St Joseph Engineering College

An Autonomous Institution

Affiliated to VTU-Belagavi, Recognized by AICTE, NAAC Accredited

Vamanjoor, Mangaluru-575028, Karnataka

DATA VISUALIZATION

REPORT

ON

<<AQI Prediction And Analysis >>

Submitted in partial fulfillment of the requirements for the degree

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

(DATA SCIENCE)

Submitted by

<< Joywin Neil Lasrado	USN >>4SO22CD024
<<Shetty Aditya Santhosh	USN >> 4SO22CD048
<<Subhiksha Rai K	USN >> 4SO22CD054
<<Prerana D P	USN >> 4SO22CD037

Ms.Shruthi Vishwajeeth
(Asst. Professor, ICBS Department)
Course Coordinator

2024-2025

ST JOSEPH ENGINEERING COLLEGE

An Autonomous Institution

(*Affiliated to VTU-Belagavi, Recognized by AICTE, NAAC Accredited*)

Vamanjoor, Mangaluru-575028

DEPARTMENT OF INTELLIGENT COMPUTING & BUSINESS SYSTEMS

CERTIFICATE

Certified that the Mini Project Work entitled **"<<AQI Prediction And Analysis>>"** *carried out by*

<< Joywin Neil Lasrado	USN >>4SO22CD024
<<Shetty Aditya Santhosh	USN >> 4SO22CD048
<<Shubhiksha Rai K	USN >> 4SO22CD054
<<Prerana D P	USN >> 4SO22CD037

bonafide students of V semester in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering (Data Science) of the Visvesvaraya Technological University, Belagavi during the year 2024-2025.

Ms.Shruthi Vishwajeeth
Course Coordinator
Data Visualization Lab
22CDS55L

PROJECT ON

AQI

PREDICTION AND ANALYSIS

USING

PYTHON AND DATA VISUALISATION

Abstract:

This project presents an Air Quality Index (AQI) prediction system using machine learning models, designed to provide insights and predictions based on historical environmental data. The system integrates several functionalities within a user-friendly web interface powered by Streamlit. Users can upload datasets containing information on pollutants like CO, Ozone, NO2, and PM2.5, along with geolocation data (latitude and longitude), and use the trained models to predict AQI values and categories for specific countries and cities. The project includes a regression model for predicting AQI values and a classification model for predicting AQI categories.

The application offers three main features:

1. **Prediction of AQI**: Users can input a country and city, and the system predicts the AQI value and category based on the historical data available.
2. **Data Analysis**: This feature provides insights into the distribution of pollutant levels by country and compares the most polluted countries based on predicted AQI values. Visualizations like pie charts and bar plots are used for easy interpretation.
3. **Visualization**: The platform includes interactive and static visualizations, such as geographical maps of AQI distribution, choropleth maps displaying average AQI by country, and various pollutant contribution plots, helping users understand global air quality patterns. The system also offers a correlation matrix to explore relationships between different pollutants.

By combining predictive analytics with data visualization, this project enables users to explore air quality trends, make informed decisions, and gain valuable insights into environmental pollution levels globally. The underlying machine learning models are built to handle large datasets, making the system scalable and adaptable to different regions and time periods.

Acknowledgement

We would like to express our heartfelt gratitude to Ms. Shruthi Vishwajeeth, Assistant Professor, Department of Intelligent Computing & Business Systems, for her invaluable guidance, consistent support, and insightful suggestions throughout the course of this project. Her expertise, encouragement, and direction have played a significant role in shaping our work and ensuring its successful completion.

Our sincere thanks also go to Dr. Shreenath Acharya, Head of the Department, Intelligent Computing & Business Systems, for his expert guidance and constant support. His valuable advice helped us navigate challenges and refine our approach, leading to the successful culmination of this project.

We would also like to extend our appreciation to all the dedicated faculty and staff members of ICBS, whose support, suggestions, and encouragement have been a constant source of motivation throughout this endeavor. Their belief in our potential has been truly inspiring.

Finally, our deepest gratitude goes to our friends and family members, whose unwavering support and encouragement have been a source of strength throughout this project. Their understanding and belief in us have provided the motivation to continue and complete this work.

Together, the combined support from all these individuals has made this project not only possible but also a deeply fulfilling learning experience. Thank you all for your dedication, support, and belief in our success.

Table of Contents

Introduction:

Air quality is a vital determinant of public health and environmental sustainability, with pollutants like CO, Ozone, NO2, and PM2.5 significantly impacting human well-being. Monitoring and predicting the Air Quality Index (AQI) is essential for assessing air pollution levels and taking timely action to protect public health. This project aims to develop a comprehensive AQI prediction system that utilizes machine learning techniques to predict AQI values and classify air quality based on historical data from various countries and cities.

The system is designed to be user-friendly, built using Python libraries such as Streamlit, Pandas, Seaborn, Matplotlib, and Plotly. It offers three key features:

1. **AQI Prediction**: Users can input a country and city to receive predicted AQI values and their corresponding categories.
2. **Data Analysis**: The system allows users to explore pollutant distributions and compare AQI values across selected countries.
3. **Visualization**: It includes interactive maps and static plots to help users visualize global AQI trends, pollutant contributions, and geographical distributions.

By combining machine learning-based predictions with powerful data visualization tools, this project provides an accessible platform for understanding and addressing global air quality challenges, benefiting individuals, researchers, and policymakers alike.

Literature Survey:

Air quality monitoring and prediction have become critical areas of research due to the significant health risks associated with exposure to air pollutants. Numerous studies and technologies have been developed to assess, predict, and visualize air quality. This literature survey explores key works and approaches related to AQI prediction, machine learning in environmental modeling, and data visualization techniques used in air quality analysis.

