**Air Quality Analysis and Prediction Using Machine Learning Techniques**

**List of Abbreviations:**

|  |  |
| --- | --- |
| **Abbreviation** | **Full Form** |
| AQI | Air Quality Index |
| AQS | Air Quality System |
| WHO | World Health Organization |
| KNN | K Nearest Neighbor |
| LSTM | Long Short Term Memory |
| ARIMA | Auto Regressive Integrated Moving Average |
| SVM | Support Vector Machine |
| ANN | Artificial Neural Network |
| **DL** | Deep Learning |
| **GRU** | Gated Recurrent Unit |
| **WKNN** | Weighted K Nearest Neighbor |
| **HL** | Hallow Learning |
| **CNN** | Convolution Neural Network |
| **PM** | Particulate Matter |

1. **Introduction**

Public health and economic influences, and the environment affect air quality. Respiratory and cardiovascular conditions are magnified by the poor air quality, productivity of agriculture is affected, and air quality is responsible for climate change. Air quality is of critical importance, therefore monitoring and predicting air quality is crucial for development of an efficient mitigation strategies and air quality policy. The goal of this project is to make use of machine learning techniques to analyse air quality data and predict the levels of pollution from that data using historical trends, temporal features, and meteorological factors. The records cover indicators and key air quality and environmental condition of different counties in Texas state in the timeline of range 2022 to 2024.

**Research questions and their Hypothesis:**

1. Can machine learning models be able to accurately predict the AQI values using pollutant and observational data?  
**H₀:** Machine learning models are not that accurate in AQI prediction.  
**H₁:** Machine learning models significantly improve AQI prediction accuracy.

2. Which pollutants have the most impact on AQI levels?  
**H₀:** All pollutants contribute equally to AQI levels.  
**H₁:** At least one pollutant has a significantly different impact on AQI.

3. Does air quality vary significantly across different months/seasons?  
**H₀:** AQI levels are the same across all months.  
**H₁:** AQI levels significantly vary across months.

4. Can AQI be accurately categorized into health bands using classification models?  
**H₀:** Classification models cannot reliably categorize AQI into WHO-based bands.  
**H₁:** Classification models can reliably categorize AQI into WHO-based bands.

5. How will AQI levels change in the coming months based on time-series forecasts?  
**H₀:** Forecast models do not capture seasonal AQI trends.  
**H₁:** Forecast models like LSTM and ARIMA can capture and predict seasonal AQI trends.

6. Are there significant differences in AQI levels between various geographic clusters (regions)?  
**H₀:** AQI distributions are the same across all geographic clusters.  
**H₁:** AQI distributions differ significantly across clusters.

1. **Literature Review**

Public health, economic production, as well as the environmental sustainability, are highly influenced by the air quality. Pollutants cause more than 6.6 million early deaths annually across the globe. This is with pollutant exposure linked with chronic illness, cardiovascular illness, as well as increased death rates. Climatic regimes along with the atmosphere quality is affected by pollutants. Examples of these pollutants are particulate matter, nitrogen dioxide (NO2), carbon monoxide (CO), ozone (O₃), and Lead (Pb). It is necessary to monitor and predict Air Quality Index (AQI). Monitoring as well as forecasting the Air Quality Index (AQI) is essential. Monitoring helps in designing mitigation strategies that can limit risks from pollutant exposure. The classical models forecasting AQI are imprecise because pollutant-meteorological variable relationships are poorly captured in a non-linear manner.

**Machine Learning-Based Approaches for AQI Prediction**

Machine learning as well as deep learning (DL) models have made great advancements in forecasting pollutant concentration with high accuracy by harnessing vast data as well as advanced algorithmic structures. The paper explores key areas in research on air pollution that included variability in terms of seasons in AQI, geospatial research, anomaly detection methods, as well as application in prediction models. The paper highlights recent research studies from 2020 onwards that reflect AI-based developments in forecasting air quality as well as policymaking.

Pollution in the air, varies greatly with seasons due to variations in the weather as well as human activities. Variations in seasons in terms of pollution in the air are influenced by; humidity, industrial emissions, temperature, as well as winds. Studies on pollution trends in Asia, North America, as well as in Europe, confirmed that seasons in winters are more likely to have more pollution in terms of AQI because stagnant atmospheric conditions keep pollutants on ground level. Summer seasons have more ozone, though, due to photoreactive chemicals because high temperatures dominate. According to the research done by Song et al. (2024) in China, there is condition known as the "Winter Haze Phenomenon" which makes pollution from PM₂. ₅ as well as from PM₁₀ more prevalent, which will intensify risks in terms of human health. In America as well, summers have more pollution from wildfires that impact multiple states in terms of AQI. The seasons in terms of Saharan Dust impact quality in terms of amount in the air over the Atlantic that impacts regions in North America as well as in South America. The awareness regarding these seasons is crucial in order to boost prediction in terms of AQI as well as in terms of pollution measures that are season-based (Jaffe et al., 2020).

