and high-dimensional nature of air quality datasets. Transformers offer a versatile and scalable solution, allowing them to learn and represent patterns in the data, contributing to more accurate and context-aware predictions in air quality forecasting.

3.5.5 Linear Regression

Linear regression is a statistical technique used to analyze and model the linear relationship between a dependent variable and one or more independent variables. The method aims to uncover the underlying pattern or trend in the data by finding the optimal linear equation that minimizes the differences between the observed values and the values predicted by the model. Widely applied in fields like economics, finance, and biology, linear regression helps researchers and analysts understand how changes in independent variables affect the dependent variable, enabling predictions and insights into the nature of the relationship.

3.5.6 Naive Bayes

Naive Bayes is a probabilistic classification algorithm grounded in Bayes' theorem. Despite its seeming simplicity and the "naive" assumption of independence among features, Naive Bayes proves effective in various applications, notably in text classification tasks like spam filtering or sentiment analysis. The algorithm calculates the likelihood of a data point belonging to each class based on its features and then assigns the class with the highest probability as the predicted class. While its independence assumption may not always align with real-world data dependencies, Naive Bayes remains computationally efficient, making it suitable for scenarios with high-dimensional data or where the independence assumption is reasonably justified.

Chapter 4 Results and Analysis

The results and evaluation of air quality prediction models involve assessing their performance metrics, such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). These metrics provide a quantitative measure of the model's accuracy in predicting air quality indices. A comprehensive evaluation considers factors like model interpretability, computational efficiency, and the ability to generalize to new, unseen data, providing a holistic understanding of the model's efficacy in real-world applications.

Mean Squared Error (MSE): Mean Squared Error is a common metric used to measure the average squared difference between predicted and actual values in a regression problem. It involves calculating the squared differences for each data point, averaging them, and provides a quantitative measure of the model's accuracy. A lower MSE indicates better predictive performance, with zero representing a perfect match between predicted and actual values.

Root Mean Squared Error (RMSE): Root Mean Squared Error is derived from MSE by taking the square root of the average squared differences between predicted and actual values. RMSE is particularly useful as it shares the same scale as the target variable, offering a more interpretable measure of prediction accuracy. Like MSE, a lower RMSE signifies improved model performance, with zero indicating a perfect match between predicted and actual values.

	Models	Mean Squared Error	Root Mean Squared Error
4	GNN	1601.863892	40.023292
0	Linear Regression	1937.551557	44.017628
5	Hybrid Model	2057.528076	45.359983
1	Naive Bayes	3840.805174	61.974230
3	LSTM	3997.251953	63.223824
2	Transformer Model	19766.738281	140.594233

Figure 4.1: MSE and RMSE of the models

The comparative analysis of air quality index (AQI) prediction models reveals significant differences in their performance. Transformer Models, despite their success in various domains, exhibit the highest Mean Squared Error (MSE) at 19766.74 and a Root Mean Squared Error (RMSE) of 140.59, indicating considerable deviations from actual AQI values. Graph Neural Networks (GNN) perform comparatively better with an MSE of 1601.86 and an RMSE of 40.02, showcasing their potential in AQI prediction. Linear Regression and Hybrid Models fall in between, with MSE values of 1937.55 and 2057.53, and RMSE values of 44.02 and 45.36, respectively. Traditional approaches such as Naive Bayes and LSTM exhibit higher errors, with Naive Bayes having an MSE of 3840.81 and an RMSE of 61.97, while LSTM shows an MSE of 3997.25 and an RMSE of 63.22. The findings emphasize the importance of selecting appropriate models for accurate AQI predictions, with GNN demonstrating promise in this environmental forecasting domain.