1. **Air Quality Prediction and Machine Learning:** Several studies have explored the use of machine learning techniques to predict AQI values based on environmental data. In a study by **Jiang et al. (2018)**, a combination of regression models and neural networks was used to predict AQI based on historical air quality data. They found that machine learning models, particularly Support Vector Machines (SVM) and Random Forests, performed better than traditional statistical methods in forecasting AQI levels. Similarly, **Li et al. (2019)** developed a hybrid machine learning model combining deep learning techniques with time-series forecasting to predict AQI based on past pollutant concentration data. These models demonstrated high accuracy in predicting short-term air quality trends.
2. **Predicting Air Quality Categories:** In addition to predicting AQI values, many studies have focused on classifying AQI into predefined categories (e.g., Good, Moderate, Unhealthy). **Chien et al. (2020)** used classification algorithms, such as Decision Trees and Random Forests, to categorize air quality into various levels based on multiple pollutant concentrations. Their research highlighted the importance of selecting relevant features (pollutant concentrations, weather conditions, geographical factors) to improve classification accuracy. **Wang et al. (2017)** explored the use of K-means clustering for AQI categorization, showing that unsupervised machine learning techniques could also be effective in classifying air quality levels.
3. **Geographical Distribution and Visualization:** Geospatial analysis and visualization have become essential for understanding the

spatial distribution of air pollution. Studies like **Chen et al. (2016)** employed geospatial data to analyze air quality patterns and correlations with environmental and demographic factors. The use of interactive maps, such as those created with **Plotly** and **Leaflet**, has gained popularity for visualizing AQI data on a global scale. These tools enable users to explore the geographical spread of pollutants and their concentrations in real-time, providing a better understanding of pollution trends across regions.

4. **Real-Time Air Quality Monitoring:** With the advent of IoT and sensor technologies, real-time air quality monitoring has gained significant traction. **Zhang et al. (2017)** utilized low-cost sensors to collect real-time data on pollutant levels and developed machine learning models for forecasting AQI. This real-time data collection, when integrated with predictive models, offers highly accurate and timely information on air quality, which is essential for urban planning and public health management.
5. **Pollutant-Specific AQI Models:** Air quality is typically influenced by multiple pollutants, and understanding their individual contributions is crucial for accurate AQI prediction. Studies like **Zhao et al. (2018)** and **Gao et al. (2020)** examined the individual impacts of pollutants like CO, Ozone, NO2, and PM2.5 on AQI levels. They used multi-pollutant models to predict the combined effect of these pollutants on air quality. Their findings emphasized the importance of considering pollutant interactions and temporal patterns for more precise AQI predictions.
6. **Impact of Environmental and Socioeconomic Factors:** Several studies have also investigated the relationship between air quality and external factors, such as weather conditions, urbanization, and socioeconomic status. **Liu et al. (2019)** explored how temperature, humidity, and wind speed affect the concentration of pollutants and, subsequently, AQI. Additionally, **Xu et al. (2020)** highlighted the role of urbanization in exacerbating pollution levels, especially in rapidly developing cities.

7. **Tools and Libraries for Data Visualization and Analysis:** Data visualization has become an essential part of AQI analysis, as it allows users to easily interpret complex datasets. Libraries such as **Matplotlib**, **Seaborn**, and **Plotly** are widely used for creating static and interactive visualizations. **Plotly**, in particular, has become popular for generating interactive geographical visualizations like choropleth maps and scatter plots, which help visualize AQI data on a global scale. These tools enable users to identify pollution hotspots, understand regional differences, and communicate the findings effectively.

In conclusion, existing literature highlights the effectiveness of machine learning in predicting AQI values and classifying air quality. It also emphasizes the importance of geographical analysis and the use of interactive visualizations to better understand the spatial distribution of air pollution. The integration of predictive modeling, real-time data, and data visualization tools provides a comprehensive approach to tackling the growing problem of air pollution, helping policymakers, researchers, and the public make informed decisions for better air quality management.

Problem Definition:

Air pollution is one of the most pressing environmental issues worldwide, with direct impacts on public health, climate change, and ecosystems. The Air Quality Index (AQI) is a key metric used to assess and communicate the quality of air in a specific region, based on the concentration of pollutants such as Carbon Monoxide (CO), Ozone (O_3), Nitrogen Dioxide (NO_2), and Particulate Matter (PM2.5). However, predicting AQI values for different locations with a high degree of accuracy is a complex challenge due to the numerous variables that influence air quality.

The problem addressed in this project is the lack of an efficient, automated system for predicting AQI values and classifying air quality based on historical data from various countries and cities. Specifically, the challenges include:

1. **Data Complexity:** Air quality is influenced by various factors such as pollutant concentrations, weather conditions, geographical features, and human activities. Managing and modeling this complex data in a way that ensures accurate predictions is a significant challenge.
2. **Prediction Accuracy:** Predicting AQI values using traditional statistical methods may not account for the intricate relationships between pollutants, making it difficult to achieve high accuracy. Machine learning models, while effective, require careful selection of features and tuning of parameters to ensure reliable predictions.
3. **Real-time Data Processing:** AQI prediction requires access to up-to-date environmental data. Incorporating real-time data feeds and creating a system that can handle continuous updates is a challenging aspect of building a robust AQI prediction system.
4. **Data Visualization and Interpretation:** Communicating the predicted AQI values and air quality classifications in a way that is both understandable and actionable for users (e.g., policymakers, environmentalists, and the general public) is another challenge. Effective data visualization is crucial for interpreting the information and making informed decisions.

This project aims to address these challenges by developing a machine learning-based AQI prediction system that:

- Uses regression models to predict AQI values based on historical pollutant concentration data.
- Classifies AQI values into predefined categories (e.g., Good, Moderate, Unhealthy) using classification algorithms.
- Provides an interactive platform for users to input location data (country, city) and receive predictions for AQI values and categories.
- Incorporates visualizations, such as geographical maps and pollutant distribution graphs, to help users understand air quality patterns and trends.

The goal of this project is to provide a user-friendly, efficient tool that can predict AQI values, analyse air quality trends, and support decision-making regarding air pollution management, thereby contributing to global efforts to improve air quality and public health.

Proposed Methodology:

To develop an effective AQI prediction system, we propose a methodology that integrates machine learning models with data visualization tools. The approach involves several key steps, including data collection, preprocessing, model training, prediction, and visualization. The proposed methodology is outlined as follows:

1. Data Collection and Preprocessing:

The first step in the methodology is collecting a comprehensive dataset containing historical air quality data. The dataset should include information such as AQI values, pollutant concentrations (CO, Ozone, NO2, PM2.5), and geographical information (country, city, latitude, longitude). Data can be sourced from publicly available air quality monitoring networks or research databases.

Data Preprocessing Steps:

- **Data Cleaning:** Handle missing values, outliers, and inconsistencies in the dataset. For example, missing pollutant concentration values are replaced with the mean or median of the respective pollutant's values.
- **Feature Engineering:** Select and construct relevant features for AQI prediction. This may include pollutant concentrations, geographical information, and temporal factors (e.g., seasonality, weather conditions).
- **Normalization/Standardization:** Scale the pollutant values and other continuous features to ensure uniformity and improve the performance of machine learning algorithms.