Legacy models traditionally utilized in forecasting air quality are statistical in form, employing multiple linear regression, as well as autoregressive integrated moving average (ARIMA). (Gao et al., 2024) in his research used ARIMA to do the forecast and the analysis. Machine learning models have been more effective in enhancing forecasting accuracy in terms of pollution.

Recent studies have extensively utilized a range of ML models, with Random Forest, which is effective in capturing non-linearity with high prediction accuracy, Support Vector Machine, which is utilized in classification of AQI with good generalizability, as well as Long Short-Term Memory networks that are effective in sequential AQI in forecasting time series. Combination models that merge multiple models, e.g., Random Forest with LSTM (Long short-term memory), have also proved effective in enhancing prediction. In the study done by Kalantari et al. (2024) he have used several machine learning models to predict air pollution and the classification. The various models he included are hallow learning (HL) models (e.g., random forest (RF), K-nearest neighbor (KNN), weighted K-nearest neighbor (WKNN), support vector machine (SVM), artificial neural network (ANN), and deep learning (DL) models (e.g., long short-term memory (LSTM), gated recurrent unit (GRU), recurrent neural network (RNN), and convolutional neural network (CNN))among all the models CNN has performed well in his model. A study in India proved that a combination with CNN with LSTM enhanced prediction in terms of concentration in PM₂. ₅ by 28% (Duan et al., 2023). Furthermore, recent studies combined deep learning with satellite-based sensing data in enhancing predictions in terms of AQI. The application of Convolutional Neural Networks in satellite images enabled more accurate spatial prediction in terms of air quality. All these advancements have not addressed challenges that have constrained research in terms of computational cost, as well as limitations in data. Real-time predictions are also a requirement (Chauhan et al., 2021). In future, studies need to ensure that they make models more interpretable as this will help in improving the use of these models in policy making when it comes to air quality.

When it comes to understand the pollution levels across the regions, Geospatial analysis has been a very important key role. Policy makers have found it easy to analyse and pinpoint the areas with high and critical levels of pollution, as well as they come up with regional mitigation strategies. This has been done by the use of Geographic Information Systems, remote sensing from satellites, and spatial interpolation methods (Kamel Boulos & Wilson, 2023). Research that has utilized GIS mapping and satellite-based information has assessed AQI differences in urban and rural settings. Results from the various studies have shown that urban areas having much greater levels of pollution due to emissions from industrial processes as well as motor vehicle emissions. Additionally, in today’s time, anomaly detection methods are being used increasingly in detecting sharp rise in air pollution (Mahmud et al., 2023). Machine learning models that, make use of anomaly detection methods such as; Isolation Forest and Autoencoders have been effective in detecting outliers in pollution. These methods are very useful in detecting high levels of outliers in pollution in real time and therefore helping policy makers to take action in good time. By combining both the geospatial data with anomaly detection methods, has greatly helped in making use of air quality monitoring networks. These networks can be used in detecting the trends in pollution in real time helping in swift decision making (Mahmood Almanor & Miklós Telek, 2023). Also Celik et al. (2011) stated that DBSCAN system has been one more best method in detecting any anomalies in AQI data

**Policy Implications and Future Research Directions**

By integrating the machine learning models in air quality forecasting has important policy relevance. By using the Ai based models’ government or the policy makers can try to reduce the emission and also the pollution can be controlled in the respective regions by following the special guidelines by formulating some strategies. In the research done by Yan et al. (2021) They have forecasted the Pollution with Machine Learning in Beijing by using these type predictions government can reduce the pollution and promote for the good air quality in next few months. Likewise, in India, AI-based air monitoring networks have been introduced in major urban cities. This introduction, provide real-time decision-making in controlling pollution (Rautela & Goyal, 2024). But still there are some of the challenges that should be addressed such as we should make sure that the data we used is consistent and need to be understand how the particular type of machine learning need to be used. By doing the above review we can summarize that its always important to do the proper preprocessing and then detect the outliers by using the various anomaly techniques. It is very important to find the factors that are influencing the AQI and then we can use many regression models such as Random Forest, Linear regression to predict exact AQI value and various classification models such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Neural Network (MLP) to Categorize AQI into health-based groups. Also, we need to forecast the data by using various techniques such as ARIMA, LSTM model to prevent the health risks and implement the proper policies to reduce the pollution.

