DETECTION AND PREDICTION OF AIR POLLUTION LEVEL

A MAIN PROJECT REPORT

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

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

of

APJ Abdul Kalam Technological University

by

SRUTHI K S

TESLIN ROSE P V

VINI SASIDHARAN

VISHNU V U

(AN ISO 9001:2008 CERTIFIED INSTITUTION)

Department of Computer Science and Engineering

Vidya Academy of Science & Technology

Thalakkottukara, Thrissur - 680 501

(http://www.vidyaacademy.ac.in)

November 2019

Department of Computer Science and Engineering
Vidya Academy of Science & Technology

Thalakkottukara, Thrissur - 680 501

(http://www.vidyaacademy.ac.in)

(AN ISO 9001:2008 CERTIFIED INSTITUTION)

Certificate

This is to certify that the Main Project Report titled **"DETECTION AND PREDICTION OF AIR POLLUTION LEVEL"** is a bonafide record of the work carried out by **SRUTHI K S ,TESLIN ROSE P V ,VINI SASIDHARAN ,VISHNU V U** of Vidya Academy of Science & Technology, Thalakkottukara, Thrissur - 680 501 in partial fulfillment of the requirements for the award of **Degree of Bachelor of Technology** in **Computer Science and Engineering** of **APJ Abdul Kalam Technological University**, during the academic year 2019-2020. The Main Project report has been approved as it satisfies the academic requirements in the respect of main project work prescribed for the said degree.

Project Guide/Supervisor

Ms. Nitha K P

Prof.,Dept. of CSE

....

Project Coordinator

....

Head of Department

Dr. Ramani Bai V

Prof.,Dept. of CSE

....

Undertaking

We, **SRUTHI K S, TESLIN ROSE P V, VINI SASIDHARAN, VISHNU V U**, hereby undertake that, the main project work entitled **"DETECTION AND PREDICTION OF AIR POLLUTION LEVEL"**, is carried out by us independently under the valuable guidance of **Ms. Nitha K P**, Asst. Prof., Dept. of Computer Science and Engineering, Vidya Academy of Science and Technology, Thalakkottukara, Thrissur, in partial fulfillment of the requirements for the award of degree of **Bachelor of Technology in Computer Science Engineering** of APJ Abdul Kalam Technological University, during the academic year 2019-2020.

SRUTHI K S(VAS16CS116)
TESLIN ROSE P V(VAS16CS119)
VINI SASIDHARAN(VAS16CS123)
VISHNU V U(LVAS16CS136)

Thrissur
November 2019

Acknowledgement

During the course of our main project work several persons collaborated directly and indirectly with us. Without their support it would be impossible for us to finish our work. That is why we wish to dedicate this section to recognize their support.

We want to start expressing our thanks to our project guide **Ms. Nitha K P**, Asst. Prof., Dept. of Computer Science and Engineering, because of his valuable advice and guidance towards this work. We received motivation, encouragement and hold up from his during the course of work.

We are grateful to express our thanks to all the faculty members of our department for their support. We articulate our gratitude to all my friends for their support and help for this work.

We thankful to **Dr. Ramani Bai V** , Head of Computer Science and Engineering Department, and our Principal **Dr. Saji C.B**, for their sole co-operation.

Last, but not the least we wish to express our gratitude to God Almighty for His abundant blessings without which this effort would not have been successful.

SRUTHI K S, TESLIN ROSE P V, VINI SASIDHARAN, VISHNU V U
B.Tech (CSE) (2016 Admissions)
Vidya Academy of Science & Technology
Thrissur - 680 501.

November 2019

Abstract

In the populated and developing countries, governments consider the regulation of air as a major task. The meteorological and traffic factors, burning of fossil fuels, industrial parameters such as power plant emissions play significant roles in air pollution. Among all the particulate matter that determine the quality of the air, Particulate matter (PM 2.5) needs more attention. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. When it's level is high in the air, it causes serious issues on people's health. However, the relationships between the concentration of these particles and meteorological and traffic factors are poorly understood.the proposed system combines meteorological features with gaseous pollutants such as SO2,NO2 etc.The proposed system employs logistic regression to detect whether a data sample is either polluted or not polluted and for prediction, efficient machine learning techniques like Random forest,Deep learning and Extreme gradient boosting are used.

Contents

List of Figures

Chapter 1

INTRODUCTION

In recent years, air pollution has become an important public health concern The regulation of air pollutant levels is rapidly becoming one of the most important tasks for the governments of developing countries,

1.1 General

In the populated and developing countries, governments consider the regulation of air as a major task. The meteorological and gasous pollutents, burning of fossil fuels, industrial parameters such as power plant emissions play significant roles in air pollution. Among all the particulate matter that determine the quality of the air, Particulate matter (PM 2.5) needs more attention. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. When it's level is high in the air, it causes serious issues on people's health. However, the relationships between the concentration of these particles and meteorological and traffic factors are poorly understood.the proposed system combines meteorological features with gaseous pollutants such as SO2,NO2 etc.The proposed system employs logistic regression to detect whether a data sample is either polluted or not polluted and for prediction, efficient machine learning techniques like Random forest,Deep learning and Extreme gradient boosting are used.

1.2 Objectives of the Work

The objective of the system is to develop a system which is more accurate and efficient, The existing system doesn't combine the meteorological features with the gaseous pollutant like SO2, NO2, The proposed system combines these features to attain optimum efficiency

1.3 Motivation for this work

PM is a complex mixture of solid and liquid particles suspended in air that is released into the atmosphere when coal, gasoline, diesel fuels and wood are burned. It is also produced by chemical reactions of nitrogen oxides and organic compounds that occur in the environment. Vegetation and livestock are also sources of PM. In big cities, production of PM is attributed to cars, trucks and coal-fired power plants.

