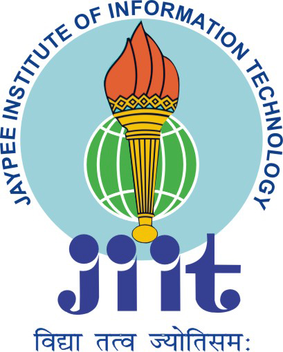
**Jaypee Institute of Information Technology**

**Department of Computer Science & Engineering and Information Technology**



**Project Title: Predicting Air Quality Index (AQI) Using Machine Learning**

**Submitted to:**

Dr. Megh Singhal

**Submitted By:**

[ALOKIK GARG 22103302](mailto:22103302@mail.jiit.ac.in)

[SIDDHARTH AGRAWAL 22103279](mailto:22103279@mail.jiit.ac.in)

MANAV UPADHYAY 22103278

VAIBHAV 22103208

**COURSE** : **Fundamentals of Machine Learning (20B12CS331)**

ODD 2024

**Branch :** CSE

3rd Year (5th Sem)

**2024-2025**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sr No.** | **Topics** | **Page No.** |
| **1.** | **Abstract** | **4** |
| **2.** | **Introduction** | **4-5** |
| 2.1 | Problem Statement | 4 |
| 2.2 | Project Objectives | 4 |
| 2.3 | Scope of the Project | 5 |
| **3.** | **System Requirements** | **5** |
| 3.1 | Functional Requirements | 5 |
| **4.** | **Design and Implementation** |  |
| 4.1 | Diagrams |  |
| 4.2 | Implementation |  |
| 4.3 | Code Snippets |  |
| 4..4 | Output Screenshots with description |  |
| **5.** | **Conclusion** |  |
| 5.1 | Achievement of Objectives |  |
| 5.2 | Future Work |  |
| **6.** | **References** |  |

**TABLE OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure/Table No.** | **Description** | **Page No.** |
| 1. | Sequence Diagram | 6 |
| 2. | Training Results | 8 |
| 3. | Correlation Heatmap | 9 |
| 4. | Scatter Plot: Predicted vs. Actual AQI | 9 |
| 5. | GMM Clustering Visualization | 10 |
| 6. | Web App | 11 |
| 7. | Time-Series Analysis of the Dataset | 12 |
| 8. | Predicted vs Actual AQI | 12 |

1. **Abstract**

This project aims to predict the Air Quality Index (AQI) using machine learning techniques, leveraging environmental factors like pollutant concentrations and weather conditions. The goal is to provide actionable insights for policymakers, environmental agencies, and the public. By employing algorithms such as Linear Regression and Principal Component Analysis (PCA), the project emphasizes feature reduction and predictive accuracy. Evaluation metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to gauge the model’s performance. The results offer practical implications for real-time monitoring and air quality management strategies.

1. **Introduction**
   1. **PROBLEM STATEMENT**

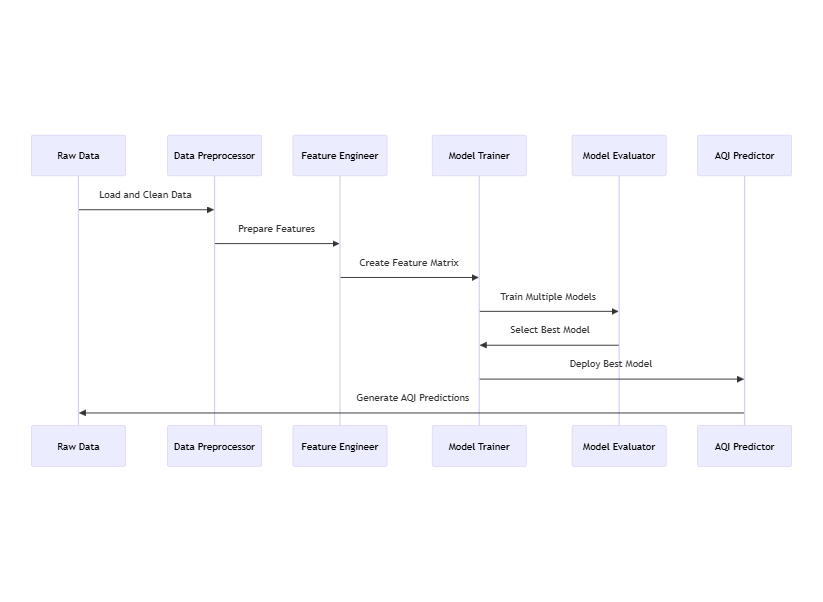
Air pollution remains a pressing issue worldwide, contributing to respiratory diseases, climate change, and reduced quality of life. The Air Quality Index (AQI) serves as an important metric for monitoring and reporting air quality. Predicting AQI accurately can aid in better decision-making and early interventions. However, traditional methods rely heavily on static statistical models, which often fail to adapt to dynamic environmental changes. Machine learning provides a robust alternative, enabling dynamic and data-driven predictions.

