**PREDICTION OF AIR QUALITY**

**by**

Pusuloor S Shruthi

A research study proposal submitted in partial fulfillment of the

requirements for the degree of Master of Engineering in Information Management.

Examination Committee: Dr. Vatcharaporn Esichaikul (Chairperson)

Dr.Attaphongse Taparugssanagorn

Dr.Teerapat Sanguankotchakorn

Nationality: Indian

Previous Degree: Bachelor of Engineering in Computer Science Engineering, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India.

Scholarship Donor:AIT Fellowship

Asian Institute of Technology

School of Engineering and Technology

Thailand

May 2020

**Abstract**

Air quality in India is a field of trouble. There has been a steady decrease in the air quality from the previous years. As indicated by the reports of WHO/UNEP (1992), a portion of the metropolitan urban areas in India are in the list of most polluted urban areas on the planet. The poor air quality straightforwardly influences the environment and majorly affects the human wellbeing. Sulfur dioxide, nitrogen dioxide, respirable suspended particulate matter, particulate matter are the main pollutants in the air which are responsible for the poor air quality. According to the statistics around 4 to 5 million people die each year due to poor air quality. Poor air quality prompts breathing issues and legitimately influences the inside respiratory organs of the body. The air quality is measured with the help of air quality index (AQI) which gives a relative measure for surrounding air concentration. This paper gives a thought regarding the air quality in a portion of the districts of India. The study likewise illuminates the impact of parameters, for example, sulfur dioxide, nitrogen dioxide, respirable suspended particulate matter, particulate matter on to air contamination and the requirement for the measures against the poor air quality.

**CHAPTER 1**

**Introduction**

In this chapter some background related to the topic along with the problem statement

around which the research revolves, objectives and scope of the research and the limitations

of this research study will be discussed.

**1.1 Overview:**

Air pollution is one of the most serious problems in the world. The Earth's climate is a blend of gases and particulate-phase substances. The most abundant of these, nitrogen (N2) and oxygen (O2), involve around 78% and 21%, separately, of climatic mass and volume \cite{godish2014air}. It refers to the contamination of the atmosphere by harmful chemicals or biological materials. There are many air pollution indicators affecting human health. Some of the important indicators are particulate matter (PM10), carbon monoxide (CO)and sulphur dioxide (SO2). Air pollution in India is a serious issue with the major sources being fuelwood and biomass burning, fuel adulteration, vehicle emission and traffic congestion. In autumn and winter months, large scale crop residue is being burnt in agriculture fields which is a low-cost alternative to room heaters and electric blankets. This is the major source for smoke, smog and particulate pollution. India has a low per capita emissions of greenhouse gases, but the country as a whole is the third largest after China and the United States. A 2013 study on non-smokers has found that Indians have 30% lower lung function compared to Europeans.

The Air Prevention and Control of Pollution Act was passed in 1981 to regulate air pollution and there have been some measurable improvements. However, the 2016 Environmental Performance Index ranked India 141 out of 180 countries. So, we have to perform analysis on Air quality data for the better understanding the cause of air pollution and affecting features.

Air Quality Index (AQI) is explained as an index or rating scale for giving the detailed day by day joined impact of surrounding air pollutants recorded in the atmosphere.\cite{kumar2011forecasting} So the Air Pollution levels can be easily understood from the AQI.

**1.2 Problem Statement:**

Air pollution is one of ecological issues that can't be disregarded. Breathing in poisonous gases for a long-term cause harms in human wellbeing. To solve the issues of air contamination, it's important to comprehend the issues and search for approaches to control it. Predicting the AQI for India by using Machine learning model can help in finding the important insights of Air Quality of India dataset.

**1.3 Objectives:**

The main objective of this research study it to develop a system model to predict the air quality in India. The sub objectives to achieve this research are as following:

1. To predict the district level air pollution of a state using different machine learning models.
2. To build a system which can predict the level of air pollution.

**1.4 Limitations and Scope:**

After data analysis, suggestions and advices can be given to control the components which are influencing Air Quality Index (AQI) in India. It is required to predict the following day's air pollution levels and for coming up with right activities and controlling strategic methods. Air quality warning systems therefore required to get early notification when the air pollutants levels may cross the air quality rules or the air quality limits. Warnings can be given to the public to take care of their health and to the traffic and environmental management so that the further effects can be controlled and minimised. And the warnings should be easily understandable and reliable to the public.

**1.5 Research Outline:**

I organize the rest of this dissertation as follows.

In Chapter literature-review, Literature review and the related works are discussed.

In Chapter methodology, Proposing the step by step process involved to achieve the objective.

In Chapter results, Experimental results are presented.

In Chapter conclusion, Final conclusion of the research and the recommendations for the future work are discussed.

**CHAPTER 2**

**LITERATURE REVIEW**

This chapter provides us a description about Air Quality Index, factors affecting air quality, pollutant index it also gives us knowledge about data mining techniques and its uses in the air quality index prediction, and some previous work related to topic.

**2.1 Air Quality:**

Air quality is a proportion of how free the air is from dirt or how polluted the air is. Observing air quality is very important because breathing polluted air is dangerous for human’s health and also to the health of environment. Air quality is measured with the help of Air Quality Index.

