PREDICTING AIR POLLUTION LEVEL USING DIFFERENT MACHINE LEARNING MODELS

A MAIN PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF DEGREE OF

MASTER OF SCIENCE IN COMPUTER SCIENCE

of

MAHATMA GANDHI UNIVERSITY
KOTTAYAM-686560

2020-2022

by

SAFAL.P.NAUSHAD
(REGISTER NO: 200011024400)

(Affiliated to Mahatma Gandhi University , Kottayam)

Department of Computer Science

Al - Ameen College
Edathala, Aluva, Ernakulam
Kerala – 683 564.
(https://www.alameencollege.org)

September 2022

Department of Computer Science
Al - Ameen College

Edathala, Aluva, Ernakulam, Kerala – 683 564.

(https://www.alameencollege.org)

(Affiliated to Mahatma Gandhi University , Kottayam))

Certificate

This is to certify that the main project report titled **“PREDICTING AIR POLLUTION LEVEL USING DIFFERENT MACHINE LEARNING MODELS ”** is a bonafide record of the work carried out by **SAFAL.P.NAUSHAD** with **REGISTER NUMBER 200011024400** of computer science, Al Ameen College Edathala in partial fulfillment of the requirements for the award of **Master of Science in Computer Science** by **Mahatma Gandhi University, Kottayam**, during the academic year 2020-2022.

Project Guide/Supervisor

Sig:...............

Head of Department

Sig:...............

Submitted for the university exam held on......................

Internal examiner

Sig:...............

External examiner

Sig:...............

Acknowledgement

I wish to record my indebtedness and thankfulness to all those who helped me prepare this report titled "**PREDICTING AIR POLLUTION LEVEL USING DIFFERENT MACHINE LEARNING MODELS** " and present it in a satisfactory way.

My sincere thanks to **Dr.Cini Kurian**, Principal, for providing me all the necessary facilities. I take this opportunity to extend my thanks to **Ms. Jyothi** the Head of **PG Computer Science** for providing critical inputs in the preparation of this report. I also thank all other faculty members in our department for their guidance.

Finally, I would like to extend my sincere gratitude to friends who have always been helpful, in preparing and presenting the report and in the discussion following the presentation.

SAFAL.P.NAUSHAD
(200011024400) (2022 Admissions)
Al - Ameen College Edathala,
Aluva, Ernakulam, Kerala – 683 564.

September 2022

Abstract

The regulation of air pollutant levels is rapidly becoming one of the most important tasks for the governments of developing countries. Among the pollutant index, Fine particulate matter (PM2.5) is a significant one because it is a big concern to people's health when its level in the air is relatively high. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. However, the relationships between the concentration of these particles and meteorological and traffic factors are poorly understood. To shed some light on these connections, some of these advanced techniques have been introduced into air quality research. These studies utilized selected techniques, such as Support Vector Machine (SVM) and Neural Network, to predict ambient air pollutant levels based on mostly weather and sometimes traffic variables. This project attempted to apply some machine learning techniques to predict PM2.5 levels based on a dataset consisting of daily weather and traffic parameters. Due to the uncertainty of the specific number PM2.5 level, I simplified the problem to be a binary classification one, that is to classify the PM2.5 level into ”High” (Greater Than 115 ug/m3) and ”low” (Less than 115 ug/m3). The value is chosen based on the Air Quality Level standard, which set 115 ug/m3 to be mild level pollution.