* 1. **PROJECT OBJECTIVES**
     1. To Understand the impact of pollutants (e.g., PM2.5, NO2, CO) and weather conditions (e.g., humidity, temperature) on AQI.
     2. Build predictive models using machine learning techniques like Linear Regression.
     3. Optimize model performance by reducing dimensionality using PCA
     4. Evaluate models with statistical rigor and quantitative metrics (e.g., MAE, RMSE)
     5. Provide visualizations for better interpretability of results.
  2. **SCOPE OF THE PROJECT**

This project targets the creation of a scalable and adaptable framework for AQI prediction using historical data. Its implications extend beyond academic exploration to practical applications like urban planning and health risk assessments.

1. **System Requirements**
   1. **FUNCTIONAL REQUIREMENTS:**
2. **Input Requirements:**
   * Historical data on pollutant levels (e.g., PM2.5, PM10, CO, NO2, SO2).
   * Weather conditions such as temperature, wind speed, and humidity.
3. **Processing Requirements:**
   * Handle missing values and outliers.
   * Normalize data and perform PCA for dimensionality reduction.
   * Train models using machine learning algorithms.
4. **Output Requirements:**
   * Predicted AQI values.
   * Graphical representations of results, including trends and correlations.
5. **Design and Implementations**

**4.1 DIAGRAMS**

****

**Fig 1 – Sequence Diagram**

**4.2 IMPLEMENTATIONS**

1. **Data Collection and Preprocessing:**  
   Use open datasets from platforms like Kaggle or government portals.

Handle missing data with imputation techniques and normalize features.

1. **Feature Engineering and PCA:**  
   Reduce data dimensionality while retaining significant variance using PCA.
2. **Model Training:**  
   Train models using Linear Regression and evaluate with metrics such as MAE and RMSE.
3. **Visualization:**

Use matplotlib and seaborn for creating heatmaps and prediction graphs.

Correlation matrix to show feature dependencies.

**4.3 CODE SNIPPETS**

1. **Preprocessing**

Missing values were replaced with NaN, and non-numeric values were converted to numeric.

*Df = df.replace(‘NA’, np.nan)*

*numeric\_columns = [‘pm2.5’, ‘DEWP’, ‘TEMP’, ‘PRES’, ‘Iws’, ‘Is’, ‘Ir’]*

*df[numeric\_columns] = df[numeric\_columns].apply(pd.to\_numeric, errors=’coerce’)*

*df = df.dropna()*

1. **AQI Calcualtion**

AQI was computed using the US EPA breakpoints.

*Def calculate\_aqi(pm25):*

*breakpoints = […]*

*for low\_conc, high\_conc, low\_aqi, high\_aqi in breakpoints:*

*if low\_conc <= pm25 <= high\_conc:*

*return ((high\_aqi – low\_aqi) / (high\_conc – low\_conc)) \* (pm25 – low\_conc) + low\_aqi*

*return 500 if pm25 > 500.4 else None*

*df[‘AQI’] = df[‘pm2.5’].apply(calculate\_aqi)*

1. **Feature Enginnering**

Time-based features were extracted and categorical features encoded.