The health effects of PM depend on several factors, including the size and composition of the particles, the level and duration of exposure, and the gender, age and sensitivity of the exposed individual. Symptoms of exposure may include persistent cough, sore throat, burning eyes and chest tightness. PM may also trigger asthma or lead to premature death, particularly in elderly individuals with pre-existing disease.3,4 In addition, people who are active outdoors are at higher risk, as physical activity increases the amounts of PM penetrating into the airways. People with disease (e.g. diabetes mellitus, malnutrition) are also at increased risk.5–7 A comprehensive review on diesel PM by Ristovski et al. was published in an earlier issue of this review series on air pollution and lung disease.

1.4 Methodologies Adopted

For the project, we evaluated several different types of detection and prediction machine learning models which will be described below. We use Logistic Regression method for detecting PM2.5 level and XGBoost,Random Forest and Deep Learning for prediction.Data processing and matching is necessary because the data is obtained from different sources.Therefore, we intend to use interpolation to estimate and fill in the missing data.Data normalization is an important step for many machine-learning estimators. A correlation matrix can be used to investigate the dependence between multiple variables at the same time

1.5 Outline of the Report

This report contains 4 chapters. Chapter 1 gives the introduction to the project work and describes the objectives of the work. Literature survey is describes in Chapter 2. System Design is explained in the chapter 3 and Methodology is well explained in chater 4.

Chapter 2

LITERATURE REVIEW

2.1 A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization

Dixian Zhu ,Changjie Cai , Tianbao Yang and Xun Zhou

In this paper, we tackle air quality forecasting by using machine learning approaches to predict the hourly concentration of air pollutants (e.g., ozone, particle matter (PM2.5) and sulfur dioxide). Machine learning, as one of the most popular techniques, is able to efficiently train a model on big data by using large-scale optimization algorithms. Although there exist some works applying machine learning to air quality prediction, most of the prior studies are restricted to several-year data and simply train standard regression models (linear or nonlinear) to predict the hourly air pollution concentration. In this work, we propose refined models to predict the hourly air pollution concentration on the basis of meteorological data of previous days by formulating the prediction over 24 h as a multi-task learning (MTL) problem. This enables us to select a good model with different regularization techniques. We propose a useful regularization by enforcing the prediction models of consecutive hours to be close to each other and compare it with several typical regularizations for MTL, including standard Frobenius norm regularization, nuclear norm regularization, and ‘2,1-norm regularization. Our experiments have showed that the proposed parameter-reducing formulations and consecutive-hour-related regularizations achieve better performance than existing standard regression models and existing regularizations.

In this paper, we have developed efficient machine learning methods for air pollutant prediction. We have formulated the problem as regularized MTL and employed advanced optimization algorithms for solving different formulations. We have focused on alleviating model complexity by reducing the number of model parameters and on improving the performance by using a structured regularizer. Our results show that the proposed light formulation achieves much better performance than the other two model formulations and that the regularization by enforcing prediction models for two consecutive hours to be close can also boost the performance of predictions. We have also shown that advanced optimization techniques are important for improving the convergence of optimization and that they speed up the training process for big data. For future work, we will further consider the commonalities between nearby meteorology stations and combine them in a MTL framework, which may provide a further boosting for the prediction.

2.2 A Deep Learning Approach for Forecasting Air Pollution in South Korea Using LSTM

Tien-Cuong Bui, Van-Duc Le, Sang K. Cha

Over the last few years, tackling air pollution is an urgent problem in South Korea. Much research is being conducted in environmental science to evaluate the severe impact of particulate matters on public health. Besides that, deterministic models of air pollutant behavior are also generated; however, these are both complex and often inaccurate. On the contrary, deep recurrent neural network reveals strong potential on forecasting outcomes of time-series data and has become more prevalent. This paper uses Recurrent Neural Networks and Long Short-Term Memory units as a framework for leveraging knowledge from time-series data of air quality and meteorological information. Finally, we investigate prediction accuracies of various configurations. This paper is a significant motivation for not only continuing researching on urban air quality but also helping the government leverage that insight to enact beneficial policies.

The goal of the presented work was to evaluate the effectiveness of encoder-decoder networks for building prediction machines with time series data. The proposed model shows significant results in prediction PM2.5 AQI of long future based on historical meteorological data. However, to enhance the accuracy of the prediction machine, the model needs to be evaluated more in the future. Finally, forecasting the status of air pollution can help governments in policy-making and resource allocation.

2.3 Prediction Model of Air Pollutant Levels Using Linear Model with Component Analysis

Arie Dipareza Syafei, Akimasa Fujiwara, and Junyi Zhang

Abstract—The prediction of each of air pollutants as dependent variable was investigated using lag-1(30 minutes before) values of air pollutants (nitrogen dioxide, NO2, particulate matter 10um, PM10, and ozone, O3) and meteorological factors and temporal variables as independent variables by taking into account serial error correlations in the predicted concentration. Alternative variables selection based on independent component analysis (ICA) and principal component analysis (PCA) were used to obtain subsets of the predictor variables to be imputed into the linear model. The data was taken from five monitoring stations in Surabaya City, Indonesia with data period between March-April 2002. The regression with variables extracted from ICA was the worst model for all pollutants NO2, PM10, and O3 as their residual errors were highest compared with other models. The prediction of one-step ahead 30-mins interval of each pollutant NO2, PM10, and O3 was best obtained by employing original variables combination of air pollutants and meteorological factors. Besides the importance of pollutants interaction and meteorological aspects into the prediction, the addition spatial source such as wind direction from each monitoring station has significant contribution to the prediction as the emission sources are different for each station.

