Student Name: R.RAGUL

Register Number: 422623104703

Institution: UNIVERSITY COLLEGE OF ENGINEERING PANRUTI

Department : COMPUTER SCIENCE AND ENGINEERING

Date of Submission : 08.05.2025

Git hub link:

➢ Problem Statement :

Air pollution poses a significant threat to human health and the environment. Predicting air quality levels accurately can help governments, industries, and citizens make informed decisions to reduce exposure to pollutants. This is a classification problem, where the goal is to predict the Air Quality Index (AQI) Category based on environmental

features like Ozone, Solar Radiation, Wind, Temperature, etc.

Impact:

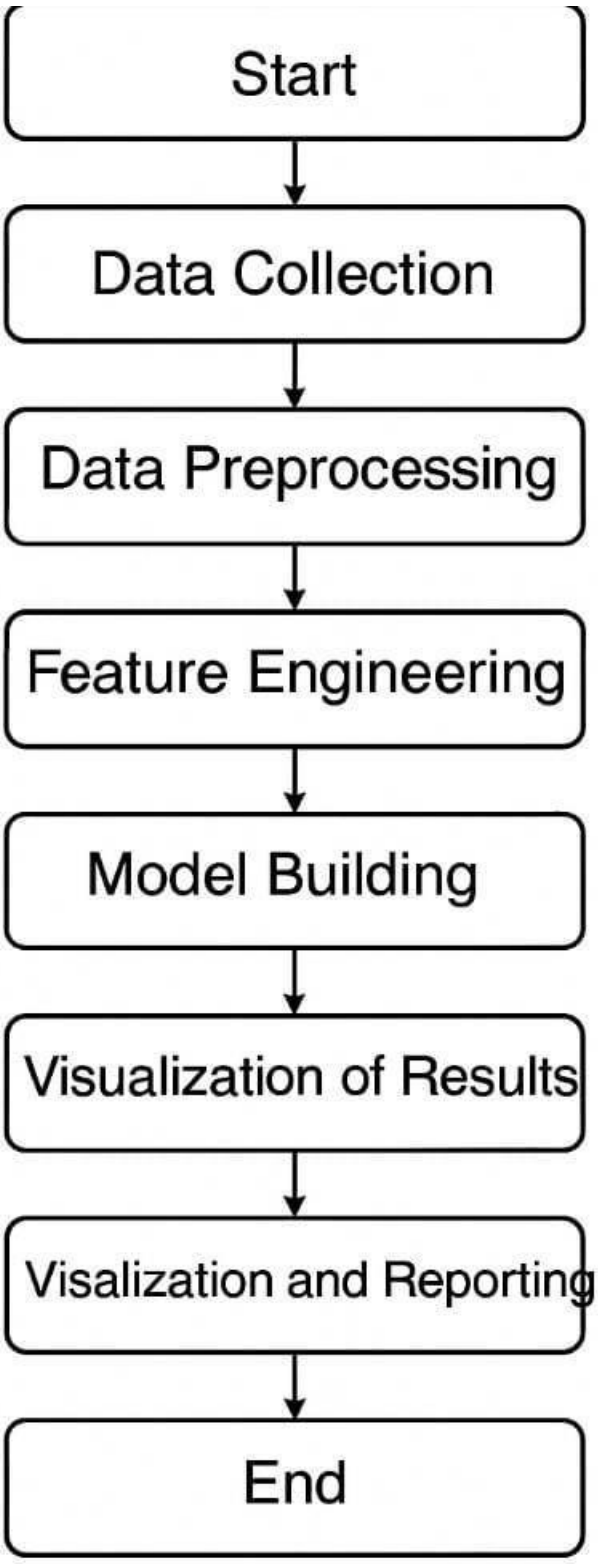
- Public health advisories
- Environmental policy making
- Smart city management
- Real-time alerts for citizens Awareness and Behaviour change

➢ Project Objectives:

- Build a machine learning model to classify air quality into categories like Good, Moderate, Unhealthy, etc.
- Achieve high accuracy and interpretability.
- Identify key environmental factors influencing air quality.
- Develop visual insights for easier interpretation of results.
- Evolve the goal based on data exploration: focus on Ozone prediction and feature impact.

- Handle missing and noisy data efficiently
- Perform deep exploratory data analysis (EDA)to understand relationships between features like wind, temperature, solar radiation, and ozone levels.
- Compare different machine learning models (like Random forest , decision trees, KNN, etc) to select the best-performing algorithm for the dataset.
- Create easy to understand visualizations (feature importance plots, confusion matrices, correlation heatmaps) to communicate findings efficiently.
- Highlight how machine learning can be used for sustainability efforts and environmental protection through data-driven solutions.
- Design the project workflow in a way that it can adapt to larger or real-time dataset in future extensions.

