Pollution detection and prediction

December 20, 2019

1 Introduction

1.1 Project area

Environmental pollution is one of the most serious problems facing humanity and other life forms on our planet today, industrial pollution contributing a major share in it. Industrial pollution is generally referred to the undesirable outcome when factories or other industrial plants emits harmful by-products and waste into the environment such as emissions to air or water bodies. The six major types of pollutants are carbon monoxide, hydrocarbons, nitrogen oxides, particulates, Sulphur dioxide and photochemical oxidants.

The Paris agreement's central aim is to strengthen the global response to the threat of climate change by keeping a global temperature rise this century well below 2 degrees Celsius above pre-industrial levels. Long term exposure to polluted air and water cause's chronic health problems making the issue of industrial pollution into a severe one .It also lowers the air quality in surrounding areas which causes many respiratory disorders affecting both lungs and heart. Not just the humans, but the marine life is greatly deteriorating and affected with the extent of increasing industrial pollution. However with effective measures, the ill effect of industrial pollution could be reduced significantly. The prevention and control of industrial pollution are highly encouraged by government worldwide. Simple things like purchasing energy-efficient equipment and products made from recycled materials for your organization. Having industrial pollution control policies in place and guiding strictly upon them.

1.2 Motivation

The Draft Environment Laws (Amendment) Bill, 2015 was published by the Ministry of Environment, Forest and Climate Change (MoEFCC) on October 7, 2015. The objectives of the Draft Bill are to provide for "effective deterrent penal provisions" and to introduce "the concept of monetary penalty for violations and contraventions".There are no effective strict rules for pollution monitoring and control in industries yet. But the government of India is in the midst of making industrial pollution control and monitoring laws more strict.

1.3 Gap Analysis

The previous work in this field included setting up monitoring stations to measure the amount of pollutants in the atmosphere .This was done using network of sensors topologically arranged as a grid. These sensors are usually placed in a large area for e.g. around a city or a large industry. The placement of sensors around a large area can reduce the accuracy of the model. This happens because of external factors like weather, noise, etc. affect the accuracy of the sensors to a large extent. Also, these models can't accurately reflect the extent of pollution a particular industry is causing due to the large area. Some previous models also predicted pollutant concentration for the future based on machine learning algorithms. Also there were some models that predicted dispersion of pollutants from a point source i.e. the area to which the pollutants will spread to. No previous models have linked the emission rate of the stack and the dispersion from the source(stack) together i.e. they have not been predicting the emission rate and calculating the spread in the same model.

The proposed model will however predict emission rate at the stack and use it as a source to calculate dispersion using air dispersion models. This will be done by placing sensors at the source (stack). Placing the sensors at the source will help the sensors take more accurate readings as the sensors will not be affected by external factors to a large extent as compared to the previous models. This will allow us to take more accurate readings and hence make more accurate predictions of future emission rates.

1.4 Objectives of the project

The main objective of the project is to monitor the pollutants emitted from an industry/factory and predict the future dispersion of these pollutants also to then output these results in the form of reports for industries/factories. To accomplish the main objective can be broken down into 3 sub-objectives which can be defined as

- Sensor module: to create a series of gadgets which can measure the emission parameters and the meteorological parameters and also that can withstand the environment conditions present near the stack/chimney and transmit this information to a server.
- Prediction module: this module uses machine learning models with data acquired from the cloud server to predict future emission parameter values.
- Air dispersion module: to create a module that simulates the movement of fluid particles (pollutant) in air using air dispersion models with meteorological data.

1.5 Dataset and study area

The objective of the prediction module is to predict the emission rate at the stack/chimney. To build and test the prediction model dummy data was gen-

erated using Gaussian distribution to randomly generate values for the feature set. The feature set is chosen on the bases on what factors will correlate to the emission rate and the type of pollutants emitted from the stack.

The feature set consists of independent variables: day, month, type of industry, size of industry, and output efficiency of industry and dependent variables: emission rate.

Definitions:

- Type of industry: what the industry/factory produces which will correlate to what gases are emitted out of the stack/chimney, this will be in the form of labeled classes.
- Size of industry: how big is the industry/factory which will correlate how much the maximum is outputted, this will be represented in the form of a scale from 1 to 10.
- Output efficiency: the amount of output it produces each day divided by the total amount of output it can ideally produce.
- Emission rate: this is defined as the amount of pollutants released from the stack per unit time.

