Decennial Analysis of Air Quality Indices Employing Machine Learning and Topological Data Analysis Clustering Techniques

Dr Umme Salma, Raghava Vigneswar V, and A S Shakthi Aswin

CHRIST(Deemed to be University) https://christuniversity.in/

Abstract. The imperative of preserving air quality amidst rapid industrialization and urbanization, notably in developing nations like India, is paramount. This research leverages six years of air pollution data from 26 cities, employing advanced machine learning and topological data analysis (TDA) techniques to analyze and predict air quality indices. Through preprocessing and correlation analysis, key features impacting air quality are identified, emphasizing the discovery of hidden patterns within the dataset. The study observes a noteworthy decline in pollutant levels during the pandemic year of 2020. Additionally, advanced methodologies, including TDA, cluster similar cities based on pollution profiles, while classification tasks using the LightGBM algorithm and XGBoost regression analysis exhibit high predictive accuracy. This integrated approach offers a comprehensive framework for understanding and predicting air quality dynamics, crucial for informed environmental management and policy formulation.

Keywords: Air quality, pollution, machine learning, topological data analysis, LightGBM, XGBoost.

1 Introduction

Air pollution presents a pressing challenge in densely populated urban areas, especially amid India's rapid industrialization and urbanization. Our study is driven by the need to better understand air pollution dynamics across Indian cities. We aim to cluster cities based on pollution profiles, develop predictive models using advanced machine learning algorithms, and assess the efficacy of integrating topological data analysis (TDA) techniques. Our objectives include employing persistence homology and Betti Curve analysis for clustering, utilizing boosting methods like XGBoost and CATBoost for prediction, and assessing the methodology's accuracy in predicting air quality indices.

Our methodology integrates cutting-edge techniques, including TDA and machine learning. We extract topological features using persistence homology and Betti Curve analysis for clustering, refining results with hierarchical clustering. Predictive models employ boosting methods like XGBoost and data preprocessing techniques such as K-nearest neighbors (KNN) imputation. We leverage

daily air quality data from Kaggle, sourced from the Center for Pollution Control Board, India, comprising twelve independent variables and the Air Quality Index as the target variable.

By combining TDA and machine learning, our research aims to provide actionable insights for policymakers and stakeholders to mitigate air pollution in Indian urban centers. This interdisciplinary approach seeks to deepen our understanding of air pollution dynamics, facilitating precise interventions for environmental sustainability.

The proposed work follows a structured approach following the introduction in section 1. The subsequent sections aim to provide a comprehensive exploration of the research topic. In section 2, a thorough literature review is conducted, delving into recent articles to elucidate the importance of the study and pinpoint existing gaps in the current body of knowledge. Following this, section 3 outlines the prerequisites necessary for understanding and engaging with the study's methodologies. Section 4 focuses on data collection and preparation processes, ensuring the robustness and reliability of the dataset under analysis. Next, section 5 shifts the spotlight to the application of Topological Data Analysis, a cutting-edge methodology employed in the study. This is followed by section 6, dedicated to Exploratory Data Analysis, which provides insights into the initial patterns and trends within the dataset. Section 7 delves into the modeling phase, employing various statistical and machine learning techniques to analyze the data and derive meaningful conclusions. The ensuing section, 8, presents the results obtained from the analyses conducted, followed by a comprehensive discussion in section 9, which interprets the findings in the context of the research objectives. Finally, section 10 concludes the study by summarizing key findings, implications, and avenues for future research.

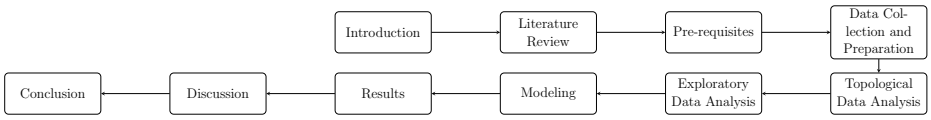

Fig. 1. Flowchart of Proposed Work

2 Literature Review

The literature review delves into recent research endeavors focused on using machine learning techniques to analyze and predict air pollution. It starts with a study on "Air Quality Index prediction using machine learning for Ahmedabad City" [11], which explores localized forecasting methods like SARIMA, SVM, and LSTM, showing potential for tailored predictive modeling to tackle urban air pollution. Then, "Deep learning-based air pollution analysis on carbon monoxide in

Taiwan" [23] discusses advanced deep learning architectures like SGRU for comprehensive carbon monoxide analysis, enhancing pollution assessment methods.

Next, the paper "Comparison of Machine Learning Algorithms for Air Pollution Monitoring System" [19] conducts a comparative analysis of regression techniques, emphasizing the need for understanding predictive models' strengths and limitations in pollution monitoring. "Applying machine learning techniques in air quality prediction" [6] focuses on forecasting particulate matter concentrations, showcasing models like Facebook Prophet's resilience amidst data limitations, underlining accurate forecasting's importance for pollution management.

A pivotal study, "Ensembled Method for Air Quality Monitoring and Control Using Machine Learning" [10] emphasizes accurate prediction models' critical importance in addressing air pollution challenges. These studies collectively contribute to developing robust monitoring systems and effective pollution control measures, enhancing environmental stewardship and public health protection.

"Detection and Prediction of Air Pollution Using Machine Learning Models" [1] underscores high accuracy in pollution detection through machine learning algorithms, paving the way for early warnings and efficient data analysis in environmental monitoring. "Estimation of Air Pollution in Delhi using Machine Learning Techniques" [21] focuses on urban air quality prediction, showcasing Support Vector Regression and Artificial Neural Networks' promise for accurate forecasting and effective environmental management.

"Multi-directional Temporal Convolutional Artificial Neural Network (MT-CAN) for Long-term PM2.5 Forecasting with Missing Values" [18] introduces a novel approach to address missing data and forecasting accuracy challenges in air pollution modeling, demonstrating superior accuracy compared to traditional methods.

The base paper "Air pollution prediction with machine learning: a case study of Indian cities" [8] meticulously examines air quality data across Indian cities, employing various machine learning techniques and ultimately identifying XGBoost as a superior model. This finding highlights machine learning's transformative potential in advancing air quality monitoring and analysis. Building upon this foundation, our research incorporates Topological Data Analysis (TDA) for clustering and advanced machine learning models for classification and regression tasks. By integrating TDA with state-of-the-art machine learning techniques, we aim to enhance the predictive accuracy and interpretability of our models, thereby contributing to more effective air quality management strategies and environmental stewardship efforts.

3 Pre-requisites

This section provides preliminary information regarding various concepts used in the article. It also highlights some important terms that are to be known for the better understanding of the proposed work. The foundational knowledge of machine learning models presented in this article is drawn from the book "Introduction to Machine Learning", authored by Ethem Alpaydin.[2]

Air Quality Index (AQI) - The AQI is a metric that defines the quality of the air. It serves as a vital tool for assessing and communicating the quality of the air in a specific location at a given time. It consolidates data on various pollutants like particulate matter, ozone, sulfur dioxide, nitrogen dioxide, and carbon monoxide into a single numerical value. The AQI typically ranges from 0 to 500, with lower values indicating better air quality and higher values signaling increased pollution levels. Different AQI categories, such as "Good", "Moderate", "Unhealthy", and "Hazardous", correlate with different levels of health concern, helping individuals and authorities understand and respond to potential risks to public health and the environment.