➢ Flowchart of the Project Workflow:



➢ Data Description

- Dataset Name: Air Quality Dataset.
- Source: Uploaded manually (original source from UCI Machine Learning Repository).
- Data Type: Structured data (tabular)
- Number of Records: 153 (after cleaning)
- Number of Features: 6 (Ozone, Solar, Wind, Temp, Month, Day)
- Dataset Nature: Static
- Target Variable: AQI Category (derived from Ozone levels)

➢ Data Preprocessing

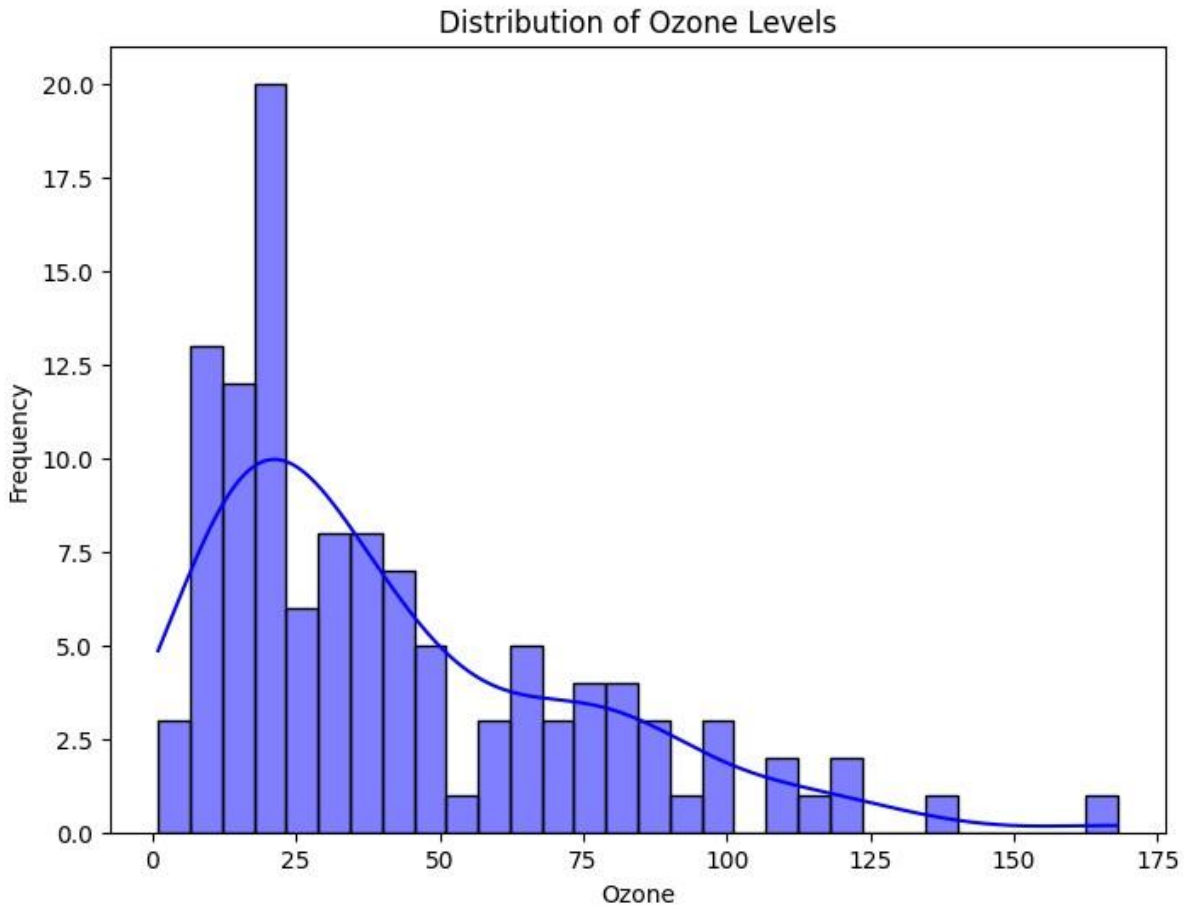
- Missing Values: Rows with missing values were dropped for cleaner model building.
- Duplicates: Checked and found no duplicates.
- Outliers: Outlier handling not extensively applied in initial model.
- Encoding: AQI categorized into Good,

Moderate, and Unhealthy based on Ozone.