1. **Methodology**

The methodology section mainly outlines about the complete workflow that have been used in this project for analysing the air-quality data. In this section, we have followed different approaches such as data loading, pre-processing the data to predictive modelling, spatial analysis and time forecasting. We have implemented this project by using the programming language python in the Jupiter notebook. We have also used many libraries such as pandas, NumPy, matplotlib etc.

**3.1 Data collection and Description:**

The dataset that we have chosen provides detailed information about the air quality and its related environmental factors. Below is a comprehensive explanation of its data structure and key attributes.

The dataset has around 140000 observations and 21 features from 2022 to 2024.

**Data source URL: https://www.epa.gov/**

***Dataset Description***

|  |  |  |
| --- | --- | --- |
| **Column** | **Variable Type** | **Description** |
| Date | Time-based | The day when the air quality measures recorded |
| Source | Categorical | The organization responsible for collecting the data |
| Site ID | Categorical (Nominal) | Unique ID for each location |
| POC (Parameter Occurrence Code) | Categorical (Nominal) | Number of instruments used to measure the pollutants |
| Daily Max 8-Hours Concentration | Numerical (Continuous) | Maximum recorded concentration for 8-hour timeframe |
| Units | Categorical (Nominal) | Units used to measure pollutants |
| Daily AQI Value | Numerical (Ordinal) | Standardized measure indicating impact of air quality on health |
| Local Site Name | Categorical (Nominal) | Name of the location |
| Daily Obs Count | Numerical (Discrete) | Number of observations recorded for a specific pollutant on that day |
| Percent Complete | Numerical (Continuous) | Percentage of valid observations recorded at the site for that day |
| AQS Parameter Code | Categorical (Nominal) | Unique code assigned to each pollutant being measured |
| AQS Parameter Description | Categorical (Nominal) | Name of the pollutant being measured (e.g., CO, PM10) |
| Method Code | Categorical (Nominal) | Specific method used to measure the pollutant |
| CBSA Code (Core Based Statistical Area Code) | Categorical (Nominal) | Unique numerical value for core-based statistical areas |
| CBSA Name | Categorical (Nominal) | Name of the core-based statistical area where the site is located |
| State FIPS Code  (State Federal Information Processing Standards Code) | Categorical (Nominal) | Federal Information Processing Standards (FIPS) code for the state |
| State | Categorical (Nominal) | Name of the state where the monitoring site is located |
| County FIPS Code | Categorical (Nominal) | FIPS code for the county where the monitoring site is located |
| County | Categorical (Nominal) | Name of the county where the monitoring site is located |
| Site Latitude | Numerical (Continuous) | Latitude coordinates of the monitoring site |
| Site Longitude | Numerical (Continuous) | Longitude coordinates of the monitoring site |

We have loaded the dataset by using the function pandas.read\_excel() in to a DataFrame which can be used for further analysis. To make sure if the data is loaded and transformed properly we have used the head() function to display the first few rows of the data.

A screenshot of a computer

AI-generated content may be incorrect.

Fig 3.1.1 Dataset preview

**3.2 Data preprocessing:**

It is essential to preprocess the data before starting with analysis and modeling. This process is performed to clean and prepare the dataset for proper exploration of data and accurate modeling by reducing the bias and errors in the results. This process includes various steps such as checking the datatypes of each column and changing the datatype if required, handling the missing values, removing the duplicates, detecting and correcting the outliers, performing the range validations and feature engineering.

***a. Handling the missing values:***

Having the data with large number of missing values will impact the analysis. So, it is very important to handle these values. And from the Fig 3.2.1 we can observe that there are several missing values in the columns CBSA Name, Method Code, CBSA Code, and Daily AQI Value.

For the categorical column CBSA Name, the missing values were handled by filling it with ‘unknown’ to retain those values without causing any introducing bias. The numerical values like Method Code, CBSA Code, and Daily AQI Value were filled by using the median of each column which would become a best choice because it wasn’t affected by the outliers unlike the mean method. For the missing AQI values which have entries like “.” Are replaced with NaN (Not A Number) and then these are handled by using the median AQI. From the figure 3.2.2 we can see that all the missing values were handled properly

*A screenshot of a computer

AI-generated content may be incorrect.*

***Fig 3.2.1*** *Missing Value Summary before handling*

*A black background with white text

AI-generated content may be incorrect.*

*Fig 3.2.2 Code for finding the missing values*

A screenshot of a computer

AI-generated content may be incorrect.

**Fig 3.2.3** *Missing Value Summary after handling*

***b. Removing the Duplicates***

The dataset has multiple duplicate columns. To ensure that each observation is unique the duplicate values were dropped.

