EXPERIMENTAL ANALYSIS OF MACHINE LEARNING MODELS FOR ANTICIPATING AIR QUALITY INDEX

A PROJECT REPORT

Submitted by

ANINDYA DAS [REG No: RA1611003010103]
NAVNEETH SREENIVASAN [REG No: RA1611003010271]

Under the Guidance of

Mrs.G.ABIRAMI
(Assistant Professor, Department of Computer Science and Engineering)

In Partial Fulfillment of the Requirements

for the Degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR- 603 203

MAY 2020

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that this B.Tech project report titled "**EXPERIMENTAL ANALYSIS OF MACHINE LEARNING MODELS FOR ANTICIPATING AIR QUALITY INDEX**" is the bonafide work of **Mr.Anindya Das and Mr.Navneeth Sreenivasan** who carried out the project work under my supervision. Certifies further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

Mrs.G.ABIRAMI
GUIDE
Assistant Professor
Department of Computer Science and Engineering

Dr.B.AMUTHA
HEAD OF THE DEPARTMENT
Department of Computer Science and Engineering

Signature of the Internal Examiner

Signature of the External Examiner

Own Work Declaration

Department of Computer Science and Engineering

SRM Institute of Science & Technology

Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/ Course :

Student Name :

Registration Number :

Title of Work :

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly references / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalised in accordance with the University policies and regulations.

DECLARATION:
I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.
If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

ACKNOWLEDGEMENT

We express our humble gratitude to **Dr. Sandeep Sancheti**, Vice Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to **Dr. C. Muthamizhchelvan**, Director, Faculty of Engineering and Technology, SRM Institute of Science and Technology, for his invaluable support.

We wish to thank **Dr. B. Amutha**, Professor & Head, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for her valuable suggestions and encouragement throughout the period of the project work.

We are extremely grateful to our Academic Advisor **Dr. A. JeyaSekar,** Associate Professor, and **Dr. R. Annie Uthra** , Associate Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for their great support at all the stages of project work.

We would like to convey our thanks to our Panel Head, **Mrs. G. K. Sandhia**, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for her inputs during the project reviews.

We register our immeasurable thanks to our Faculty Advisors**, Prof. K. Senthil Kumar**, Assistant Professor and **Mrs. K. R. Jansi**, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to our guide, **Mrs. G. Abirami**, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for providing us an opportunity to pursue our project under her mentorship. She provided us the freedom and support to explore the research topics of our interest. Her passion for solving the real problems and making a difference in the world has always been inspiring.

We sincerely thank the staff and students of the Computer Science and Engineering Department, SRM Institute of Science and Technology, for their help during our research. Finally, we would like to thank our parents, our family members and our friends for their unconditional love, constant support and encouragement.

Anindya Das - RA1611003010103

Navneeth Sreenivasan - RA1611003010271

B-Tech Year-IV (Semester – VIII)

Department of Computer Science and Engineering

SRM Institute of Science and Technology

ABSTRACT

WHO assesses that almost seven million individuals expire annually due to contact with minute specks of dirty dust in its 2018 annual report. Experience towards these nice specks of powder in contaminated air indicates to ailments for instance heart sickness, stroke, lung melanoma, continuing obstreperous respiratory ailments and lung infections, together with pneumonia. In this project, we gauge the air quality in a specific zone by utilizing Machine Learning strategies like MLR, SVR, DTR in addition to RFR. We then find out which ML model performs better at predicting the air quality accurately. Then we compare these ML models by judging their performance on various error metrics such as Coefficient of Determination (R2), MAE, RMSE and RMSLE. The trial results demonstrated that the performance of SVR was the least favourable. MLR and DTR both performed satisfactorily well. RFR performed the best among all the regression models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

MLR	Multiple Linear Regression
SVR	Support Vector Regression
DTR	Decision Tree Regression
RFR	Random Forest Regression
OGD	Open Government Data
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
RMSLE	Root Mean Square Logarithmic Error