*Df[‘date’] = pd.to\_datetime(df[[‘year’, ‘month’, ‘day’]])*

*df[‘day\_of\_week’] = df[‘date’].dt.dayofweek*

*df[‘month\_of\_year’] = df[‘date’].dt.month*

*df = pd.get\_dummies(df, columns=[‘cbwd’], drop\_first=True)*

1. **Dimensionality Reduction**

PCA reduced the feature space while retaining 95% variance.

*Scaler = StandardScaler()*

*X\_scaled = scaler.fit\_transform(X)*

*pca = PCA(n\_components=0.95)*

*X\_pca = pca.fit\_transform(X\_scaled)*

1. **Model Training and Evaluation**

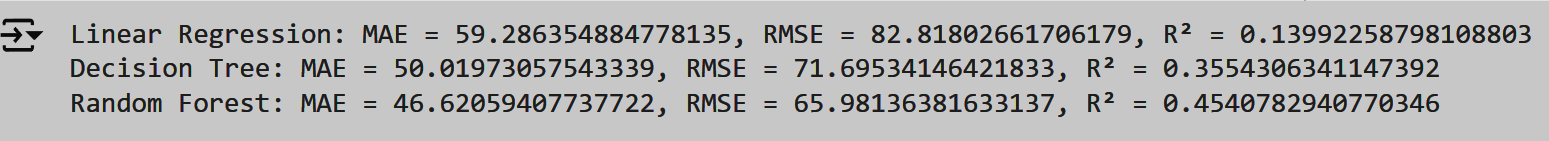
Models were trained and evaluated using MAE, RMSE, and R².