KNN Imputer, - It is a machine learning technique used for imputing missing values in datasets. It works by replacing missing data points with the average of their nearest neighbors' values, where "nearest" is determined by a distance metric such as Euclidean distance. This approach leverages the principle that similar data points tend to have similar values, effectively filling in the gaps in the dataset while preserving its underlying structure and relationships.

Topological Data Analysis (TDA) - It is a field of data analysis that focuses on extracting and understanding the underlying topological structure of datasets. TDA techniques aim to capture geometric and qualitative features of data, such as connectivity, shape, and clustering patterns, which may not be evident from traditional statistical methods. One prominent TDA method is the Mapper Algorithm, which constructs a topological graph representation of the dataset by partitioning it into overlapping regions and then summarizing each region with a concise descriptor. This approach helps reveal global and local patterns in the data, facilitating interpretation and decision-making in various domains. Fundamental concepts of TDA are given in "Introductory Topological Data Analysis" by Dayten Sheffar [20] and TDA techniques have been applied in diverse applications. For instance, in the study of the aggregation process of β-Amyloid peptides, TDA generated a significant amount of data points, providing insights into the aggregation process. [5] Additionally, TDA has been utilized in the analysis of network monitoring data, such as in the case of Darknet analysis, showcasing its versatility in handling complex high-dimensional data .[12]

Persistent Homology - It is a mathematical tool used in TDA to analyze the evolution of topological features across different scales or resolutions. It characterizes the presence and lifespan of topological features, such as connected com-

ponents, holes, and loops, in a dataset by examining how they persist through various levels of granularity. Persistent homology provides a robust framework for capturing and quantifying the intrinsic structure of complex datasets, enabling the identification of meaningful patterns and relationships that may be obscured by noise or dimensionality.

Hierarchical Clustering Dendrogram - It is a visualization technique commonly used in hierarchical clustering, a method for grouping similar data points into clusters based on their pairwise distances. In hierarchical clustering, the data points are progressively merged into clusters, forming a hierarchical structure or dendrogram that illustrates the relationships between clusters and their subclusters. The height of each node in the dendrogram represents the distance at which clusters are merged, with shorter distances indicating greater similarity between clusters. Hierarchical clustering dendrograms provide insight into the organization and hierarchy of clusters within a dataset, aiding in the interpretation and exploration of complex data structures.

Gaussian Naive Bayes - It is a simple probabilistic classifier based on Bayes' theorem with the assumption of independence between features. It works by calculating the probability of a given sample belonging to each class based on the probability distribution of features in each class. Despite its simplicity and the "naive" assumption of feature independence, Gaussian Naive Bayes can perform well in many classification tasks, especially when the features are approximately normally distributed and the class separation is clear.

Support Vector Machine (SVM) - It is a powerful supervised learning algorithm used for classification and regression tasks. SVM works by finding the optimal hyperplane that separates data points of different classes in a high-dimensional space. It aims to maximize the margin, or the distance between the hyperplane and the nearest data points of each class, thus promoting robust generalization to unseen data. SVM can handle both linear and nonlinear classification problems through the use of different kernel functions, which map the input data into higher-dimensional feature spaces.

Boosting Techniques - It refer to a family of ensemble learning methods that combine multiple weak learners, typically decision trees, to create a strong learner. Boosting algorithms iteratively train weak learners on subsets of the data, with each subsequent learner focusing more on the instances that the previous ones misclassified. This process allows boosting algorithms to gradually improve the model's performance by emphasizing the difficult-to-classify instances. Popular boosting techniques include AdaBoost, Gradient Boosting Machines (GBM), XGBoost, CatBoost, and LightGBM.

XGBoost (Extreme Gradient Boosting) - It is an efficient and scalable implementation of gradient boosting for classification and regression tasks.

It enhances traditional gradient boosting with several optimizations, such as parallelization, regularization, and handling missing values. XGBoost builds an ensemble of decision trees sequentially, with each tree learning from the residuals of the previous ones. It employs a novel split-finding algorithm to efficiently search for the best split points, leading to faster training and higher accuracy compared to traditional gradient boosting methods.

CatBoost - It is another gradient boosting library designed for high performance and accuracy, particularly in handling categorical features. CatBoost incorporates several advanced techniques, including ordered boosting, oblivious trees, and symmetric trees, to effectively handle categorical variables without the need for pre-processing or one-hot encoding. By optimizing the handling of categorical features and employing regularization strategies, CatBoost can achieve competitive performance while mitigating overfitting.

LightGBM - It is a gradient boosting framework developed by Microsoft that focuses on efficiency, scalability, and high performance. It introduces novel techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce training time and memory usage while improving accuracy. LightGBM employs a leaf-wise tree growth strategy and histogram-based algorithms to speed up training without sacrificing model quality, making it particularly well-suited for large-scale datasets and resource-constrained environments.

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) - They are specialized types of recurrent neural networks (RNNs) designed to address the vanishing gradient problem and capture long-range dependencies in sequential data. LSTM and GRU architectures include gated mechanisms, such as forget gates, input gates, and output gates, which regulate the flow of information within the network over time. These gates enable LSTM and GRU models to retain relevant information over long sequences while selectively updating and forgetting less important information, making them suitable for tasks like natural language processing, time series prediction, and speech recognition.

4 Data Collection

The dataset utilized in this study was sourced from Kaggle, specifically obtained through data extraction from the CPCB.in website. Originating from the Central Pollution Control Board (CPCB) of India, this dataset represents a comprehensive compilation of air quality data spanning the years 2015 to 2020. As the principal regulatory authority overseeing pollution-related matters nationwide, the CPCB operates under the purview of the Ministry of Environment, Forest

and Climate Change, Government of India.

Table 1. Variables Table

Variable Name	Role	Type	Description	Missing Values
User	Feature	Categorical	Identifier of the user	0
City	Feature	Categorical	City of observation	0
Date	Feature	Date	Timestamp of data collection	0
PM2.5	Feature	Numerical	Particulate Matter 2.5	4598
PM10	Feature	Numerical	Particulate Matter 10	11140
NO	Feature	Numerical	Nitric Oxide	3582
NO_2	Feature	Numerical	Nitrogen Dioxide	3585
NOx	Feature	Numerical	Nitrogen Oxides	4185
NH_3	Feature	Numerical	Ammonia	10328
CO	Feature	Numerical	Carbon Monoxide	2059
SO_2	Feature	Numerical	Sulfur Dioxide	3854
O_3	Feature	Numerical	Ozone	4022
Benzene	Feature	Numerical	Benzene	5623
Toluene	Feature	Numerical	Toluene	8041
Xylene	Feature	Numerical	Xylene	18109
AQI	Target	Numerical	Air Quality Index	4681
AQI_Bucket	Target	Categorical	AQI Bucket	4681

Titled "Air Quality Data in India (2015 - 2020)", this dataset encompasses a diverse array of features, furnishing detailed insights into air quality metrics observed across various cities in India. Among its attributes are timestamps denoting data collection instances alongside the corresponding city of observation. Additionally, the dataset encompasses measurements of numerous pollutants, including $PM_{2.5}$, NO_2, NH_3, CO, O_3, NO_x, PM_{10}, Toluene, Benzene, Xylene, and SO_2. Furthermore, it provides the Air Quality Index (AQI), a numerical indicator representing the overall air quality status at specific locations and times, derived from pollutant concentrations. The AQI is further stratified into distinct categories delineating the severity of air pollution, ranging from Good to Severe. This dataset offers a comprehensive repository of air quality metrics, facilitating the analysis and prediction of AQI levels. The inclusion of a diverse range of pollutants enables a multifaceted assessment of air quality, crucial for understanding the environmental and health implications of pollution.