CHAPTER 1

INTRODUCTION

1.1 Air Pollution

Air pollution ensues when hazardous or inflated quantities of materials counting the vapors (CO2, SO2, CO, CH4, NO, Radon, CFC etc), particles (both organic as well as synthetic), besides biological particles are carried into Globe's weather. WHO assesses that about seven million individuals expire annually due to contact with minute specks of dirty dust in its 2018 annual report World Health Organization (WHO) (2014) [16]. It could lead to sicknesses, aversions and even losses of people; it may likewise hurt additional existing life forms, perhaps living beings and food produces, besides, it can harm the external in addition to the manmade environment (Fig 1.1). Both animal travel and organic actions could create air contamination. Open air contamination alone makes 2.1 to 4.21 million unexpected losses every year. As per specified via the 2014 WHO report, air impurity in 2012 triggered the fatality of about seven million beings all over the biosphere, a measurement normally echoed by dint of the International Energy Agency World Health Organization (WHO) (2014) [16].

Figure 1.1: Air Pollution

Figure 1.2: Causes of Air Pollution

1.2 Causes of Air Pollution

Fresh and toxic-free environment is important for species' all-round development. The air we inhale assumes a significant job in our physical and mental advancement. Fresh air provides oxygen which helps our brain to function properly cleans our lungs, helps in the transportation of blood, and so on. Many earlier humans or tribes use to understand this fact and protect our natural ecosystem. They would worship the rivers, trees and many other natural resources. However since the industrial revolution and human modernization began the concern for our environment started to decline pretty ghastly. Deforestation started in a large scale in order to create space for new industries. With the expansion in the number of enterprises the measure of toxins being transmitted into the air additionally expanded, and with the quantity of timberlands diminishing there weren't adequate trees to assimilate these poisons and convert them into oxygen and other organic materials. This unevenness prompts air pollution.

The presence or introduction of harmful substance into the air is known as air pollution (Fig 1.2). It occurs when substances like carbon dioxide (CO2), carbon monooxide (CO), chlorofluorocarbons (CFC), etc. are introduced into earth's atmosphere in an excessive quantity. This can cause sensitivities, sicknesses besides mortalities of individuals; it may even injure other existing faunae, let's say, mortals in addition with food crops, and so on. Many countries have taken dire measures to overcome air pollution (Fig 1.3).

Figure 1.3: Results of Air Pollution

1.3 Machine Learning

ML refers to the reasonable study of designs along with quantifiable replicas that Personal Computer structures use to produce a specific project starved of exploiting fast instructions, dependent on samples in addition to AI. Machine Learning intentions create a technical archetype in need of test facts, called as "training data", in an attempt to resolve estimates or picks deprived of being clearly adapted to complete the task. Machine Learning intentions are operated in an extensive variety of operations, e.g., electronic mail filtering and Personal Computer vision, where it is troublesome or not good to figure up a customary manipulation for successfully achieving the task.

1.4 Regression

Most of viable ML utilizes supervised learning, which remains the area where you need input factors (X) and an output factor (Y) and we apply a calculation to yield the plotting size commencing from idea to the produce Y = f(X).

Supervised learning topics may be furthermore amassed into Regression and Classification problems. A Classification question stays the fact where the output variable is a class, for instance, "green" and "white" or else "disease" and "no disease". A Re-

gression question is the fact where the output variable remains a candid or consistent worth, for instance, "income" or else "weight".

In statistical analysis, study of regression is loads of factual events for assessing the networks amongst a poor variable (frequently termed the 'result variable') and a minimum of a unique free factors (often named 'predictors' or 'covariates'). Regression investigation is essentially utilized for two theoretically particular purposes. To start with, regression examination is usually exploited for anticipation in addition to predicting, where the aforementioned employment has substantial insurance in the arena of ML. Second, in certain situations regression investigation may be employed to assemble underlying associates among the predictors plus needy variables.

1.5 Types of Regression

Multiple Linear Regression (MLR) is one of the straightforward forms of regression. It is a method wherein the response variable is repeated in nature. The connection between the needy variable and independent factors is thought to be linear in essence.

SVM can also remain exploited by way of one Regression technique, possessing altogether the primary features that depict the manipulation. The Support Vector Regression utilizes indistinguishable standards from SVM for characterization. It consists of Linear and Non-Linear SVR.