**2.1.1 Air Quality Index:**

Air quality is estimated with the Air Quality Index, or AQI. AQI is just like a thermometer that runs from 0 to 500 degrees. Rather than indicating changes in the temperature, it shows how clean or polluted the air is.\cite {Airquality}

**2.1.2 What is in the Air?**

The Air in the atmosphere mostly contains two gases which are very essential for the life on earth: Nitrogen and Oxygen. However, air also contains many other gases and particulates in smaller amounts. Air quality index tracks five major air pollutants. They are:

1. Ground level ozone
2. Carbon monoxide
3. Sulfur dioxide
4. Nitrogen dioxide
5. Airborne particles, or aerosols

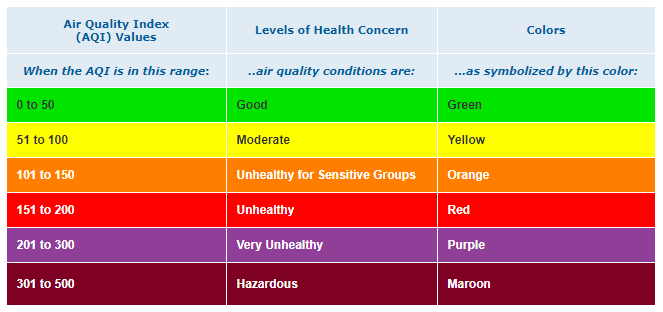
Environmental Protection Agency calculates AQI for these five dangerous air pollutants. EPA is an independent agency of the Unites States for the protection of environment to protect public health. Ground level ozone and Airborne particles are main two elements which cause greatest risk for human health. These are also main part in smog. Smog is a type of air pollution, due to which visibility reduces.\cite{Airqualitymeasure}.

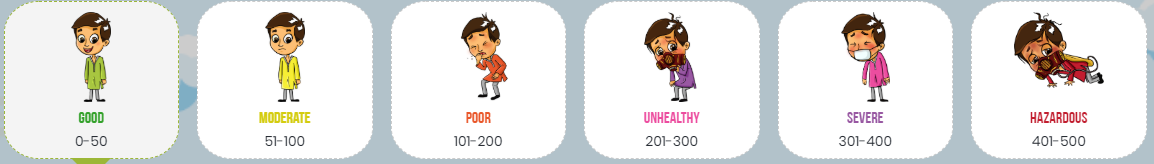
A Picture showing smog over downtown Los Angeles.



**2.1.3 Understanding AQI**

The main purpose of AQI is to help everyone to understand what air quality means to the health. To make it more obvious, the AQI is separated into six classes:





Every classification relates to a dissimilar level of wellbeing concern. The six levels of health concern and the meaning are as following:

* "Good" AQI is 0 to 50. Air quality is viewed as acceptable, and the air pollution is practically zero hazard.
* "Moderate" AQI is 51 to 100. Air quality is adequate. However, for few pollutants can show effect on few people's health who are very sensitive. They may experience respiratory symptoms.
* "Unhealthy for Sensitive Groups" AQI is 101 to 150. Although general people are not effected much at this AQI level, but it is risk for people who are suffering from lung and heart diseases and for older people and children.
* "Unhealthy" AQI is 151 to 200. At this level of AQI most of the people may suffer with adverse health effects, but people who are more sensitive will face more serious effects.
* "Very Unhealthy" AQI is 201 to 300. This will trigger everyone that everyone may face serious effects.
* "Hazardous" AQI greater than 300.This would trigger emergency conditions. The entire population is bound to be effected.

**2.1.4 Factors affecting air quality:**

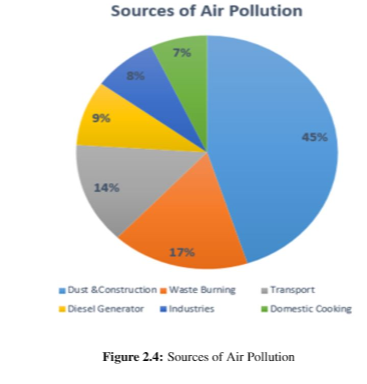
The amount and kind of pollutants that are released into the air play a major role in determining the degree of air pollution in a specific area. However, other factors are involved, mainly:

* Topography (terrain), such as mountains and valleys;
* Weather, such as wind, temperature, air turbulence, air pressure, rainfall and cloud cover; and
* The physical and chemical properties of pollutants.

Major outdoor air pollutants in cities include ozone (O3), particle matter (PM), sulfur dioxide (SO2), carbon monoxide (CO), nitrogen oxides (NOx), volatile organic compounds (VOCs), pesticides, and metals, among others. Increased mortality and morbidity rates have been found in association with increased air pollutants (such as O3, PM and SO2) concentrations.

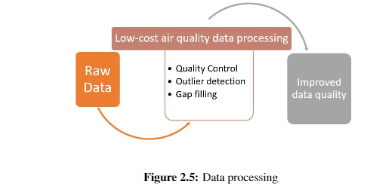
Although O3 precursor (such as VOCs, NOx, and CO) emissions have significantly decreased since the late 1970s, O3 levels in India have not been in compliance with standards set by the Environmental Protection Agency (EPA) to protect public health. Particle size is critical in determining the particle deposition location in the human respiratory system. PM2.5, referring to particles with a diameter less than or equal to 2.5\( \mu M\) , has been an increasing concern, as these particles can be deposited into the lung gas-exchange region, the alveoli. Meteorological conditions, including regional and synoptic meteorology, are critical in determining the air pollutant concentrations.

Increasing wind speed could either increase or decrease the air pollutant concentrations. For instance, when the wind speed was low (weak dispersion/ventilation), the pollutants associated with traffic were found at the highest concentrations. However, strong wind speeds might form dust storms by blowing up the particles on the ground \cite{argyriou2010spectral}. High humidity is usually associated with high concentrations of certain air pollutants (such as PM, CO and SO2) but with low concentrations of other air pollutants (such as NO2 and O3) because of various formation and removal mechanisms.



**2.2 Data Mining Concepts and Techniques:**

In order to implement a successful data mining solution, the user must analyze and formalize their objective. The problem objective guides the user to the appropriate paradigm of learning algorithm. If the objective is to identify hidden groups in data or identify associations between key variables in the data, the users are interested in knowledge discovery and will want to select a clustering or association mining algorithm \cite{corani2005air}. Alternatively, the objective might be to induce a predictive model that can classify samples as belonging to a particular category, such as poor air quality, or a real-valued outcome, such as the air quality index.