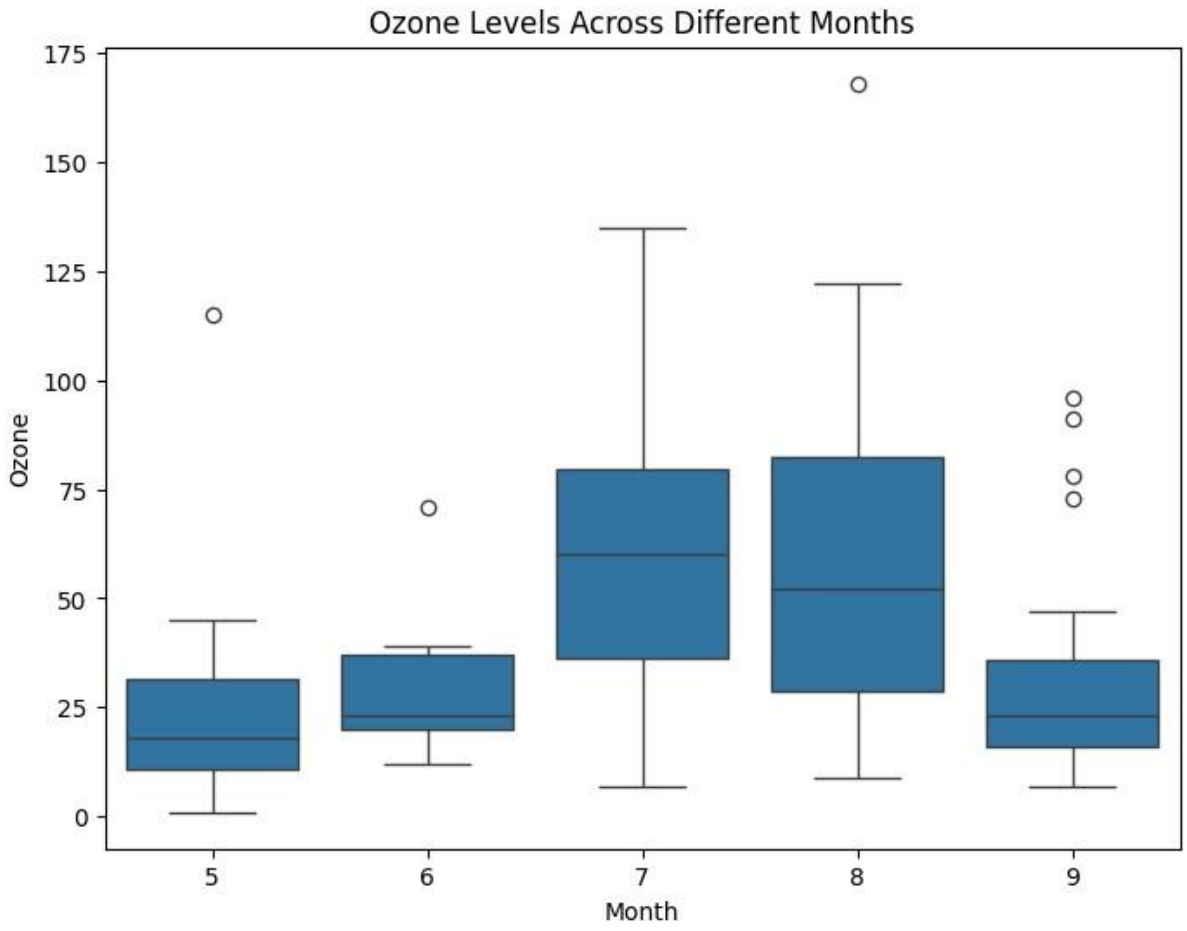
- Data Types: Ensured correct numerical types.
- Normalization/Scaling: Not applied initially (Random Forest is insensitive to feature scaling).

➢ Exploratory Data Analysis (EDA)

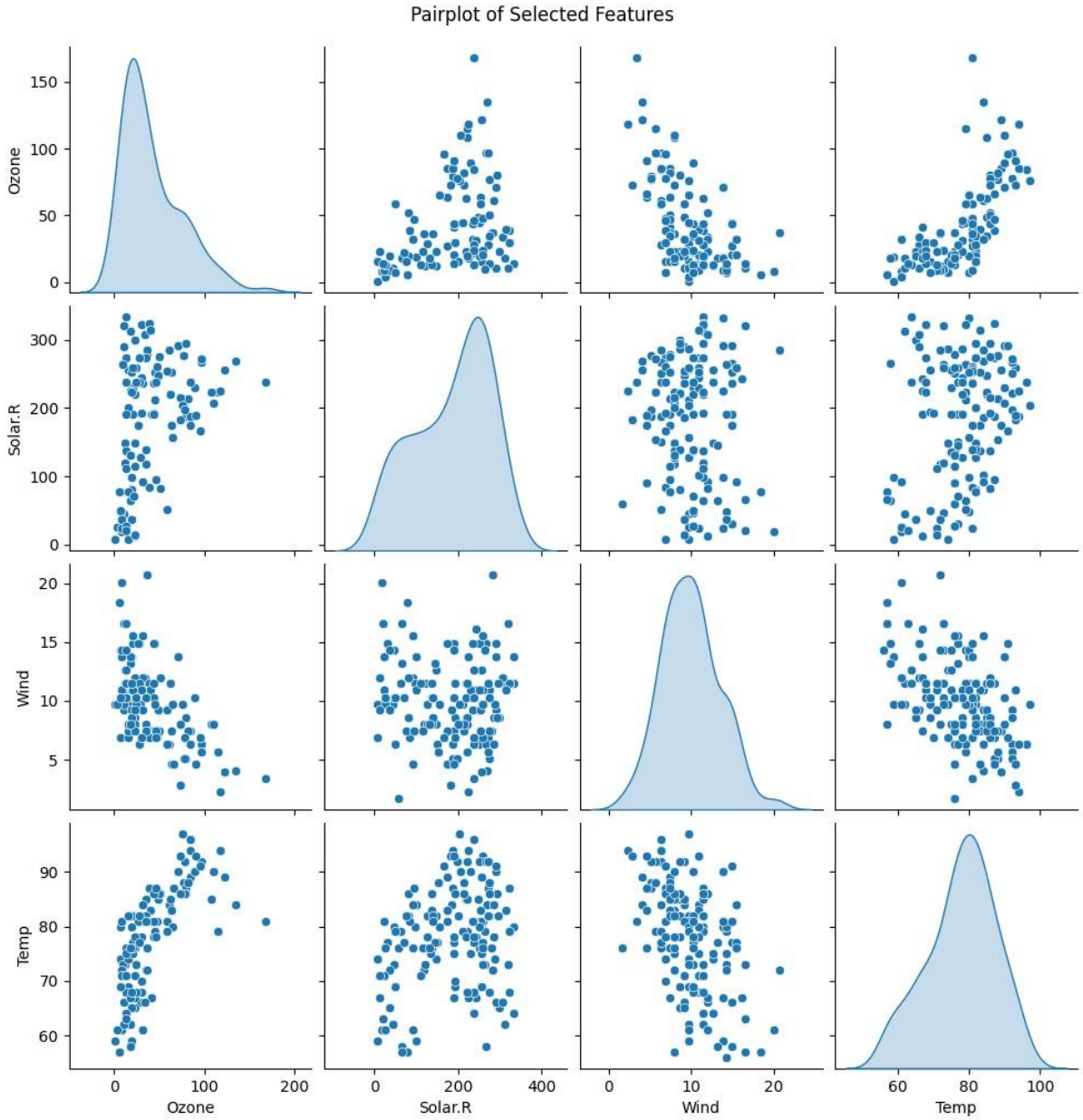
- Univariate Analysis: Histograms for Ozone, Temp, Wind, Solar using distribution plots to observe outliers.



