**AIR POLLUTION ANALYSIS**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

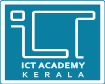
**Team-9**

**JISHNU P T, SHAHINA T A**

**SANITHA MARY KURIAN**

**SHAH SHAMSUDHEEN C N**

**SREEJITH K**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Feb 2022**

**Table of Contents**

1.Problem Definition

2.Introduction

3.Literature Survey

4.Understanding the dataset

5.Exploratory data analysis

6.Preprocessing

7.Modelling

8.Hyperparameter tunning

9.Web Hosting

**Abstract**

Air pollution is an important environmental risk factor in many propagation diseases. Regulation of air quality is an important task of the government in developing countries for ensuring people’s health and welfare. Air pollution differs from place to place and depends on multiple pollutant sources such as industrial emissions, heavy traffic congestions, temperature, pressure, wind, humidity and burning of fossil fuels etc. Analyzing and protecting air quality has become one of the most required activities for the government in almost all the industrial and urban areas today.The air quality index (AQI) is a measure of how air pollution affects one's health within a short time period. The purpose of the AQI is to help people know how the local air quality impacts their health. AQI for five major air pollutants Ground-level ozone ,Particle pollution/particulate matter (PM2.5/pm 10),Carbon Monoxide,Sulfur dioxide,Nitrogen dioxide are calculated.The index classifies air quality simply as good, satisfactory, moderately polluted, poor, very poor, and severe.Here a machine learning model is built to predict the air quality index based on past air quality data.

**1.PROBLEM DEFINITION**

**1.1 Overview**

The project focuses on the various techniques used for prediction or modelling of Air Quality Index (AQI)

**1.2 Problem Statement**

Air pollution can be defined as the pollutants contamination into the air which is highly injurious to human health and the whole planet. Several million deaths are reported because of the pollution related diseases. The purpose of predicting air quality is to estimate the air pollution level by considering meteorological data and pollutant concentrations data. AQI is the index value used to define the contamination level of the air and to inform people about the air pollution level.

**2. INTRODUCTION**

Especially after the COVID-19 pandemic, which is great misfortune for human race and nothing like this were happened in past decades or in history. Which is hardly cause for recession and also heavily caused human health. Since this each and every one has been much concerned about health. Alike human health was threatened by increasing in the air pollution. Air pollution can be measured or calculated by AQI index, these AQI index can measured using 5 major air pollutants with other air pollutants. In records, WHO claims that maximum people inhale polluted air. The main causes of air pollution are exhaust from vehicles, industries and factories, power plants, mining and agricultural activities, the burning of fossil fuels, wildfires and volcanoes. The purpose of predicting air quality is to estimate the air pollution level by considering meteorological data and pollutant concentrations data. AQI is the index value used to define the contamination level of the air and to inform people about the air pollution level.

**Dataset Description**



**Particulate Matter ( PM2.5 and PM10)**

Particulate matters(PM2.5 and PM10) are a mixture of solid particles and liquid droplets found in the air. These particles come in many sizes and shapes and can be made up of hundreds of different chemicals. Some are emitted directly from a source, such as construction sites, unpaved roads, fields, smokestacks or fires.

**Nitrogen Oxides(NO , NO2 , NOx)**

Nitrogen oxides are a mixture of gases that are composed of nitrogen and oxygen. NOx is produced from the reaction of nitrogen and oxygen gases in the air during combustion, especially at high temperatures. In areas of high motor vehicle traffic, such as in large cities, the amount of nitrogen oxides emitted into the atmosphere as air pollution can be significant.

**Ammonia (NH3)**

Ammonia pollution is pollution by the chemical ammonia (NH3) – a compound of nitrogen and hydrogen which is a byproduct of agriculture and industry.

**Carbon Monoxide(CO)**

Carbon monoxide (CO)—a colorless, odorless, tasteless, and toxic air pollutant—is produced in the incomplete combustion of carbon-containing fuels, such as gasoline, natural gas, oil, coal, and wood.

**Sulphur Dioxide (SO2)**

Sulfur dioxide is a gas. It is invisible and has a nasty, sharp smell. It irritates the nose, throat, and airways to cause coughing, wheezing, shortness of breath, or a tight feeling around the chest. The effects of sulfur dioxide are felt very quickly and most people would feel the worst symptoms in 10 or 15 minutes after breathing it in.

**Ozone (O3)**

Ground-level ozone is a colorless and highly irritating gas that forms just above the earth's surface. It is called a "secondary" pollutant because it is produced when two primary pollutants react in sunlight and stagnant air. These two primary pollutants are nitrogen oxides (NOx) and volatile organic compounds (VOCs).