The prediction module is also used on the collected metrological data to predict the air velocity and direction, this is then applied to calculate the dispersion of pollutants in air.

The feature set of this prediction model consists of independent variables: day, month, ambient temprature, ambient pressure, moisture content and dependent variables: wind velocity and wind direction.

2 Litrature survey

Sr no.	Name of Author	Title of paper	Description
1	Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie	Air Quality Prediction: Big Data and Machine Learning Approaches	• This paper reports recent literature study, reviews and compares current research work on air quality evaluation based on big data analytics, machine learning models

Sr no.	Name of Author	Title of paper	Description
2	Chen Xiaojun1, Liu Xianpeng2, Xu Peng3	IOT- Based Air Pollution Monitoring and Forecasting System	• It was also observed that adding more meteorological factors, the prediction performance is greatly improved. • They used past 5 years data for training the model. Which shows large sample data can improve performance and train the model well. • The overall experiment used IOT and neural network for air pollution monitoring and forecasting.
3	Temesegan Walelign Ayele , Rutvik Mehta	Air pollution monitoring and prediction using IoT	• The proposed work on an air pollution monitoring and prediction system enables to monitor air quality with the help IoT devices • For predicting the LSTM is implemented. It has a quick convergence and reduces the training cycles with a good accuracy

Sr no.	Name of Author	Title of paper	Description
4	Zhongshan Yang, Jian Wang	A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction	• In this paper, fuzzy comprehensive evaluation was used to determine the main air pollutants and evaluate the level of air pollution. • The experimental results showed that the proposed model has the best accuracy and stability compared to the other five individual models and five combined
5	Dixian Zhu 1,*, Changjie Cai 2, Tianbao Yang 1 and Xun Zhou 3	A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization	• They have developed efficient machine learning methods for air pollutant prediction. • They have formulated the problem as regularized MTL and employed advanced optimization algorithms for solving different formulations.
6	Mahmoud Reza Delavar 1,* , Amin Gholami 2, Gholam Reza Shiran 3, Yousef Rashidi 4 , Gholam Reza Nakhaeizadeh 5, Kurt Fedra 6 and Smaeil Hatefi Afshar 2	A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran	• A comparative study of machine learning methods including NARX, ANN and SVR has been employed for air pollution prediction and the NARX finally selected as the optimum one.

Sr no.	Name of Author	Title of paper	Description
7	Xiuwen Yi1,2, Junbo Zhang2,1,+, Zhaoyuan Wang1,2, Tianrui Li1, Yu Zheng2,1,3	Deep Distributed Fusion Network for Air Quality Prediction	• They proposed a DNN-based approach to predict air quality. • This approach achieves a higher accuracy in both general cases and sudden changes
8	QU Hongquan , PANG Liping	A Dynamic Method to Estimate Source Emission Rate and Predict Contaminant Concentrations	• This paper develops a dynamic method to estimate source emission rate and predict contaminant concentrations in an enclosed space. • Based on a variable structure model ofconcentration, this method uses EKF algorithm in combination with least squares method to realize state prediction and parameter estimation at the same time. • This method could realize to track and real-time predict contaminant concentration, and identify source emission rate accurately and efficiently.

Sr no.	Name of Author	Title of paper	Description
9	N. Li and S. Thompson	A Simplified Non-Linear Model Of NOx Emissions In A Power Station Boiler	• This paper has presented a model of NO emissions for a power plant boiler. It is modelled from the extended Zeldovich mechanism and require only a few physical parameters obtained from experiments. • A set of new test data is used to compare the simulated values with real measurements. It is shown that good results are obtained from the model with real plant input variables. • The model can also be used in other applications such as for optimising boiler operation and combustion control system design.