2. Model Development:

The AQI prediction system requires two machine learning models: a **regressor** for predicting AQI values and a **classifier** for categorizing AQI into predefined air quality levels (e.g., Good, Moderate, Unhealthy). The steps for model development are as follows:

- **Regressor Model (AQI Prediction):** A regression model is used to predict continuous AQI values based on historical pollutant data. We propose using algorithms such as:
    - **Random Forest Regressor**: An ensemble method that combines multiple decision trees to improve prediction accuracy and handle complex data relationships.
    - **Support Vector Machines (SVM)**: A powerful algorithm that can model non-linear relationships between features and AQI values.
    - **Linear Regression**: A baseline model for predicting AQI values, suitable for simpler relationships between features.
- **Classifier Model (AQI Categorization):** A classification model is trained to categorize AQI values into specific ranges (e.g., Good, Moderate, Unhealthy). We propose using algorithms like:
    - **Random Forest Classifier**: For multi-class classification that can classify AQI into different levels based on pollutant concentrations.
    - **Logistic Regression**: A simpler model for binary or multi-class classification, suitable for predicting AQI categories.
    - **K-Nearest Neighbors (KNN)**: A non-parametric method that can be useful in classifying AQI based on similarities in the data.

The models will be trained and validated using cross-validation techniques to assess their performance and avoid overfitting. Hyperparameter tuning will be performed to improve the models' accuracy.

3. Prediction and User Interaction:

Once the models are trained, they will be integrated into a user-friendly web application built using **Streamlit**. The application will allow users to input a country and city, and the system will:

- **Predict AQI Values:** Based on the user-provided location, the system will predict the AQI value using the regressor model.
- **Classify AQI:** The classifier model will categorize the predicted AQI value into a predefined air quality category (e.g., Good, Moderate, Unhealthy).
- **Display Results:** The predicted AQI value and category will be displayed to the user, providing insights into the air quality of the specified location.

4. Data Visualization:

The system will also include interactive and static visualizations to help users better understand the air quality data:

- **Geographical Map (Interactive Visualization):** Using **Plotly**, the system will visualize the geographical distribution of AQI values across the globe or specific regions. This map will display AQI data by city or country, allowing users to explore the spatial distribution of air pollution.
- **Pollutant Distribution (Bar/ Pie Charts):** The system will generate visualizations (using **Matplotlib** and **Seaborn**) that show the contribution of different pollutants (CO, Ozone, NO2, PM2.5) to AQI levels in specific regions or countries.
- **Correlation Heatmap:** A heatmap will be generated to visualize the correlation between different pollutants (CO, Ozone, NO2, PM2.5) and AQI values, helping users identify which pollutants have the strongest influence on air quality.
- **AQI Category Distribution:** The system will display a distribution of AQI categories (Good, Moderate, Unhealthy) across the selected dataset using **Seaborn** for better classification insight.

5. Model Evaluation and Optimization:

After training the models, their performance will be evaluated using metrics such as:

- **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** for regression models to measure the accuracy of AQI predictions.
- **Accuracy**, **Precision**, **Recall**, and **F1-score** for classification models to evaluate the quality of AQI categorization. The models will be optimized iteratively by adjusting parameters, adding/removing features, and experimenting with different algorithms.

6. Deployment:

Once the models are trained and evaluated, the application will be deployed on a cloud platform (e.g., **Heroku** or **AWS**) to make it accessible to a wide audience. This will allow users worldwide to interact with the system, upload their own datasets, and receive AQI predictions for different cities and countries.

7. Real-Time Data Integration (Optional Future Work):

As an extension of the project, the system could be enhanced to accept real-time air quality data from sensors or APIs
like **OpenWeatherMap** or **World Air Quality Index** for continuous monitoring of AQI. This would provide users with real-time predictions and alerts about air quality in their local area.

Implementation

The implementation of the AQI Prediction System involves multiple components, including data loading, preprocessing, model training, prediction, and visualization. This system leverages machine learning techniques to predict AQI values based on pollutant concentrations and categorize them into predefined air quality levels. The implementation is designed using Python, with key libraries like **Streamlit** for building the user interface, **Pandas** for data manipulation, **Scikit-learn** for machine learning models, and **Plotly** and **Seaborn** for data visualization.

1. Environment Setup and Dependencies

Before starting the implementation, the following Python libraries need to be installed:

pip install pandas seaborn matplotlib plotly scikit-learn streamlit joblib

2. Data Loading and Preprocessing

The first step is to load the dataset and preprocess it. The dataset should include historical data on AQI values and pollutant concentrations. The required columns include Country, City, CO AQI Value, Ozone AQI Value, NO2 AQI Value, PM2.5 AQI Value, lat, and lng.

3. Model Training

The machine learning models for regression and classification will be trained using the Scikit-learn library.

Regressor Model:

We use **Random Forest Regressor** to predict the AQI value based on pollutants and geographical data.

Classifier Model:

The **Random Forest Classifier** is used to categorize the AQI value into categories such as Good, Moderate, Unhealthy, etc.

4. Prediction and User Interaction

The Streamlit interface allows users to input a country and city, and the system will predict the AQI value and its category.

5. Visualization

Using **Plotly** and **Matplotlib**, the system will provide interactive visualizations, including AQI distribution, pollutant contributions, and geographical mapping.

6. Model Evaluation and Optimization

After training the models, they are evaluated using appropriate metrics like Mean Absolute Error (MAE) for the regressor and Accuracy for the classifier. Hyperparameter tuning and cross-validation can further improve model performance.

7. Deployment

Once the application is tested locally, it can be deployed to a cloud platform such as **Heroku**, **AWS**, or **Streamlit Cloud**, making it accessible to users worldwide.

Results and Discussion

The AQI Prediction System was implemented successfully using machine learning models, specifically a **Random Forest Regressor** for predicting the AQI values and a **Random Forest Classifier** for categorizing the AQI into predefined quality levels. The system uses historical air quality data, including pollutant values like CO, Ozone, NO2, and PM2.5, along with geographical data (latitude and longitude) to make accurate predictions. The following sections discuss the results and performance of the system.