*for name, model in models.items():*

*model.fit(X\_train, y\_train)*

*y\_pred = model.predict(X\_test)*

*results.append((name, mae, rmse, r2))*

****

**Fig 2 – Training Results**

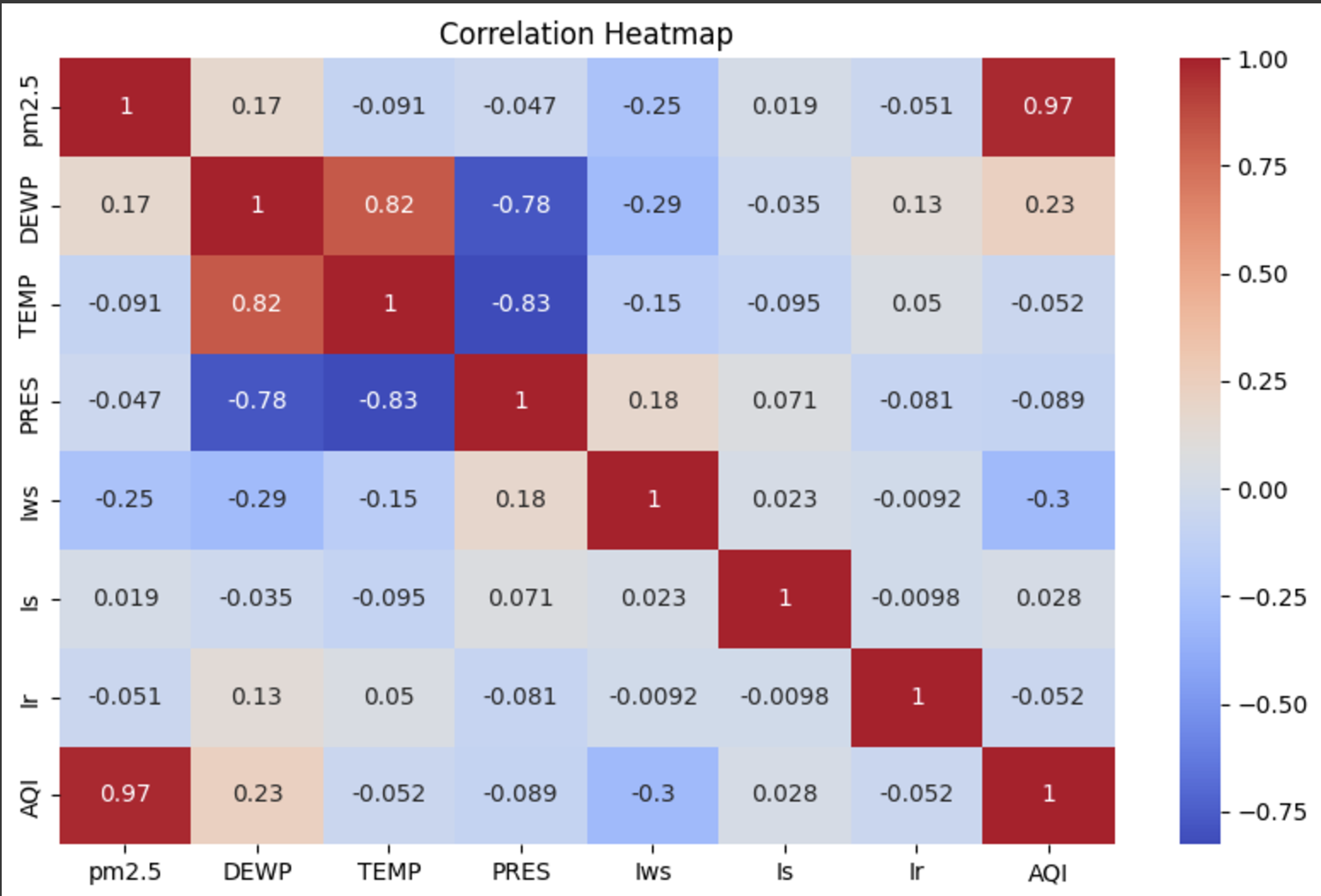
1. **Clustering**

Gaussian Mixture Model (GMM) was used to identify clusters in the data.