There is a concern of adverse effect to humans health due to high concentration of pollutants which exceed the standard value. These events occur often and people should get alerted when this happens, thus making the short-term prediction of pollutant become crucial. Linear models with original variables, ICs, and PCs extracted from six pollutants (NO, NO2, O3, SO2, CO, PM10 and meteorological factors (wind speed, solar gradiation, humidity and temperatures) were employed to predict 30-mins ahead of NO2, PM10, and O3. In addition, we include serial error correlation computation in the model for model accuracy. As expected, the presence of NO has positive correlation with NO2, aside with CO, wind speed and solar gradiation. Furthermore, it was shown that meteorologica factors have high role in the formation of O3. The faster wind speed will reduce the concentration of NO2 while on the opposite will increase the concentration of O3. This pattern is also found for humidity. Since PM10 is relatively inert particle gas with less than 10um, using the 30-mins data we obtained, no significant correlation was found with other variables.

2.4 Industrial Air Pollution Prediction Using Deep Neural Network

Yu Pengfei, He Juanjuan, Liu Xiaoming, and Zhang Kai

In this paper, a deep neural network model is proposed to predict industrial air pollution, such as PM2.5 and PM10. The deep neural network model contains 9 hidden layers, each layer contains 45 neurons. The output of the hidden layer neurons is calculated using the ReLU activation function, which can effectively reduce the gradient elim- ination effect of the deep neural network. Twelve air pollutant indicators from industrial factories are collected as the input data, such as CO, NO2, O3, and SO2. About 180,000 real industrial air pollution data from Wuhan City are used to train and test the DNN model. Furthermore, the performance of our approach is compared with the SVM and Artifi- cial neural network methods, and the comparison result shows that our algorithm is accurate and competitive with higher prediction accuracy and generalization ability.

In this paper, we proposed a deep neural network model to predict industrial air pollutant. About 180,000 real industrial air pollution data from 2016 to 2018 in Wuhan are used to train, validate and test the model. We use ReLU non- linear activation function instead of the traditional Sigmoid activation function, effectively improve the training speed of the network, eliminating the gradient disappearance or gradient explosion phenomenon. We use the Batch Normal- ization method to improve the training accuracy and convergence speed of the network. Through the DropOff technology, effectively prevent the depth of the neural network over-fitting, improve the test data prediction accuracy. The per- formance of our approach is compared with the SVM and BP neural network, and the comparison result shows that our algorithm is accurate and competitive with higher prediction accuracy and generalization ability.

2.5 Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU

QING TAO, FANG LIU , (Member, IEEE), YONG LI , (Senior Member, IEEE), DENIS SIDOROV, (Senior Member, IEEE)

Air pollution forecasting can provide reliable information about the future pollution situation, which is useful for an efficient operation of air pollution control and helps to plan for prevention. Dynamics of air pollution are usually reflected by various factors, such as the temperature, humidity, wind direction, wind speed, snowfall, rainfall, and so on, which increase the difficulty in understanding the change of air pollutant concentration. In this paper, a short-term forecasting model based on deep learning is proposed for PM2.5 concentration, and the convolutional-based bidirectional gated recurrent unit (CBGRU) method is presented, which combines 1D convnets (convolutional neural networks) and bidirectional GRU (gated recurrent unit) neural networks. The case is carried out by using the Beijing PM2.5 data set in UCI Machine Learning Repository. Comparing the prediction results with the traditional ones, it is proved that the error of the CBGRU model is lower and the prediction performance is better.

The results are compared with traditional mechine learning models and conventional deep learning models. The results show that the proposed method can be suitable and compet- itive on the PM2.5 data time series forecasting. To be more specific, compared with shallow machine learning models, such as DTR, SVR and GBR, deep learning-based methods exhibited better prediction performance. Furthermore, com- pared with GRU, bidirectional GRU has lower error value, which indicates that the use of bidirectional GRU can improve the prediction effect. This is because the bidirectional GRU processes the time series chronologically and antichrono- logically, it captures patterns that may be ignored by one- direction GRUs, improving feature learning capabilities in time series. In addition, compared with the other benchmark models, the accuracy of the CBGRU model is significantly improved, which shows that the convnets can help the GRU to obtain better prediction performance, because convnets uses its local feature learning ability and subsampling ability to obtain a sequence pattern that is more conducive to GRU processing.

Chapter 3

SYSTEM DESIGN

System design is the process of designing the elements of a system such as the architecture, modules and components, the different interfaces of those components and the data that goes through that system.