***c. Detecting and handling the outliers***

In the dataset, we have found that there are some of the outliers present in the numerical column such as concentration, and AQI values. There were some of the negative values present in the air pollutant levels which are not realistic, so we have replaced these negative values with the median of that column as median is less affected by the extreme values. Also, we have identified these extreme high values by using a 99th percentile as a cut-off. Any value which is higher than this is capped at the 99th percentile to reduce the effect of this outliers without removing the useful data. So that by doing this process, we have ensured that our dataset has an accurate balance for the analysis and modeling.

A screenshot of a computer program

AI-generated content may be incorrect.

**Fig 3.2.4** *Handling the outliers*

***d. Range Validations of Latitude and Longitude***

To maintain a valid range of latitude and longitude, we considered doing the range validation. since the latitude range must be between 25.837 and 36.500 and the longitude must be between -106.645 and -93.508, we have eliminated the rows which are present out of this range by considering them as invalid.

A computer screen with colorful text

AI-generated content may be incorrect.

**Fig 3.2.4** *Validating the ranges of Latitude and Longitude*

1. ***Data Formatting and Feature Engineering***

We have converted the date column to a datetime format by using the pd.to\_datetime() from text, because this makes easy while analysing the time trends.

After that, we have created a new column called month, which is created by using df['Date'].dt.month. We have done this to understand how the air quality is changing across different months.

1. ***Converting Data types:***

We have converted the Daily AQI Value from string to numeric value by using pd.to\_numeric().

**3.3 Exploratory Data Analysis:**

Eda is a very important step which helps us to understand the patterns in the data and gain insights before performing the predictive modeling. In this section we use various plots and charts to explore air quality based on various parameters.

***Distribution of Daily AQI Values***

From the below figure 3.3.1, which shows the histogram with the curve representing the overall distribution of AQI values we can observe that most of the AQI values are lying between 0 and 40 indicating good air quality and a long tail on the right side suggest that occasionally there are high AQI values which may have caused due to the pollution spikes.

A graph of a distribution of daily aq values

AI-generated content may be incorrect.

**Fig 3.3.1** *Distribution of Daily AQI Values*

***Monthly AQI Distribution***

To explore the seasonal trends, we have created a box plot containing the AQI values for each month. The figure 3.3.2 shows that the AQI values are varying from month to month.

A graph showing different colored squares

AI-generated content may be incorrect.

**Fig 3.3.2** *Monthly AQI Distribution*

From the above plot, we can observe that AQI values are generally higher in the summer months that is from June to August due to the heat produced in those months and because of the ozone buildup and the AQI is lower in the winter months that is from December to February. We can also observe the outliers which are present above the boxesare the unusual high AQI which may have caused due to some uncertain events like forest fires etc.

***Geospatial Distribution of Average AQI***

To see the distribution of AQI across the various locations, we have plotted the average AQI values for each location based on its latitude and longitude values.

A diagram of a geospatial distribution of average aq

AI-generated content may be incorrect.

**Fig 3.3.3** *Geospatial Distribution of Average AQI*

From the above scatter plot which has color shading to indicate pollution levels, we can observe that the red points indicate the higher average AQI, blue point shows the places with cleaner air. This will help us to identify the pollution hotspots geographically.

***Frequency of Recorded Pollutants (AQS Parameters)***

The below bar chart shows us the air pollutants which were measured most frequently in the data set.

From the figure, we can see that ozone and nitrogen dioxide are the most frequently recorded pollutants and carbon monoxide, PM 10 and lead are the less recorded pollutants

***A graph of different colored squares

AI-generated content may be incorrect.***

**Fig 3.3.4** *Frequency of Recorded Pollutants*

***Pollutant Trends Over Time***

From the figure 3.3.5, we can see how AQI values related to different pollutants have changed over the time. We can observe Ozone and PM 10 show large fluctuations and seem to be mostly influenced by AQI and carbon monoxide and nitrogen dioxide have stayed stable over time.

***A graph of different colored lines

AI-generated content may be incorrect.***

**Fig 3.3.5** *Pollutant Trends Over Time*

***Impact of Pollutants on AQI***

To compare the impact of each pollutant on the AQI, we have plotted the average AQI values caused by each pollutant. The below figure 3.3.6 shows us that Ozone have contributed most higher AQI values followed by PM 10 and lead.

***A graph of different pollutants

AI-generated content may be incorrect.***

**Fig 3.3.6** *Pollutant Trends Over Time*

***Correlation Between Pollutants and AQI***

With the help of the correlation matrix, we can check how each numerical variable is related with AQI value

From the figure 3.3.7, we can see that concentration doesn’t have any correlation with AQI, percentage complete and observation count has light positive relationship with AQI.