5 Data Preparation

In this section, we outline the procedures employed to preprocess the dataset in preparation for subsequent analysis and modeling endeavors. This preparatory phase encompasses the mitigation of missing values, rectification of data skewness, and adaptation of target variables for regression and classification tasks.

Initially, the dataset presented instances of missing values, predominantly within the AQI (Air Quality Index) feature, with a total of 4681 missing values identified. To uphold data integrity and ensure the robustness of ensuing analyses, the missing values within the AQI feature were removed from the dataset.

Furthermore, to address missing values within other features, a K-Nearest Neighbors (KNN) imputation technique was utilized. This imputation approach involved estimating missing values based on AQI values associated with features such as PM2.5, PM10, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, and Xylene. By leveraging related AQI data for imputation, the dataset underwent augmentation, thereby facilitating subsequent analyses and modeling endeavors.

Upon scrutiny, it was observed that all features in the dataset exhibited positive skewness, deviating from a normal distribution. To address this skewness and achieve symmetrical distributions, Box-Cox Transformation was applied to all features, including PM2.5, PM10, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, and Xylene. This transformation aids in stabilizing variance and enhancing distribution symmetry.

Additionally, to prepare the dataset for regression tasks involving the AQI feature, Min-Max scaling was employed to normalize values within a specified range, thereby facilitating the regression modeling process. Similarly, for the AQI_BUCKET feature intended for classification, Label Encoding was performed to convert categorical labels into numerical representations, ensuring compatibility with classification algorithms.

By addressing missing values and mitigating data skewness through transformation techniques, the dataset is now primed for subsequent analysis and modeling tasks. These preprocessing steps significantly enhance the dataset's quality and reliability, establishing a robust foundation for predictive modeling and analysis.

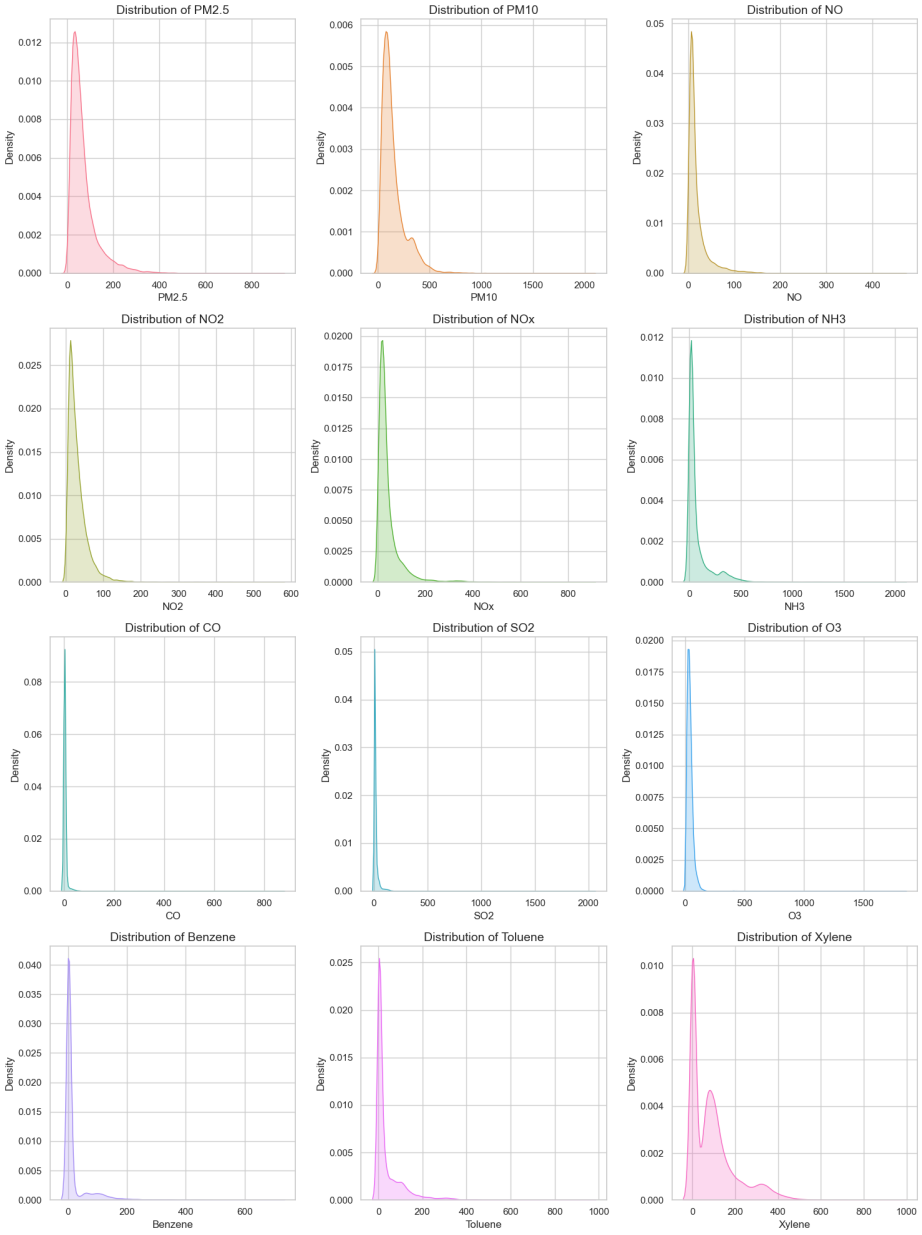

Fig. 2. Plots to understand the Distribution

6 Topological Data Analysis

Topological Data Analysis (TDA) emerges as a potent approach for dissecting intricate datasets, harnessing principles from algebraic topology to derive meaningful insights. Within this chapter, we delve into two pivotal techniques under TDA: Mapper and Betti curves.

6.1 Mapper Algorithm

Mapper serves as a method to streamline the intricate topological and geometric intricacies inherent in datasets. In our analysis, we specifically applied Mapper to scrutinize the AQI (Air Quality Index) feature, employing distinct parameters, including `clusterer=sklearn.cluster.DBSCAN(eps=0.1, min_samples=100)` and `cover=km.Cover(35, 0.4)`, to instantiate Mapper. Consequently, we derived 11 nodes delineating distinct clusters within the dataset.

Fig. 3. Cluster

The Mapper Diagram, as referenced in 3, elucidates the topological configuration of the data, with nodes symbolizing diverse clusters or cohorts. The highlighted node signifies the "Severe" class, accentuating its prominence within the dataset. A cursor positioned on the yellow node at the terminus of the graph accentuates its pertinence within the analysis.

This graphical depiction embodies a hierarchical clustering scheme of data points, each correlating to specific periods and associated measurements or ob-

servations. It furnishes several pivotal insights: firstly, it unveils temporal clustering, spotlighting cohorts of periods sharing akin characteristics or patterns within the dataset. Secondly, the hierarchical arrangement unveils relationships and disparities within the data, with more extensive clusters undergoing further subdivision into smaller entities. Thirdly, node dimensions reflect the relative significance or prevalence of particular periods or patterns. Additionally, inter-cluster distances denote the degree of similarity or dissimilarity between periods, facilitating outlier detection or pattern discernment. Finally, color-coded nodes may denote diverse attributes or traits linked with the periods, augmenting data interpretability. Overall, this graphical rendition proffers invaluable insights into temporal patterns and relationships within the dataset, fostering trend identification, anomaly detection, and comprehension of underlying structures.