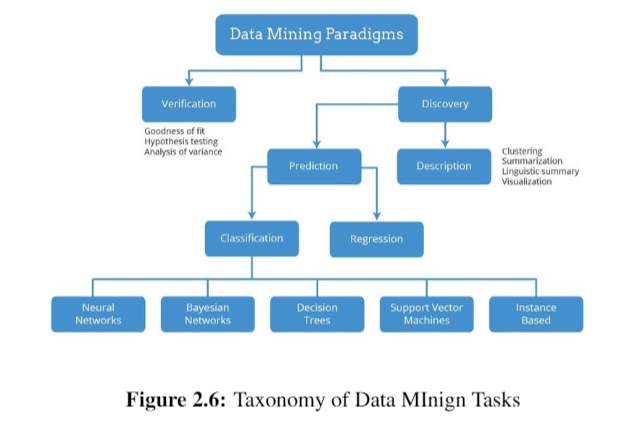
Data Mining follows several sequences in several steps. The steps are as follows:

* Data Cleaning: Removal of inconsistency Data
* Data Integration: Union of different data sources in to a single source.
* Data Selection: Selection of data according to the requirements.
* Data Transformation: Changing into different types of data by consolidating.
* Data Mining: It’s a process of extracting the data patterns.
* Pattern Evaluation: Identifying the different patterns based on interestingness measures.
* Knowledge Presentation: Visualization is used to present knowledge.

**2.2.1 Data Mining Process:**

This is used to describe the large data sets in order to complete the project to measure and manage the data for the prediction purpose. This should follow the some of the steps which are shown below:

1. Data Understanding: The data in the data set should be studied and have to understand that which is consistent, and which is not consistent and to know about the missing values.
2. Data Preparation: This data preparation takes almost 90\% of the total project time to prepare the data. This prepared dataset is the final dataset which is used for the data mining. For preparation of dataset first the data has to be selected, cleaned and constructed into a desired format.
3. Modelling: Firstly, the techniques have to be selected for modelling of the data. Next, the dataset has to separate in to testing and training sets later testing data should be evaluated for getting the accuracy.
4. Evaluation: In this phase, the results of models must be evaluated. By seeing the results the decision must be made to go into the next step.
5. Deployment: The information obtained from the above all these steps are made into a report and has to be shown for the stakeholders.



**2.2.2 Prediction using Data Mining Techniques:**

Classification and Regression are two major prediction problems which are usually dealt with Data mining and machine learning**.**

* Classification: It is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes i.e. discrete values. In classification, data is categorized under different labels according to some parameters given in input and then the labels are predicted for the data.

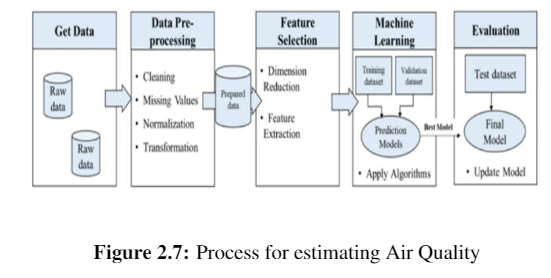
The derived mapping function could be demonstrated in the form of “IF-THEN” rules. The classification process deal with the problems where the data can be divided into binary or multiple discrete labels.

* Regression: Regression is the process of finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can also identify the distribution movement depending on the historical data. Because a regression predictive model predicts a quantity, therefore, the skill of the model must be reported as an error in those predictions.

**2.3 Machine Learning:**

Due to advancement in technology, artificial intelligence-based algorithms are being widely used for prediction purposes, especially for air quality forecasting. A machine learning approach takes into account multiple parameters for prediction unlike a pure statistical model. Artificial Neural Networks (ANNs) have appeared to be the most widely used technique for prediction of air quality. Other studies have presented use of hybrid models or mixed models based on neural networks for prediction.

Another study conducted in Jugder, used six meteorological factors for predicting PM2.5 concentrations. K. Hu et al. \cite{natsagdorj2003analysis}, designed a machine learning model HazeEst for predicting the air quality. Here, first the system was assessed using seven different regression models and finally SVR was selected as the final prediction model. Similar research was carried out in Gauteng, South Africa for prediction of ozone concentrations at ground-levels using ANNs and multiple linear regression techniques. Another efficient machine learning method that is used is Extreme Learning Machine (ELM) which is a non-linear machine learning algorithm. Here, randomized neural networks are used to forecast O3, NO2 and PM2.5 concentrations based on this non-linear methods using the data from six stations spread across Canada. Following are the few algorithms which will used for the prediction.



**2.3.1 Decision Tree:**

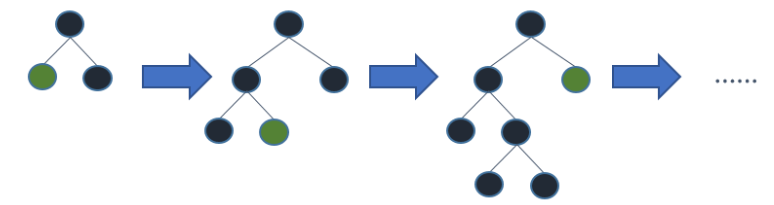
Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. It is a simple representation for classifying examples.