**A screenshot of a graph

AI-generated content may be incorrect.**

**Fig 3.3.7** *Correlation Matrix Between Pollutants and AQI*

**3.4 Anomaly & Clustering Analysis**

After performing the exploratory data analysis, we have used various advanced techniques like anomaly detection and clustering to find the unusual behavior that has been occurred in the Air quality data.

***Anomaly detection using isolation forest:***

To detect a sudden spike at the AQI level that is the abnormal pollution events that has been occurred in the air, we have used a machine learning technique called as isolation forest. By using this algorithm, we can find the outliers in the large data sets by identifying the points that behave unusual from the others.

A graph of air quality

AI-generated content may be incorrect.

**Fig 3.4.1** *Air quality anomalies detected by isolation forest*

From the above figure 4.4.1, we can see there are two types of dots one is the red dot and the other one is the blue dot. Red dots indicate the normal AQI values, whereas the blue dots indicate the anomalies

From these plots we can conclude that the anomaly points represent unusual events such as wildfires, industrial pollution or weather related conditions. Over 550 anomalies are detected in the data.

***Spatial clustering using DB Scan***

To explore how different locations behave geographically, we have used a cluster called DB scan, which is density based spatial clustering of applications with noise. This method groups the various monitoring locations which have similar type of air quality patterns and based on their latitude and longitude

A screen shot of a graph

AI-generated content may be incorrect.

**Fig 3.4.2** *Spatial clustering of air quality monitoring sites using DBSCAN*

Here, each color represents a site and each color stands for a unique cluster ID and the sites marked in the grey are the outliers different from all other locations indicating that these AQI levels are different from the other nearby sites, which possible due to some local pollution sources.

In this analysis, the model has created 68 distinct clusters, and this helps us to identify the regional similarities in the Air quality.

**3.5 Modeling**

In this section we have developed various machine learning models to achieve the two main goals:

* Predicting the actual AQI value
* Classifying AQI into air quality categories
* Predicting the future AQI trends using the Time series Forecasting

We have used various supervised, unsupervised and time series modeling techniques such as Random Forest Regression, Linear Regression, XGBoost, LSTM, ARIMA, and DBSCAN Clustering to gain insights from the dataset.

We have prepared the dataset by splitting the dataset into 80% training and 20% testing sets.

***Predicting the actual AQI value:***

Firstly, we are predicting the numerical AQI values based on various factors which helps us estimate the daily pollution levels. This will help the authorities to issue the alerts and plan the preventive actions in advance.

We have used the models like Linear Regression, Random Forest and XGBoost Regression to predict the AQI values.

**Linear Regression:** Linear Regression model assumes the linear relationship between the inputs and the output. This will be used as a baseline model. We have considered the Daily AQI Value as the target variable and the features such as Daily Max 8-hour Concentration, Percent Complete and Daily Observation Count are used as the key features.

**Random Forest Regression:** Random forest is an ensemble learning method which builds the multiple decision trees and then combine its output to improve the accuracy and helps in reducing the over fitting. Like linear regression model, we have used the same target variable and key features.

**XGBoost Regression:** The next regression technique which we used to predict the AQI value is the XGBoost regression. This is a boosting technique which is known for its accuracy and speed. Similar to the linear regression and Random forest we have used the same target variable.

Then, we have compared these models by evaluating them using:

* Root Mean Square Error (RMSE)
* Mean Absolute Error (MAE)
* R² Score

***Classifying AQI into air quality categories***

In this part, we have categorized the AQI values into various bands using the different classification models. These categories are based in the WHO health standards. This approach will help to ease the communication with people about understanding the levels of pollution.

From the table below, we can find the ranges and their categories for the specific type of range.

|  |  |
| --- | --- |
| **AQI Range** | **Category** |
| 0–50 | Good |
| 51–100 | Moderate |
| 101–150 | Unhealthy |
| 151+ | Hazardous |

For this classification, we have used 4 classification models.

* Support Vector Machine (SVM)
* Decision Tree
* K-Nearest Neighbors (KNN)
* Neural Network

**Support Vector Machine (SVM):** SVM is a supervised learning algorithm, which will find the best boundary between the different types of classes. This algorithm works well with the different types of data like high dimensional data and can also handle the non-linear relationships by using the kernel functions. To handle the imbalances, we have used the class weight balancing, and we use the RBF radial basis function kernel.

**Decision Tree:** Decision trees will split the data into the various branches based on the feature threshold, and these are good at capturing the non-linear class boundaries. To ensure the fairness in the predictions, we have used balanced class weighting.

**K-Nearest Neighbors (KNN):** KNN is a simple model which classify is based on the majority class of nearby points. It is very useful for the small and medium datasets. We have used five neighbours, which is K=5 with distance-based weighting.