Sr no.	Name of Author	Title of paper	Description
10	S.H. Yu, Y.S. Koo, E.Y. Ha and H.Y. Kwon	PM-10 Forecasting using Neural Networks Model	• In this paper, the study on the factors to affect the PM-10 pollution and develop a PM-10 prediction model using MLP neural network model was done. • Especially, neural model has an advantage that there doesn't need to analyze the input data before the data are used, like regression model. • To improve the performance of the model, it needs to shorten the learning period from year to quarter month and to learn and predict PM-10 with multiple networks according to the PM-10 levels.
11	Moustafa.Elshafei, Mohamed A.Habib, Mansour Al-Dajani	Prediction Of Boilers Emission Using Polynomial Networks	• The paper provided an efficient polynomial network solution to the problem of on-line monitoring of NOx emission from industrial boilers. • The effect of six variables were studied using 3D CFD simulation model and used by polynomial networks for prediction of NOx and 02 in the exhaust flue.

Sr no.	Name of Author	Title of paper	Description
12	S. Raza, R. Avilaand J. Cervantes	A 3D Lagrangian Particle Model For The Atmospheric Dispersion Of Toxic Pollutants	• The Lagrangian Monte Carlo particle dispersion models work very efficiently for the atmospheric dispersion of effluents. • In order to incorporate the effect of vertical wind shear the modiffed dispersion coeffcient should be used with the Gaussian plume model.
13	H. Kaplan And N. Dinar	A Lagrangian Dispersion Model For Calculating Concentration Distribution Within A Built-Up Domain	• In this paper a diagnostic model for calculating concentration distribution of a passive scalar in a built-up area was presented. • The model requires measurements of the wind velocity and direction at a reference height above the obstacles. • The model is able to predict 3-d concentration distributions and to identify concentration accumulation at specific points. the model succeeds in predicting concentration distribution quantitatively and qualitatively and can be used to study many air pollution phenomena.

Sr no.	Name of Author	Title of paper	Description
14	By Peter De Haan' And Mathias W. Rotach	A novel approach to atmospheric dispersion modelling: The Puff-Particle Model	• In the present paper, an approach to model dispersion is presented which aims at combining the advantages of puff models and particle models. • The resulting model type is called Puff-Particle Model (PPM). In the PPM, a few hundred puffs are simulated in three-dimensional space, as compared to many thousand particles usually required in pure particle models. • The concept of the PPM is very simple: while puff growth is described by the concept of relative dispersion (thus accounting for eddies smaller than the puff), the effect of meandering (i.e. the variation between the trajectories of different puffs) due to larger eddies (larger than the actual puff size) is simulated by introducing puff-centre trajectories derived from particle trajectories from a particle model.

3 METHODOLOGY

Our overall setup consists of network of sensors that will be mounted in a specific industry, the data collected from these sensors will be stored on the server. These sensors measure the air parameters in terms of ambient air as well as stack emission. On this data we apply various machine learning algorithms for prediction of emission rate. The air dispersion models are then applied on the

predicted emission rate to calculate the dispersion of pollutants from the source that is at the stack level. The entire system is basically divided into two broad categories

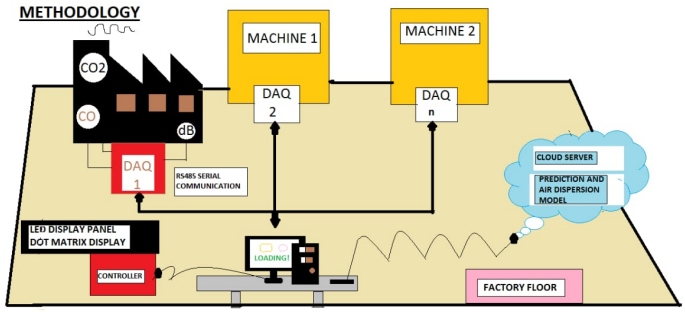

Figure 1: Methodology flow diagram

1. IOT: The IOT process flow begins with measurement of meteorological air parameters using sensors, which is divided into ambient air parameters and stack emission parameters. We considered these parameters to predict the value of V (velocity of wind) and Q (emission rate). This data from the sensors is uploaded and stored on the cloud setup.