**2.3.2 Random Forest:**

Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach. It can be used for both classification and regression tasks. Usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

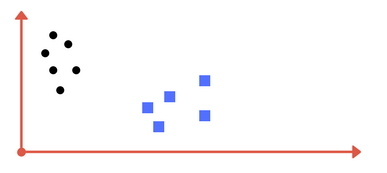
**2.3.3 Light GBM:**

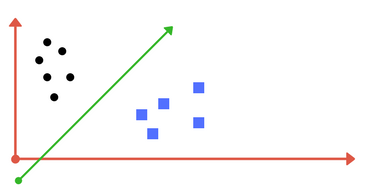
Light GBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification and many other machine learning tasks.Since it is based on decision tree algorithms, it splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise. So when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms. Also, it is surprisingly very fast, hence the word ‘Light’.



**2.3.4 Support Vector Machines:**

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. It is a discriminative classifier formally defined by a separating hyper plane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. In two-dimensional space this hyper plane is a line dividing a plane in two parts where in each class lay in either side. It fairly separates the two classes. Any point that is left of line falls into black circle class and on right falls into blue square class. Separation of classes. That’s what SVM does. It finds out a line/ hyper-plane (in multidimensional space that separate outs classes).





**2.3.5 Naive Bayes:**

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

**2.4 Related work:**

Many previous works have been proposed to apply machine learning algorithms for air quality predictions.

Jianzhou Wang aimed to develop an early warning system for the prediction of air quality and assessment modules which helps to emphasize the uncertainty of quality of air, so that information can be extracted to improve the performance of early warning system. This air quality assessment described in this paper helps in understanding and analysing the levels of different pollutants. The six air pollutants which are considered as inputs are carbon monoxide, sulfur dioxide, nitrogen dioxide and particulate matter smaller than 10 and 2.5 $\mu$ m.In this study, back propagation neural network combined with data preprocessing techniques are used to predict. Chengdu and Hangzhou which are located in china are considered. Singular Spectrum Analysis algorithm is also used to improve the BP neural network performance. \cite{wang2017developing}

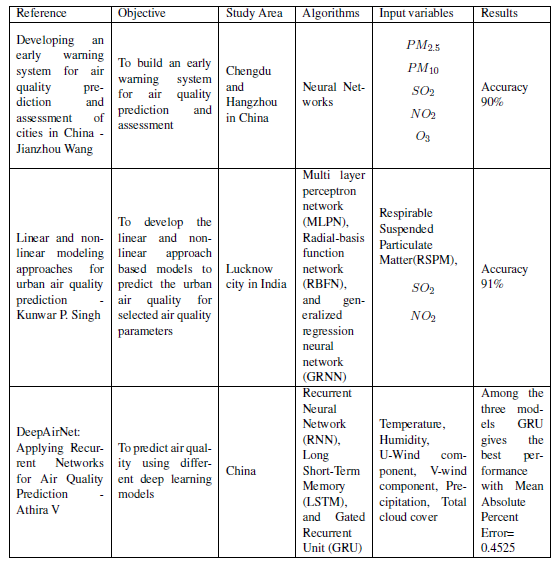
Kunar P. Singh aimed to develop linear and non linear based models i.e partial least squares regression (PLSR), multivariate polynomial regression (MPR), artificial neural network (ANN) for the prediction of urban air quality of the city Lucknow, in India using set of five independent variables . And the second objective is to compare predictive and generalization abilities using Multi layer perceptron network (MLPN), Radial-basis function network (RBFN), and generalized regression neural network (GRNN) models. Using Kennard-Stone approach data was partitioned into training, validation and test datasets. Here predictive and generalisation capabilities of linear (PLSR) and the non-linear (MPR,MLPN,GRNN) models for Sulfur dioxide, Nitrogen dioxide and Respirable Suspended Particulate Matter (RSPM) were compared. ANN performed relatively better than MPR and PLSR.The study futhur concluded that ANN's three variants (MLPN, RBFN, GRNN) were able to catch the non linearity complexities in data and showed good predicitve and generalisation capabilities. Therefore, ANN models is one of the useful tools in air quality predictions. \cite{singh2012linear}

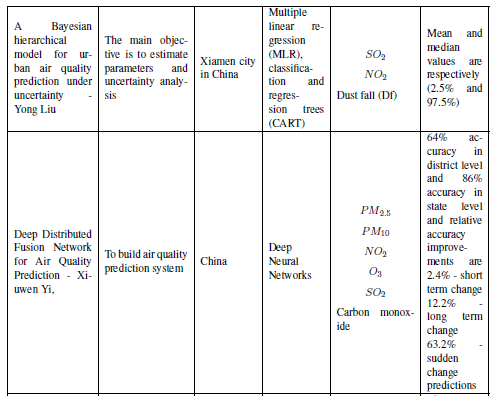
Athira V aimed to apply deep learning model to predict the air quality i.e using Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU).The data is collected from 1498 stations from China National Environmental Monitoring Centre and is collected from April 1, 2012 to September 1, 2017.RMSE, MAPE values are noted for all the three models. Among the three models, GRU gave the best performance than the RNN and LSTM.\cite{athira2018deepairnet}

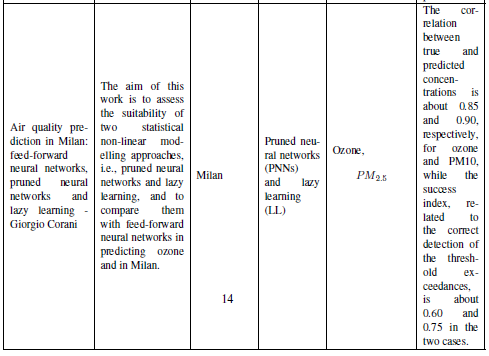
Yong Liu aimed to develop a Bayesian hierarchical model for urban air quality prediction in Xiamen city, China. Air quality data of Huang is considered. And the air pollutants which are considered as inputs are Nitrogen dioxide, Sulfur dioxide and dust fall. Correlation Analysis, classification and regression trees, Hierarchical cluster analysis, discriminant analysis was used to determine air quality model structure and prior distributions of model parameters. To measure the relationship between pollutant concentrations and driving variables, a multiple linear regression equation is proposed. The mean and median values are 2.5\% and 97.5\%. And the average relative errors between the mean values and observed data of Nitrogen dioxide, Sulfur dioxide and dust fall are 6.81\%, 6.79\%,3.52\%.\cite{liu2008bayesian}