The highlighted node in 3 delineates the dispersion of data points across assorted cities, particularly accentuating those aligning with the severe class of air quality. Among the enumerated cities, Ahmedabad emerges with the highest count of 73 data points, trailed by Delhi with 12, Gurugram with 7, Lucknow with 4, Patna with 3, Hyderabad with 1, and Jorapokhar with 1. This distribution underscores the disproportionate concentration of severe air quality occurrences within select urban locales. The prevalence of severe air quality conditions, notably in cities like Ahmedabad and Delhi, underscores the imperative nature of addressing air pollution in densely inhabited metropolitan areas. These revelations advocate for targeted interventions and policy formulations aimed at ameliorating air pollution levels, bolstering public health, and advocating for sustainable urban development practices. Moreover, such insights can guide stakeholders and policymakers in prioritizing resources and devising efficacious strategies to combat the root causes of air pollution, thereby enhancing overall air quality standards within these urban vicinities.

6.2 Persistent Homology and Clustering using Betti Curve

Betti curves play a crucial role in comprehending the topological characteristics embedded within a dataset. By generating persistence homology diagrams for each city, we extracted their respective Betti curves, as depicted in 4. Further analysis was conducted by clustering the cities through hierarchical clustering, yielding the cluster representation showcased in 6. Wasserstein distances between the Betti curves of individual cities served as a metric for gauging dissimilarity.

The clustering process, informed by Betti curves analysis, facilitated the categorization of cities into three distinctive clusters, each offering valuable insights into the underlying structures of air pollution patterns. Cluster 1 encompassed cities such as Ahmedabad, Aizawl, and Chennai, hinting at potential similarities in the distribution and concentration of pollutants, likely stemming from shared pollution sources or environmental contexts.