Contents

List of Figures

Chapter 1

INTRODUCTION

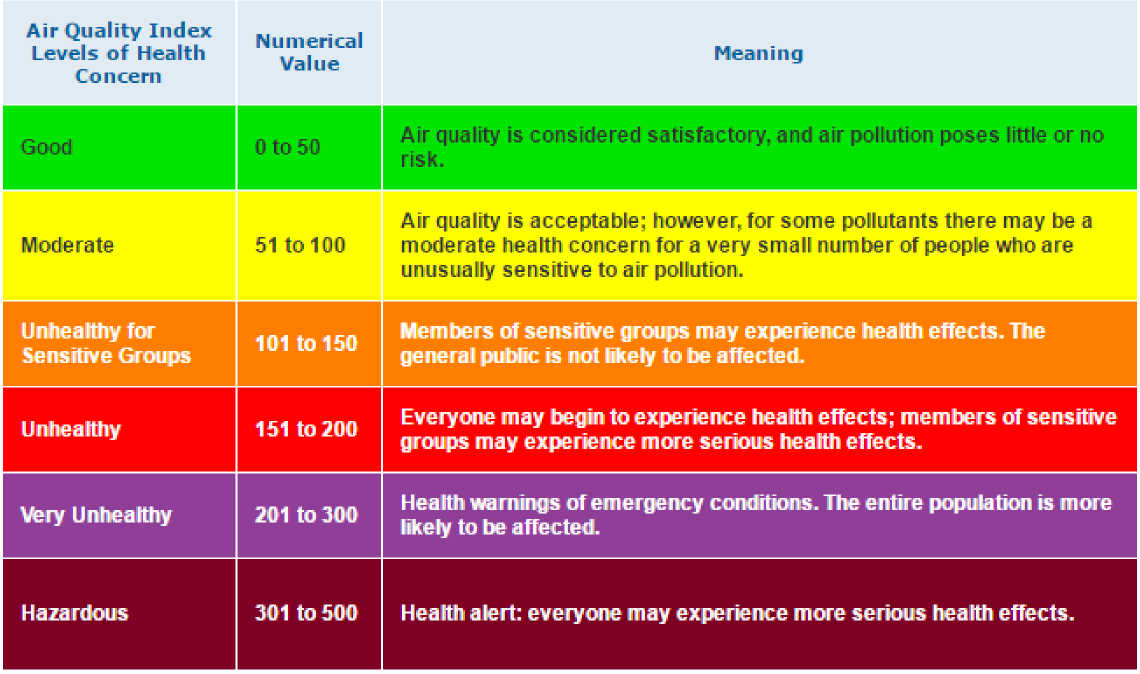
Predict PM2.5 level and detect air quality based on a data set consisting of daily atmospheric conditions in a specific city.

1.1 General

Particulate matter can be either human-made or naturally occur.Some examples include dust, ash and sea-spray. Particulate matter is emitted during the combustion of solid and liquid fuels, such as for power generation, domestic heating and in vehicle engines. Particulate matter varies in size (i.e. the diameter or width of the particle). PM2.5 refers to the mass per cubic meter of air of particles with a size (diameter) generally less than 2.5 micrometers. Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. Different machine learning models have been applied to detect air pollution and predict PM2.5 levels based on a data set consisting of daily atmospheric conditions

1.2 Objective

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it.



Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health warnings of emergency conditions. The entire population is more likely to be affected.
Hazardous	301 to 500	Health alert: everyone may experience more serious health effects.

Figure 1.1: Air Quality Index

Chapter 2

FEATURE SELECTION

A variety of meteorological, traffic and industrial parameters affect the air pollution level. After taking consideration of the data availability and importance 5 Features are selected.

2.1 Data Overview

In order to identify and forecast key parameters affecting air quality and propose appropriate preventive strategies and policies, it is essential to systematically collect data characterizing air quality. The data includes two parts: training data set and test data set. Training data set has 322 observation points and the test data has 55 points. Each point represents the meteorological and traffic condition of a specific day in Beijing City. The total data set covers 47 days in 2014 and 330 days in 2013. The data comes from China Meteorological Data Sharing Service System, Beijing Transportation Research Center and US Embassy in Beijing. As mentioned before, the output data was labeled as one or zero. One refers to high pollution level and zero refers to low pollution level. The total number labeled as zero is 103, while the remaining 274 points are labeled as 0.

2.2 Parameters

- Temperature
- Wind Speed
- Relative Humidity
- Traffic Index
- Air Quality Of Previous Day

2.2.1 Temperature

Temperature affect air quality because of temperate inversion: the warm air above cooler air acts like a lid, suppressing vertical mixing and trapping the cooler air at the surface. As pollutants from vehicles, fireplaces, and industry are emitted into the air, the inversion traps these pollutants near the ground.