Xiuwen Yi aimed to address these challenges by proposing a DNN based air quality prediction. Here air quality of next 48 hours for each monitoring system is predicted for 300+ Chinese cities by air quality data, meteorology data and weather forecast data. Air pollution prediction system is deployed as "cloud+client" framework, where real time data is continuously collected by the cloud to make predictions and the public air quality information is available. The data is compared with ten baseline, they achieved a higher accuracy for general and sudden changes. And the relative accuracy improvements on short-term, long term and sudden change predictions are 2.4\% , 12.2\%, 63.2\% respectively.\cite{yi2018deep}

Giorgio Corani aimed to assess the suitability of two statistical non-linear modelling approaches (pruned neural networks and lazy learning) and to compare with the feed forward neural network (FFNN) for predicting ozone and particulate matter 10 micrometers in Milan. Lazy learning is found to be the best performing approach for such as mean absolute error and correlation. Best approaching for detecting threshold exceedances is Pruned neural networks.\cite{corani2005air}







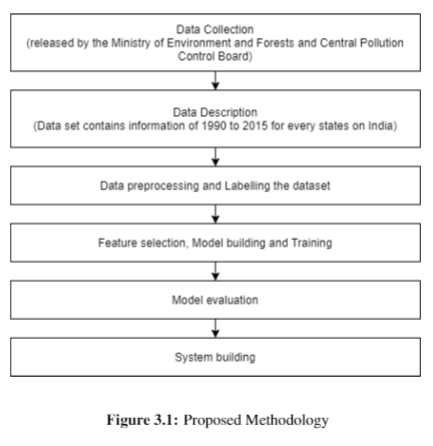
**CHAPTER 3**

**METHODOLOGY**

In this chapter, a step by step flow of different steps i.e. the proposed methodology used to complete the research and achieve results are discussed. This chapter will also help the to understand better about the process of research implementation.

**3.1 Proposed Methodology:**

The main aim of this study is to identify the trends, causes and the demographic insights of the pollution levels across the states of India. The proposed methodology consists of six steps : 1. Data collection, 2. Data description, 3. Data prepossessing and labelling of data, 4. Feature selection, Model building and Training, 5. Model evaluation and System building.

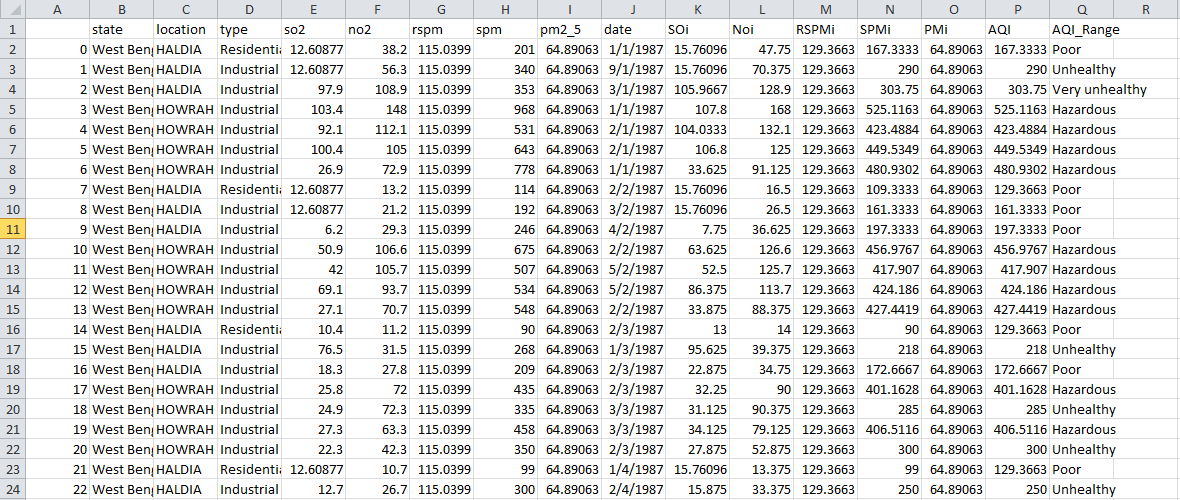


**3.2 Data Collection:**

This data is clean version of the Historical Daily Ambient Air Quality Data released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).

Reference -http://www.data.gov.in

**3.3 Data Description:**

****

Dataset consists of 22463 rows and 17 attributes. The dataset contains the information of 1990 to 2015 for every states sulfur dioxide, suspended particulate matter, respirable suspended particulate matter information, nitrogen dioxide etc.

Attribute informationof the data are are follows:

* Unnamed 0: This is unique id given to each recoed of the data.
* State: As we are doing the analysis for the West Bengal. So it has same value across the data.
* Location: Represents the different locations of the west bengal.
* Type: It represent the type due to which air quality has increased or decreased.
* so2: The amount of sulphur dioxide measured.
* no2: The amount of nitrogen dioxide measured.
* Rspm: RSPM measured.
* Spm: SPM measured.
* pm2\_5: Value of particular matter measured.
* date: Date on which record is measured.
* SOi: Soi measured.
* NOi: Noi measured.
* RSPMi: RSPMi measured.
* SPMi: SPMi measured.
* PMi: Pmi measured.
* AQI: Real values which represents the AQI ( Regression target attribute).
* AQI\_Range: Categorical values which represents the AQI Range (Classification target attribute).