**Benzene,Toluene,Xylene**

Benzene, toluene, xylene, and formaldehyde are well-known indoor air pollutants, especially after house decoration. They are also common pollutants in the working places of the plastic industry, chemical industry, and leather industry.

In the proposed work, air pollutant concentration dataset has been done to predict the AQI.

The work done is structured into 1.Data collection 2.Understanding the dataset 3.Exploratory data analysis 4.Data preprocessing 5.Building ML models 6.Hyper parameter tunning.

**3. LITERATURE SURVEY**

Pollution is the largest environmental cause of disease and premature death in the world today. Diseases caused by pollution were responsible for an estimated 9 million premature deaths in 2015—16% of all deaths worldwide[2]. Seven million people are killed every year in the universe by the diseases related to air pollution. From the world health organization statistics, it is revealed that maximum people inhale polluted air [3]. The main causes of air pollution are exhaust from vehicles, industries and factories, power plants, mining and agricultural activities, the burning of fossil fuels, wildfires and volcanoes. Air pollution appears on account of the release of dangerous suspended particulates like carbon monoxide (CO), sulphur dioxide (SO2), particulate matter10 (PM10), particulate matter 2.5 (PM2.5), ozone (O3) in to the atmosphere. The purpose of predicting air quality is to estimate the air pollution level by considering meteorological data and pollutant concentrations data. AQI is the index value used to define the contamination level of the air and to inform people about the air pollution level.[4]

**4.UNDERSTANDING THE DATASET**

The data used in this project is collected from the Central Pollution Control Board (CPCB), Government of India from 2015 to 2020.We collected the data from kaggle. Air Quality Index (AQI) is a number that measures the pollution level of the ambient air and it is measured based on 6 air pollutants: Sulfur dioxide (SO2), Nitrogen dioxide (NO2), Carbon monoxide (CO), Ozone (O3), PM2.5(particulate matter with diameter less than 2.5 µm) and PM10 (particulate matter with diameter less than 10 µm). AQI is classified into 6 levels of severity based on the value.

We have uploaded the dataset to python environment and done some basic codes.There are 29531 rows and 16 columns in the dataset.Under these columns 13 are float type and 2 are objective type.There is no duplicate entries in all 16 columns.By using describe function it is crystal clear that our data set contains samples from 26 different main cities in India. Also, AQI contains 6 different categories.Using the statistic table we simply interpreted that each columns have entirely different means, which is because of the values scales of columns are entirely different. The other thing most of all columns contain missing values,maximum values are way too high compared to its quantiles and mean,in almost all the columns mean is greater than the median.

After checking for null values we found that except the 'City' and 'Date' columns all other columns contain atleast 2000 missing values.That means 14 columns contain atleast one missing [value.By](http://value.by/) investigating the percentage of missing we found that Xylene column contain more missing values(61.32%). The unique values for 'City' is 26.The dataset contains air quality and AQI (Air Quality Index) of 26 cities across India.The unique values for 'AQI\_Bucket' is 6.The air quality index is classified into six buckets;Good,Poor,Very poor,Moderate and Satisfactory. This dataset is to predict the air quality of states on a daily [basis.In](http://basis.in/) total most of our data is incomplete.

**5.EXPLORATORY DATA ANALYSIS AND ITS RESULTS**

The preliminary analysis of data to discover relationships between measures in the data and to gain an insight on the trends, patterns, and relationships among various entities present in the data set with the help of statistics and visualization tools is called Exploratory Data Analysis (EDA). Exploratory data analysis is cross-classified in two different ways where each method is either graphical or non-graphical. And then, each method is either univariate, bivariate or multivariate.

**Univariate Analysis**

In Univariate analysis there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.Univariate data can be described through:

**1.Boxplot**

Boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles. It interpret the presence of outliers in data. From the box plots of our dataset we can conclude that all of pollutants contains outliers, among these SO2 contains most outliers.Also in these case all, outliers of each variables/pollutants are lies above the maximum limit. Which means values of each pollutants are consistently and exponentially increasing from time to time. In such time series data or data of same variable collected over time will have some shifts or shocks, but all time it can’t be consider as outliers.

**2.Histogram**

Using the histogram we can find conclusion about the distribution of that variable.We can say that around 15,000 to 25,000 data’s of each pollutants are lies in between value of 0 – 100.By accounting the above statement, it is crystal clear that all pollutants had a right/positive skewed distribution, that implies mode always less than median & median is always less than mean.Measure of dispersion in such data is meaningless.