3.1 Basic Model

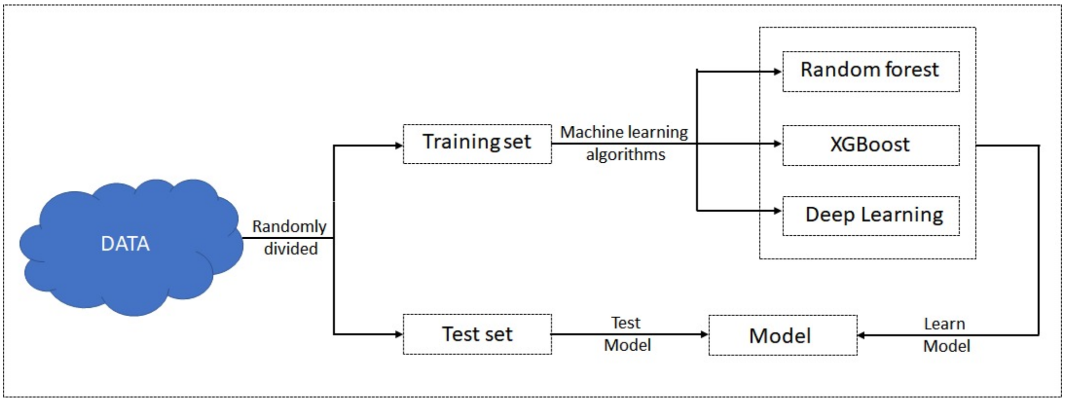


Figure 3.1: Basic model

The data, which is a combination of both meteorological features and gaseous pollutants is randomly divided into two : Training set and Test set. Usually,it is divided as 70% for training and 30% for test set. Learn models using training set and test the model with test set,from that we select the most accurate model. During training process,the training set is trained using three machine learning algorithms.The algorithms used are : Random Forest, XGBoost, Deep Learning.

3.2 Architecture

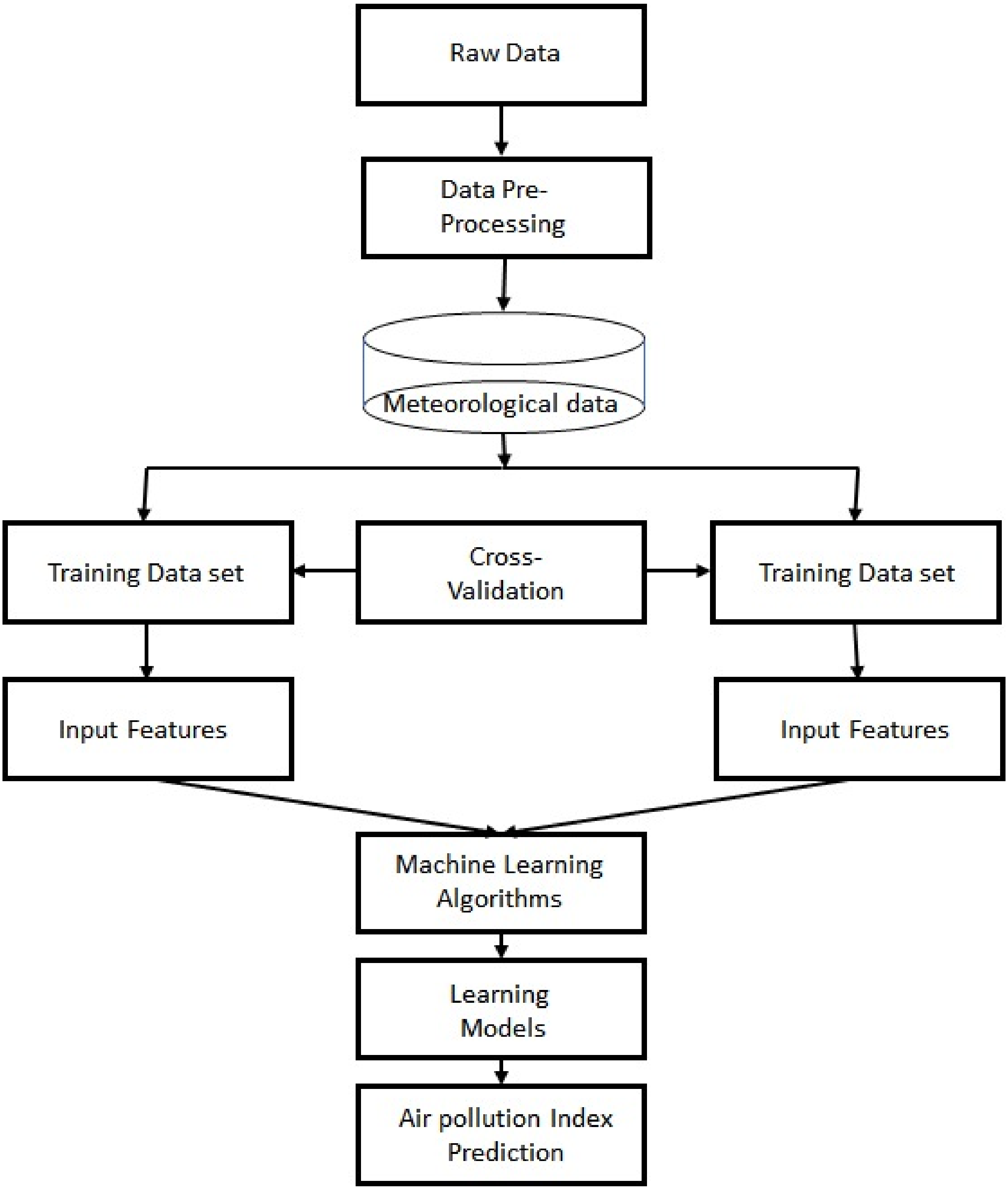


Figure 3.2: Architecture of the system

Initially the raw dataset is cleaned using several data preprocessing techniques. The data preprocessing techniques used are Interpolation, Normalization. After the missing values are removed ,and the data is normalized to a commom range, cross validation is performed on the data set. The dataset is divided into k folds,and training and testing is performed on the dataset. Correlation matrix of the dataset is formed ,and the features with positive correlation with PM2.5 are selected. After the data model is is created, the data is trained and tested using machine learning algorithms i.e Random Forest, XG-Boost,Deep Learning respectively.The most accurate model is selected and used for the air pollution index prediction.

