**Comparative Analysis Study for Air Quality Prediction in**

**Smart Cities Using Regression Techniques**

**Alternate Title:** An evaluation of Regression Methods for Accurate Air Quality Prediction.

**Aim**:

 To systematically compare and evaluate the performance of various regression techniques for Air Quality prediction in smart cities, aiming to identify the most effective model that can contribute to accurate and timely forecasting.

**Description:**

 In smart cities, air pollution has detrimental impacts on human physical health and the quality of living environment. Therefore, correctly predicting air quality plays an important effective action plan to mitigate air pollution and create healthier and more sustainable environments. Monitoring and predicting air pollution is crucial to empower individuals to make informed decisions that protect their health. This research presents a comprehensive comparative analysis focused on air quality prediction using three distinct regression techniques- Random Forest regression, linear regression, and Decision Tree regression. The main goal of this study is to discern the most effective model by considering a range of evaluation criteria, including Mean Absolute Error and R2 measures. Moreover, it considers the crucial aspects of minimizing prediction errors and enhancing computational efficiency by evaluating the regression models within two frameworks.

The findings of this study underscore the superiority of the Decision Tree regression approach over the other models, demonstrating its exceptional accuracy with a high R2 score and a minimal error rate. Moreover, integrating cloud computing technology has resulted in substantial improvements in the execution time of these approaches. This technology enhancement significantly affects the overall efficiency of the air quality prediction process. By leveraging distributed computing resources, real-time air quality forecasting becomes feasible, enabling timely decision-making and proactive measures to address air pollution episodes effectively.

**Existing System:**

This accelerated experimentation, training, and deployment of the models, enhancing their practical applicability in real-world applications. The result showed that Decision Tree is the most suitable with respect to accuracy. But the accuracy is less. Then now we create a new system for better Air Quality prediction. So now we move on to the proposed system.

**Proposed System:**

This study provides to create an effective prediction model using different regression of ML methods to predict Air Quality. First of all, the datasets are collected, and then the preprocessing is accomplished via the missing values imputation. Feature selection for supervised models using SelectKbest. This feature selection is techniques where we choose those features in our data that contribute most to the target variable. In other words we choose the best predictions for target variable. Then we are using Decision Tree, Logistic regression and Random Forest Algorithm for prediction accuracy. Decision Tree gives best results with respect to high accuracy. Compare to existing system our new system gives results are more Accuracy.

**Module Description:**

* Data Pre-Processing
* Algorithm Implementation
* Prediction

**Data Pre-Processing:**

The dataset used in this study encompasses a comprehensive collection of 103,205 records. Those are collected from the publicly available repositories: Kaggle and Github. Datasets are collected, and then the preprocessing is accomplished via the missing values imputation. The Missing Value Imputation (MVI) method is used to impute the missing values of the dataset.AQI is a numerical value within a defined range, typically from 0 to 500. A higher value of AQI indicates poorer air quality and the existence of harmful air pollutants.

Each pollutant has specific constraints and specific averaging periods to ensure accurate assessment such as the period is 8-hour maximum for Q3 and 24-hour average concentrations for SO2, PM10, CO, NO2, and PM2.5. To calculate the AQI, the concentrations of these air pollutants are categorized into sub-indices. These sub-indices were defined based on predefined ranges that help to give the level of air quality, ranging from ‘‘good’’ to ‘‘Severe’’. And in the dataset there is unwanted columns are presented; these columns are removed and get a neat and clean dataset to predict more accuracy.

**Algorithm Implementation:**

 The Regression Algorithms to produce the best results. Feature selection for supervised models using SelectKbest. This feature selection is techniques where we choose those features in our data that contribute most to the target variable. In other words we choose the best predictions for target variable. We are using Decision Tree, Logistic regression and Random Forest Algorithm to predict Air Quality by using dataset in ML. On an analysis conducted within various algorithms, the Decision Tree was found to provide highest efficiency. Then, the regressors are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error.

* Decision Tree
* Logistic Regression
* Random Forest

**Selected Algorithm:**

 Decision Trees (DT) are a non-parametric supervised learning method used for [classification](https://scikit-learn.org/stable/modules/tree.html#tree-classification)The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

**Prediction:**

The Air Quality is depends on the Air Quality Index range 0 to 500, that’s are classified “Good”, “Satisfactory” , “Moderate”, “Unhealthy”, “Very Unhealthy” and “Severe”. Standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficacy of this model.Preprocessed data are trained and input given by the user goes to the trained dataset and predict the results that the accurate classification.

**Future Enhancement:**

For future work recommendations, we explore diverse machine-learning approaches for predicting air quality and air pollution in smart cities. Additionally, investigating the effect of meteorological data, including temperature, pressure, humidity, and wind speed, further enhances AQI and air pollution prediction accuracy. This endeavor provides valuable insight into identifying air quality levels and contributes to more effective air quality management approaches.

**Software Requirements:**

* Operating System : Windows 10 (64 bit)
* Software : Python 3.7
* Tools : Anaconda (Jupyter Note Book IDE)

**Hardware Requirements:**

* Hard Disk : 500GB and Above
* RAM : 4GB and Above
* Processor : I3 and Above

**Architecture Diagram:**

Dataset

Preprocessing

User Input

Trained model

Machine Learning

Predicting Air Quality Index

Satisfactory

Good

Moderate

Unhealthy

Very Unhealthy

Severe