**Neural Network (MLP):** MLP is a feed forward, artificial neural network with one or more hidden layers. It is used to learn complex patterns by adjusting the weights during the training. We have used one hidden layer with hundred neurons, ReLU activation and 500 iterations.

We have evaluated all these classifiers by using the accuracy precision, recall, F1 score and confusion matrix.

***Time-Series Forecasting of AQI Trends***

By forecasting the AQI trends, we can predict the future pollution levels based on historical observations. For doing this, we have used two models:

* LSTM
* ARIMA

**ARIMA:** ARIMA (autoregressive integrated moving average) is a traditional time series model that forecasts the future values based on the past patterns**.** In this, we have grouped the data into the continuous daily time series and then the patterns (P,D,Q) were selected as (5,1,5) to the experimentation and then the model forecasted the AQI values for the next six months

**LSTM:** LSTM (Long short-term memory) is a deep learning model, which is mainly suitable for tasks like time series

For the historical daily average, AQI values were aggregated and sorted logically, and then seven day look back window was applied, allowing the model to use the previous weeks’ AI values to predict the next day. Before training, the AQI values were normalized by using MinMaxScaler to bring all the values in the same range to improve the model performance

1. **Results and interpretation:**

In the results and interpretation section. We will be presenting the findings obtained through our machine learning models and time series forecasting. Along with this, we will also answer our 10 research questions.

***Predicting the actual AQI value:***

Among the three regression models which were used for predicting the AQI prediction, random forest has performed very well by accurately capturing the non-linear relationships between the different features and the main target variable AQI, whereas the linear regression model has performed the worst.

**Table 4.1**

*Validation metrics for Regression Models*

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **MAE** | **R² Score** |
| Linear Regression | 19.34 | 14.99 | 0.04 |
| Random Forest | 1.54 | 0.19 | 0.999 |
| XGBoost | 1.5 | 0.2 | 0.994 |

From the above table, we can see that random forest has got the R Square score of 0.99 with less errors, followed by the XGBoost.

***Classifying AQI into air quality categories***

We have categorised the AQI values into four health related classes as suggested by WHO. for health-related classes are good, moderate, unhealthy, and hazardous.

A screenshot of a computer

AI-generated content may be incorrect.

**Fig 4.1** *Classification of air quality categories*

The four models, SVM decision tree, KNN and neural network MLP were trained and evaluated to classify.

From the below table, we can find the performance evaluated by using the precision, recall and F1 score

**Table 4.2**

*Evaluation metrics of all classifiers*

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1-Score** |
| SVM | 0.98 | 0.46 | 0.61 |
| Decision Tree | 1 | 1 | 1 |
| KNN | 0.98 | 0.97 | 0.997 |
| Neural Network | 0.92 | 0.93 | **0.98** |

Confusion matrix analysis:

Confusion matrix analysis can help us to visualise how accurately each classes predicted by using the model and what type of misclassifications are occurred.

A screen shot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect. A screenshot of a computer screen

AI-generated content may be incorrect.

**Fig 4.2** *Confusion matrix for 4 classifiers*

From all four-confusion matrix, we can see that SVM has failed to classify the moderate and unhealthy categories effectively by showing the poor recall and the decision tree model gave almost perfect predictions for all classes with very minimal misclassifications. The KNN model also showed an excellent overall performance with a few errors. It have effectively handled all AQI categories, but not as best as decision trees. Neural network performed good, but its recall was slow for moderate and unhealthy. This is mainly due to the data imbalance. Therefore, neural networks and decision trees have achieved the best balance of precision and recall across all AQI categories, whereas SVM under-performed and Canon did well, but with minimal miss classifications.

***Time-Series Forecasting of AQI Trends***

To forecast the future AQI trends, we have used to time series models

1. LSTM
2. ARIMA

**LSTM forecasting:**

The LSTM model was trained on seven sequences and predicted the AQI for the next day. It achieved an RMSE of 5.57, which indicates that it is strong and good for short-term prediction.

This forecast has revealed that the AQI trends will rise slightly during the summer months, and then there is a dip in the other seasonal cycles.

A graph showing a red and green line

AI-generated content may be incorrect.

**Fig 4.3** *Forecasting of AQI values using LSTM*

**ARIMA**

We have used ARIMA model to forecast the long term AQI trends, but it was very less responsive to the short-term spikes and produced a stable six-month forecast. We can observe this in the below fig.

A graph showing a blue line

AI-generated content may be incorrect.

**Fig 4.3** *AQI forecasted by ARIMA*

**From our analysis we have answered the 10 research questions below:**

**1. Can machine learning models be able to accurately predict the AQI values using pollutant and observational data?  
H₀:** Machine learning models are not that accurate in AQI prediction.  
**H₁:** Machine learning models significantly improve AQI prediction accuracy.

