Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Name : Savinay Pandey
Roll No : 99
Moodle Id : 21102125
Div : B.E. B-1

Name : Karan Patel
Roll No : 102
Moodle Id : 21102092
Div : B.E. B-2

Subject : Machine Learning

Experiment No : 1 (Case Study)

TITLE :
Air Quality Index (AQI) Prediction Using Machine Learning Regression Techniques.

INTRODUCTION :

The Air Quality Index (AQI) is a crucial indicator of air pollution levels, directly impacting public health and environmental quality. Understanding and predicting AQI allows governments and organizations to take pre-emptive actions against pollution. Traditional AQI prediction methods often struggle to accommodate the complex, non-linear relationships between various pollutants. This case study explores how machine learning, particularly regression techniques, can be leveraged to accurately predict AQI based on multiple environmental factors.

By applying machine learning models, we aim to uncover the intricate relationships between pollutants and AQI, providing a robust framework for environmental monitoring and public health protection.

OBJECTIVES:

1. **To explore and analyze the dataset:**
    - Conduct a thorough examination of the pollutant data and its relationship with the Air Quality Index (AQI).
    - Visualize data distributions and correlations to understand the underlying patterns.
2. **To implement various machine learning regression models:**
    - Develop and train multiple regression models, including Linear Regression, Decision Trees, Random Forest, and XGBoost, to predict AQI.
    - Compare the predictive capabilities of each model.

3. **To optimize model performance through parameter tuning:**
    - Fine-tune the hyperparameters of each model to enhance accuracy and minimize errors in AQI prediction.
4. **To generate insights and recommendations based on model results:**
    - Analyze the impact of different pollutants on AQI predictions.
    - Provide actionable recommendations for using machine learning models in real-world AQI monitoring and forecasting.

MODULES:

1. **Data Understanding and Preparation:**
    - **Exploratory Data Analysis (EDA):**
        - Visualize pollutant distributions and their relationships with AQI.
        - Generate heatmaps and correlation matrices to identify strong predictors.
    - **Data Cleaning:**
        - Confirm the absence of missing values.
        - Normalize or scale features if necessary to improve model performance.

2. **Model Implementation:**

- **Linear Regression:**
    - A straightforward approach to modeling the linear relationship between pollutants and AQI. This model serves as a baseline for comparison.
- **Decision Trees:**
    - A non-linear model that splits the dataset based on pollutant levels, effectively capturing complex patterns and interactions between variables.
- **Random Forest:**
    - An ensemble technique that builds multiple decision trees and merges their outputs for a more accurate and stable prediction. It helps reduce overfitting and improves generalization.
- **XGBoost:**
    - A powerful gradient boosting model that iteratively corrects errors made by previous models. Known for its high performance in predictive modeling, XGBoost is employed to achieve optimal AQI predictions.

3. **Model Training and Evaluation:**

- **Data Splitting:**
    - Divide the dataset into training (80%) and testing (20%) sets to validate model performance.
- **Model Training:**

- Train each regression model on the training data, adjusting hyperparameters to optimize accuracy.

- **Evaluation Metrics:**
  - Use metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to assess model performance. Lower values indicate better model accuracy and reliability.

4. **Comparative Analysis and Insights:**

- **Performance Comparison:**
  - Compare the predictive accuracy of each model to determine the best-performing technique for AQI prediction.
- **Impact Analysis:**
  - Analyze how different pollutants impact AQI prediction accuracy. Identify which pollutants are the most significant predictors.
- **Model Selection:**
  - Based on the evaluation metrics and comparative analysis, select the most effective model for real-world AQI prediction.

MATERIALS AND METHODS:

- **Data Collection Procedure:**
  - The dataset was sourced from environmental monitoring stations, containing historical records of pollutant concentrations and corresponding AQI values.
  - Additional contextual data (e.g., weather conditions, traffic patterns) could be integrated to enhance model accuracy in future studies.
- **Data Analysis Procedure:**

1. Data Exploration: Perform initial data analysis to understand pollutant distributions and relationships.
2. Model Implementation: Implement regression models, fine-tune parameters, and train the models on the dataset.
3. Model Evaluation: Validate models using test data, compare performance, and refine models for better accuracy.
4. Insights Generation: Identify key insights and implications of the findings, providing actionable recommendations for environmental monitoring.

- **DATASET DESCRIPTION:**

The dataset utilized in this study comprises 8 key attributes:

- **Independent Variables:**
  - **PM2.5-AVG:** Average concentration of Particulate Matter 2.5 microns.
  - **PM10-AVG:** Average concentration of Particulate Matter 10 microns.
  - **NO2-AVG:** Average concentration of Nitrogen Dioxide.

    - **NH3-AVG:** Average concentration of Ammonia.
    - **SO2-AVG:** Average concentration of Sulfur Dioxide.
    - **OZONE-AVG:** Average concentration of Ozone.
- **Dependent Variable:**
    - **air_quality_index:** A composite index representing the overall level of air pollution.

DURATION OF STUDY:

The project was carried out over an 8-week period, structured as follows:

- **Weeks 1-2:** Data understanding, preparation, and exploratory analysis.
- **Weeks 3-4:** Implementation of regression models.
- **Weeks 5-6:** Training, evaluation, and fine-tuning of models.
- **Week 7:** Comparative analysis and insights generation.
- **Week 8:** Compilation of findings, report preparation, and review.

REFERENCES:

1. Schmidt, A., Kabir, M.W.U., & Hoque, M.T. (2022). *Machine Learning Based Environmental Monitoring.*
2. Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice.* OTexts.
3. Längkvist, M., Karlsson, L., & Loutfi, A. (2014). *A Review of Unsupervised Feature Learning and Deep Learning for Time-Series Modeling.* Pattern Recognition Letters, 42, 11-24.
4. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* JMLR, 12, 2825-2830.
5. Breiman, L. (2001). *Random Forests.* Machine Learning, 45(1), 5-32.
6. Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
7. Qi, X., & Zhang, L. (2020). *Prediction of Air Quality Based on Machine Learning Algorithms.* Environmental Science and Pollution Research, 27(35), 44245-44259.