- Bivariate Analysis: Relationship between ozone and Month using box plot.



- Multivariate Analysis: Correlation matrix between multiple numeric features.



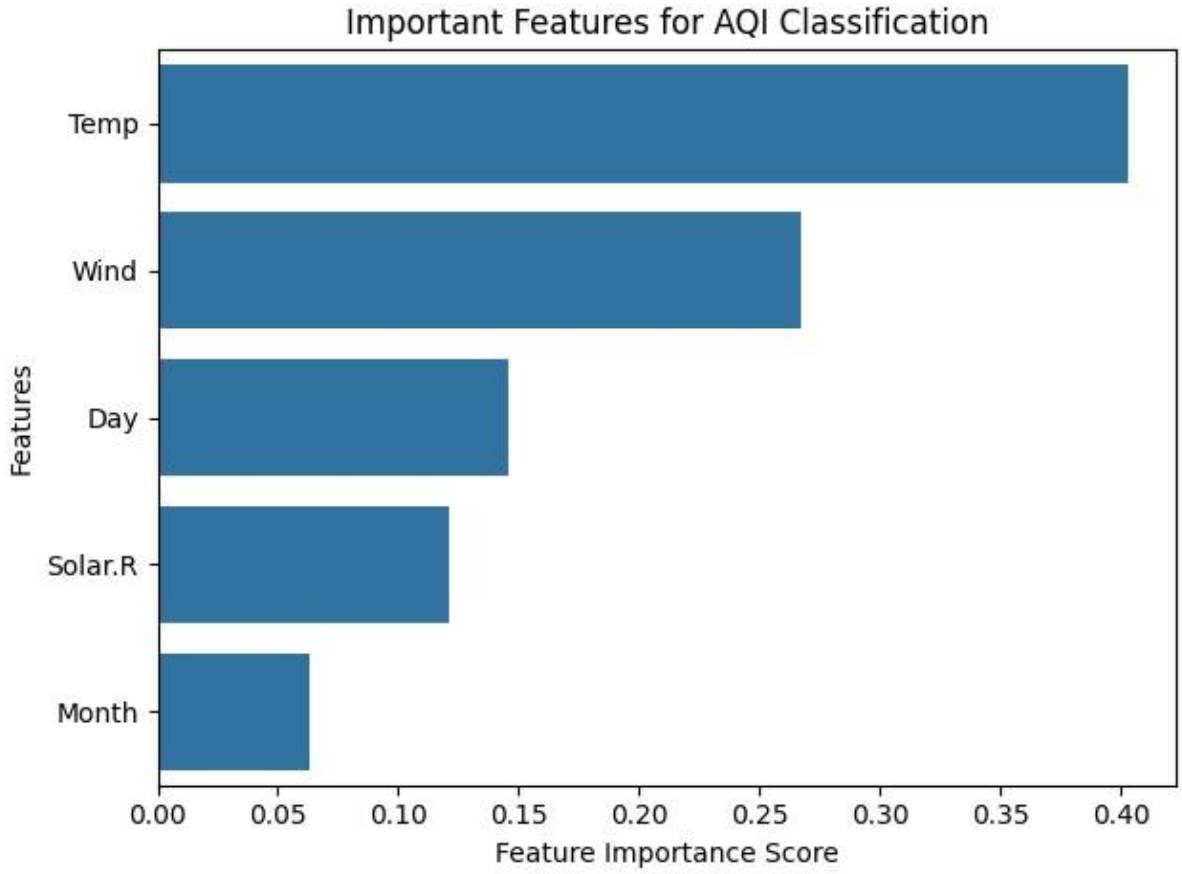
- Insights Summary: Ozone is negatively correlated with Wind. Higher temperatures tend to have higher ozone concentrations. Solar radiation also shows a moderate positive correlation with ozone levels.