Yes, The Machine Learning models were abled to predict the AQI values with at most accuracy. And among all the models such as Random Forest, Linear Regression, Support Vector Machine the Random Forest Model outperforms with lowest Root Mean Squared Error, Mean Absolute Error and with the highest R2 score. Whereas the Linear Regression performed poor because of no linear relationship exists between the target and the independent variables. And these results support the alternative hypothesis (H1) that the machine learning models have significantly improved AQI prediction accuracy.

**2. Which pollutants have the most impact on AQI levels?**  
**H₀:** All pollutants contribute equally to AQI levels.  
**H₁:** At least one pollutant has a significantly different impact on AQI.

The analysis of pollutant impact was done in the Exploratory Data Analysis phase, where the analysis revealed that Ozone (O3) has the highest contribution towards the AQI fluctuations then followed by the other pollutant PM10, Lead. And other pollutants such as Carbon Monoxide and Nitrogen Dioxide (NO2) has an average impact on the AQI.

A screenshot of a computer program

AI-generated content may be incorrect.

Fig 4.4 One-way ANOVA test for AQI Variation Among Pollutants

And also the ANOVA test that was performed by us shows that there is significant difference in the AQI values across the pollutants (p<0.00001), by which we have to reject the null hypothesis (H0) and need to accept that some pollutants have a significant impact than the other pollutants.

**3. Does air quality vary significantly across different months/seasons?  
H₀:** AQI levels are the same across all months.  
**H₁:** AQI levels significantly vary across months.

Yes, the air quality shows some variations in the seasons and months. Boxplot of AQI by months revealed that in the months ranging from June to September has higher Median AQI which is most likely due to the ozone formation due to heat and the other aspects such as wildfires whereas in the winter months there is lower AQI values. The Kruskal-Wallis test that we performed are supporting these observations. With the p-value <0.05, rejecting the null hypothesis.

A screen shot of a computer

AI-generated content may be incorrect.

Fig 4.5 *Kruskal-Wallis Test for Seasonal Variation in AQI Across Months*

**4. Can AQI be accurately categorized into health bands using classification models?**  
**H₀:** Classification models cannot reliably categorize AQI into WHO-based bands.  
**H₁:** Classification models can reliably categorize AQI into WHO-based bands.

Yes, the classification models have performed well in categorizing the AQI values into several different bands based on WHO such as Good Moderate and Unhealthy. And the Decision Tree, SVM, KNN classifiers all have achieved a good accuracy. Which confirms that the classification models can be categorized into bands based on AQI. And here we are supporting the Alternative Hypothesis.

**5. How will AQI levels change in the coming months based on time-series forecasts?**  
**H₀:** Forecast models do not capture seasonal AQI trends.  
**H₁:** Forecast models like LSTM and ARIMA can capture and predict seasonal AQI trends.

We have performed the time series forecasting by using models such as LSTM and ARIMA. The LSTM model have shown excellent results in short term forecasting with the performance of RMSE = 5.67 by capturing the seasonal trends effectively. Whereas the ARIMA model showing the stable seasonal patterns. These forecasting results supports the Alternative hypothesis that the forecasting models can be captured and predicts the AQI trends over the time.

**6. Are there significant differences in AQI levels between various geographic clusters (regions)?  
H₀:** AQI distributions are the same across all geographic clusters.  
**H₁:** AQI distributions differ significantly across clusters.

Yes, we can find the significance difference across the different geographic clusters and are identified by the DBSCAN algorithm. And we have confirmed that this model will perform both the one-way ANOVA and the Kruskal-Wallis test which showed p-values <0.00001, showing the AQI levels are not evenly distributed across regions. Thus, we reject the null hypothesis and conclude that AQI varies significantly across the different geographical regions.

A screenshot of a computer program

AI-generated content may be incorrect.

Fig 4.6 Kruskal-Wallis Test for AQI Variation Across Geographic Clusters

**Summary:**

* With the help of accurate AQI predictions and the classification, city and different government agencies can issue the health warnings and take a preventive measure in advance and can restrict some factory pollution levels or traffic control during the highest pollution time
* Regions with high AI values can be in priority for taking some strict regulations by adding more sensors or public awareness campaigns.
* From our insights, we can see that air quality is declining in the summer months, therefore many policies can be applied during the seasons and can help in reducing the pollution levels. Also, different healthcare providers be prepared for any respiratory related issues during these months.
* The anomaly detection model can help the higher authorities to respond to the sudden pollution events like wildfires, industrial leaks, and sensor malfunctions
* The clustering analysis shows which areas will face the similar type of air quality issues, so the strategies can be shared to the cities or states in these clusters and different strategies can be used for another cities.
* The forecasting models should be used by the planner to visualise the trends in the future, which will be used in assisting the environmental impact and can be helpful for sustainability goals.