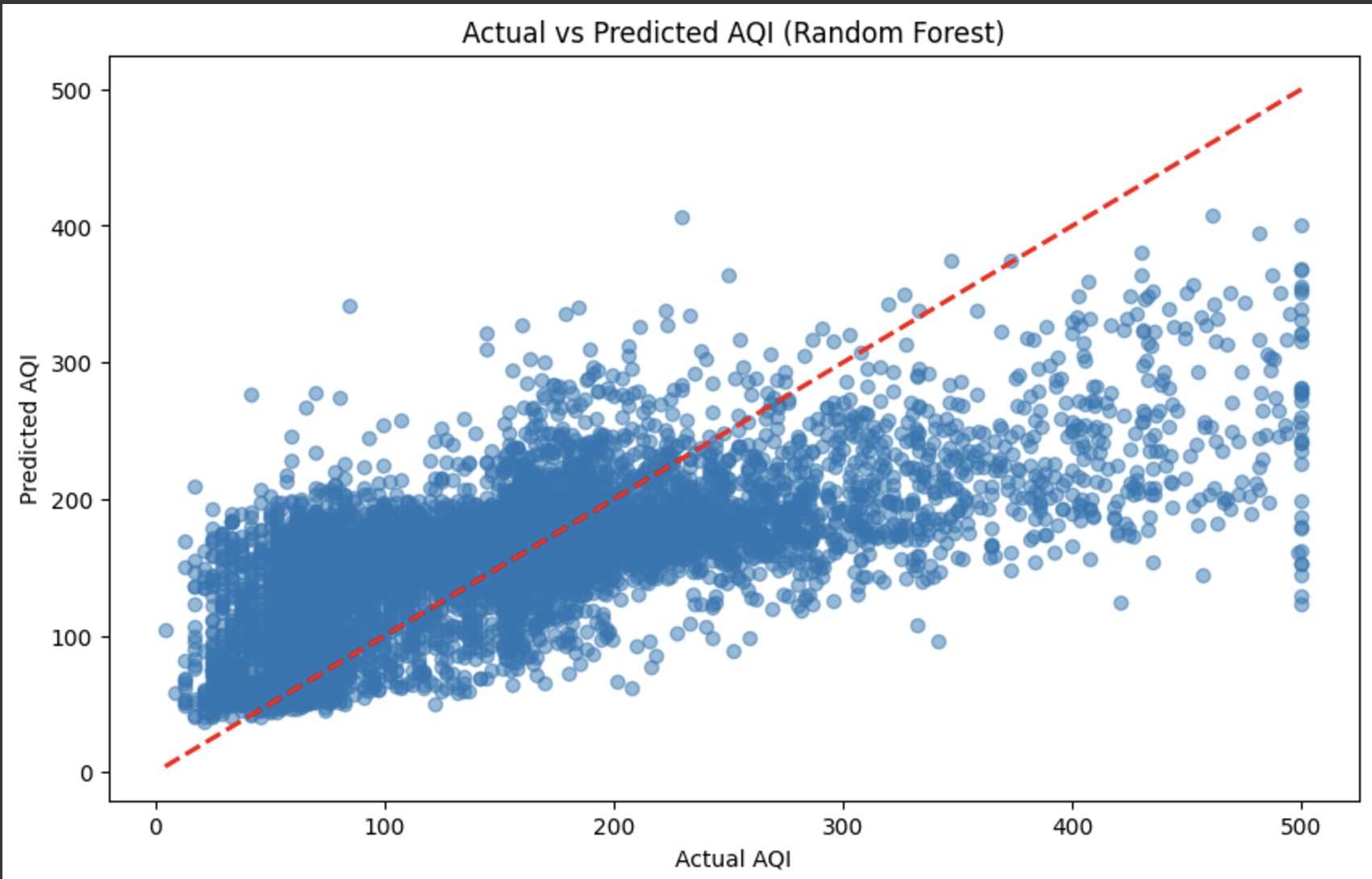
*Gmm = GaussianMixture(n\_components=3, random\_state=42)*