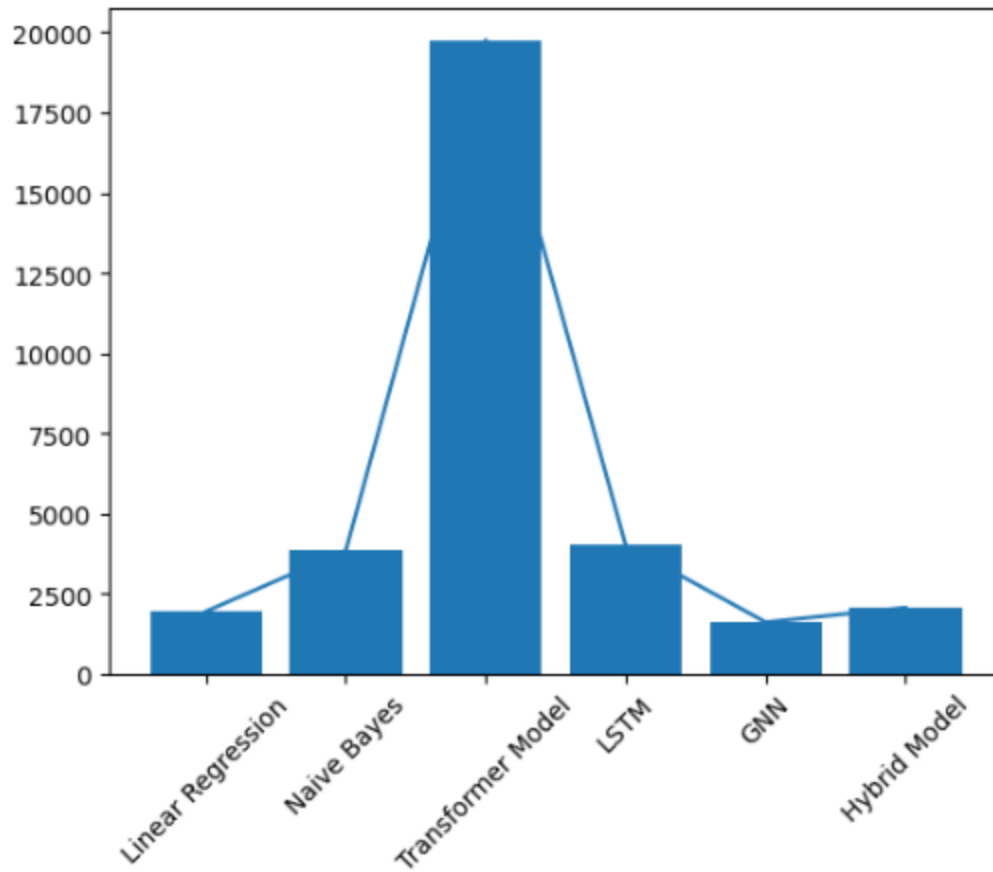


Figure 4.2: Bar chart for MSE of the models

The figure above show the bar chart for mean square error of the models , here transformer has highest mse and GNN has lowest mse.