**Limitations and areas of improvement:**

* This analysis has focused more on the observation and pollutant levels, but by including the factors like wind, speed, temperature, and humidity could have improve the prediction and provide us the deep insights
* Some regions may have more monitoring sites, whereas some regions have less than the others which would add some bias in the clustering
* While the models have used to concentration and observation metrics, they did not include all pollutants like SO2 due to limited data.

**References**

* Afifa, Arshad, K., Hussain, N., Ashraf, M. H., & Saleem, M. Z. (2024). Air pollution and climate change as grand challenges to sustainability. *Science of the Total Environment*, *928*, 172370. <https://doi.org/10.1016/j.scitotenv.2024.172370>
* Anwar, M. N., Shabbir, M., Tahir, E., Iftikhar, M., Saif, H., Tahir, A., Murtaza, M. A., Khokhar, M. F., Rehan, M., Aghbashlo, M., Tabatabaei, M., & Nizami, A.-S. (2021). Emerging challenges of air pollution and particulate matter in China, India, and Pakistan and mitigating solutions. *Journal of Hazardous Materials, 416*, 125851. <https://doi.org/10.1016/j.jhazmat.2021.125851>
* Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş. (2011). *Anomaly detection in temperature data using DBSCAN algorithm.* In 2011 International Symposium on Innovations in Intelligent Systems and Applications (pp. 91-95). IEEE. <https://doi.org/10.1109/INISTA.2011.5946052>
* Chauhan, R., Kaur, H., & Alankar, B. (2021). Air Quality Forecast using Convolutional Neural Network for Sustainable Development in Urban Environments. *Sustainable Cities and Society*, *75*, 103239. https://doi.org/10.1016/j.scs.2021.103239
* Duan, J., Gong, Y., Luo, J., & Zhao, Z. (2023). Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer. *Scientific Reports*, *13*(1). <https://doi.org/10.1038/s41598-023-36620-4>
* Gao, W., Xiao, T., Zou, L., Li, H., & Gu, S. (2024). Analysis and Prediction of Atmospheric Environmental Quality Based on the Autoregressive Integrated Moving Average Model (ARIMA Model) in Hunan Province, China. *Sustainability*, *16*(19), 8471–8471. <https://doi.org/10.3390/su16198471>
* Jaffe, D. A., O’Neill, S. M., Larkin, N. K., Holder, A. L., Peterson, D. L., Halofsky, J. E., & Rappold, A. G. (2020). Wildfire and prescribed burning impacts on air quality in the United States. *Journal of the Air & Waste Management Association*, *70*(6). <https://doi.org/10.1080/10962247.2020.1749731>
* Kamel Boulos, M. N., & Wilson, J. P. (2023). Geospatial techniques for monitoring and mitigating climate change and its effects on human health. *International Journal of Health Geographics*, *22*(1). <https://doi.org/10.1186/s12942-023-00324-9>
* Kalantari, E., Gholami, H., Malakooti, H., & others. (2024). Machine learning for air quality index (AQI) forecasting: Shallow learning or deep learning? *Environmental Science and Pollution Research, 31,* 62962–62982. <https://doi.org/10.1007/s11356-024-35404-1>
* Mahmood Almansoori, & Miklós Telek. (2023). *Anomaly Detection using combination of Autoencoder and Isolation Forest*. 25–30. <https://doi.org/10.3311/wins2023-005>
* Mahmud, K., Mitra, B., Mohammed Sakib Uddin, Al-Ekram Elahee Hridoy, Aina, Y. A., Ismaila Rimi Abubakar, Syed Masiur Rahman, Mou Leong Tan, & Muhammad Muhitur Rahman. (2023). Temporal assessment of air quality in major cities in Nigeria using satellite data. *Atmospheric Environment: X*, 100227–100227. <https://doi.org/10.1016/j.aeaoa.2023.100227>
* Rautela, K. S., & Goyal, M. K. (2024). Transforming air pollution management in India with AI and machine learning technologies. *Scientific Reports*, *14*(1). <https://doi.org/10.1038/s41598-024-71269-7>
* Song, W., Wang, M., Zhao, Y., Bo, Y., Yao, W., Chen, R., Wang, X., Wang, X., Li, C., & He, K. (2024). Low-condensation diesel use contributes to winter haze in cold regions of China. *Environmental Science and Ecotechnology*, *22*, 100456. <https://doi.org/10.1016/j.ese.2024.100456>
* Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., & Li, F. (2021). Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Systems with Applications*, *169*, 114513. <https://doi.org/10.1016/j.eswa.2020.114513>