PREDICTION OF AIR QUALITY USING SUPERVISED MACHINE LEARNING

Use Case :

- Classify about different types of air pollution.
- Specify the AQI for India report.
- Possibility of health impacts.
- Predict by accuracy.
- Best quality air prediction in the real time world.

Implementation:

- Implemented by machine learning approach by user interface of GUI application
- Multiple datasets from different sources would be combined to form a generalized dataset, and then decision tree machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.
- Data validation process
- To train a model by given dataset using sklearn package
- Accuracy results of decision tree algorithms
- GUI based patient login / Registration page
- GUI based prediction results of air quality

Challenges Faced:

Getting the real time dataset from the meteorological department and expected accuracy in real time.

Used Packages:

- sklearn - The **sklearn** library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
- NumPy - Used for data manipulation process.
- Pandas - it offers data structures and operations for manipulating numerical tables and time series.
- matplotlib - by matplotlib data visualization is a useful way to help with identifying the patterns from a given dataset.

Algorithm Used:

- K-nearest neighbor -
    - The **k-nearest neighbors** (**KNN**) **algorithm** is a simple, supervised machine learning **algorithm** that can be used to solve both classification and regression problems. It's easy to implement and understand.

- Support vector machine -
    - In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.
- Logistic regression -
    - **Logistic regression** is basically a supervised classification **algorithm**. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X. Contrary to popular belief, **logistic regression** IS a **regression** model.
- Random forest -
    - Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

Final Outcome:

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on a public test set is a higher accuracy score. This application can help India meteorological department in predicting the future of air quality and its status and depends on that they can take action.