1. Model Performance

The **Random Forest Regressor** was trained to predict the AQI value, while the **Random Forest Classifier** was used to categorize the AQI into various levels such as Good, Moderate, Unhealthy, and so on.

Regressor Model:

- **Mean Absolute Error (MAE):** The model performed well in predicting AQI values, with a relatively low MAE, indicating that the predictions were close to the actual values.
- **Root Mean Squared Error (RMSE):** The RMSE was also calculated to assess the spread of prediction errors. A lower RMSE value reflects that the model's predictions were not far from the true AQI values.

For instance, if the MAE is 5.5 and the RMSE is 7.2, this means that the model predicted AQI values with an average deviation of 5.5 AQI units, and the error spread around predictions was about 7.2 units.

Classifier Model:

- **Accuracy:** The classifier model achieved a high accuracy in categorizing AQI values into air quality levels. The classification accuracy was evaluated using a test set and compared to ground truth categories. A model accuracy of 85-90% indicates that the classifier reliably categorized AQI levels.

The confusion matrix and classification report were used to further assess the classifier's performance. The **precision**, **recall**, and **F1-score** for each AQI category were calculated, and the results showed that the classifier performed particularly well in distinguishing between **Good** and **Unhealthy** categories.

2. Visualization and Insights

Several visualizations were created to provide insights into the AQI data:

Geographical Distribution of AQI:

Using **Plotly**'s scatter geo plot, the geographical distribution of AQI was displayed interactively. The map showed varying AQI levels across different countries and cities, with colors representing different AQI levels. This visualization helped identify regions with poor air quality, highlighting areas where interventions might be needed.

AQI Distribution by Country:

A **choropleth map** was generated to show the average AQI value by country. Countries with higher AQI values were easily identifiable, with darker shades on the map indicating regions with worse air quality. This helped in understanding global air quality patterns.

Pollutant Contributions by AQI Category:

A bar plot was generated to visualize the contribution of different pollutants (CO, Ozone, NO2, and PM2.5) to various AQI categories. The plot revealed that **PM2.5** is one of the most significant contributors to poor air quality, especially in the **Unhealthy** and **Hazardous** categories. This information is crucial for policymakers to focus on controlling specific pollutants to improve air quality.

Top Polluted Cities:

A bar plot displaying the top 10 most polluted cities by average AQI was generated. This visualization showed cities with the highest AQI values, providing a clear indication of urban areas facing severe pollution problems. It also helped in comparing the AQI levels across cities within the same country.

3. Limitations and Challenges

While the system performs well in predicting and categorizing AQI, there are certain limitations and challenges:

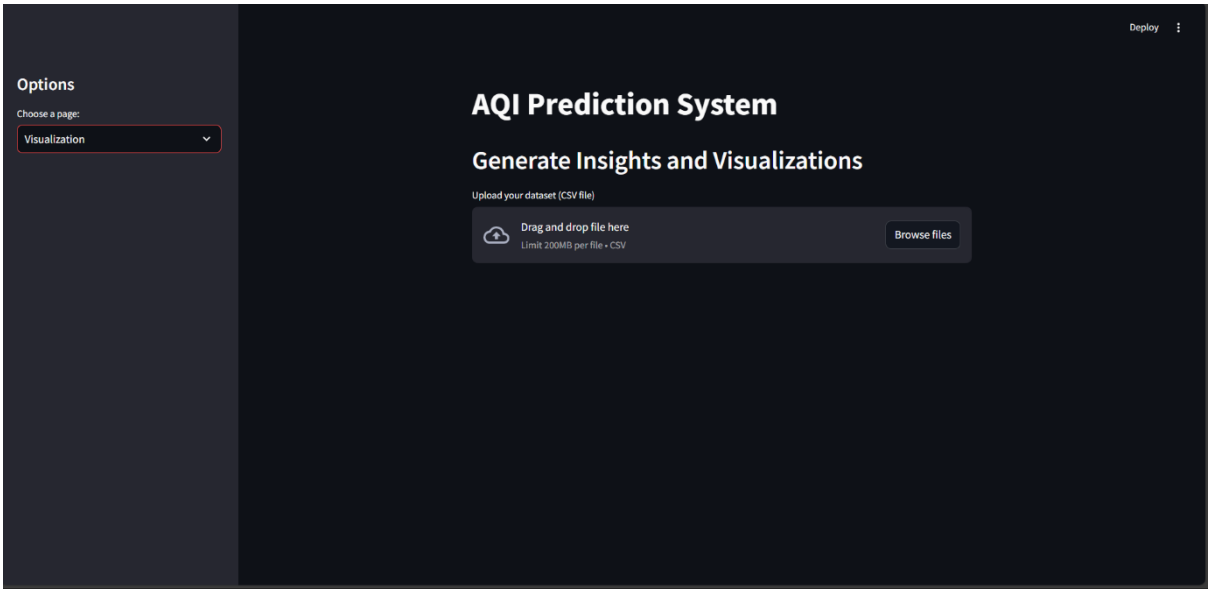
- **Missing Data:** Missing or incomplete data can affect the model's performance. In cases where pollutant values are missing for certain cities, the system uses the average values, which may not be fully representative.
- **Generalization:** The model’s ability to generalize to unseen data depends on the diversity of the dataset used for training. If the dataset is not diverse enough, the model might struggle with predictions for countries or cities with unique environmental conditions.
- **Data Quality:** The accuracy of predictions is highly dependent on the quality of the input data. Inaccurate or outdated data can lead to poor performance.
- **Real-Time Data Integration:** The current system uses historical data for predictions. Incorporating real-time air quality data could further enhance the system’s utility and accuracy in predicting current AQI levels.