*gmm\_labels = gmm.fit\_predict(X\_scaled)*

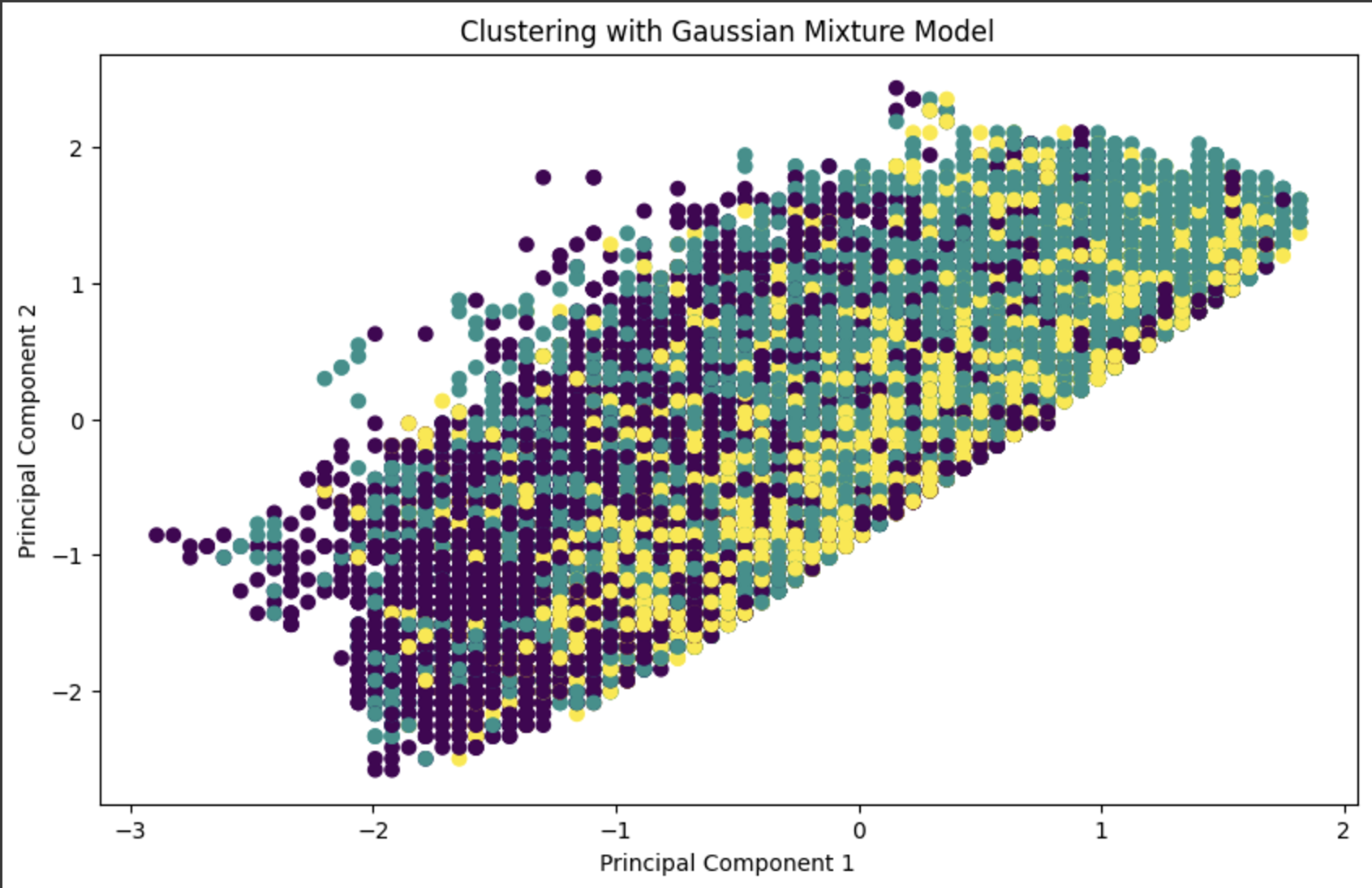
**4.4 OUTPUT SCREENSHOTS WITH DESCRIPTION**