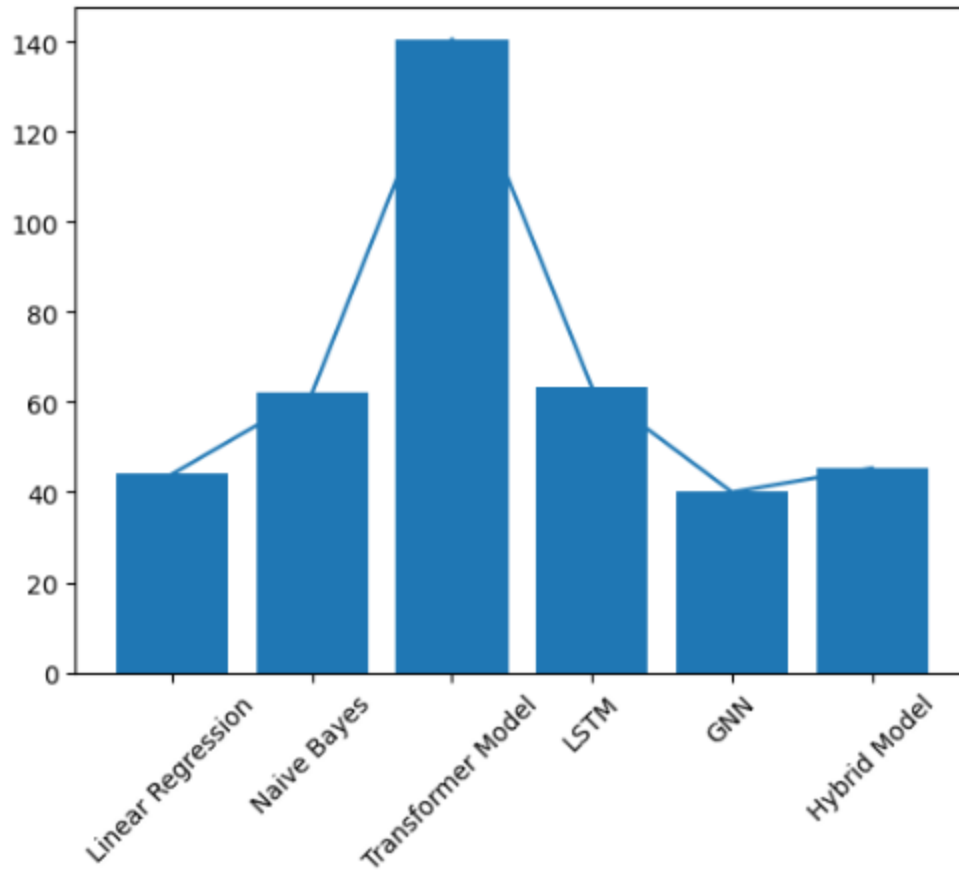


Figure 4.3: RMSE of all models

The figure above show RMSE of all models, here transformer has highest RMSE and GNN has lowest RMSE.

Chapter 5 Conclusion and Future Scope

In conclusion, our exhaustive examination of air quality index (AQI) prediction models provides a nuanced understanding of their capabilities and limitations, offering valuable insights for environmental forecasting and decision-making. The Transformer Model, while versatile across domains, encounters significant challenges when applied to AQI prediction. The pronounced Mean Squared Error (MSE) of 19766.74 and Root Mean Squared Error (RMSE) of 140.59 underscore its struggle to capture the intricate dynamics inherent in AQI data. This highlights the limitations of the Transformer architecture in modeling the complex relationships among environmental factors influencing air quality.

Conversely, Graph Neural Networks (GNN) emerge as a beacon of promise in AQI prediction. With a markedly lower MSE of 1601.86 and an RMSE of 40.02, GNNs demonstrate a notable ability to harness graph structures and relationships, potentially capturing the interconnected nature of environmental variables. This emphasizes the importance of accounting for spatial and temporal dependencies within AQI data, elements that traditional models may overlook.

Linear Regression and Hybrid Models occupy intermediary positions, finding a delicate balance between simplicity and predictive accuracy. Linear Regression, a conventional method, yields an MSE of 1937.55 and an RMSE of 44.02, while the Hybrid Model, amalgamating diverse techniques, achieves an MSE of 2057.53 and an RMSE of 45.36. These models showcase their utility in scenarios where interpretability is valued alongside reasonable predictive performance.

In stark contrast, traditional approaches like Naive Bayes and Long Short-Term Memory (LSTM) networks exhibit higher prediction errors. Naive Bayes, celebrated for its simplicity and efficiency, presents an MSE of 3840.81 and an RMSE of 61.97, revealing a limited ability to

discern nuanced patterns in AQI data. Similarly, LSTM, designed to handle temporal dependencies, grapples with the intricacies of AQI dynamics, yielding an MSE of 3997.25 and an RMSE of 63.22.

These findings underscore the critical importance of selecting models tailored to the unique characteristics of environmental datasets. GNNs, with their inherent capacity to capture spatial dependencies, emerge as a promising avenue for improving AQI predictions. The choice of an optimal model should align with specific application requirements, considering factors such as interpretability, computational efficiency, and the nature of available data. As environmental challenges intensify, refining AQI prediction models remains paramount for effective mitigation strategies and informed policy formulation.

In essence, this comparative analysis not only enriches the discourse on AQI prediction but also serves as a compass guiding researchers and policymakers towards judicious model selection, thereby fostering advancements in environmental science and public health.