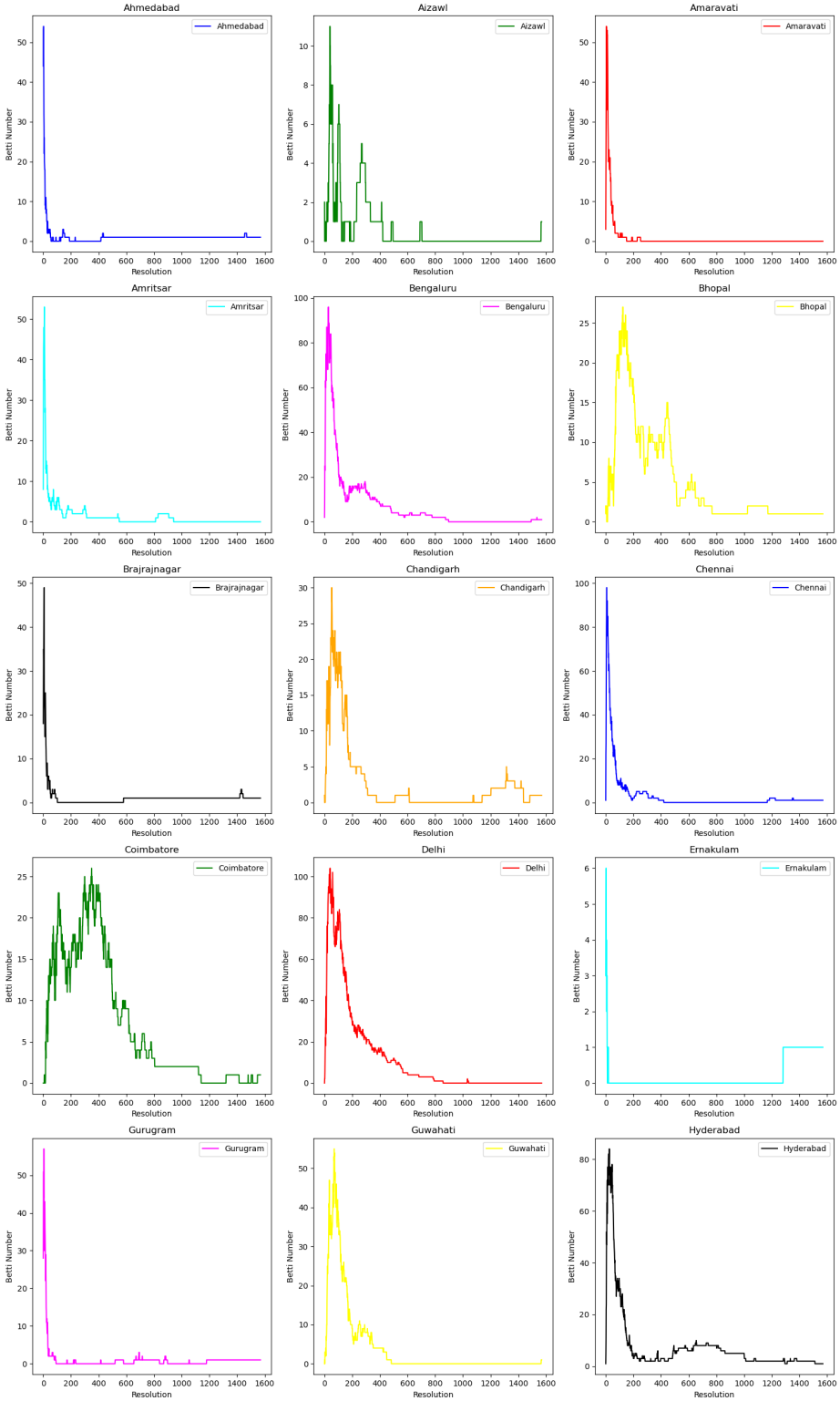

Fig. 4. Betti Curve for different cities

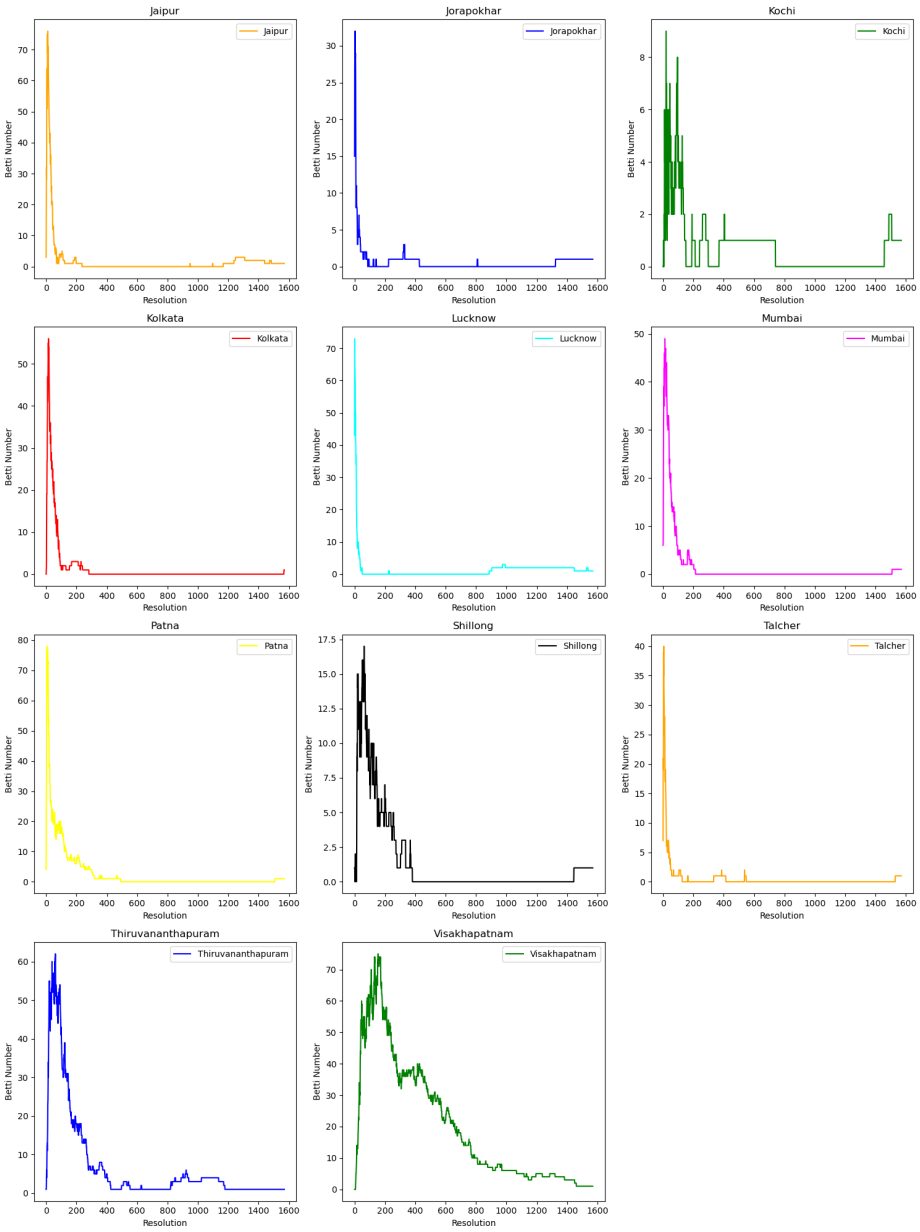

Fig. 5. Betti Curve for different cities

Table 2. Clusters Identified Through Betti Curves Analysis

Cluster	**Cities**
Cluster 1	Ahmedabad, Aizawl, Amaravati, Amritsar, Bhopal, Brajrajnagar, Chandigarh, Chennai, Guwahati, Ernakulam, Gurugram, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher
Cluster 2	Visakhapatnam
Cluster 3	Bengaluru, Delhi, Coimbatore, Hyderabad, Thiruvananthapuram

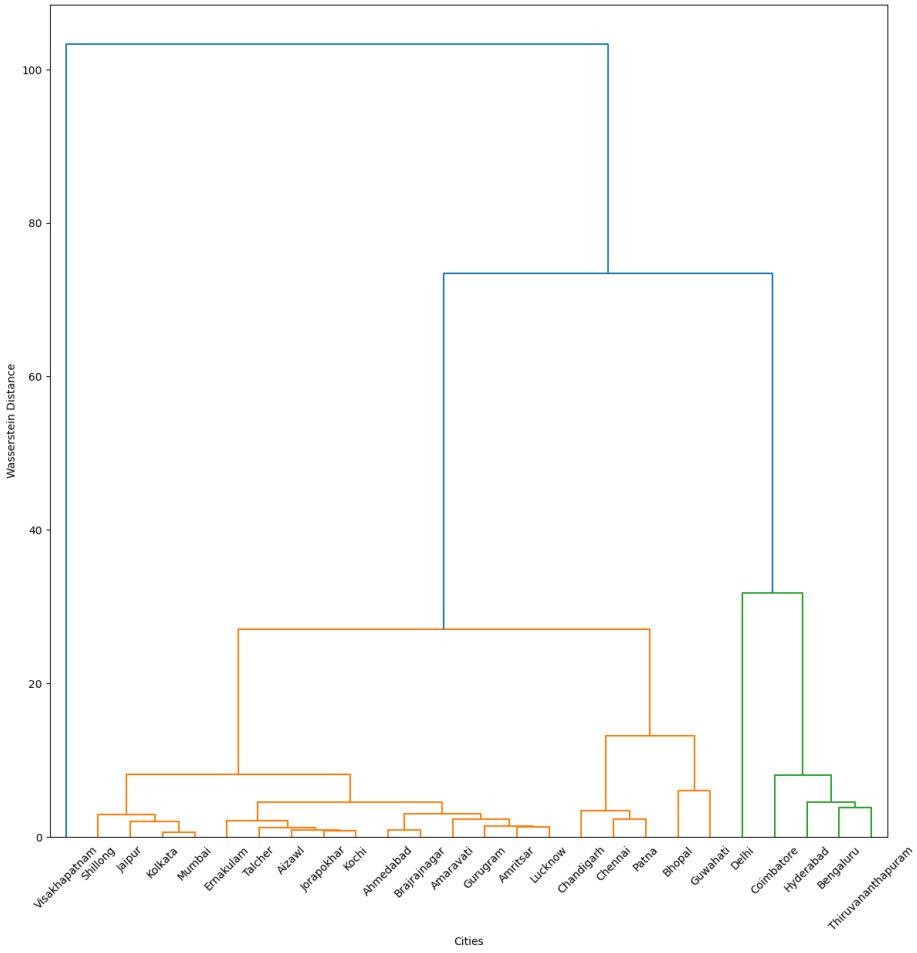

Fig. 6. Clustering according to Wasserstein Distances

This suggests avenues for collaborative endeavors aimed at addressing common air quality challenges prevalent among these cities. Conversely, Cluster 2

featured Visakhapatnam in isolation, highlighting its distinct air pollution profile attributed to unique local circumstances or pollution origins. Meanwhile, Cluster 3, comprising cities like Bengaluru and Delhi, showcased congruence's in pollutant distribution and concentration, potentially influenced by akin regulatory frameworks or pollution mitigation measures.

These observations underscore the significance of tailoring air quality management strategies to encompass both shared attributes and divergences among cities, acknowledging nuances extending beyond mere topological structures. While implementing uniform policies for cities with analogous air pollution profiles may yield benefits, it is imperative to acknowledge that such resemblances do not necessarily guarantee uniformity in actual pollutant levels or composition. Policymakers must take into account additional factors such as geographical, demographic, and industrial disparities to devise tailored interventions that effectively address the multifaceted complexities of air pollution.

Leveraging Mapper and Betti curves analyses equips policymakers with deeper insights into the dynamics of air pollution, fostering more informed decision-making processes and enabling the implementation of interventions aimed at enhancing air quality and safeguarding public health.

7 Exploratory Data Analysis

In this chapter, we undertake exploratory data analysis to glean insights from the dataset and comprehend the interrelationships among variables. Outliers present a potential to distort statistical analyses and modeling procedures, inducing skewness in data distribution and compromising parameter estimation accuracy. Their existence may lead to erroneous interpretations and diminish the predictive models' efficacy. Utilizing Boxplot analysis, as depicted in **??**, we pinpoint outliers, acknowledging their considerable impact on analysis and modeling endeavors. To ensure the integrity and dependability of our analyses, these outliers are substituted with the third quartile values of the respective features, thereby mitigating their influence on subsequent analyses.

The significance of correlation resides in its capacity to quantify the magnitude and direction of associations among various variables within a dataset. Through the computation of a correlation matrix, as illustrated in 8, we garner insights into the relationships among variables, essential for discerning patterns and making well-informed decisions in data analysis. Specifically, in our investigation of air quality data, we scrutinize the correlation of the Air Quality Index (AQI) with diverse pollutants such as PM2.5, PM10, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, and Xylene.