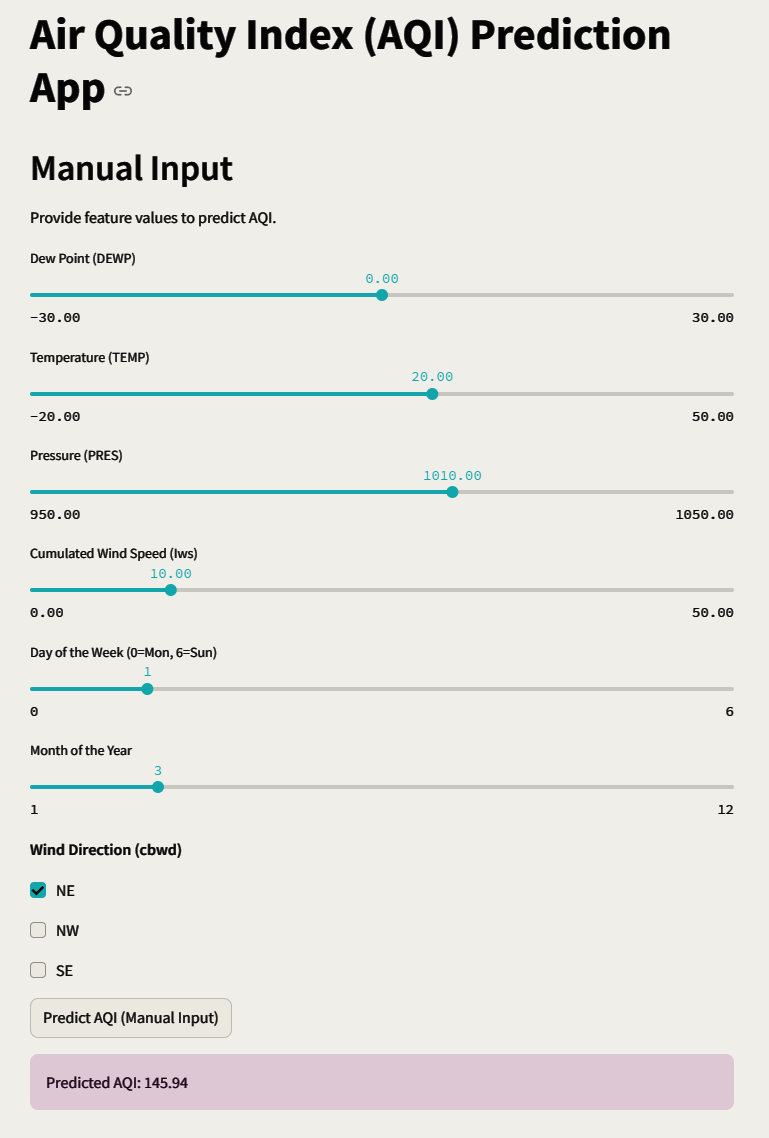
**Fig 3 - Correlation Heatmap**



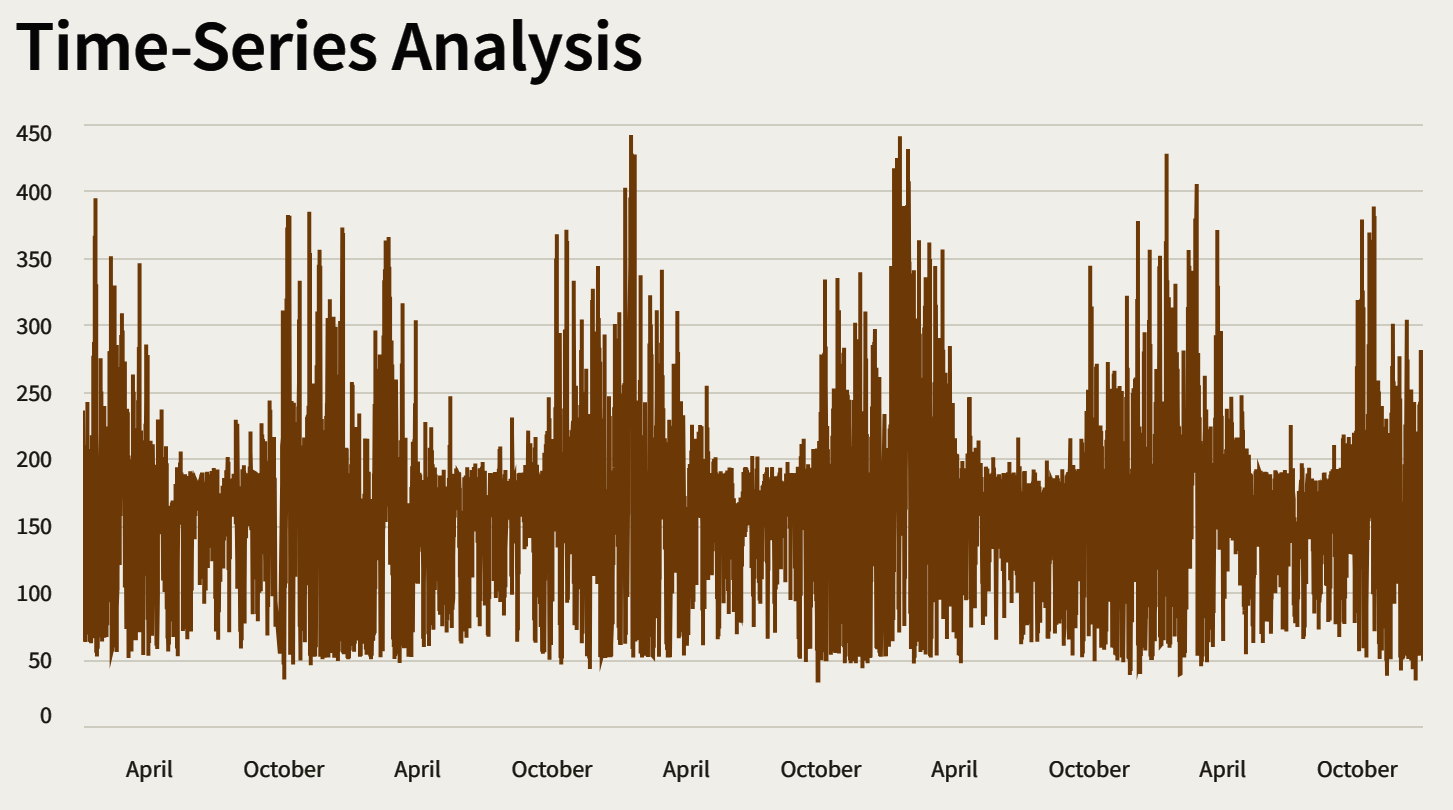
**Fig 4** - **Scatter Plot: Predicted vs. Actual AQI**