5.1 Future Scope

The future scope of air quality prediction models lies in the continuous refinement and advancement of existing methodologies. Integrating cutting-edge techniques, such as incorporating more sophisticated graph neural network architectures and exploring novel hybrid models, could further enhance the accuracy and interpretability of predictions. Additionally, leveraging advancements in data assimilation techniques and integrating real-time sensor data could provide more timely and precise predictions, contributing to proactive environmental management.

The integration of emerging technologies like Internet of Things (IoT) devices and satellite imagery presents an exciting avenue for expanding the spatial and temporal coverage of air quality monitoring. Furthermore, collaboration between interdisciplinary research fields, including meteorology, environmental science, and machine learning, holds promise for developing comprehensive models that consider a broader range of influencing factors. The future also entails addressing data challenges, ensuring the availability of high-quality and diverse datasets for model training and validation. Ultimately, ongoing research endeavors should focus on creating robust, scalable, and adaptable models to meet the evolving demands of air quality prediction in the face of dynamic environmental conditions.

References

- Chen, L., Xu, J., Wu, B., Qian, Y., Du, Z., Li, Y. and Zhang, Y., 2021. Group-aware graph neural network for nationwide city air quality forecasting. arXiv preprint arXiv:2108.12238.
- Chen, X., Hu, Y., Dong, F., Chen, K. and Xia, H., 2023. A Multi-graph Spatial-temporal Attention Network for Air-quality Prediction. *Process Safety and Environmental Protection*.
- Chen, Y., Liang, C., Liu, D., Niu, Q., Miao, X., Dong, G., Li, L., Liao, S., Ni, X. and Huang, X., 2023. Embedding-graph-neural-network for transient NO_x emissions prediction. *Energies*, 16(1), p.3.
- Deng, C., Liu, L., Wang, C. and Chen, Z., 2023, May. Air Quality Prediction Based on Graph Attention Network. In *2023 4th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)* (pp. 364-369). IEEE.
- Feng, H. and Zhang, X., 2023. A novel encoder-decoder model based on Autoformer for air quality index prediction. *Plos one*, 18(4), p.e0284293.

Feng, Y., Kim, J.S., Yu, J.W., Ri, K.C., Yun, S.J., Han, I.N., Qi, Z. and Wang, X., 2023. Spatiotemporal informer: A new approach based on spatiotemporal embedding and attention for air quality forecasting. *Environmental Pollution*, 336, p.122402.

Ghobadi, F. and Kang, D., 2022. Improving long-term streamflow prediction in a poorly gauged basin using geo-spatiotemporal mesoscale data and attention-based deep learning: A comparative study. *Journal of Hydrology*, 615, p.128608.

Han, J., Liu, H., Zhu, H., and Xiong, H., 2023. Kill Two Birds with One Stone: A Multi-View Multi-Adversarial Learning Approach for Joint Air Quality and Weather Prediction. *IEEE Transactions on Knowledge and Data Engineering*.

Huang, W., Li, T., Liu, J., Xie, P., Du, S. and Teng, F., 2021. An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability. *Information Fusion*, 75, pp.28-40.

Irwan, D., Ali, M., Ahmed, A.N., Jacky, G., Nurhakim, A., Ping Han, M.C., AlDahoul, N. and El-Shafie, A., 2023. Predicting Water Quality with Artificial Intelligence: A Review of Methods and Applications. *Archives of Computational Methods in Engineering*, pp.1-20.

Li, P., Zhang, T. and Jin, Y., 2023. A Spatio-Temporal Graph Convolutional Network for Air Quality Prediction. *Sustainability*, 15(9), p.7624.

Liao, H., Yuan, L., Wu, M. and Chen, H., 2023. Air quality prediction by integrating mechanism model and machine learning model. *Science of The Total Environment*, 899, p.165646.

Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S. and Duan, C., 2021. Statistical approaches for forecasting primary air pollutants: a review. *Atmosphere*, 12(6), p.686.

Limperis, J., Tong, W., Hamza-Lup, F. and Li, L., 2023. PM 2.5 forecasting based on transformer neural network and data embedding. *Earth Science Informatics*, pp.1-14.

