MACHINE LEARNING – PROJECT

TITLE : AQI PREDICTION USING RIDGE REGRESSION

INTRODUCTION

Air pollution has become one of the most critical environmental issues affecting human health and quality of life. The Air Quality Index (AQI) is a standardized indicator used to represent air pollution levels based on various pollutant concentrations. Accurate prediction of AQI helps government agencies, environmental organizations, and the public take timely preventive measures. In this project, Ridge Regression is used to predict AQI values using multiple air pollutant parameters while reducing overfitting caused by correlated features.

PROBLEM STATEMENT :

To build a machine learning model that accurately predicts AQI values based on air pollutant concentrations while handling multicollinearity and reducing overfitting using Ridge Regression.

OBJECTIVES :

- To analyze air pollution data and its impact on AQI.
- To apply Ridge Regression for AQI prediction.
- To reduce overfitting caused by correlated pollutant features.
- To evaluate model performance using appropriate metrics and visualizations.

PROBLEM TYPE :

This project is a supervised machine learning regression problem. The model learns from labeled historical data where pollutant concentrations are input features and AQI is the continuous target variable.

DATASET :

Selected Features (Input Variables):

- PM2.5: Fine particulate matter with diameter ≤ 2.5 micrometers

- PM10: Particulate matter with diameter ≤ 10 micrometers
- NO: Nitric Oxide concentration
- NO2: Nitrogen Dioxide concentration
- CO: Carbon Monoxide concentration
- O3: Ozone concentration

Target Variable:

- AQI: Air Quality Index (numerical value representing pollution level)

ALGORITHM USED :

Ridge Regression Algorithm

Ridge Regression is a supervised machine learning algorithm used for regression problems. It is an extension of Linear Regression that includes L2 regularization. The algorithm minimizes the sum of squared errors while adding a penalty proportional to the square of the magnitude of model coefficients.

Mathematical Representation:

$\text{Loss} = \Sigma(y - \hat{y})^2 + \lambda\Sigma w^2$

Where:

- y is the actual AQI value
- $\hat{y}$ is the predicted AQI value
- λ (alpha) is the regularization parameter
- w represents model coefficients

This regularization helps reduce overfitting and handles multicollinearity among pollutant features.

METHODOLOGY :

Step 1: Dataset Collection

The air quality dataset is collected from Kaggle. It contains daily pollutant concentration values such as PM2.5, PM10, NO, NO_2, CO, and O_3 along with the corresponding Air Quality Index (AQI).

Step 2: Dataset Loading

The collected dataset is loaded into the Python environment using the Pandas library and stored as a DataFrame for analysis and processing.

Step 3: Data Preprocessing

Data preprocessing is performed to improve data quality:

- Column names are cleaned by removing extra spaces.
- Missing and null values are handled.
- Inconsistent or invalid records are removed.

Step 4: Feature Selection

Relevant pollutant variables influencing AQI are selected as input features:
PM2.5, PM10, NO, NO_2, CO, and O_3.

Step 5: Target Variable Identification

Air Quality Index (AQI) is selected as the target variable to be predicted.

Step 6: Feature and Target Separation

The dataset is divided into:

- Input features (X) – selected pollutant values
- Target variable (y) – AQI values

Step 7: Dataset Splitting

The data is split into training and testing datasets using an 80:20 ratio to ensure unbiased model evaluation.

Step 8: Feature Scaling

Standardization is applied to the input features using StandardScaler so that all features contribute equally to the model.

Step 9: Model Selection

Ridge Regression is chosen as the prediction model due to its ability to handle multicollinearity and reduce overfitting using L2 regularization.

Step 10: Model Training

The Ridge Regression model is trained using the scaled training data to learn the relationship between pollutant levels and AQI.

Step 11: Model Prediction

The trained model is used to predict AQI values for the testing dataset.

Step 12: Model Evaluation

Model performance is evaluated using Mean Squared Error (MSE) to measure the accuracy of AQI predictions.

TOOLS AND TECHNOLOGIES :

□ Programming Language: Python

□ Libraries: Pandas, NumPy, Scikit-learn, Matplotlib

□ Development Environment: Jupyter Notebook (Anaconda)

RESULTS AND PREDICTION :

The Ridge Regression model successfully predicts the Air Quality Index (AQI) based on air pollutant data.
For a sample input, the predicted AQI value indicates the current air quality level and its impact on health.

Based on the prediction:

- Low AQI → Air quality is good and safe.
- Moderate AQI → Air quality is acceptable with minor health concern.
- High AQI → Air quality is unhealthy and requires precaution.

REAL-LIFE APPLICATIONS :

This project is useful in real life because:

- It predicts the Air Quality Index (AQI) based on pollutant concentrations.
- It helps government agencies monitor air pollution levels efficiently.
- It supports public health awareness by identifying unhealthy air conditions.
- It assists urban planners in controlling pollution sources.
- It enables early warnings for sensitive groups such as children and elderly people.

CONCLUSION :

Ridge Regression effectively predicts Air Quality Index by handling multicollinearity among air pollutants and reducing overfitting. The model provides stable and accurate AQI predictions, making it suitable for real-world air quality monitoring and environmental decision-making systems.