**3.3.1 Why we need these details?**

Sulphur dioxide, nitrogen dioxide ,suspended particulate matter , respirable suspended particulate matter, particulate matter 2.5 values are the parameters used to measure the quality of air based on the number of particles present in it. Using these values, we will identify the air quality over the period in different states of India.

**3.4 Data Preprocessing:**

The data set which is initially obtained is inconsistent data i.e., many missing values, null values, so this data preprocessing is done for acquiring the consistent data. Therefore, that it could be easy for the algorithms that are going to be implemented. Some of the features in the data set are in string and categorical format. Machines can’t understand the string or categorical values so these should be changed into numerical formats, for changing them into numerical formats one-hot encoding and binary encoding are used in this research project.

* One hot encoding : Categorical features are transformed to a format which works better with classification and regression by using one end coding for the better prediction. For each unique category a separate column is created using 0 and 1.
* Binary encoding : In binary encoding it can only be applied to a attribute with only two unique values. So for this type of column, only one new column is created, which will contain either 1 or 0.

Data cleaning and missing value imputation, data transformation, one hot encoding, normalization or standardization etc. these preprocessing tasks will be performed.At first data analysis is done to know the basic information of the data , for example to know the level of concentrations of the gases in various states with the help of visualisation. And then data cleaning is done to know the number of null values, missing values for each column so that unwanted data can be dropped.

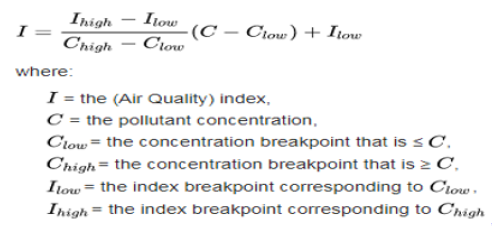
**3.5 Feature Selection:**

Statistics shows that, population of India in year 2018 was around 1,358,137,719. India is one of the developing nations; therefore there is need for the perfection in field such as education, agriculture, proper sanitation, curability of diseases, traffic control and safety, maintenance of good air quality, etc. Emission of these gaseous pollutants has been identified as a major source of air pollution. Among the total population caused in cities like Delhi, Mumbai and Kolkata, 70\% of pollution is caused due to the pollutants majorly SO2, NO2, RSPM and PM10. \cite{AQI}

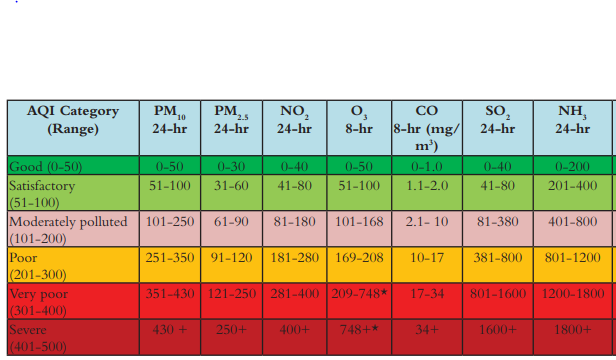
The Data set consists of 13 attributes. In which State, SO2, NO2, RSPM, SPM, PM2.5,type are selected for the feature selection, because these are the attributes we need to predict the Air quality index value. These attributes we will use to fit the machine-learning model by using different classification algorithms and regression models for making better prediction of Air quality index of India. The data set we will split into training and testing data, where training data used 80\% of the data set and testing data used 20\% of the total data set.

**3.6 Computing the Air quality index:**

To convert the concentration to Air quality index, the equation mentioned below is used:



The Central Pollution Control Board along with State Pollution Control Boards has been operating National Air Monitoring Program (NAMP) covering 240 cities of the country having more than 342 monitoring stations. The proposed Air quality index will consider eight pollutants (PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb) for which short-term (up to 24-hourly averaging period) National Ambient Air Quality Standards are prescribed. The Air quality index values and corresponding ambient concentrations (health breakpoints) as well as associated likely health impacts for the identified eight pollutants are as follows:



**3.7 Model Building:**

In this project, we will compare various machine-learning models to find which model works better for predicting the Air quality index (Air Quality Index).Machine learning models used in this project; we are using for regression such as:

1. Random Forest
2. Ridge Regression
3. XGBoost

For the classification problem we are using:

1. K-Nearest Neighbor
2. Logistic Regression
3. Gradient Boosting

**3.8 Model Evaluation and System Development:**

The research we will use machine-learning algorithms to build the models & there is a need to find the best fitting model, which gives the best results by using the accuracy metrics. Error metrics used In this project. We will use the following error metrics to evaluate and compare our models:

1. Coefficient of determination (R\*2)
2. Root Mean Square Error (RMSE)
3. Mean absolute error (MAE)
4. Root Mean Squared Logarithmic Error (RMSLE)

We will use the following error metrics to evaluate and compare our models for classification task:

1. Confusion Matrix
2. Accuracy

**3.9 Experiments:**

We have perfomed the above 3 algorithms on the dataset for regression and have obained the following results for each model.

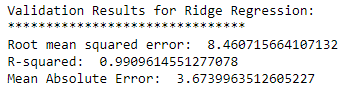


Figure: Results obtained for Ridge regression

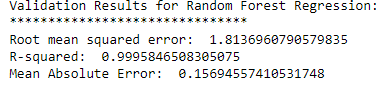


Figure: Results obtained for Random Forest regression

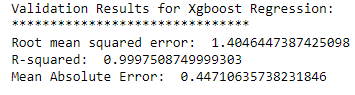


Figure: Results obtained for XGBoost regression

Of the all above 3 models the XGBoost has performed well. So we will use it for the final prediction on the web application.

Also for classificaion problem we have perfomed the above 3 algorithms on the dataset and have obained the following results for each model.