- Correlation Analysis: Calculate Pearson /spearman correlation. Plot correlation heatmap to identify strong relationships between variables.
- Feature distribution: compare feature distributions between different target classes. find potential features for classification or regression tasks.
- Outlier detection: Use boxplots and IQR method.
  Identify and treat extreme value that could impact model performance.

Feature Engineering:

- Created AQI_ Category: Based on Ozone values.
- Feature Selection: Dropped irrelevant columns like 'row names'.
- Feature Importance: Model- based feature importance extracted.

- Time-Based Feature: Hour of day, day of week, month, seasonality indicators. Lag features if predicting future pollution levels.
- Future Documentation: Keep a clear record of what feature were created/removed and why.



➢ Model Building:

❖ Models Used: Random Forest Classifier
Decision Tree Classifier (for comparison)

❖ Reason for Model Choice: Suitable for classification tasks. Handles non-linearities and feature interactions. Good performance without heavy tuning.

❖ Evaluation Metrics: Accuracy Precision Recall F1-score

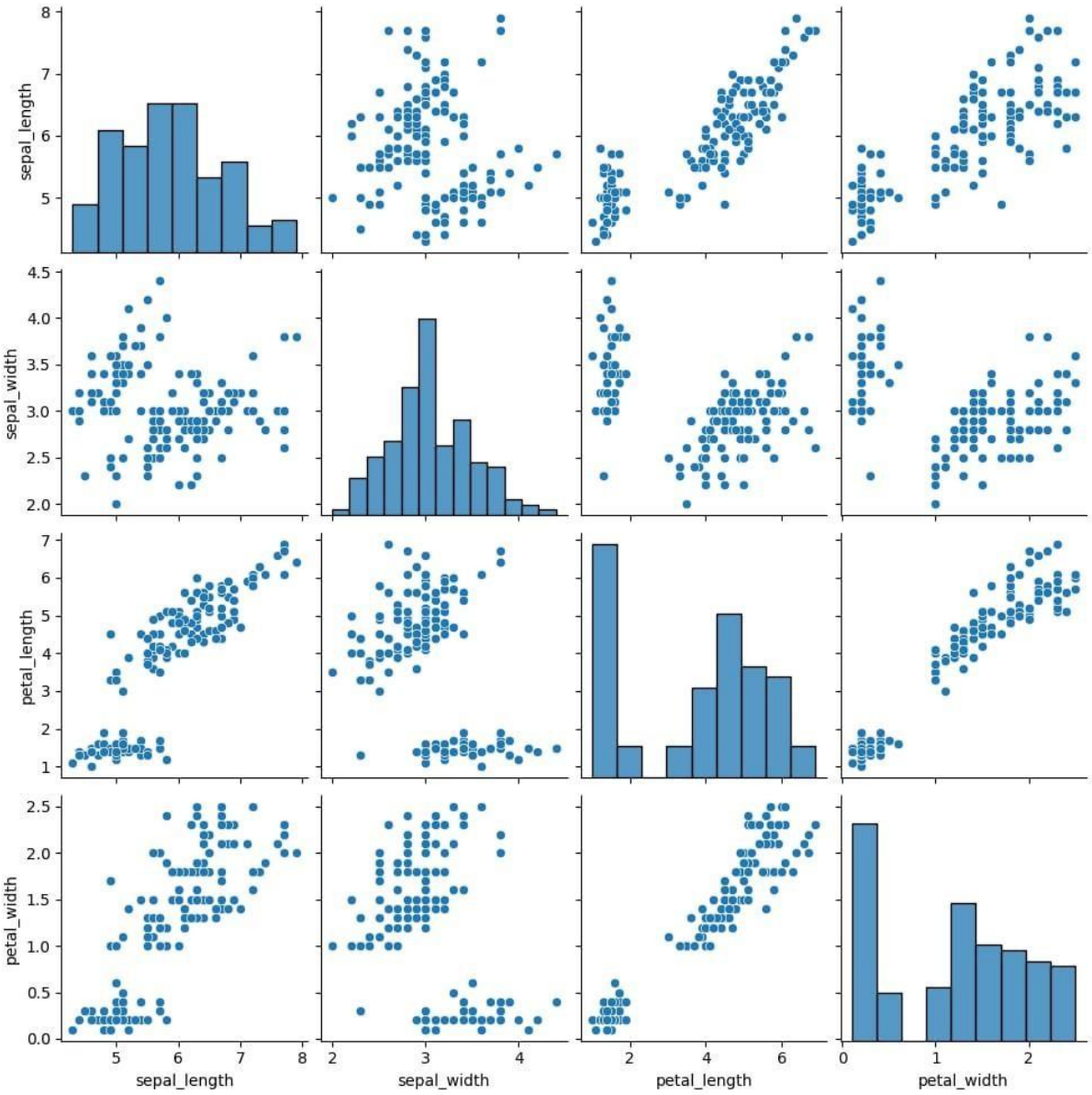
❖ Train-Test Split: 80% training, 20% testing

➢ Visualization of Results & Model Insights

➢ Pair chart:

❖ Pair plot help in understanding the relationship between different pairs numerical variables in a dataset.