**3.Line Plot**

Usually line plot/chart are used to determine frequencies in each variables.Here we can see that minimum number of data is in Aizawl and most numbers of data’s is in Ahmedabad, Bengaluru, Chennai, Delhi, Hyderabad, Lucknow and Mumbai.May be less number of data in Aizawl is due to constraints set by investigator such as samples will collected proportionally to the population of respective place, for example; in Delhi population is 1.9 Crore but in Aizawl population is around 4 lakhs

**4.Bar** **Plot**

Histogram tells about the distributions while Bar plots compares variables.Here the bar plot is used to compare the frequencies of over years.It is clear that this poses a increasing trend till year 2019, but in 2020 it is exponentially decreased.May be it is because of in every year population is increasing and in 2020 due to the pandemic investigator may have faced difficulty to conduct survey.Using the bar plot of AQI buckets (Categorical Variable), we can conclude that we have only minimum Extreme Sever situation & extremely good quality air.Most of times situation of quality of air is Moderate (101-150) and Satisfactory (51-100).Also we can conclude that average AQI bucket situation in these cities is Moderate.Also around 30% of AQI index where above 150.Less than 14% of AQI index are lesst than 51

**Bivariate Analysis**

Generally, it is performed to find the relationship between each variable in the dataset. That is dependent variable of interest and independent variable. Using two variables and finding the relationship between them.Since instead of 2 variables in MULTIVARIATE it will be more 2 variables.

**Correlation**

Correlation is simply find existing the linear relationship between the independent and dependent variable.Using the plot we can infer that among the pollutants, AQI & PM10 are positively linearly related and it is almost perfectly correlated.And AQI & Benzene are less linearly related, almost no relation among this variables

Using these correlation matrix we can conclude that Benzene is not need in these data, since Benzene is only one pollutant which has almost less linear relation ship with other pollutants

**Scatter Plot**

In Correlation Matrix we have the measure of the relationship between variables, but in scatter plot displays the strength of relationships.In correlation matrix we have evidence that AQI & PM10 strong relationship. But in scatter plot we can’t so sure about these things.In scatter plot we can infer that where the variables accumulating and at which range has high density and so and so.

**Line Chart**

Line chart can be used in both Single and multi-variable analysis, it will dependent upon how much variables are taking into account.Here we use line chart for each pollutants for 5 years, also showing the range of values.Using the plots, we can infer that PM10, PM2.5 & AQI showing a decreasing trend over years. All other pollutants expect NO2 showing a ‘Zig-Zag’ trend, in case NO2 the plot implies a curve which was initially low and get peak at 2018 and in 2020 sink into the bottom

**6.PREPROCESSING**

It is technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training machine learning models. The quality of the data should be checked before applying machine learning. Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

* **Accuracy**: To check whether the data entered is correct or not.
* **Completeness**: To check whether the data is available or not recorded.
* **Consistency:** To check whether the same data is kept in all the places that do or do not match.
* **Timeliness**: The data should be updated correctly.
* **Believability**: The data should be trustable.
* **Interpretability**: The understandability of the data.

**Major Tasks in Data Preprocessing:**

Different techniques in data preprocessing include,

1.Missing value handling

In our dataset number of columns with at least one missing value is 14.Deleting the rows containing missing values may lead to parting away with useful information or patterns. Also imputing the level of pollutants by mean/median level of pollutants may not necessarily be an appropriate strategy. So we used the k-nearest neighbor (KNN) imputer and it has proven to be generally effective. In KNN imputer missing values in datasets can be imputed with the help of values of observations from the k-Nearest Neighbours in the dataset.Neighboring points for a dataset are identified by certain distance metrics, generally euclidean distance.kNN method identify ‘k’ samples in the dataset that are similar or close in the space.Using these ‘k’ samples to estimate the value of the missing data points.Each sample’s missing values are imputed using the mean value of the ‘k’-neighbors found in the dataset.

2.Outlier handling

Outliers are those data points which differs significantly from other observations present in given dataset.In our dataset distribution is positively skewed and most of the outliers are present on the right side of the distribution.Transforming variables can eliminate outliers. These transformed values reduces the variation caused by extreme values.We used log transformation method in which it replaces each variable x with a log(x).  It transforms skewed data to approximately conform to normality.

3.Encoding

The machine learning models use mathematical equations. So categorical data is not accepted so we convert it into numerical form.

* Label Encoding

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form.For AQI bucket column label encoding is used.

* One-Hot Encoding

A one hot encoding is a representation of categorical variables as binary vectors.This first requires that the categorical values be mapped to integer values.Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.For city column this technique is used.