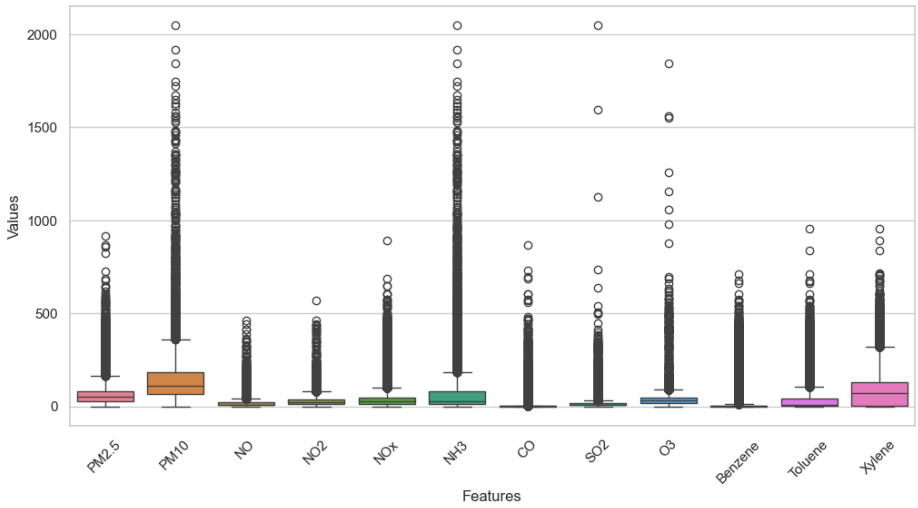

Fig. 7. Box Plot

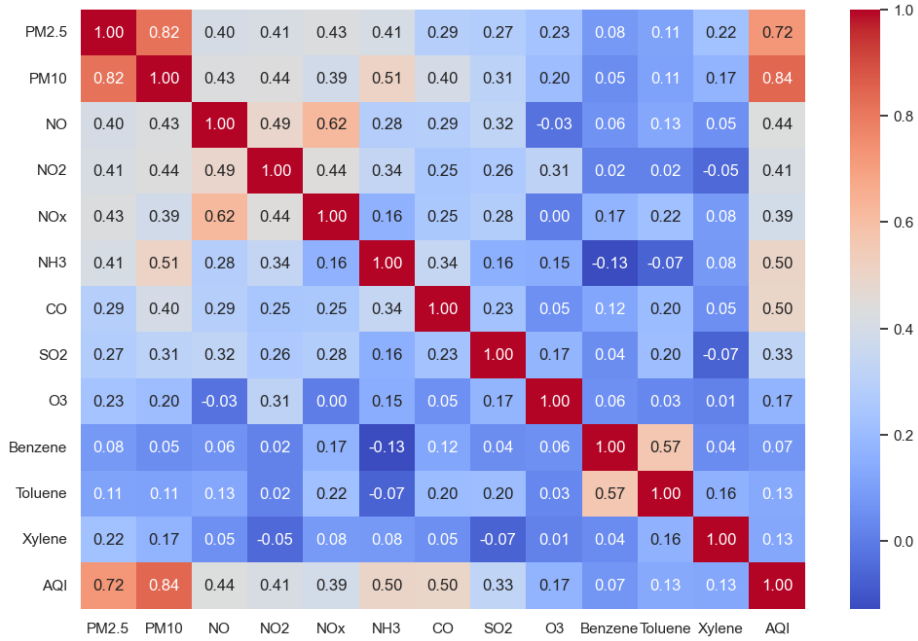

Fig. 8. Correlation Matrix

7.1 Time Series Analysis

Time series analysis of the Air Quality Index (AQI) serves as a pivotal tool in comprehending the temporal dynamics of air pollution levels, aiding in the identification of patterns and trends over time. Through this analytical approach, we can discern how AQI values fluctuate across various time intervals, enabling the detection of recurring patterns, trends, and seasonality in air quality. In our investigation, we employed time series analysis techniques to visualize the changes in AQI values over time, aiming to uncover discernible patterns or trends. The insights garnered from this analysis unveiled notable temporal fluctuations in AQI levels.

Specifically, our examination revealed a tendency for higher AQI indices at the onset of each year, with the highest recorded values observed between 2015 and 2016. Conversely, towards the latter part of 2018 and extending into 2019, we observed the lowest AQI levels. Notably, this period coincided with the emergence of the COVID-19 pandemic. This observation underscores the influence of external factors, including environmental policies and significant events like the COVID-19 pandemic, on air quality dynamics.

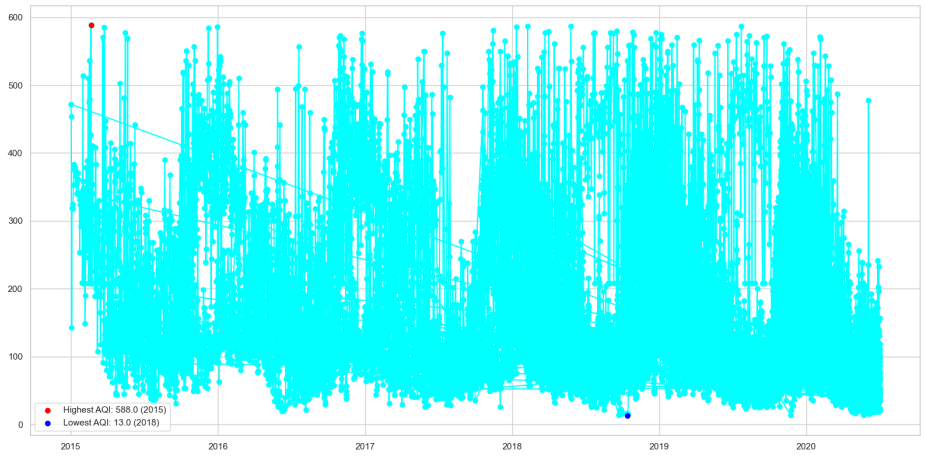

Fig. 9. Time Series plot of Air Quality Index over time

By scrutinizing AQI trends over time, we gain invaluable insights into the temporal patterns of air pollution. Such insights facilitate informed decision-making and the formulation of effective pollution management strategies. Understanding the temporal variations in AQI levels enables policymakers and stakeholders to devise proactive measures aimed at mitigating air pollution and safeguarding public health, especially during critical periods of heightened pollution levels or external disruptions.

8 Modelling

In this chapter, we delve into the intricacies of the modeling techniques employed to address two critical tasks: classifying the AQI_Bucket feature and predicting the AQI index. We recognize that accurate classification and prediction are paramount for effective air quality management and public health intervention strategies.

Acknowledging the class imbalance within the AQI_Bucket feature, we proactively tackled this issue using the Synthetic Minority Over-sampling Technique (SMOTE). By synthesizing new instances for the minority class, SMOTE effectively balanced the class distribution, thereby enhancing the performance of our classifiers. This step was crucial to ensure that our models were not biased towards the majority class, thus improving their ability to accurately classify AQI levels.

Furthermore, to safeguard against overfitting and assess the generalization ability of our models, we employed cross-validation. This robust technique involved partitioning the dataset into multiple subsets, enabling us to iteratively train and evaluate the models on different data partitions. By averaging the performance metrics across these iterations, we obtained reliable estimates of the models' performance and their ability to generalize to unseen data.