- Pair plots are helpful in detecting between numerical features. strong correlation can be seen if the scatter plots show linear.
- Pair plots help identify potential issues of multicollinearity, where two or more variables are highly correlated.



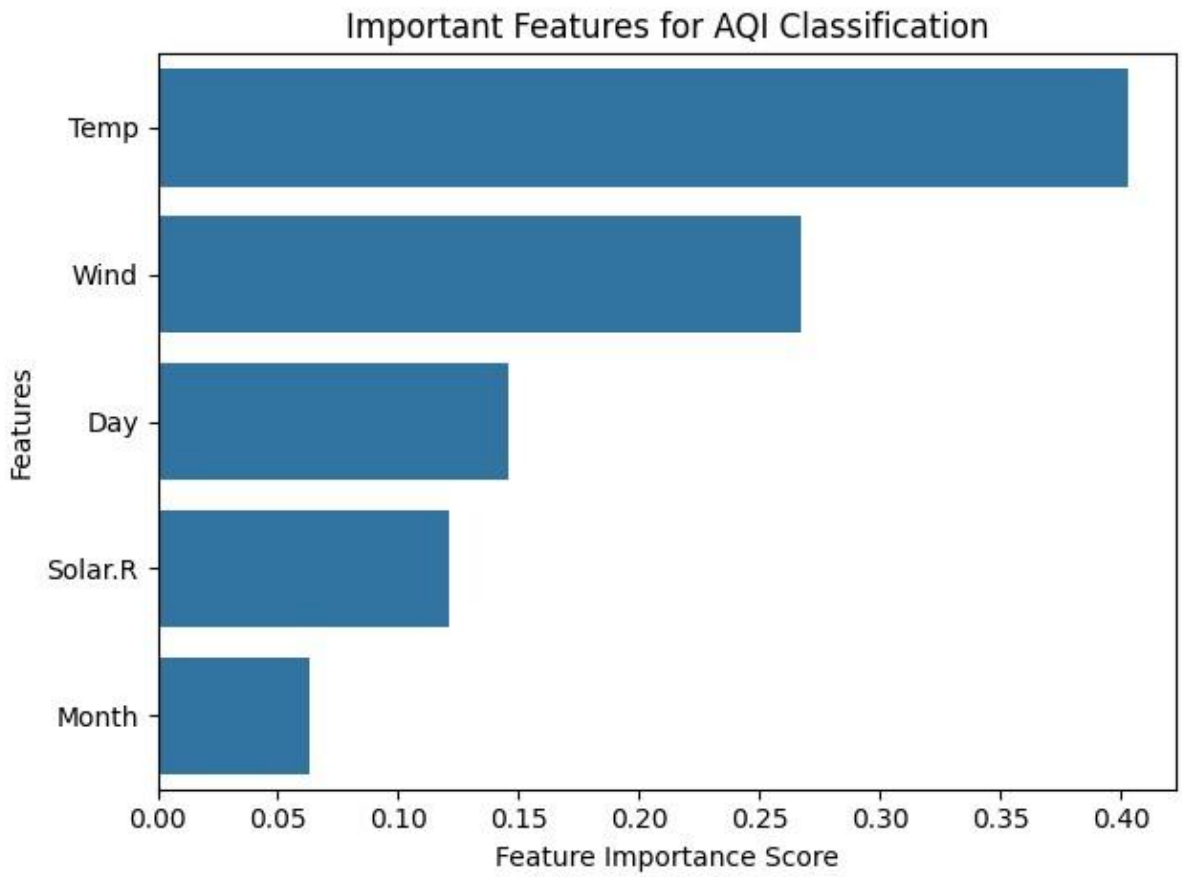
➢ Bar chart

- ❖ Bar plot showing which features influence AQI category prediction the most.

- ❖ Bar charts can show which features (E.g: Temperature, Wind, Solar, Radiation) are most influential in predicting ozone levels.

- ❖ Bar charts can be used to compare average ozone levels across different temperature ranges, wind speed categories.