2.2.2 Wind Speed

Wind speed plays a big role in diluting pollutants. Generally, strong winds disperse pollutants, whereas light winds generally result in stagnant conditions allowing pollutants to build up over an area.

2.2.3 Relative Humidity

Humidity could affect the diffusion of contaminant.

2.2.4 Traffic Index

The large number of cars on the road cause high level of air pollution and traffic jam may increase the pollutants concentration from vehicles. The definition of traffic index is a index reflecting the smooth status of traffic. The index range is from 0 to 10. 0 represents smooth and 10 represents sever traffic jam.

2.2.5 Air Quality Of Previous Day

The air pollution level is influenced by the condition of the previous day to some extent. If the air pollution level of the previous day is high, the pollutants may stay and affect the following day.

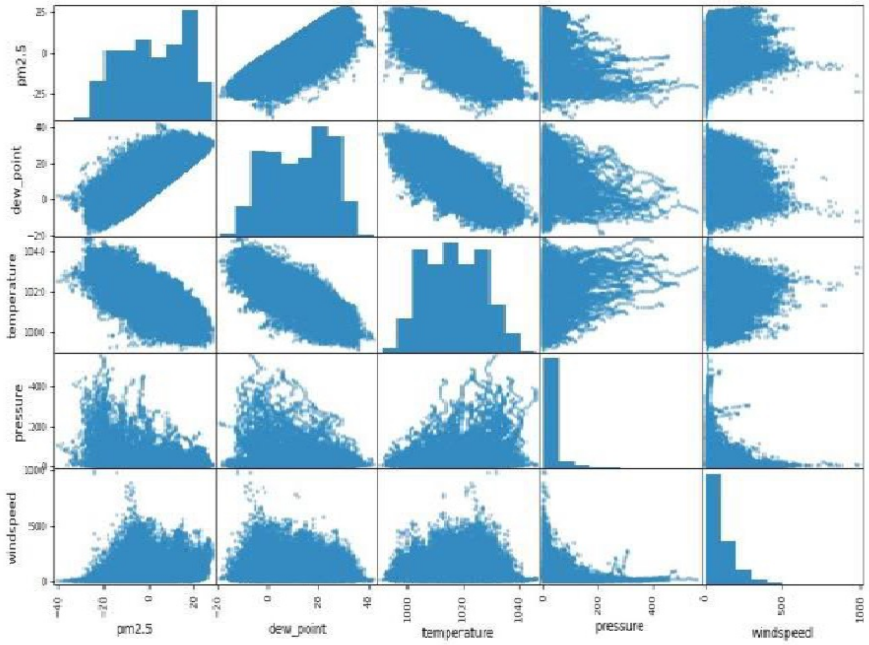


Figure 2.1: Scatter plot for the relation among attributes

Chapter 3

METHODOLOGY

3.1 Two primary Phases

3.1.1 Training

Estimating the parameters for the machine learning is called training the data. The observations in the training set form the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables.

3.1.2 Testing

Evaluating how well the machine learning method work is called testing. The test set is a set of observations used to evaluate the performance of the model using some performance metric. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.

3.2 Supervised Learning Algorithms

This prediction is a binary classification problem, so the following three supervised learning algorithms were used:

- Logistic Regression
- Naive Bayes Classification
- Support Vector Machines

3.2.1 Logistic Regression

The output is a Generalized Linear Model. For this model, the prediction value is range for 0 to 1. In order to get the label, the values were converted to zero and one. Logistic regression is the algorithm employed to detect a user-defined sample to be polluted or not. Logistic regression is the appropriate regression model to conduct analysis when the dependent variable is dichotomous. For example, here, the data set gets classified into two classes I.E, Polluted or Not Polluted. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to explain the relationship between one or more independent variables and one dependent binary variable.