3.3 Data pre-processing

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis.Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

3.3.1 Normalization

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

3.3.2 Interpolation

Interpolation is the process of deriving a simple function from a set of discrete data points so that the function passes through all the given data points (i.e. reproduces the data points exactly) and can be used to estimate data points in-between the given ones.

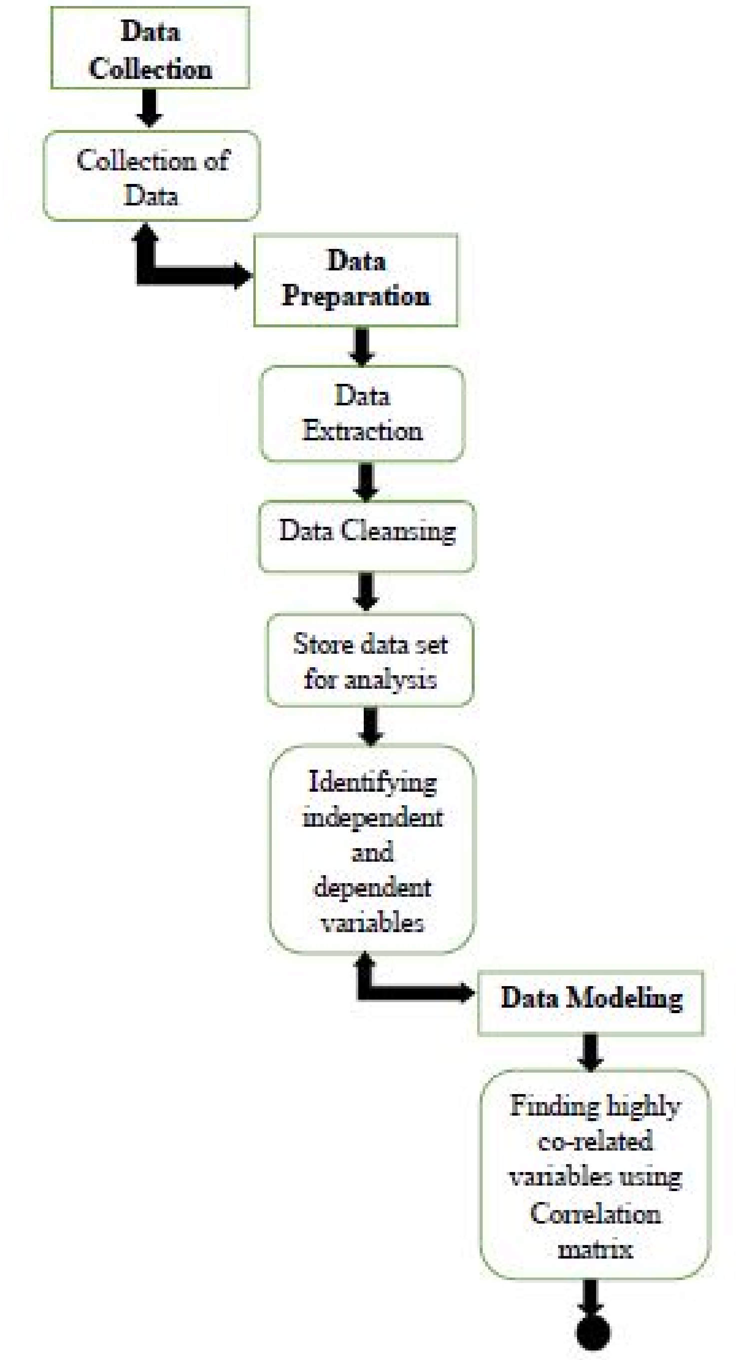


Figure 3.3: Steps involved

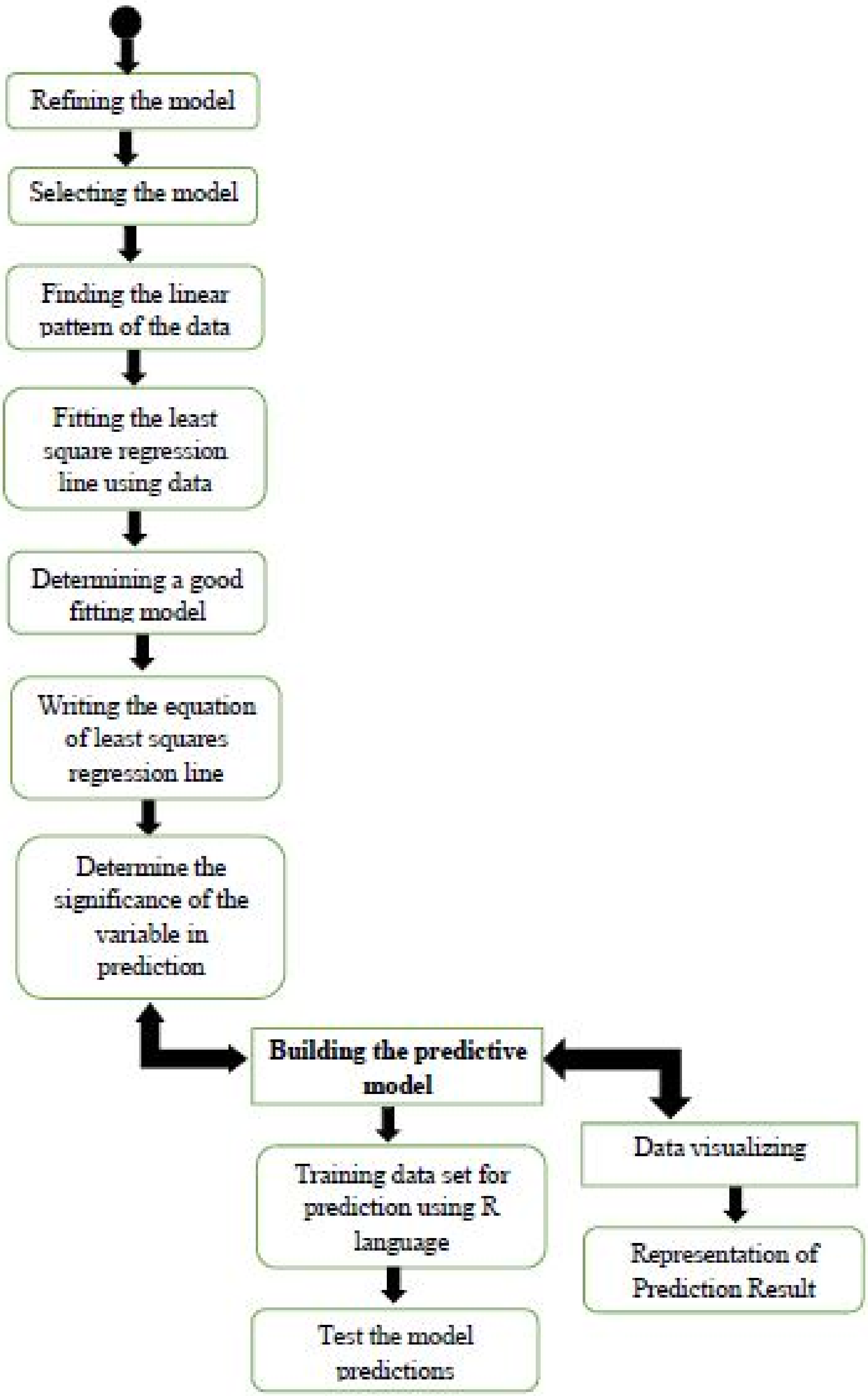


Figure 3.4: Steps involved

Chapter 4

METHODOLOGIES FOR THE PROJECT

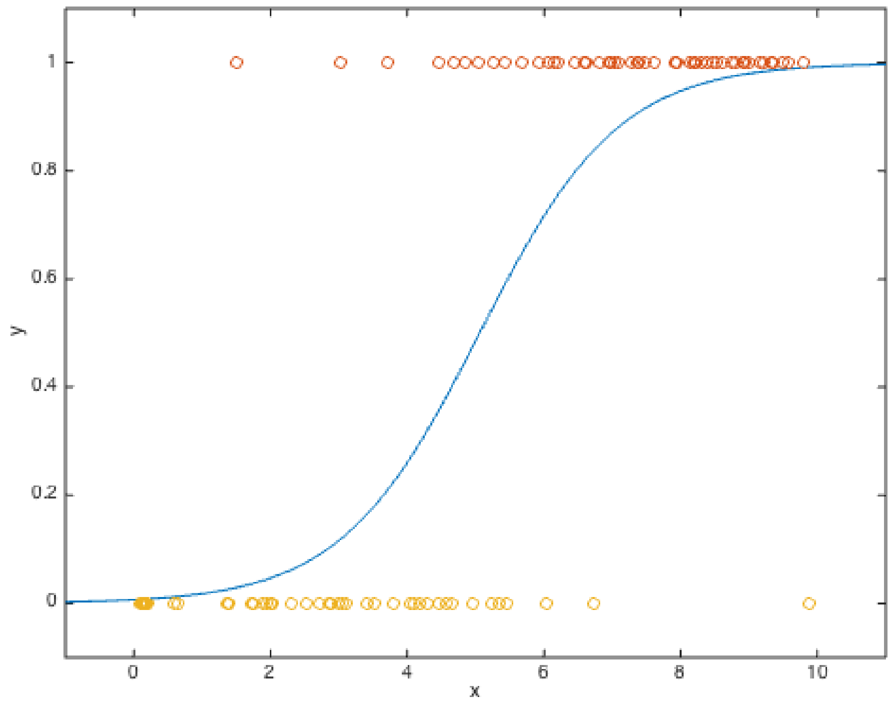
4.1 Detection Methods

4.1.1 Logistic regression

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 (binary or has two classes). 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. Logit function is used to generate log odds of an attribute

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

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.



4.2 Prediction methods

4.2.1 Random Forest Modeling

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.

Random Forests are trained via the bagging method. Bagging or Bootstrap Aggregating, consists of randomly sampling subsets of the training data, fitting a model to these smaller data sets, and aggregating the predictions. This method allows several instances to be used repeatedly for the training stage given that we are sampling with replacement. Tree bagging consists of sampling subsets of the training set, fitting a Decision Tree to each, and aggregating their result. In the Random Forests algorithm, each new data point goes through the same process, it visits all the different trees in the ensemble, which are were grown using random samples of both training data and features. Depending on the task at hand, the functions used for aggregation will differ. For Classification problems, it uses the mode or most frequent class predicted by the individual trees, whereas for Regression tasks, it uses the average prediction of each tree.