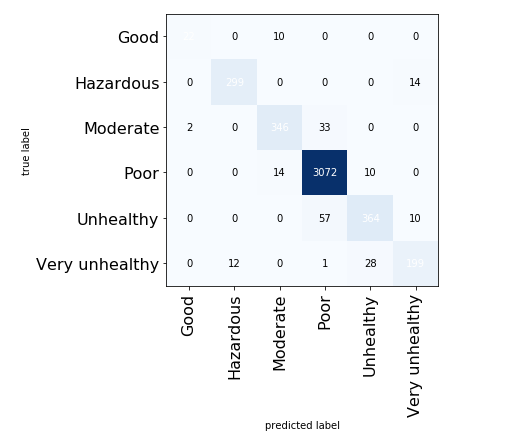


Figure: Results obtained for KNN

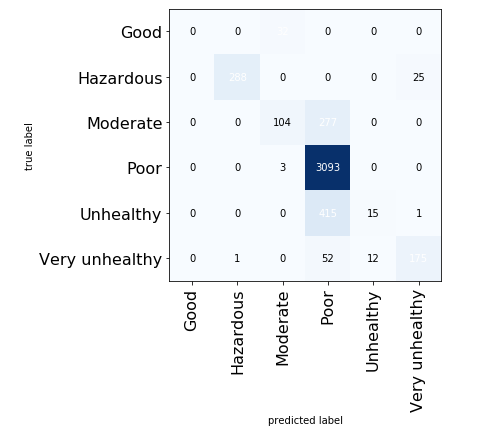


Figure: Results obtained for Logistic Regression

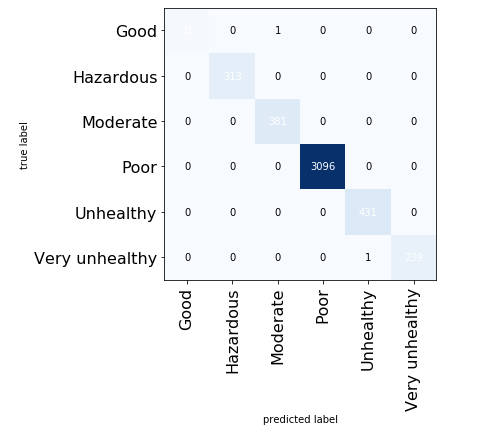


Figure: Results obtained for Gradient Boosting

Of the all above 3 models the Gradient Boosting has performed well. So we will use it for the final prediction on the web application.

**3.10 System Development:**

Flask will be used to develop the Air quality index of India prediction system. Flask is micro web-framework written in Python. The flash package can be downloaded from Python Package Index (PPI).It is written in Python. By using flask, we will deploy our model as a web application. Our project takes data from harmful air particles from other sources of air pollution and predicts the air quality index. The following are the snippets for the web application.