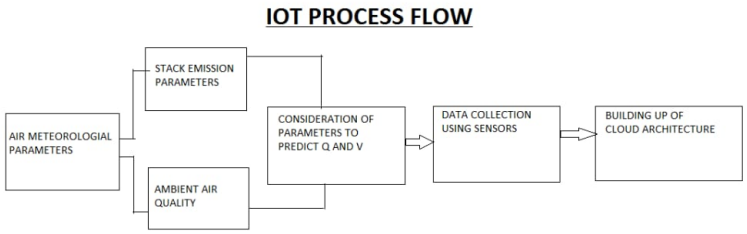

Figure 2: IOT process flow

1. MACHINE LEARNING: The machine learning process flow begins with the generation of dummy data in bulk using python. The dummy data was not truly random, it was correlated with various meteorological air parameters so that the machine could be trained well. It was observed that by adding more metrological factors, the prediction performance is greatly

improved and large sample data can also improve performance and train the model well. Next, machine learning algorithms were implemented on the created dummy data to predict the value of Q-emission rate and V-velocity of wind. For this purpose, the data was divided into training set (80%) and test set (20%). If the error reduces for training as well as test data, the process is continued. Else if the error reduces only for training data, but increases for test data then the process is terminated. This could greatly increase chances of generalizability of the algorithm. The performance check was then conducted on the predicted emission rate .The mean square error was measured in each case to check for accuracy. Optimization of the various algorithms was done in such a way so as to reduce the error as minimum as possible thereby increasing the accuracy of prediction. On this basis, the best algorithm was selected. To calculate the extent of pollution spread we used Gaussian dispersion model since it was the most optimal model in terms of computing power. In this way the entire process of prediction of pollution and calculation of its spread is done.

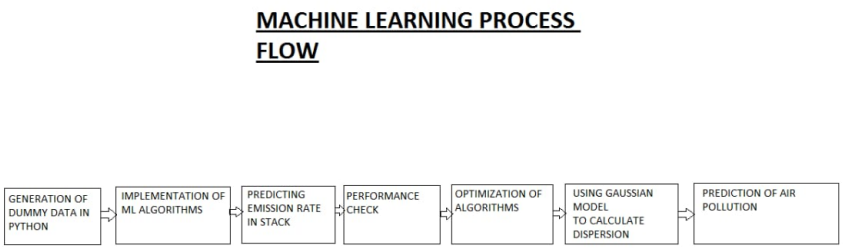

Figure 3: Machine Leaning process flow

3.1 MACHINE LEARNING ALGORITHMS

The various machine learning algorithms that were applied are as follows:

1. Kth Nearest neighbors (KNN)-This is a supervised machine learning algorithm which assumes that similar things exist in close proximity (similar things are near to each other). The value of K indicates the nearest neighbors that can be taken for consideration. Its purpose is to use a database which is separated into several classes to predict classification on a new sample point. This point is classified by a majority vote of its neighbors, it is assigned to the class most common to its K nearest neighbors It can

also be used for regression wherein the output of the point is the average or median value of its nearest neighbors. This is a non-parametric technique of classification since it does not make any assumptions based on the underlying data distribution. Also, in KNN there is lack of generalization which means that the training data required is very less.

2. Support vector Regression (SVR)-SVR is a supervised machine learning algorithm which can be used for both classification and regression problems. In this algorithm each data point is plotted in n dimensional space after which classification is performed by finding the hyperplane that will differentiate the classes very well. The hyperplane that is considered can be a linear separator of any dimension (line, plane, hyperplane). The training points are used in the decision function and are called support vectors.

3. Random Forest-Random forest is a method that operates by constructing multiple decision trees during training phase. The decision of the majority of the trees is chosen by the random forest as the final decision .The decision tree is a decision support tool. It uses a tree like graph to show possible consequences. When a training data set is considered with targets and features, some set of rules are formulated, these rules are used to perform predictions. It identifies the most important features out of all the available features in the training dataset. The lager the no. of trees the more accurate the results will be. Random forest classifier handles missing values and overfitting problem doesn't exist.

4. Multi linear regression (MLR)-It is a technique that uses several explanatory variables to predict the outcome of response variable. A linear relationship is modelled between these independent and response variable (dependent).Prediction about one variable is done based on the information about the other variable. In this case the independent variables should not be too highly correlated with each other.

5. Neural Network-it is an algorithm that mimics the way the human brain operate. A neuron in a neural network is a mathematical function that collects and classifies information according to some specific architecture. A neural network contains layers of interconnected nodes .Each connection is associated with a weight that is multiplied with the input value .Each neuron has an activation function that defines the output of the neuron which is used to introduce non linearity in the network model.

These Machine learning algorithms were implemented using python and the mean square error of each of these was measured to check for accuracy.