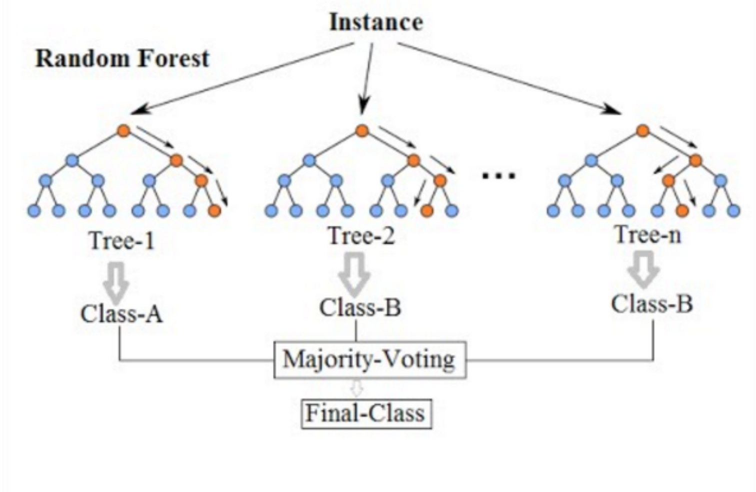


Figure 4.1: Random Forest

4.2.2 Extreme gradient boosting

XGBoost has been widely used in many fields to achieve state-of-the-art results on some data challenges (e.g., Kaggle competitions), which is a high effective scalable machine learning system for tree boosting. XGBoost is optimized under the Gradient Boosting framework and developed by Chen and Guestrin , which is designed to be highly efficient, flexible and portable. The main idea of boosting is to combine a series of weak classifiers with low accuracy to build a strong classifier with better classification performance. If the weak learner for each step is based on the gradient direction of the loss function, it can be called the Gradient Boosting Machines.

This is an ensemble method that seeks to create a strong classifier (model) based on "weak" classifiers.Weak and strong refer to a measure of how correlated are the learners to the actual target variable. By adding models on top of each other iteratively, the errors of the previous model are corrected by the next predictor, until the training data is accurately predicted or reproduced by the model.

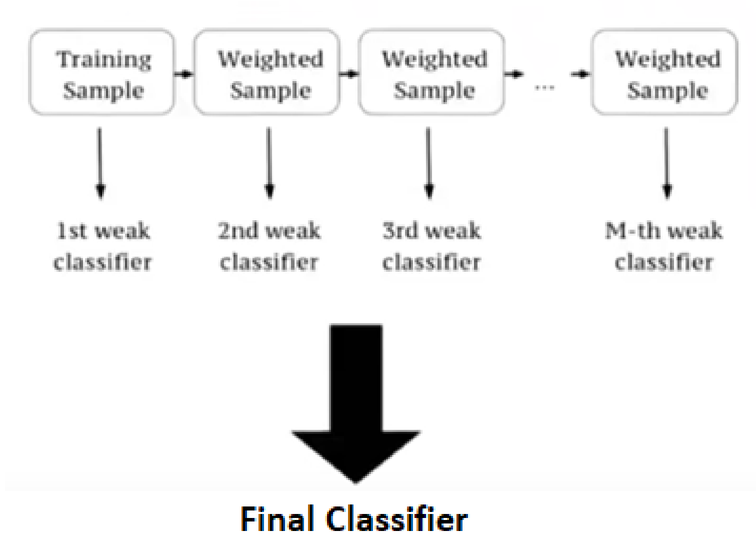


Figure 4.2: Extreme Gradient Boosting

4.2.3 Deep Learning

Deep learning is one of the machine learning methods that is based on its ancestor—the Artificial Neural Network (ANN). The most beautiful thing about Deep Learning is that it is based upon how we, humans, learn and process information. Everything we do, every memory we have, every action we take is controlled by our nervous system which is composed of neurons! The simplest of the ANNs can be created from three layers of "neurons". The input layer, the hidden layer and the output layer. Information flows from the input layer, through the hidden layer to the output layer and then out.