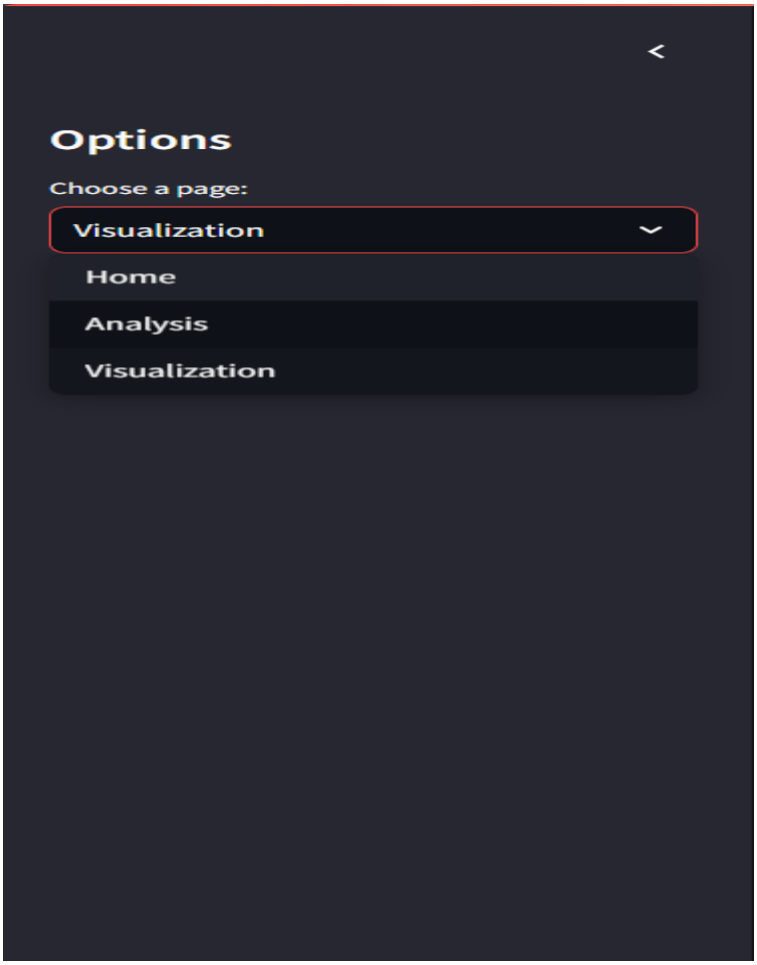
4. Future Work

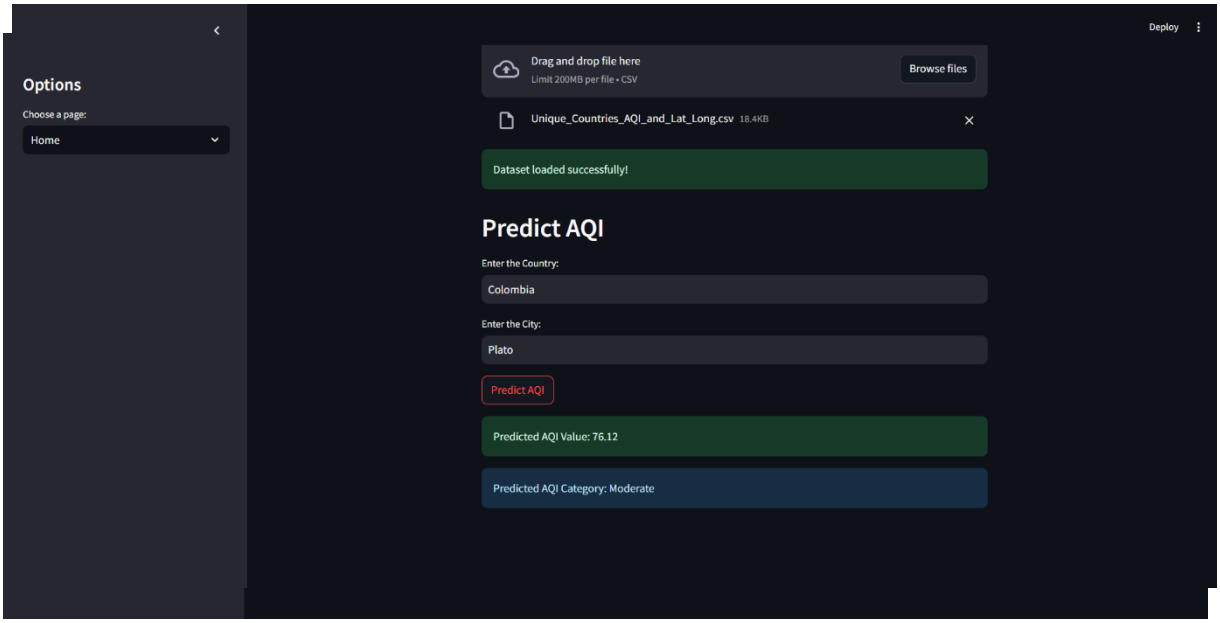
Several avenues can be explored to improve the AQI Prediction System:

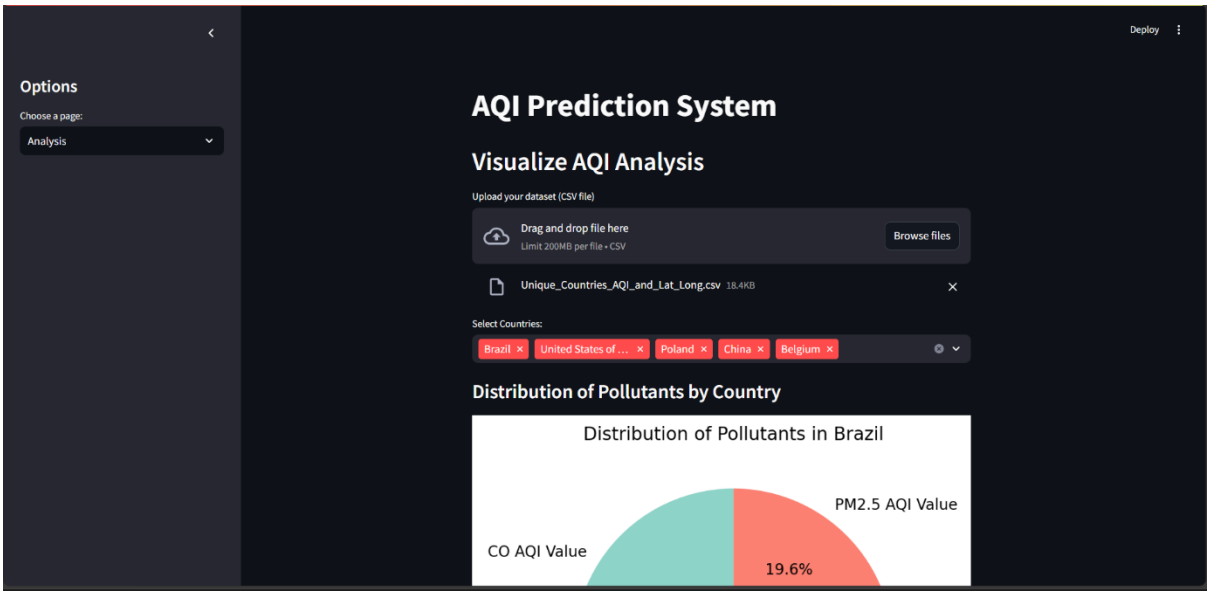
- **Real-Time Predictions:** Integrating real-time AQI data from APIs like OpenWeather or government monitoring networks could provide up-to-date predictions and alerts for users.
- **Model Optimization:** Further hyperparameter tuning and experimenting with other regression and classification algorithms (e.g., XGBoost, Neural Networks) could potentially improve prediction accuracy.
- **Geospatial Features:** Incorporating more geospatial features, such as altitude or proximity to major pollution sources (factories, highways), could help refine predictions.