The train-test split of the dataset was another pivotal step in our modeling process. By allocating 70% of the data for training and reserving 30% for testing, we ensured a balanced approach to model evaluation. The training set served as the foundation for training our classifiers and regression models, allowing them to learn from the underlying patterns and relationships within the data. Subsequently, the testing set provided an independent dataset for evaluating the models' performance, offering valuable insights into their effectiveness in real-world scenarios.

For the classification task of the AQI_Bucket feature, we employed a diverse set of machine learning algorithms, including XGBoost, CatBoost, LightGBM, Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM). Each algorithm was carefully chosen for its ability to handle classification tasks efficiently and its suitability for handling non-linear relationships between features.

On the other hand, for the regression task of predicting the AQI index, we adopted a combination of machine learning and deep learning algorithms. This included XGBoost, CatBoost, LightGBM, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). These models were selected based on their capacity to capture complex temporal patterns in the data, essential for accurate time-series forecasting.

The features selected for both classification and regression tasks were meticulously chosen based on their known impact on air quality and their availability in the dataset. Features such as PM2.5, PM10, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, and Xylene were deemed crucial for capturing the diverse factors influencing air pollution levels.

By leveraging this comprehensive suite of modeling techniques and features, our aim is to provide robust and accurate analyses of air quality, empowering stakeholders with the insights needed for informed decision-making and effective pollution control measures.

9 Results and Discussion

In our analysis of predictive models, we employ various metrics tailored to the task at hand. For classification tasks, metrics like accuracy, precision, recall, and F1 score are standard, while regression tasks rely on metrics such as mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and R-squared. These metrics serve as critical benchmarks for evaluating the performance of predictive models across different domains.

The classification analysis involved assessing the performance of multiple models—XGBoost, CatBoost, LightGBM, Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM)—using cross-validation and test data. Table 3 presents the performance metrics of these models during cross-validation.

Table 3. Cross-validation results for classification models.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Naive Bayes	0.7072	0.7082	0.7072	0.7072
SVM	0.7404	0.7397	0.7404	0.7394
CatBoost	0.9696	0.9696	0.9696	0.9696
XGBoost	0.9716	0.9716	0.9716	0.9715
LightGBM	**0.9737**	**0.9737**	**0.9737**	**0.9737**

As observed in Table 3, LightGBM demonstrated the highest accuracy, precision, recall, and F1 score during cross-validation, indicating its robust performance in classifying AQI levels. Additionally, XGBoost and CatBoost displayed strong performance, with high accuracy scores.

The test results, summarized in Table 4, further confirmed the superior performance of LightGBM, which maintained the highest accuracy, precision, recall, and F1 score among the classification models.

Table 4. Test results for classification models.

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.6778	0.6103	0.6935	0.6362
SVM	0.7120	0.6573	0.7270	0.6774
XGBoost	0.9408	0.9327	0.9397	0.9361
CatBoost	0.9435	0.9301	0.9412	0.9355
LightGBM	**0.9450**	**0.9356**	**0.9387**	**0.9371**

In summary, the classification models, particularly LightGBM, XGBoost, and CatBoost, exhibited robust performance in accurately classifying AQI levels. However, Gaussian Naive Bayes (GNB) and Support Vector Machine (SVM) showed comparatively lower metrics, suggesting potential areas for improvement.

Moving to the regression analysis, Tables 3 and 4 present the results on both the training and test sets, respectively.

Table 5. Regression Results on Train Set

Model	MAE	MSE	RMSE	R-Squared
GRU	0.01719	0.00099	0.03149	0.94799
LSTM	0.01142	0.00054	0.02324	0.97165
LightGBM	0.01043	0.00037	0.01922	0.98063
CatBoost	0.00876	0.00026	0.01616	0.98629
XGBoost	**0.00655**	**0.00012**	**0.01082**	**0.99386**

Table 6. Regression Results on Test Set

Model	MAE	MSE	RMSE	R-Squared
GRU	0.01723	0.00107	0.03278	0.94439
LightGBM	0.01345	0.00076	0.02760	0.96056
LSTM	0.01248	0.00075	0.02744	0.96103
CatBoost	0.01200	0.00064	0.02522	0.96706
XGBoost	**0.01186**	**0.00080**	**0.02836**	**0.95835**

The regression results indicated that XGBoost consistently outperformed other models, demonstrating superior accuracy in predicting the AQI index.

However, despite the promising performance of the models, several limitations must be acknowledged, including data quality issues, overfitting concerns, and challenges in generalization. Ongoing refinement and validation of the models are essential to ensure their reliability and effectiveness in supporting decision-making processes related to air quality management.

9.1 Limitations

While this study provides valuable insights into air quality prediction using a variety of machine learning and deep learning models, it's important to acknowledge several limitations. Firstly, the accuracy and reliability of predictions are heavily dependent on the quality of the input data. Despite efforts to source air quality data from reputable sources, the dataset contains numerous missing or erroneous values, which could potentially impact the models' performance. Additionally, the selection of features used in the models may affect their predictive capabilities. While specific air pollutant features were chosen for prediction, other relevant factors such as weather conditions or socio-economic variables were not included, which could influence air quality dynamics. Furthermore, the complexity of models like XGBoost, LSTM, CatBoost, GRU, and LightGBM may present challenges in terms of interpretability and computational resources. Despite attempts to address overfitting using techniques like cross-validation and regularization, imputing missing values using KNN with respect to AQI could make the models susceptible to overfitting. Additionally, the generalizability of the models may be limited to specific regions or time periods, given the variability of air quality conditions across different contexts. Moreover, while common evaluation metrics like MAE, MSE, RMSE, and R-squared were employed, they may not fully capture the nuances of air quality prediction. Finally, it's crucial to recognize the probabilistic nature of predictive models and their limitations in accounting for all possible scenarios or external factors. Therefore, continuous refinement and validation of the models, along with careful consideration of their policy implications, are essential to ensure their reliability and effectiveness in supporting decision-making processes related to air quality management.

9.2 Future Scope

This study focused on air quality prediction using machine learning and deep learning models, but there are several areas for future exploration and improvement. Firstly, integrating additional features such as meteorological data and geographical attributes could offer a more comprehensive understanding of air quality dynamics, potentially enhancing the accuracy and robustness of predictive models. Exploring advanced modeling techniques like ensemble methods and hybrid models could help capture complex relationships within air quality data. Temporal and spatial analysis could provide insights into seasonal variations and localized pollution hotspots, guiding targeted intervention strategies. Real-time monitoring and forecasting systems, integrated with IoT devices and remote sensing technologies, can enable proactive measures to mitigate pollution levels. Incorporating uncertainty quantification techniques can provide probabilistic estimates of air quality predictions, supporting decision-making under uncertainty. Additionally, fostering collaborative data-sharing initiatives and citizen science engagement could enrich existing datasets and promote public awareness and advocacy for air quality issues. Overall, future research in this domain holds promise for advancing our understanding of air quality dynamics and improving environmental health outcomes.

10 Conclusion

In this study, we embarked on a thorough exploration of air quality prediction using a combination of machine learning and deep learning methodologies. By analyzing various research articles and datasets, we gained valuable insights into the different approaches and techniques employed in this field. Utilizing citation information and references, we established a strong foundation for our investigation, drawing upon the expertise and findings of numerous researchers. Our exploration covered a wide range of topics, including predictive modeling techniques, evaluation metrics, and data preprocessing. By synthesizing information from diverse sources and methodologies, we developed a nuanced understanding of the complexities and challenges associated with air quality prediction.