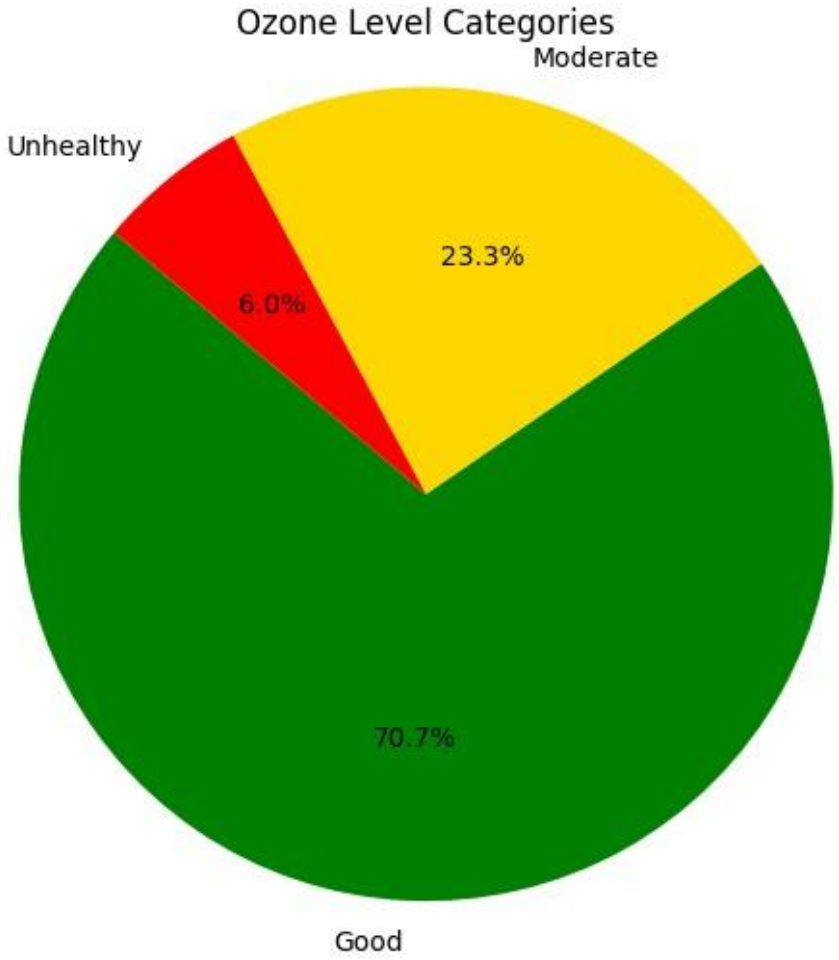
- ❖ Bar charts can show the accuracy, precision and recall scores of different models (e.g: Random forest).



➢ Pie Chart:

- ❖ A pie chart shows proportions or percentages of different categories relative to the whole dataset. Showing distribution of AQI Model.
- ❖ A pie chart offer a simple and intuitive visual-easy for any audience to quickly understand the distribution.

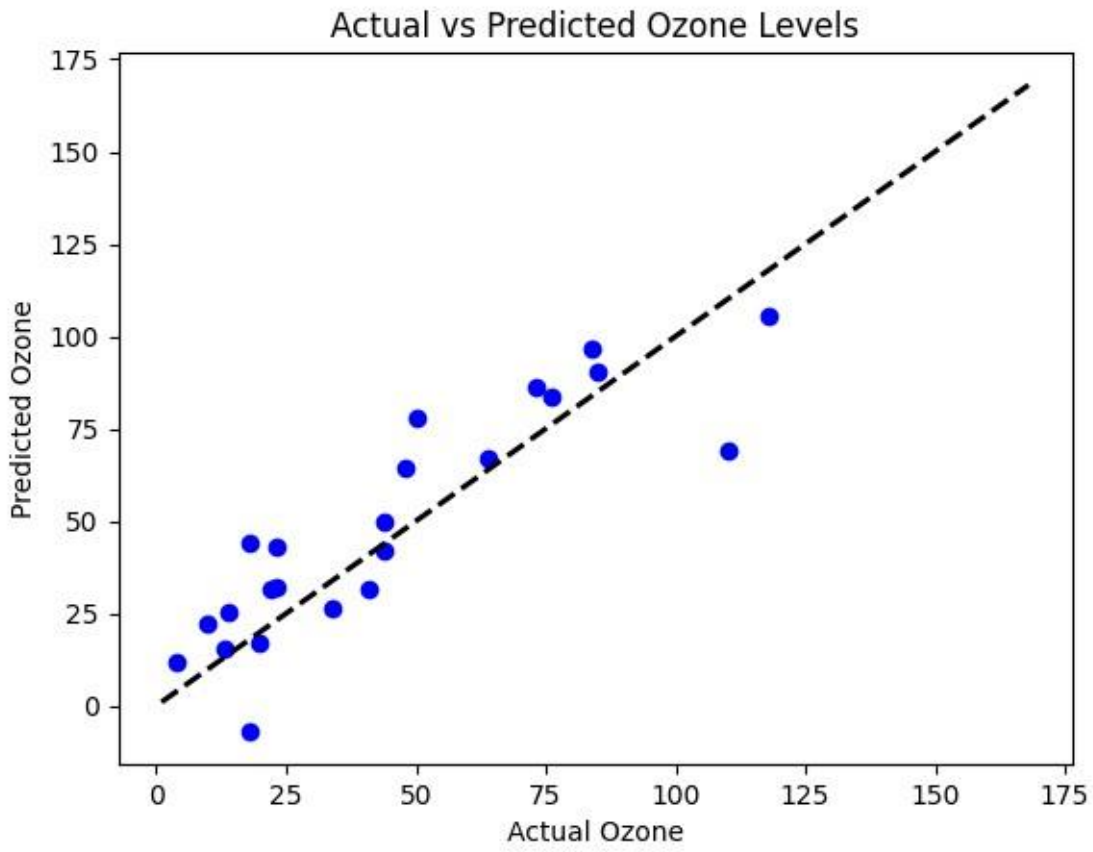
- Helps in comparing the size of different groups (like “Good”, “Moderate”, “Unhealthy”) air quality categories.