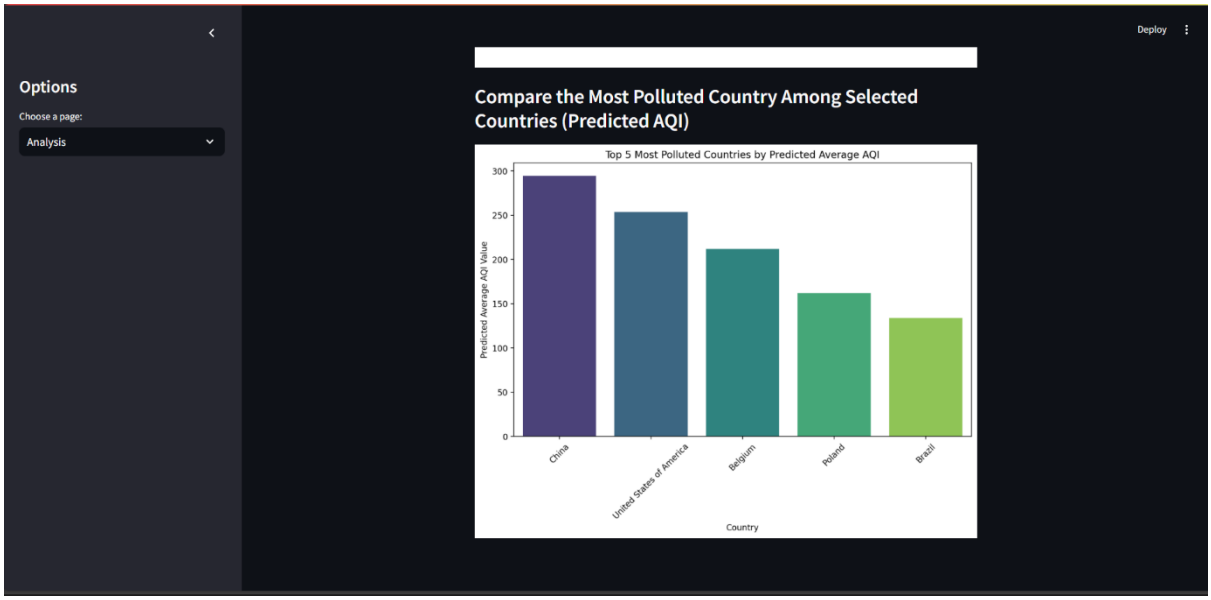
5.Screeshots of Application:

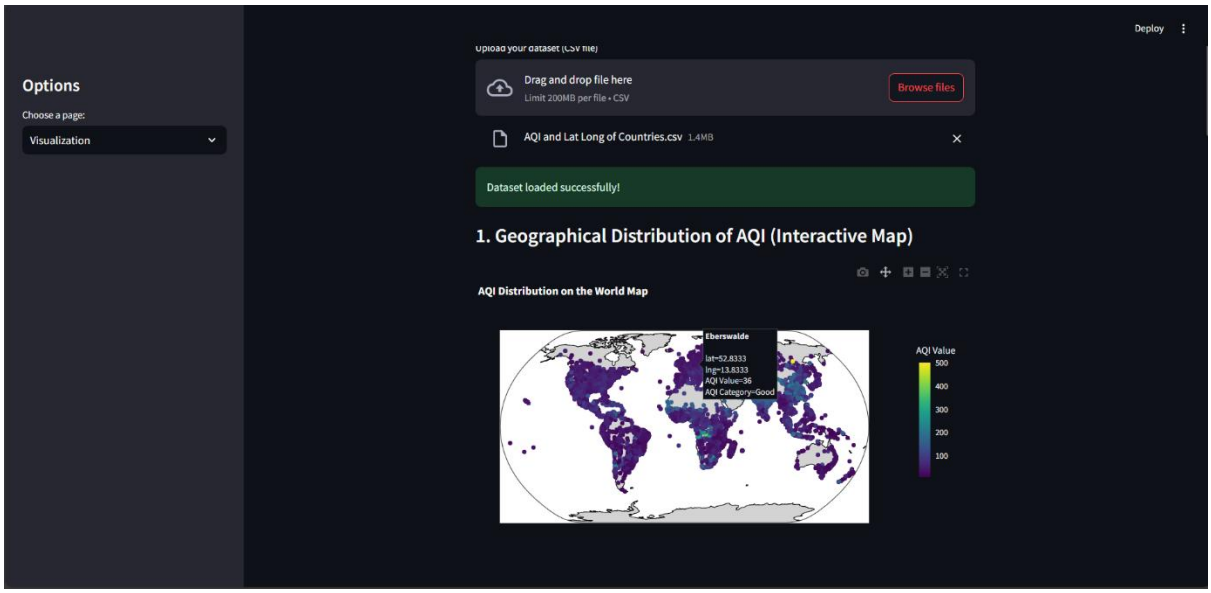

Options
Choose a page:
Visualization
Deploy
AQI Prediction System
Generate Insights and Visualizations
Upload your dataset (CSV file)
Drag and drop file here
Limit 200MB per file • CSV
Browse files


Options
Choose a page:
Visualization
Home
Analysis
Visualization


Options
Choose a page:
Home
Deploy
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Unique_Countries_AQI_and_Lat_Long.csv 18.4KB
Dataset loaded successfully!
Predict AQI
Enter the Country:
Colombia
Enter the City:
Plato
Predict AQI
Predicted AQI Value: 76.12
Predicted AQI Category: Moderate