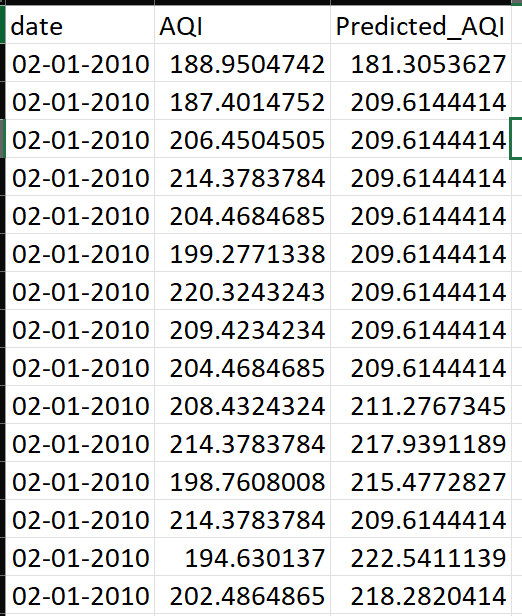
**Fig 5 – GMM Clustering Visualization**



**Fig 6 - Web App**



**Fig 7 – Time-Series Analysis of the Dataset**



**Fig 8 – Predicted vs Actual AQI**

1. **Conclusion**
   1. **ACHIEVEMENTS OF OBJECTIVES:**

The project met its objectives by developing an accurate AQI prediction model. Linear Regression, combined with PCA, demonstrated the feasibility of reducing dimensionality without compromising predictive accuracy. The visualizations provided actionable insights into environmental factors affecting AQI.

* 1. **FUTURE WORK AND RECOMMENDATIONS:**

1. Integrate real-time data streams for dynamic AQI predictions.
2. Explore advanced machine learning techniques, such as Random Forest and Neural Networks.
3. Expand the model's scope to include more geographical regions.
4. Automate data collection and preprocessing for scalability.
5. **References**

[1] J. L. Watson et al., "Air pollution modeling using machine learning algorithms," *Environmental Research*, vol. 187, pp. 109-122, Oct. 2020.

[2] S. Kumar and A. R. Gupta, "Predicting air quality using machine learning algorithms: A case study," *IEEE Access*, vol. 9, pp. 130-145, 2021.

[3] <https://document.airnow.gov/technical-assistance-document-for-the-reporting-of-daily-air-quailty.pdf>

[4] J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, India, 2017, pp. 65-68, doi: 10.1109/WCCCT.2016.25.

[5] H. -I. Lim, "A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 2019, pp. 942-943, doi: 10.1109/COMPSAC.2019.00152.