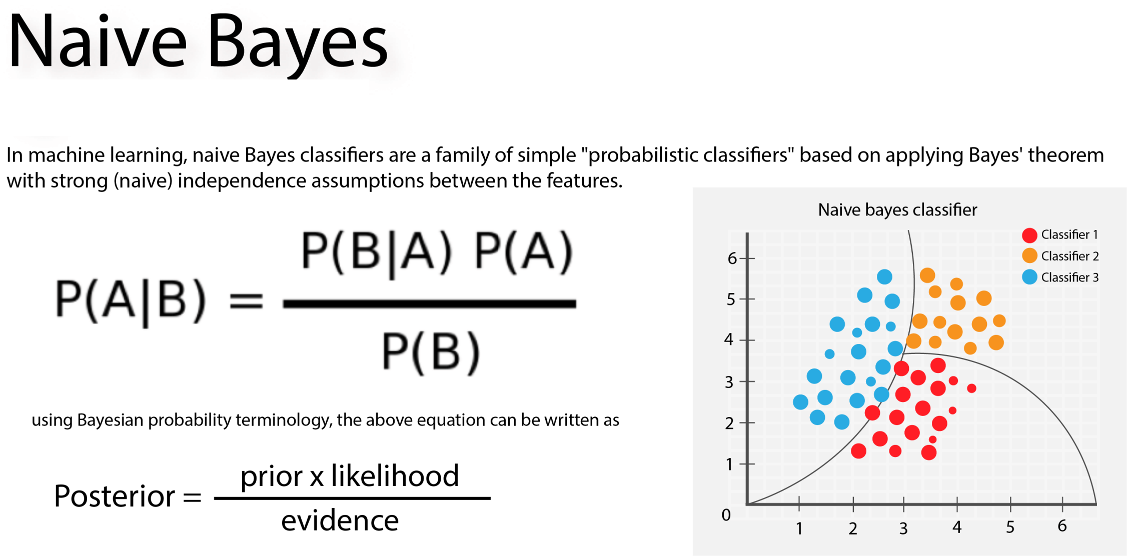
$$\text{Logit(p)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{m}$$

Figure 3.1: Logit Function

Logit function is used to generate log odds of an attribute that signifies the probability of the attribute. Log odds are an alternate way of expressing probabilities, which simplifies the process of updating them with new evidence. Based on logit function, the system classifies the training data to be either 0 (not polluted) or 1 (polluted) and verifies its accuracy using the test data. The result of the user input is also 0/1 and not the PM2.5 level

3.2.2 Naive Bayes Classification

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.



Naive Bayes

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

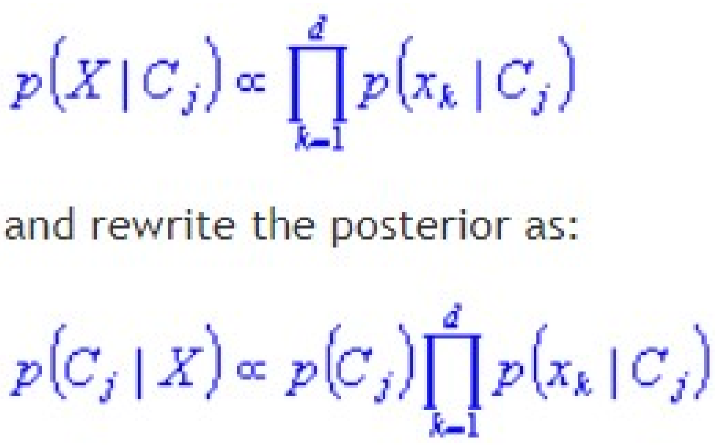
Figure 3.2: Naive Bayes Classifier

Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. Given a set of variables, X = x1,x2,x...,xd, we want to construct the posterior probability for the event Cj among a set of possible outcomes C = c1,c2,c...,cd. In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable. Using Bayes' rule:

$$p(C_j \mid x_1, x_2, \ldots, x_d) \propto p(x_1, x_2, \ldots, x_d \mid C_j)\, p(C_j)$$