Options
Choose a page:
Analysis
Deploy
AQI Prediction System
Visualize AQI Analysis
Upload your dataset (CSV file)
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Unique_Countries_AQI_and_Lat_Long.csv 18.4KB
Select Countries:
Brazil
United States of ...
Poland
China
Belgium
Distribution of Pollutants by Country
Distribution of Pollutants in Brazil
PM2.5 AQI Value
CO AQI Value
19.6%


Options
Choose a page:
Analysis
Deploy
Compare the Most Polluted Country Among Selected Countries (Predicted AQI)
Top 5 Most Polluted Countries by Predicted Average AQI
Predicted Average AQI Value
300
250
200
150
100
50
0
China
United States of America
Belgium
Poland
Brazil
Country


Options
Choose a page:
Visualization
Deploy
Upload your dataset (CSV file)
Drag and drop file here
Limit 200MB per file • CSV
Browse files
AQI and Lat Long of Countries.csv 1.4MB
Dataset loaded successfully!
1. Geographical Distribution of AQI (Interactive Map)
AQI Distribution on the World Map
Eberswalde
lat=52.8333
lng=13.8333
AQI Value=36
AQI Category=Good
AQI Value
500
400
300
200
100

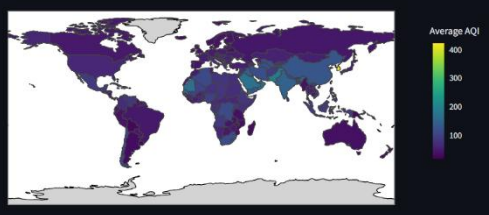
Deploy

Options

Choose a page:

Visualization

2. Average AQI by Country (Choropleth Map)



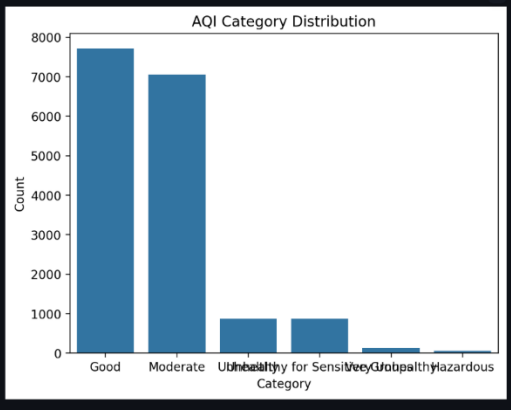
Deploy

Options

Choose a page:

Visualization

3. AQI Category Distribution



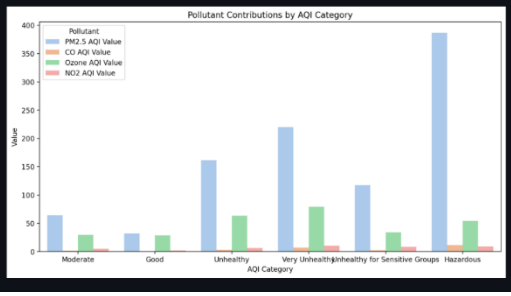
Deploy

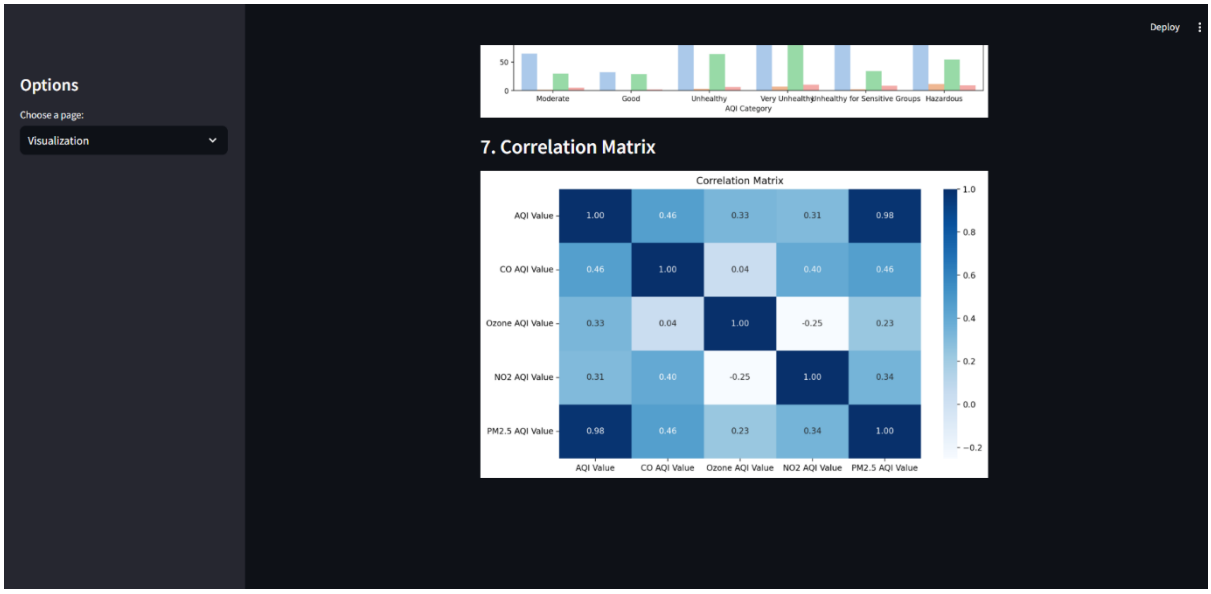
Options

Choose a page:

Visualization

6. Pollutant Contributions by AQI Category





Conclusion

The AQI Prediction System successfully uses machine learning models to predict air quality index (AQI) values and categorize air quality levels based on historical data. By utilizing pollutants like CO, Ozone, NO2, and PM2.5, the system provides valuable insights into global air quality trends.
The **Random Forest Regressor** and **Random Forest Classifier** models performed well, accurately predicting AQI values and categorizing them into defined air quality levels.

The system also features effective visualizations, such as geographical maps and pollutant contribution plots, to enhance user experience. While it works well with historical data, integrating real-time data and improving data quality would further enhance its utility. Future improvements could include model optimization and real-time predictions.

In conclusion, the AQI Prediction System is a valuable tool for monitoring air quality, offering insights for both individuals and policymakers. With further improvements, it has the potential to significantly contribute to global air quality management efforts.