3.2 GAUSSIAN DISPERSION MODEL

In Gaussian dispersion model, the concentration of pollution downwind from a source is treated as spreading outward from the stack. A point source was

considered somewhere in the air where a pollutant is released at a constant rate Q (kg/s). The wind is blowing continuously in a direction x (measured in meters from the source) with a speed U (m/s). The plume spreads as it moves in the x direction such that the local concentrations C(x,y,z) (kg/m3) at any point in space form distributions which have shapes that are "Gaussian" or "normal" in planes normal to the x direction. The total concentration at point P is given as

$$C(x,y,z) = \frac{Q}{U}\frac{1}{2\prod \sigma_y \sigma_z} e^{\left(\frac{-y^2}{2\sigma_y^2}\right)} \left[e^{\left(\frac{-(z-H)^2}{2\sigma_z^2}\right)} + e^{\left(\frac{-(z+H)^2}{2\sigma_z^2}\right)} \right]$$

The parameters σ_y and σ_z (m) are the standard deviations of these Gaussian distributions, which indicate the spread of the plume in the y and z directions, respectively. They increase with the distance x from the source.

H = effective height of plume center-line (m) hs = height of source above ground (m) Δh = initial plume rise (m) z = coordinate measured vertically from the ground to a point in the plume (m) Maximum concentration occurs when

$$\sigma_z = \frac{H}{\sqrt{2}}$$

There are many formulae and semi-empirical expressions available for determining σy and σz under different conditions of atmospheric stability. A reasonable approximation in regions near to the source when the source is elevated above the ground (such as at the top of a chimney) is

$$\sigma_y = I_y \times x$$

$$\sigma_z = I_z \times x$$

where $I_y$and Iz are the turbulent wind speed fluctuations (turbulence intensities) in the y and z directions

$$I_y = \frac{0.88}{\ln\left(\frac{h_s}{z_0}\right) - 1}$$

$$I_z = \frac{0.50}{\ln\left(\frac{h_s}{z_0}\right) - 1}$$

z_0 is the aerodynamic roughness representing different topographic ground conditions Hence, the concentration is equal to the rate of emission from the source divided by the wind speed and then multiplied by the shaping function.

4 Results and Disscussion

4.1 Machine Learning module

Different machine learning models were studied and implemented on the same dataset and the results were compared. The following machine learning models were implemented:

- Random forest The random forest model was tuned on its parameters (n_estimators, min_sample_leaf, max_depth, min_sample_split, max_features)