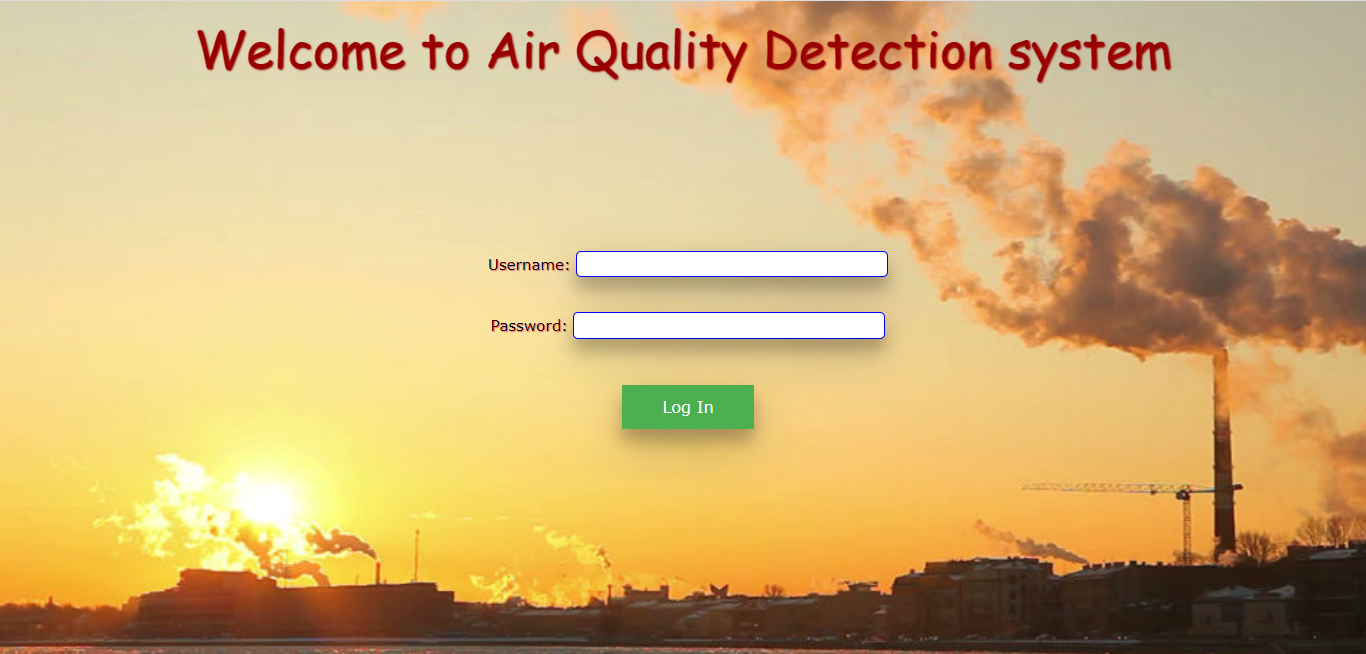


Figure: Web Application login page

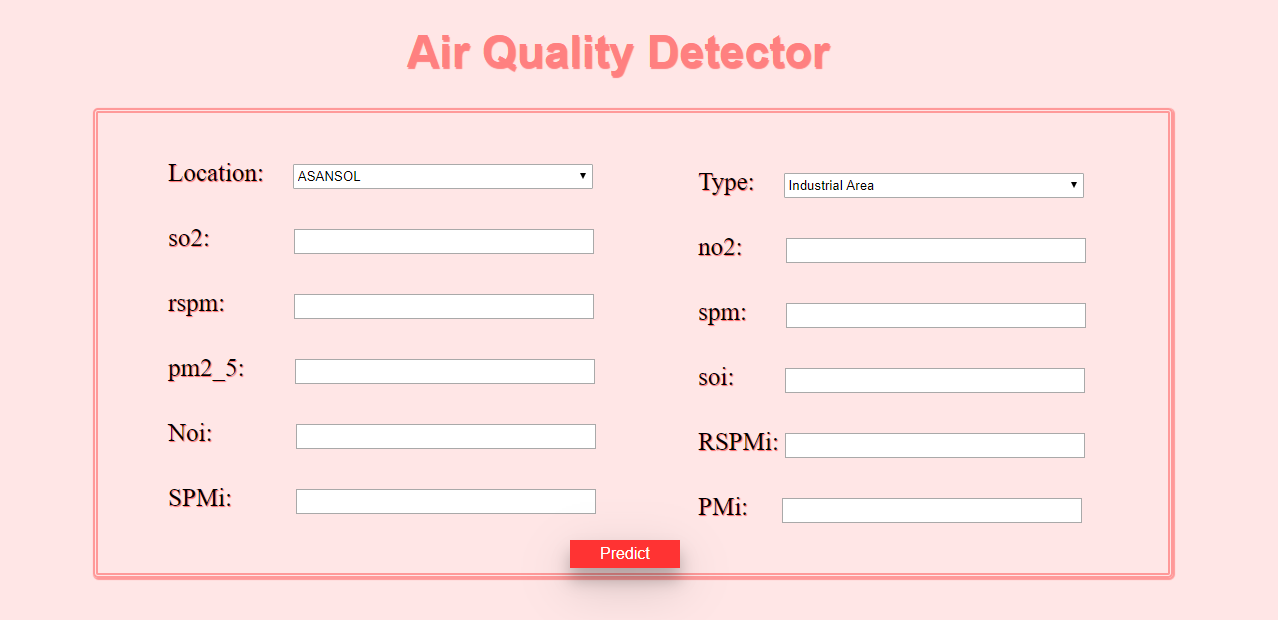


Figure: Web Application Main page

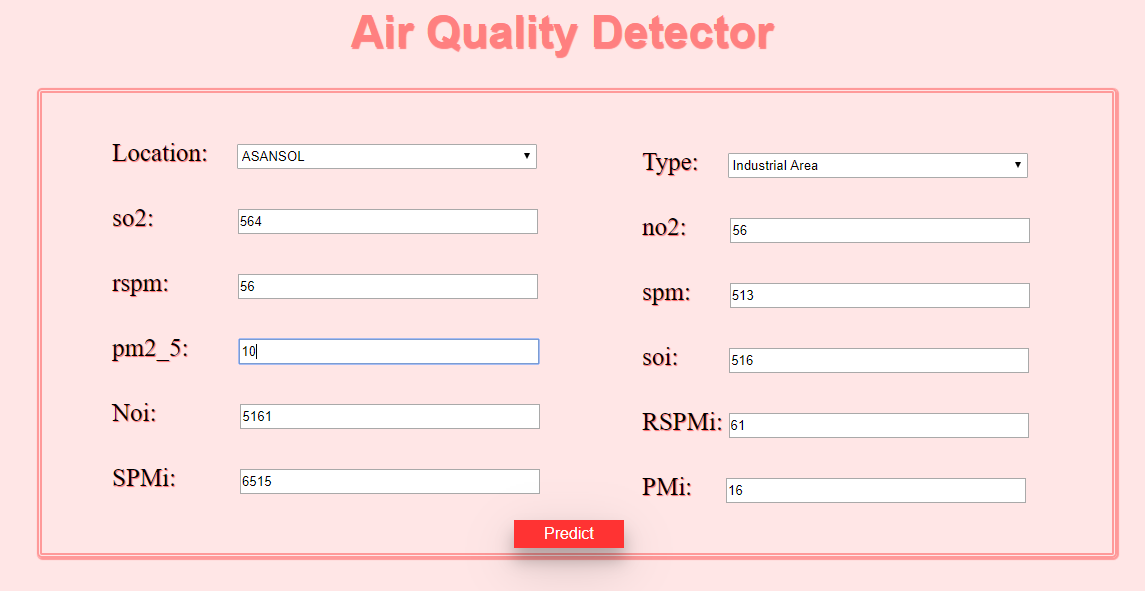


Figure: Web Application Main page

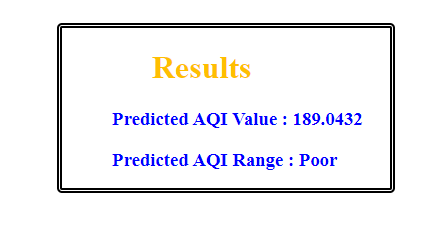


Figure: Web Application Result page

**CHAPTER 4**

**CONSLUSION**

We have successfully built our application with the use of Machine Learning. We have used the different ML in order to automate or predict on the data. We have used various evaluation metrics in order to compare the performance of the different models. Finally based on the good performing we are able to do web deployment for this application.

So, we have also built a web application where user can login with the given credentials and fill the all the information. Once the information is sent on server then our developed model will be running on that data and makes the predictions. Those predictions again sent back to the web for the user and user can check the results.