Figure 3.3: Bayes rule

where p(Cj — x1,x2,x...,xd) is the posterior probability of class membership, i.e., the probability that X belongs to Cj. Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent we can decompose the likelihood to a product of terms:



$$p(X \mid C_j) \propto \prod_{k=1}^{d} p(x_k \mid C_j)$$

and rewrite the posterior as:

$$p(C_j \mid X) \propto p(C_j) \prod_{k=1}^{d} p(x_k \mid C_j)$$

Figure 3.4: Bayes rule

3.2.3 Support Vector Machine

Support Vector Machine(SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiate the two classes very well

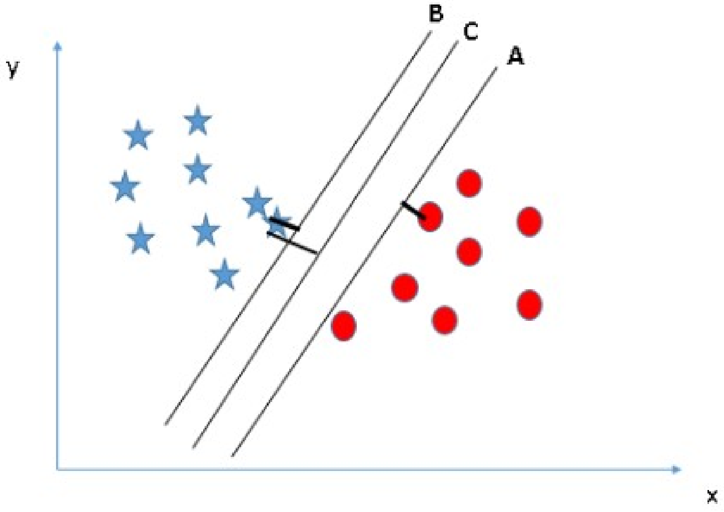


Figure 3.5: Support Vector Machine

The main objective in SVM is to find the optimal hyperplane to correctly classify between data points of different classes (Figure 2). The hyperplane dimensionality is equal to the number of input features minus one (eg. when working with three feature the hyperplane will be a two-dimensional plane).

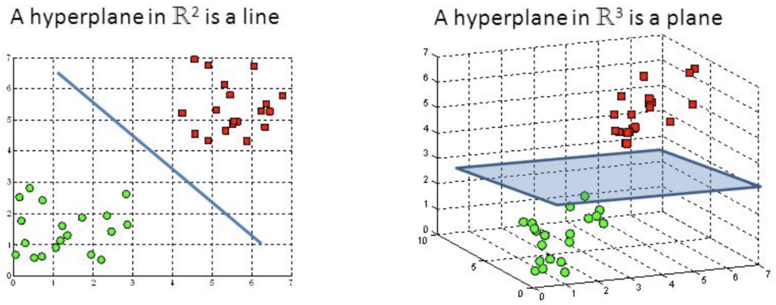

A hyperplane in $\mathbb{R}^2$ is a line
A hyperplane in $\mathbb{R}^3$ is a plane

Figure 3.6: Hyper Plane in R2 and R3

Data points on one side of the hyperplane will be classified to a certain class while data points on the other side of the hyperplane will be classified to a different class. The distance between the hyperplane and the first point (for all the different classes) on either side of the hyperplane is a measure of sure the algorithm is about its classification decision. The bigger the distance and the more confident we can be SVM is making the right decision. The data points closest to the hyperplane are called Support Vectors.

3.3 Result Analysis

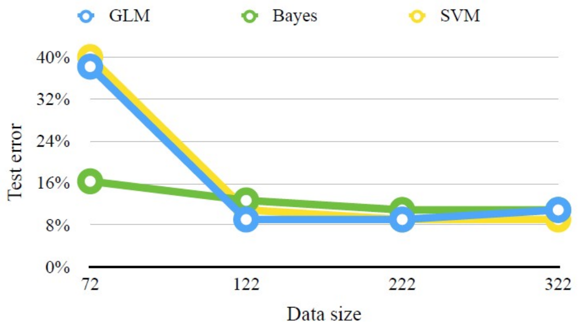


Figure 3.7: Algorithm Comparison

When compared to other machine learning models applied on the data set. The overall test error for Logistic regression is 10.91, which is the same as it for Bayes. SVM has the lowest test error, 9.09. After changing the data size and repeat training the model the test error of Bayes classifier doesn't change much with data size, however Logistic regression and SVM have large test error change with data size. Further more, the test error for SVM has the decline trend if the data size increases further.