4.Dimentionality Reduction

Feature reduction, also known as dimensionality reduction, is the process of reducing the number of features in a resource heavy computation without losing important information. Reducing the number of features means the number of variables is reduced making the computer’s work easier and faster. In our dataset year is extracted from the date’s column and dropped date.Also different pollutants are classified into two groups as vehicular pollution contents and industrial pollution contents

5.Feature selection

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. Multicollinearity can be detected via various methods,the most common one –**VIF (Variable Inflation Factors)**. VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. VIF starts at 1 and has no upper limit ,If VIF = 1, no correlation between the independent variable and the other variables,VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others.In our dataset when checked for VIF the value for different variables ranges from 1.4 to 4.7 and so we assume moderate multicollinearity

6.SMOTE

Oversampling with SMOTE

Since it was an imbalanced data we had done the oversampling with SMOTE. For over-sampling techniques, SMOTE (Synthetic Minority Oversampling Technique) is considered as one of the most popular and influential data sampling algorithms in ML and data mining. With SMOTE, the minority class is over-sampled by creating “synthetic” examples rather than by over-sampling with replacement. After this we got the data with equally sampled.

**7. MODELLING AND RESULTS**

The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data. In this paper, we are using machine learning to solve the air pollution analysis problem.

**SVM**

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features.

**Accuracy** - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. For our model accuracy is 68.82%.

**Recall**(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class. We have got recall of 0.68 which is good for this model as it’s above 0.5.

**Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. We have got 0.68precision which is pretty good.

**F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. In our case, F1 score is 0.68.

**RANDOM FOREST**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems. Accuracy is: 0.78,Precision is: 0.78,Recall is: 0.78,F1 score is: 0.78

**LOGISTIC REGRESSION**

Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. Accuracy obtained is 0.49.

**DECISION TREE**

A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node)

holds a class label.

Here accuracy is 0.71,Precision is: 0.71,Recall is: 0.71 and F1 score is: 0.71

**GRADIENT BOOSTING**

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.Here accuracy is 0.75,Precision is: 0.75,Recall is: 0.75 and F1 score is: 0.75.

**XGBoost**

[XGBoost](https://xgboost.ai/)is a decision-tree-based ensemble Machine Learning algorithm that uses a [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. Here accuracy is 0.78,Precision is: 0.78,Recall is: 0.78 and F1 score is: 0.78.

Our model worked better for Xgboost

**8.HYPERPARAMETER TUNNING AND RESULTS**

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.

**GRADIENT BOOSTING**

For our model accuracy score increased to 0.76 after tunning.

**RANDOM FOREST**

Random forest is tunned by GridSearchCV method that, instead of sampling randomly from a distribution, evaluates all combinations we define.Accuracy improved to 0 .79.

**XGBoost**

Xgboost tunning done by RandomizedSearchCV for our model. XGBoost combined with RandomSearchCV is very strong in AQI index. We have compared it with the simple RandomForest with GridSearchCV at it outperforms it anyway, because RF have accuracy 0.7805 and XGB has accuracy 0.7808. And after tuning XGB and RF accuracy has been reduced and test and train accuracy for XGB is greater than Random forest after tunning.

**9.WEB HOSTING USING FLASK**

Flask is a micro [web framework](https://en.wikipedia.org/wiki/Web_framework) written in [Python](https://en.wikipedia.org/wiki/Python_(programming_language)). It is classified as a [microframework](https://en.wikipedia.org/wiki/Microframework) because it does not require particular tools or libraries.[[2]](https://en.wikipedia.org/wiki/Flask_(web_framework)#cite_note-2) It has no [database](https://en.wikipedia.org/wiki/Database) abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

Flask makes the process of designing a web application simpler. It lets us focuson what the **users are requesting and what sort of response to give back.**

**WEBSITE LINK**

<http://unn1.pythonanywhere.com/>

**CONCLUSION**

The proposed approach predicts the air quality index of different states. In this proposed work, different machine learning models like xgboost, decision tree, support vector machine,logistic regression and random forest were evaluated for the given data set. Optimization of hyper parameters for different machine learning models has been carried out. The comparison was made by varying the parameters of machine learning models and results were recorded for the same. Xgboost model gave good results for predicting the air quality index

**References**

1.https://www.kaggle.com/hirenvora/citydaycsv

2.Acosta NJR, Fuller R., Landrigan P. J. The Lancet Commission on pollution and health 2017

3. World Health Organization. Air pollution details. 2018

4. “Optimization of hyper parameters in machine learning techniques for air quality predictive analysis”Basamma Umesh Patil, Ashoka D. V, Ajay Prakash B. V, International Journal on Information Technologies & Security, № 3 (vol. 13), 2021