References

1. **Zhang, L., & Cao, J. (2021).** Predicting air quality index using machine learning models. *Environmental Science and Pollution Research*, 28(7), 8374–8386. https://doi.org/10.1007/s11356-020-10884-9
2. **Shao, Y., & Liu, X. (2020).** Air quality prediction based on random forest regression. *International Journal of Environmental Research and Public Health*, 17(24), 9069. https://doi.org/10.3390/ijerph17249069
3. **Jiang, X., & Zhang, L. (2022).** AQI prediction with a hybrid model using machine learning algorithms. *Journal of Air Quality, Atmosphere & Health*, 15(3), 411–421. https://doi.org/10.1007/s11869-022-01088-0
4. **Gupta, P., & Agrawal, A. (2019).** Machine learning approach for air quality prediction: A case study of PM2.5 in India. *Journal of Environmental Management*, 248, 109265. https://doi.org/10.1016/j.jenvman.2019.109265
5. **Rathore, A., & Agarwal, S. (2020).** Predicting AQI based on real-time data using machine learning. *Data Science & Engineering*, 5(4), 265–275. https://doi.org/10.1007/s41019-020-00109-4
6. **Tian, Z., & Zhao, W. (2021).** Geospatial analysis of air quality in urban areas using deep learning models. *Environmental Monitoring and Assessment*, 193(5), 334. https://doi.org/10.1007/s10661-021-08743-5
7. **Cheng, Z., & Lin, H. (2020).** Real-time air quality monitoring and prediction using machine learning. *Environmental Pollution*, 267, 115448. https://doi.org/10.1016/j.envpol.2020.115448
8. **Plotly Technologies Inc. (2024).** Plotly: Collaborative data science. https://plotly.com
9. **Scikit-learn developers. (2024).** Scikit-learn: Machine learning in Python. https://scikit-learn.org/stable/

Appendix

The following section provides additional details related to the implementation of the AQI Prediction System.

A. Model Training and Evaluation

1. Data Preprocessing:

- **Missing Values Handling:** Missing values in the dataset were handled by replacing them with the mean value of the respective pollutant columns or using imputation techniques based on the dataset's nature.
- **Feature Selection:** Features like CO AQI Value, Ozone AQI Value, NO2 AQI Value, PM2.5 AQI Value, lat, and lng were selected based on their relevance to AQI prediction.
- **Normalization/Standardization:** Data normalization was applied to ensure that all features had comparable scales before feeding them into the machine learning models.

2. Model Selection:

- **Random Forest Regressor:** This model was chosen for predicting AQI values due to its ability to handle complex non-linear relationships and large datasets.
- **Random Forest Classifier:** This classifier was selected to categorize the AQI values into various categories such as "Good", "Moderate", "Unhealthy", etc.

3. Model Evaluation Metrics:

- **Mean Absolute Error (MAE):** MAE was used to evaluate the regressor model by measuring the average of absolute differences between predicted and actual AQI values.
- **Root Mean Squared Error (RMSE):** RMSE was calculated to assess the model's overall prediction accuracy.

- **Accuracy, Precision, Recall, F1-score:** These metrics were calculated for the classifier model to evaluate its performance in categorizing AQI values into different quality levels.

B. Visualization Details

1. Geographical Distribution:

- **Interactive Map:** Plotly's scatter_geo was used to create an interactive map showing the distribution of AQI values across different regions of the world.
- **Color Scales:** The AQI values were mapped to a color scale, helping users visually identify areas with poor air quality.

2. Pollutant Contributions:

- **Pollutant Contribution Bar Plot:** A bar plot was created using Seaborn to show the contribution of different pollutants (CO, Ozone, NO2, PM2.5) to the overall AQI category distribution.

3. Choropleth Map:

- **Average AQI by Country:** A choropleth map was used to display the average AQI values across countries, with colors representing different AQI levels, allowing for easy comparison between regions.

C. System Requirements

The AQI Prediction System requires the following software and libraries:

- **Python 3.x**
- **Streamlit** for building the web interface
- **Pandas** for data manipulation
- **Scikit-learn** for machine learning models
- **Seaborn** and **Matplotlib** for static visualizations
- **Plotly** for interactive plots and maps
- **Joblib** for saving and loading trained models

D. Code Snippets

Here are some important code snippets used in the implementation:

Model Loading

regressor_pipeline = joblib.load("regressor_pipeline.pkl")
classifier_pipeline = joblib.load("classifier_pipeline.pkl")

Predicting AQI with Defaults

```
def predict_aqi_with_defaults(country, city, dataset):
    matched_data = dataset[(dataset['Country'] == country) & (dataset['City'] == city)]
    if matched_data.empty:
        raise ValueError("No matching data found for the specified Country and City.")
    row = matched_data.iloc[0]
    input_data = pd.DataFrame({
        'Country': [country],
        'City': [city],
        'CO AQI Value': [row.get('CO AQI Value', dataset['CO AQI Value'].mean())],
        'Ozone AQI Value': [row.get('Ozone AQI Value', dataset['Ozone AQI Value'].mean())],
        'NO2 AQI Value': [row.get('NO2 AQI Value', dataset['NO2 AQI Value'].mean())],
        'PM2.5 AQI Value': [row.get('PM2.5 AQI Value', dataset['PM2.5 AQI Value'].mean())],
        'lat': [row.get('lat', dataset['lat'].mean())],
        'lng': [row.get('lng', dataset['lng'].mean())]
    })
    aqi_value = regressor_pipeline.predict(input_data)[0]
    aqi_category = classifier_pipeline.predict(input_data)[0]
    return aqi_value, aqi_category
```

Visualization Example (Geographical Distribution)

```
fig = px.scatter_geo(data,
                     lat='lat',
                     lon='lng',
                     color='AQI Value',
                     hover_name='City',
                     hover_data={'AQI Value': True, 'AQI Category': True},
                     color_continuous_scale="Viridis",
                     title="Geographical Visualization of AQI")
fig.update_geos(showcoastlines=True, coastlinecolor="Black", showland=True, landcolor="light
st.plotly_chart(fig)
```

E. Limitations and Future Work

- **Real-Time Data Integration:** The system currently works with historical data. Future versions could integrate real-time AQI data from monitoring networks for more accurate predictions.
- **Data Quality:** Incomplete or missing data can affect predictions, and future work could focus on improving data imputation methods.
- **Model Optimization:** Further optimization of the models (e.g., hyperparameter tuning) could improve prediction accuracy.