From hierarchical clustering analysis to temporal deep learning neural networks, we delved into cutting-edge methodologies used for data analysis and forecasting. Additionally, through the examination of hierarchical clustering graphs and the distribution patterns of air quality data across cities, we uncovered crucial insights into temporal patterns, clustering structures, and regional variations in air quality levels. These findings not only enhance our understanding of air pollution dynamics but also emphasize the importance of implementing targeted interventions and policy measures to address air quality challenges in urban environments.

Through this project, we have laid the groundwork for future research endeavors aimed at advancing our knowledge of air quality prediction and contributing to the development of effective pollution control strategies.

References

[1] Nayana D K Praveen Gandhi Vidyavastu Aditya C R Chandana R Deshmukh. "Detection and Prediction of Air Pollution using Machine Learning Models". In: *International Journal of Engineering Trends and Technology (IJETT)* 59(4) (May 2018), pp. 204–207. ISSN: 2231-5381. DOI: https://doi.org/10.14445/22315381/IJETT-V59P238. URL: www.ijettjournal.org.

[2] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[3] Tania Septi Anggraini et al. "Machine learning-based global air quality index development using remote sensing and ground-based stations". In: *Environmental Advances* 15 (2024), p. 100456. ISSN: 2666-7657. DOI: https://doi.org/10.1016/j.envadv.2023.100456. URL: https://www.sciencedirect.com/science/article/pii/S266676572300114X.

[4] Madhav Badami. "Transport and Urban Air Pollution in India". In: *Environmental management* 36 (Sept. 2005), pp. 195–204. DOI: 10.1007/s00267-004-0106-x.

[5] M. Deleanu et al. "Unraveling the speciation of -amyloid peptides during the aggregation process by taylor dispersion analysis". In: *Analytical Chemistry* 93 (16 2021), pp. 6523–6533. DOI: 10.1021/acs.analchem.1c00527.

[6] Ekaterina Gladkova and Liliya Saychenko. "Applying machine learning techniques in air quality prediction". In: *Transportation Research Procedia* 63 (2022). X International Scientific Siberian Transport Forum — TransSiberia 2022, pp. 1999–2006. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2022.06.222. URL: https://www.sciencedirect.com/science/article/pii/S235214652200477X.

[7] Sarath K. Guttikunda, Rahul Goel, and Pallavi Pant. "Nature of air pollution, emission sources, and management in the Indian cities". In: *Atmospheric Environment* 95 (2014), pp. 501–510. ISSN: 1352-2310. DOI: https://doi.org/10.1016/j.atmosenv.2014.07.006. URL: https://www.sciencedirect.com/science/article/pii/S1352231014005275.

[8] Pande B. P. Kumar K. "Air pollution prediction with machine learning: a case study of Indian cities". In: *International Journal of Environmental Science and Technology* 20 (2023), 5333–5348. ISSN: 1735-2630. DOI: https://doi.org/10.1007/s13762-022-04241-5. URL: https://rdcu.be/dDYcD.

[9] Hyunjung Lee et al. "Air pollution assessment in Seoul, South Korea, using an updated daily air quality index". In: *Atmospheric Pollution Research* 14.4 (2023), p. 101728. ISSN: 1309-1042. DOI: https://doi.org/10.1016/j.apr.2023.101728. URL: https://www.sciencedirect.com/science/article/pii/S130910422300082X.

[10] S John Livingston et al. "An ensembled method for air quality monitoring and control using machine learning". In: *Measurement: Sensors* 30 (2023), p. 100914. ISSN: 2665-9174. DOI: https://doi.org/10.1016/j.measen.2023.100914. URL: https://www.sciencedirect.com/science/article/pii/S2665917423002507.

[11] Nilesh N. Maltare and Safvan Vahora. "Air Quality Index prediction using machine learning for Ahmedabad city". In: *Digital Chemical Engineering* 7 (2023), p. 100093. ISSN: 2772-5081. DOI: https://doi.org/10.1016/j.dche.2023.100093. URL: https://www.sciencedirect.com/science/article/pii/S277250812300011X.

[12] C. Marc, A. Lahmadi, and F. Jérôme. "Topological analysis and visualisation of network monitoring data: darknet case study". In: (2016). DOI: 10.1109/wifs.2016.7823920.

[13] Anamika Pandey et al. "Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019". In: *The Lancet Planetary Health* 5.1 (2021), e25–e38. ISSN: 2542-5196. DOI: https://doi.org/10.1016/S2542-5196(20)30298-9. URL: https://www.sciencedirect.com/science/article/pii/S2542519620302989.

[14] AM Patankar and PL Trivedi. "Monetary burden of health impacts of air pollution in Mumbai, India: implications for public health policy". In:

Public Health 125.3 (2011), pp. 157–164. DOI: 10.1016/j.puhe.2010.11.009.

[15] Gowda Parameshwara Prashanth. "India's air pollution: the need for city-centric plans and regulations". In: *The Lancet Planetary Health* 5.4 (2021), e185. ISSN: 2542-5196. DOI: https://doi.org/10.1016/S2542-5196(21)00032-2. URL: https://www.sciencedirect.com/science/article/pii/S2542519621000322.

[16] Gokulan Ravindiran et al. "Impact of air pollutants on climate change and prediction of air quality index using machine learning models". In: *Environmental Research* 239 (2023), p. 117354. ISSN: 0013-9351. DOI: https://doi.org/10.1016/j.envres.2023.117354. URL: https://www.sciencedirect.com/science/article/pii/S0013935123021588.

[17] Gobithaasan R.U. *Unboxing Topological Data Analysis*. LY2020007498. Intellectual Property Corporation of Malaysia (MyIPO), 2020.

[18] K. Krishna Rani Samal, Korra Sathya Babu, and Santos Kumar Das. "Multi-directional temporal convolutional artificial neural network for PM2.5 forecasting with missing values: A deep learning approach". In: *Urban Climate* 36 (2021), p. 100800. ISSN: 2212-0955. DOI: https://doi.org/10.1016/j.uclim.2021.100800. URL: https://www.sciencedirect.com/science/article/pii/S2212095521000304.

[19] Tushar Sethi and R. C. Thakur. "Comparison of Machine Learning Algorithms for Air Pollution Monitoring System". In: *AI and IoT-Based Intelligent Automation in Robotics*. John Wiley Sons, Ltd, 2021. Chap. 19, pp. 305–322. ISBN: 9781119711230. DOI: https://doi.org/10.1002/9781119711230.ch19. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119711230.ch19. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119711230.ch19.

[20] Dayten Sheffar. "Introductory Topological Data Analysis". In: *arXiv preprint arXiv:2004.04108* (2020).

[21] Chavi Srivastava, Shyamli Singh, and Amit Prakash Singh. "Estimation of Air Pollution in Delhi Using Machine Learning Techniques". In: *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*. 2018, pp. 304–309. DOI: 10.1109/GUCON.2018.8675022.

[22] Ravi Yadav et al. "COVID-19 lockdown and air quality of SAFAR-India metro cities". In: *Urban Climate* 34 (2020), p. 100729. ISSN: 2212-0955. DOI: https://doi.org/10.1016/j.uclim.2020.100729. URL: https://www.sciencedirect.com/science/article/pii/S2212095520303291.

[23] Cheng-Hong Yang et al. "Deep learning-based air pollution analysis on carbon monoxide in Taiwan". In: *Ecological Informatics* 80 (2024), p. 102477. ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2024.102477. URL: https://www.sciencedirect.com/science/article/pii/S1574954124000190.