3.4 Performance Analysis

the prediction of air pollution level with the ground data set, The best algorithm (SVM) gave the 0.722 precision, 1.000 recall and 0.839 FMeasure value. It is relatively accurate and is an acceptable result for practical use. However, compared with results from some literature, the predicting performance (F-Measure value) for this data set is not very good. Also, the advantage of SVM are not shown obviously.

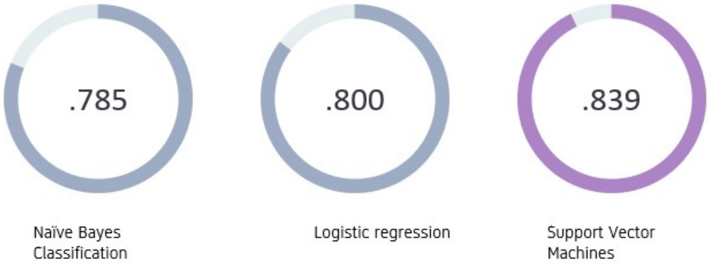


Figure 3.8: Performance Analysis(F-measure)

Chapter 4

NEW PROPOSAL

The data set in this project is not large enough. Air quality is a long-term formed problem and it is better to use a large data covering a variety of years and locations. Furthermore, beside the meteorological and traffic factors, industrial parameters such as power plant emissions also play significant roles in air pollution. Using Extreme Gradient Boosting(Egbooster) Or Auto regression Algorithm Can make More accurate prediction over Support vector machine

4.1 Autoregression

Autoregressive models and processes are stochastic calculations in which future values are estimated based on a weighted sum of past values. Autoregressive models and processes operate under the premise that past values have an effect on current values, which makes the statistical technique popular for analyzing nature, economics, and other time-varying processes. Multiple regression models forecast a variable using a linear combination of predictors, whereas autoregressive models use a combination of past values of the variable.

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

Figure 4.1: Equation of AR

4.2 Extreme Gradient Boosting

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way.

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

$$argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) = argmin_\gamma \sum_{i=1}^{n} (y_i - \gamma)^2$$

Figure 4.2: Equation of XGBoost

Chapter 5

CONCLUSION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The results show that machine learning models (logistic regression and autoregression) can be efficiently used to detect the quality of air and predict the level of PM2.5 in the future. The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities.

Bibliography

[1] Dan wei: Predicting air pollution level in a specific city [2014]

[2] Pandey, Gaurav, Bin Zhang, and Le Jian. quot; Predicting sub-micron air pollution indicators: a machine learning approach.quot ; Environmental Science: Processes amp; Impacts 15.5 (2013): 996-1005.

[3] Athanasiadis, Ioannis N., et al. ”Applying machine learning techniques on air quality data for real-time decision support.” First international NAISO symposium on information technologies in environmental engineering (ITEE’2003), Gdansk, Poland. 2003.

[4] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu:Detection and Prediction of Air Pollution using Machine Learning Models

[5] https://en.wikipedia.org/wiki/Support-vector machine

[6] https://en.wikipedia.org/wiki/Logistic regression

[7] https://en.wikipedia.org/wiki/Particulates

[8] https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[9] https://machinelearningmastery.com/logistic-regression-for-machine-learning/

[10] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[11] https://en.wikipedia.org/wiki/Naive Bayes classifier

Department of Computer Science
Al - Ameen College
Edathala, Aluva, Ernakulam, Kerala – 683 564.
(https://www.alameencollege.org)