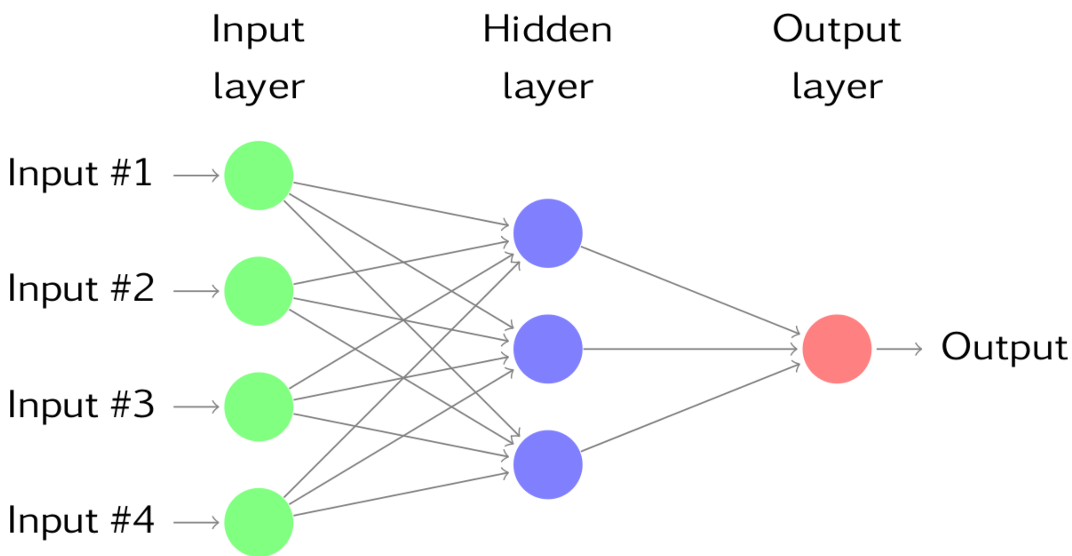


Figure 4.3: Simple Artificial Neural Network

When a neural network is being trained. it is provided with a set of inputs as well as their corresponding outputs. It runs the inputs through the neurons on each of the layers of the network, and using the parameters above, each neuron transforms the input in some way and forwards it to the next layer and so on. The result that it receives on the output layer is then compared to the outputs supplied above and it checks how far apart the two are and accordingly adjusts the parameters on each of the neurons through special algorithmsdesigned to bring the actual and produced outputs as close to each other as possible. It learns to adjust its weights and threshold values to arrive at the correct output. This is what we call as “learning” for the artificial neural network. This process is repeated a (very high) number of times until the produced and expected outputs are as close as possible. That completes the training.when new inputs are supplied to the neural network, we can confidently say that the predicted outputs of the network will be fairly close to the actual outputs. Such ANNs can be used in predicting based upon certain features and classifying objects and images.

Such neural networks which consist of more than three layers of neurons (including the input and output layer) are called as Deep Neural Networks. And training them is called as Deep Learning.

Bibliography

[1] Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, Jingjie Wang,"Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation" *1352-2310/ 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license*,

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

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

[4] Dixian Zhu, Changjie Cai, Tianbao Yang and Xun Zhou: A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. Big data and cognitive computing [2018].

[5] José Juan Carbajal-Hernándezab Luis P.Sánchez-Fernándeza Jesús A.Carrasco-OchoabJosé Fco.Martínez-Trinidadb: Assessment and prediction of air quality using fuzzy logic and autoregressive models: Center of Computer Research – National Polytechnic Institute, Av. Juan de Dios Bátiz S/N, Gustavo A. Madero, Col. Nueva. Industrial Vallejo, 07738 México D.F., Mexico1. (2012) Doi :https://doi.org/10.1016/j.atmosenv.2012.06.004

[6] Sachit Mahajan, Ling-Jyh Chen, Tzu-Chieh Tsai : An Empirical Study of PM2.5 Forecasting Using neural network. IEEE Smart World Congress, At San Francisco, USA [2017]

[7] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. 2018. Deep Distributed Fusion Network for AirQuality Prediction. In Proceedings of the 24th

ACM SIGKDD International Conference on Knowledge Discovery &38; Data Mining (KDD ’18). 965–973.

[8] Rouzbeh Shad, Mohammad Saadi Mesgari, Arefeh Shad, et al. 2009. Predicting air pollution using fuzzy genetic linear membership kriging in GIS. Computers, environment and urban systems 33, 6 (2009), 472–481

[9] World Health Organization. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulphur dioxide.Global Update 2005.Summary of risk assessment. Google Scholar. 2005.

[10] Songgang Zhao, Xingyuan Yuan, Da Xiao, Jianyuan Zhang, Zhouyuan Li. (2018) ”AirNet:a machine learning dataset for air quality forecasting”

[11] Díaz-Robles L, Ortega J, Fu J, Reed G, Chow J, Watson J, Moncada-Herrera J. (2008) ”A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile.” Atmospheric Environment 42: 8331-8340

[12] Liang X, Li S, Zhang S, Huang H, Chen S. (2016) ”PM2.5data reliability, consistency, and air quality assessment in five Chinese cities.” Journal of Geophysical Research: Atmospheres 121: 10,220-10,236

[13] Y. Zhang, Y. He, and J. Zhu, “Research on forecasting problem based on multiple linear regression model PM2.5,” J. Anhui Sci. Technol. Univ., vol. 30, no. 3, pp. 92–97, 2016.

[14] K. R. Baker and K. M. Foley, “A nonlinear regression model estimating single source concentrations of primary and secondarily formed PM2.5,” Atmos. Environ., vol. 45, no. 22, pp. 3758–3767, 2011.

[15] J. B. Ordieresa, E. P. Vergara, R. S. Capuz, and R. E. Salazar, “Neural network prediction model for fine particulate matter (PM2.5) on the US– Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua),” Envi- ron. Model. Softw., vol. 20, no. 5, pp. 547–559, 2005.

[16] Bhaskar, B.V., Rajasekhar, R.V.J., Muthusubramaian, P., Kesarkar, A.P., 2008. Measurement and modeling of respirable particulate (PM10) and lead pollution over Madurai, India. Air Quality, Atmosphere and Health 1, 45e55.

[17] EPA, 1998. National Air Quality and Emissions Trends Report 1997. Environmental Protection Agency 454:R98e016. Environmental Protection Agency, Oûce of Air Quality Planning and Standards, Research Triangle Park. EPA, 1999. Air quality index reporting: final rule. Federal Register. Part III, 40 CFR Part 58.

[18] Gorunescu, F., 2011. Data Mining Concepts, Models and Techniques, Intelligent System Reference Library. Springer-Verlag, Heidelberg. http://dx.doi.org/10.1007/978-3-642-19721-5.

[19] Singh, K.P., Gupta, S., Kumar, A., Shukla, S.P., 2012. Linear and nonlinear modeling approaches for urban air quality prediction. Science of the Total Environment 426, 244e255.

[20] Swamy, M.N., Hanumanthappa, M., 2012. Predicting academic success from student enrolment data using decision tree technique. International Journal of Applied Information Systems 4, 1e6.

Department of Computer Science and Engineering
Vidya Academy of Science & Technology
Thalakkottukara, Thrissur - 680 501
(http://www.vidyaacademy.ac.in)