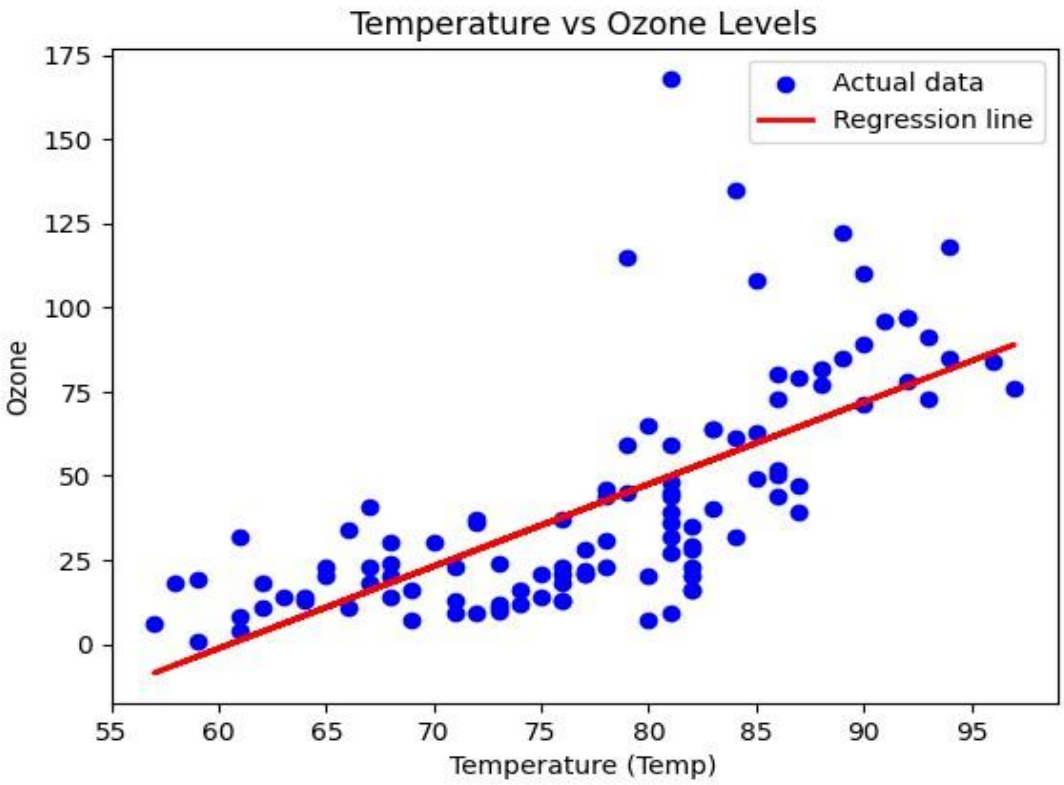


Scatter Plot:

- Temperature is a strong predictors feature of air quality.
- High ozone levels during warm weather can cause poor air quality.

- ❖ Assess Model Accuracy Visually: If most points are close to the dashed line, it means good predictive performance.
- ❖ If predictions are consistently above below line.
- ❖ In a scatterplot of temperature vs ozone.





➢ Correlation of heatmaps

- ❖ Correlation heatmaps help you select important features or remove redundant ones before modelling.
- ❖ A correlation heatmap visually shows the strength and direction of the two variables.

- Dark Red/Blue colours indicate strong correlations. Light colours (close to white) show weak or no correlation.
- Positive correlation (values near+1) usually appears red/orange colour.
- Negative correlation (values near 1) usually appears blue.

Tools and Technologies Used:

- Programming Language: Python
- IDE/Notebook: Google Colab
- Libraries: pandas (data manipulation)

  NumPy (numerical operations)

  matplotlib and seaborn (visualizations)

  scikit-learn (modelling and evaluation)

- Deployment Tools (Optional): Stream lit (if deploying as a simple web app)

Team members an roles:

Name	Role	RESPONSIBILITY
A.ABIBRISCKILLA	**Project Leader**	Oversee project development, coordinate team activities, ensure timely delivery of milestones, and contribute to documentation and final presentation.
R. RAGUL	Data manger	Collect data from APIs (e.g., Twitter), manage dataset storage, clean and preprocess text data, and ensure quality of input data.
T. JAYASUDHA	EDA and Visualization	Build sentiment and emotion classification models, perform feature engineering, and evaluate model performance using suitable metrics.
N.DAWOODKHAN	**Data Analyst / Visualization Lead**	Conduct exploratory data analysis (EDA), generate insights, and develop such as word cloud, emotion and sentiment dashboards.

T. KIRUBAKARAN	**Model Evaluation and deployment Assistant**	Evaluation Model Performing using metrics like accuracy and classification report- assisting with deployment readness summarizing model results and suggesting improvements.