Project Title: Air Quality Assessment in Tamil Nadu

Project Overview:

Engaging in the analysis of air quality and forecasting particulate matter levels such as RSPM (Respirable Suspended Particulate Matter) or PM10 (Particulate Matter with a diameter of 10 micrometers or less) requires a thoughtful selection of algorithms and methodologies suited to the intricacies of environmental data and time series modeling.

Model Selection:

When embarking on an air quality analysis project with the aim of predicting particulate matter levels like RSPM and PM10, it becomes paramount to meticulously choose algorithms and techniques tailored to the distinctive attributes of environmental data. Your dataset is an amalgamation of diverse variables, including location, type of location, date, SO2, NO2 levels, and the RSPM/PM10 values. These variables encompass both numerical and categorical characteristics, with the additional complexity of temporal elements associated with date information. To achieve precise predictions for RSPM/PM10 levels within this dataset, it is prudent to consider algorithms adept at handling mixed data types in regression tasks. Below, we illuminate a selection of algorithmic choices that exhibit particular promise for this specifc context.

1. Random Forest and Gradient Boosting (e.g., XGBoost, LightGBM): Random Forest and gradient boosting techniques, like XGBoost and LightGBM, are exceptionally versatile and skilled at managing the fusion of numerical and categorical variables within your dataset. They excel in capturing the intricate interplays and relationships among these features. Proper preprocessing of categorical variables is essential, which may involve methods such as one-hot encoding or label encoding, to render them compatible with these algorithms.

2. *Linear Regression*: Linear regression can function as a fundamental benchmark model for your air quality prediction project. By appropriately incorporating categorical attributes, for instance, through the utilization of one-hot encoding, and potentially exploring polynomial regression to accommodate nonlinear associations, you can create a straightforward yet valuable starting point for your analytical journey.

XGBoost:

- Predictive Performance: XGBoost consistently achieves high accuracy across diverse datasets thanks to its ensemble approach, which uncovers complex relationships in the data.

- Regularization: It effectively prevents overfitting through L1 and L2 regularization techniques, enabling the model to generalize better to new data.

- Handling Missing Data: XGBoost has built-in support for managing missing values, reducing the need for extensive data preprocessing and imputation.

- Feature Importance: XGBoost provides valuable insights into which variables have the most significant impact on the target variable. This aids in feature selection and enhances data understanding.

- Efficiency and Scalability: XGBoost is optimized for speed and can handle large datasets efficiently. This makes it suitable for real-world applications with big data, and its parallel processing capabilities accelerate model training.

- In conclusion, XGBoost is the top choice for predictive modeling, especially in complex datasets like air quality analysis. It excels in handling diverse data types, mitigating overfitting, and offering valuable insights, making it a valuable tool for environmental monitoring and air quality predictions.

Team Members:

1. **Hematharshini E** – 2021115041

   Email - hemae2512@gmail.com

2. **Sabitha S** – 2021115087

   Email - sabitha.suresh15@gmail.com

3. **Sandhya Shankar** – 2021115090

   Email - sandhya.shankar.2002@gmail.com

4. **Sanmitha V.S** – 2021115092

   Email - sanmithasadhishkumar@gmail.com

5. **Akash P** – 2021115314

   Email - akashpanneer2004@gmail.com