**Student Name:** [Dhivyatharshini D]

**Register Number:** [712523243008]

**Institution:** [PPG INSTITUTE OF TECHNOLOGY]

**Department:** [B.tech AI&DS 2nd year]

**Date of Submission:** [25-04-2025]

# Github Repository Link:

[*https://github.com/dharshiniDuraimanickam/NM\_DHIVYATHARSHINI\_DS*](https://github.com/dharshiniDuraimanickam/NM_DHIVYATHARSHINI_DS)

## Problem Statement

Air pollution affects public health and quality of life. This project uses machine learning models to predict air quality and classify pollution levels from sensor data. The interactive dashboard helps users visualize trends, explore correlations, and predict carbon monoxide (CO) levels, assisting stakeholders in making informed decisions to improve air quality and reduce health risks.

## Project Objectives

* + Predict air quality (CO levels) using machine learning regression models.
  + Classify air quality into health-relevant categories (Good, Moderate, Bad) using machine learning classification models.Compare machine learning models: Logistic Regression, Random Forest, XGBoost, SVM, and KNN.
  + Perform exploratory data analysis (EDA) to visualize pollution trends and correlations.
  + Deploy the model using Streamlit for interactive predictions and visualizations.

## Flowchart of the Project Workflow

Upload Dataset  
↓  
Data Preprocessing  
↓  
Exploratory Data Analysis (EDA)  
↓  
Feature Engineering  
↓  
Model Building & Training (Regression & Classification)  
↓  
Model Evaluation & Comparison  
↓  
Visualization of Results  
↓  
Deployment using Streamlit

## Data Description

* + **Dataset:** air\_quality.csv
  + **Source:** Provided by the institution
  + **Type:** Structured, tabular data
  + **Size:** ~8000 rows
  + **Static or Dynamic:** Static dataset
  + **Target Variable:** CO(GT) (for regression) / AirQuality (for classification)

## Data Preprocessing

 Dropped unnecessary columns: Date, Time

 Cleaned numeric data by replacing commas with periods and converting strings to floats

 Replaced invalid values (e.g., -200) with NaN and dropped rows with missing data

## Exploratory Data Analysis (EDA)

 Plotted time series for CO(GT) levels and air quality trends over time

 Created a correlation heatmap to identify relationships between features

 Identified strong predictors of air quality, including CO(GT), PT08.S1(CO), NOx(GT), and T (Temperature)

 Used visual tools like scatter plots, count plots, and heatmaps to explore feature interactions

## Feature Engineering

 Applied label encoding for categorical variables (if any)

 Removed non-informative features like Date and Time

 Scaled numerical features (e.g., using StandardScaler) to improve model performance

 Created a new target variable for air quality classification based on CO(GT) levels

# Model Building

* + Trained two types of models:

**Regression**: Random Forest Regressor for predicting CO(GT) levels

**Classification**: Random Forest Classifier for classifying air quality into Good, Moderate, or Bad

* + Used metrics like Mean Squared Error (MSE), R² Score, Accuracy, Precision, Recall, F1-Score, and ROC-AUC to compare models
  + The Random Forest Classifier performed well for classification, while the Random Forest Regressor excelled in regression tasks

# Visualization of Results & Model Insights

 Used **SHAP** to explain the most important features contributing to the prediction of CO(GT) and air quality classification

 Visualized model results using:

* Scatter plots for Actual vs Predicted values (for regression)
* Confusion Matrix (for classification)
* SHAP beeswarm and bar plots
* Feature importance plots to highlight key variables influencing predictions

# Tools and Technologies Used

 **Programming Language**: Python

 **Libraries**: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, SHAP

 **Deployment & Visualization**: Streamlit

 **IDE**: VS Code, Jupyter Notebook

1. **Team Members and Contributions**

|  |  |
| --- | --- |
| Name | Contributions |
| [Banumathi. V] | Team Leader & Data Cleaning EDA |
| [ Sakina. E ] | EDA (Exploratory Data Analysis) |
| [Dhivya Tharshini.D] | Feature Engineering |
| [Yogadharshini.R] | Model Evaluation |
| [Kamalakaviya. J ] | Deployment |