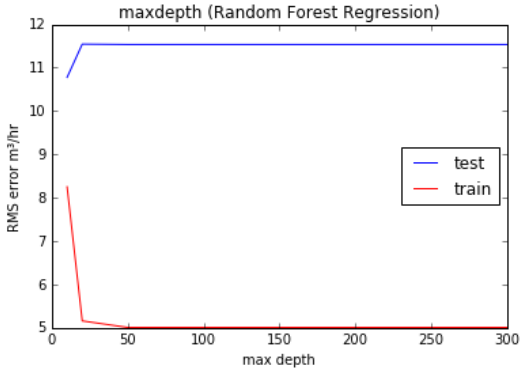

Figure 4: parameter tuned max depth

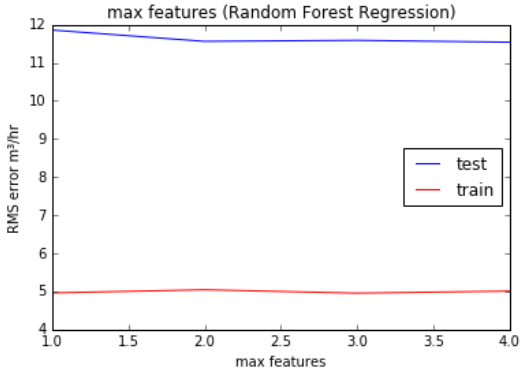

Figure 5: parameter tuned max features

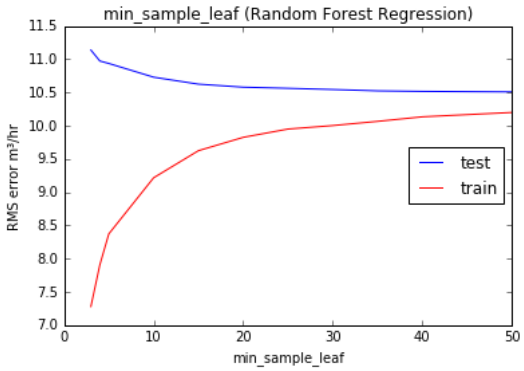

Figure 6: parameter tuned min sample leaf

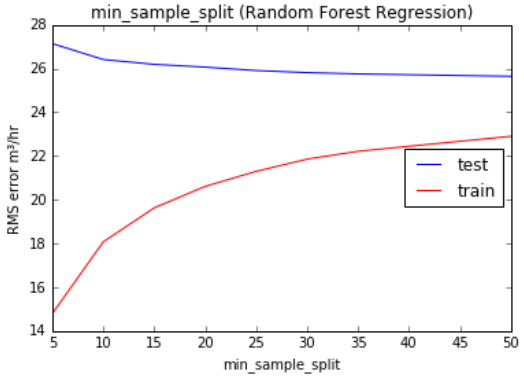

Figure 7: parameter tuned min sample split

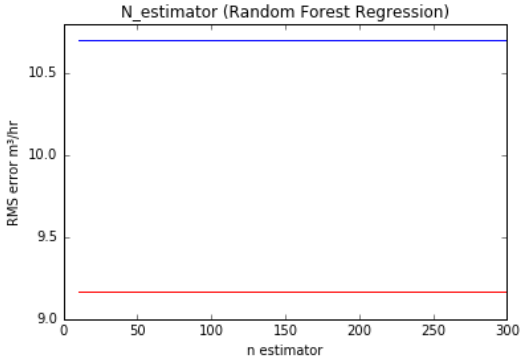

Figure 8: parameter tuned n estimators

The results show that the model can be tuned and the error ranged from 24 to 5 m3/hr

- Multi-linear perceptron regression

Here the model is tuned on activation function, number of neurons and type of solvers

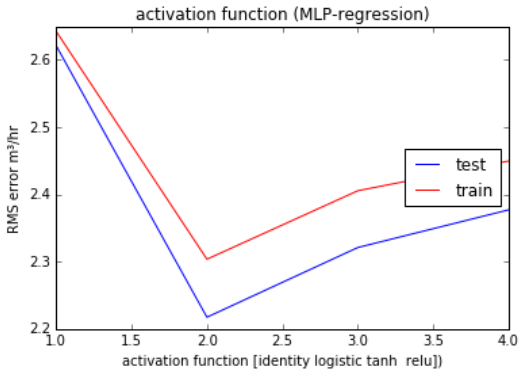

Figure 9: parameter tuned activation function

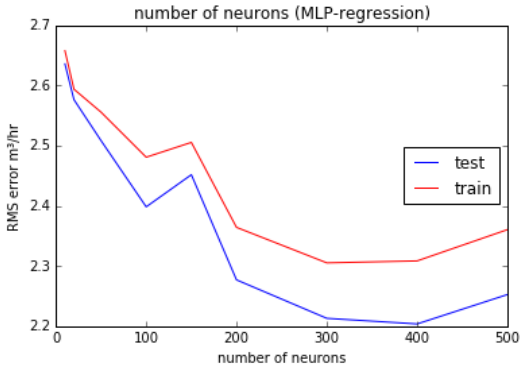

Figure 10: parameter tuned number of neurons

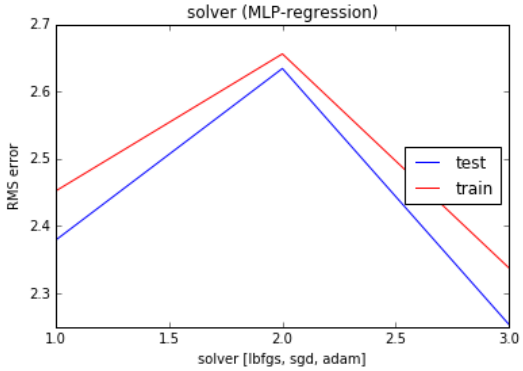

Figure 11: parameter tuned solver

The results show that by parameter tuning we get the rms error ranging from 2.2 to 2.7 m3/hr