Ma, Z., Luo, W., Jiang, J., Wang, B., Ma, Z., Lin, J. and Liu, D., 2023. Spatial and temporal characteristics analysis and prediction model of PM2.5 concentration based on SpatioTemporal-Informer model. *Plos one*, 18(6), p.e0287423.

Mengara Mengara, A.G., Park, E., Jang, J. and Yoo, Y., 2022. Attention-based distributed deep learning model for air quality forecasting. *Sustainability*, 14(6), p.3269.

Mitreska Jovanovska, E., Batz, V., Lameski, P., Zdravevski, E., Herzog, M.A. and Trajkovik, V., 2023. Methods for Urban Air Pollution Measurement and Forecasting: Challenges, Opportunities, and Solutions. *Atmosphere*, 14(9), p.1441.

Naz, F., McCann, C., Fahim, M., Cao, T.V., Hunter, R., Viet, N.T., Nguyen, L.D. and Duong, T.Q., 2023. Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas. *IEEE Access*.

Oliveira Santos, V., Costa Rocha, P.A., Thé, J.V.G. and Gharabaghi, B., 2023. Graph-Based Deep Learning Model for Forecasting Chloride Concentration in Urban Streams to Protect Salt-Vulnerable Areas. *Environments*, 10(9), p.157.

Ragab, M.G., Abdulkadir, S.J., Aziz, N., Al-Tashi, Q., Alyousifi, Y., Alhussian, H. and Alqushaibi, A., 2020. A novel one-dimensional CNN with exponential adaptive gradients for air pollution index prediction. *Sustainability*, 12(23), p.10090.

Schulte, N., Li, X., Ghosh, J.K., Fine, P.M. and Epstein, S.A., 2020. Responsive, high-resolution air quality index mapping using model, regulatory monitor, and sensor data in real-time. *Environmental Research Letters*, 15(10), p.1040a7.

Singh, A., Kumar, R. and Hasteer, N., 2020, October. Comparative Analysis of Classification Models for Predicting Quality of Air. In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) (pp. 7-11). IEEE.

Tariq, S., Tariq, S., Kim, S., Woo, S.S. and Yoo, C., 2023. Distance adaptive graph convolutional gated network-based smart air quality monitoring and health risk prediction in sensor-devoid urban areas. *Sustainable Cities and Society*, 91, p.104445.

Wang, W., An, X., Li, Q., Geng, Y.A., Yu, H. and Zhou, X., 2022. Optimization research on air quality numerical model forecasting effects based on deep learning methods. *Atmospheric research*, 271, p.106082.

Wang, Z., Yang, Y. and Yue, S., 2022. Air quality classification and measurement based on double output vision transformer. *IEEE Internet of Things Journal*, 9(21), pp.20975-20984.

Yu, M., Masrur, A. and Blaszczak-Boxe, C., 2023. Predicting hourly PM_{2.5} concentrations in wildfire-prone areas using a SpatioTemporal Transformer model. *Science of The Total Environment*, 860, p.160446.

Zhang, H., Srinivasan, R., Yang, X., Ahrentzen, S., Coker, E.S. and Alwisy, A., 2022. Factors influencing indoor air pollution in buildings using PCA-LMBP neural network: a case study of a university campus. *Building and Environment*, 225, p.109643.

Zhang, Z., Zhang, S., Zhao, X., Chen, L. and Yao, J., 2022. Temporal difference-based graph transformer networks for air quality PM_{2.5} predictions: a case study in China. *Frontiers in Environmental Science*, 10, p.924986.

Zhang, Z. and Zhang, S., 2023. Modeling air quality PM_{2.5} forecasting using deep sparse attention-based transformer networks. *International Journal of Environmental Science and Technology*, pp.1-16.

Zhao, P., Li, C., Rahaman, M.M., Xu, H., Yang, H., Sun, H., Jiang, T. and Grzegorzec, M., 2022. A comparative study of deep learning classification methods on a small environmental microorganism image dataset (EMDS-6): from convolutional neural networks to visual transformers. *Frontiers in Microbiology*, 13, p.792166.