- Support vector regression

This model is only tuned on its kernel property

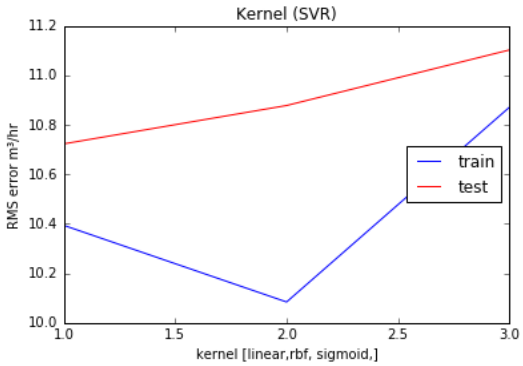

Figure 12: parameter tuned kernel

The results show that the model can be tuned and the error ranged from 10 to 10.8 m3/hr

- K-nearest neighbor

This was tuned on its weight and algorithm parameter, which are used for clustering

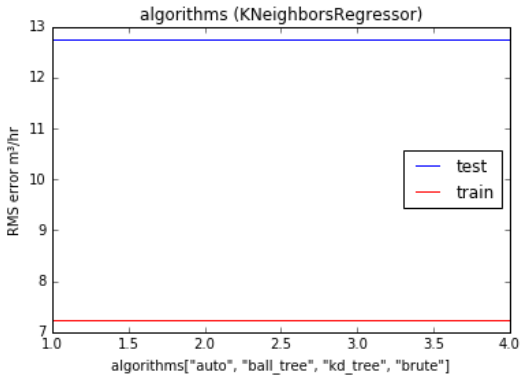

Figure 13: parameter tuned alogorithm

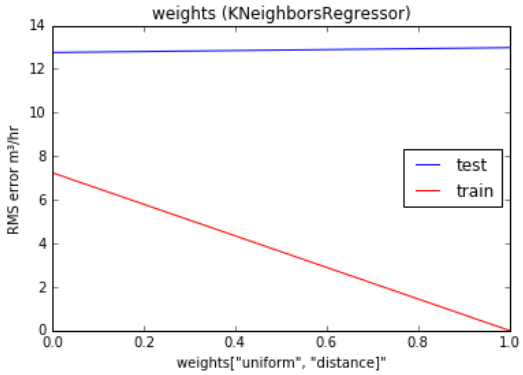

Figure 14: parameter tuned weights

The results show that the model can be tuned and the error ranged from 3 to 14 m3/hr

- Multi-linear Regression

This algorithm is an extension of linear regression with added support of higher dimensions in feature space

Figure 15: predicted and actual data using samples form the test data

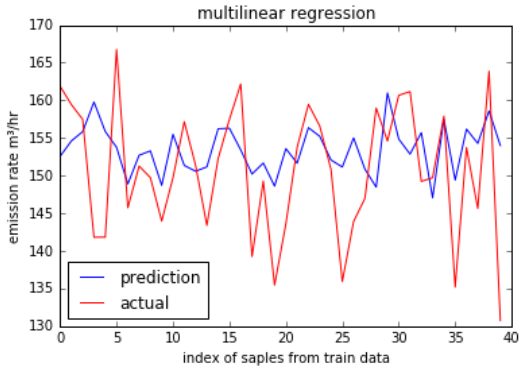

Figure 16: predicted and actual data using samples form the train data

The results show that the model predicts with an error of 10 m3/hr

4.2 Air Dispersion module

To calculate the concentration of the pollutant after emission from the stack, Gaussian air dispersion model was used for simulation, this was implemented using python programing language

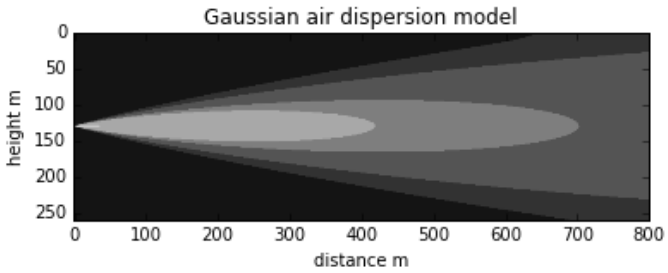

Figure 17: Gaussian air dispersion model

the white region represents regions with high concentration and black with low concentration