Decision Tree fabricates regression or else classification models as a diagram construction. It splits up a dataset towards miniature and more miniature subdivisions whereas concurrently an associated decision tree is by and large slowly fashioned. The conclusive product is a tree with Decision nodules and Leaf nodules. A Decision node has at least two divisions. The highest decision node in a tree is called Root (parent) node.

Random forests also called as Random Decision Forests remain an aggregate knowledge strategy in lieu of classification, regression in addition to diverse errands which exist via evolving a vast quantity of decision trees at training time and giving out the mean forecast of the individual trees at testing time [17]. The trees in random forests

perform in a match. There is no communication amidst these trees whilst constructing the trees. The normal expectation of the N decision trees is the Random Forest Prediction. Random Forest algorithm can foresee for both continous (genuine esteemed) and categorical information.

1.6 Error Metrics

The Coefficient of R2 Determination signified as R2, stays the extent of the difference in the needy variable that is anticipated from the experimental factors.

Mean Absolute Error (MAE) is a proportion of distinction amongst 2 constant factors.

Root Mean Square Error (RMSE) signifies the root of the sample moment in the variations stuck between prophesied values and experiential values.

In case of Root Mean Square Logarithmic Error (RMSLE), we take the log of the forecasts and real qualities. So fundamentally, what changes is the variance that we are estimating.

CHAPTER 2

LITERATURE SURVEY

2.1 Literature Survey

Experimental Analysis of ML Models aimed at Anticipating AQI is a project in which the calculation of AQI in a particular area is done using machine learning regression models. The machine learning regression models used at this point are MLR, SVR, DTR along with RFR. The error metrics used are Coefficient of R2 Determination, MAE, RMSE, and RMSLE. The project uses Decision Tree Regression instead of using Binary Neural Networks due to its advantages over Neural Networks.

Liu. B, Shi. C, Li. J, Li. Y, Lang. J, Gu. R (2019) [1] saw that RFR played out the greatest with MAE and RMSE. The evidence utilized contained clamor which could have brought around a poor showing of Support Vectors. However, the solid equipment reliance of neural systems attached to the processor besides the obscure time term of BPNNs made this less proficient. Decision trees stand quicker in addition to the fact of being extra interpretable than BPNNs.

Keller.C.A, Evans.M.J, Kutz.J. N Pawson.S (2017) [2] functioned with the maximum relevant Spatial-Temporal Relations and a mixture of numerous neural networks together with ANN and CNN in addition to a long-short term memory toward abstracting Spatial-Temporal Relations. It also uses Dynamic Time Warping and long-short term memory besides it similarly includes trends from multiple locations. But this scheme contains evidences of missing data, absent memory, unsupervised learning and overfitting owed to their usage of CNN (Convolutional Neural Network). It furthermore stays much slower and extra complicated associated to further ML prototypes such as per Decision Tree models, which don't cause any problems in the circumstances of misplaced statistics.

Ganesh.S, Modali.S.H, Palreddy.S.R and Arulmozhivarman.P (2017) [3] showed that Random Forest Algorithm allowed them towards forecasting better as soon as the

Awareness change of atmospheric species because of chemistry devoid of the necessity to invoke the chemical integrated. They predicted the prototype consuming standard chemistry model and ML designs and displayed that 50 percentage of these concentrations prophesied by means of ML consist of an error not as much as 1 ppbv. Their work included a sparse dataset which has partial memory and restricted processing power which makes the system run very slow. DTR and RFR are further faster to train matched with the usage of sparse dataset.

Adaptive Deep Learning based Air Quality Prediction Model using the most relevant Spatial-Temporal Relations Soh.P, Chang.J and Huang.J (2017) [4] used Air Quality index which hinges on the attention of impurities like SO2, NO2, CO2 etc. It uses ML models such as SVR and LR. It gives means as well as ways for refining the outcomes of ML representations which has not accomplished fit. But it uses comparatively less number of training and testing data in case of Delhi. This project has only considered a portion of records from Delhi, one of the most polluted cities of the world. In comparison, almost 10000 more records of information from Houston was taken for this work.

Ghaemi. Z, Alimohammadi. A Farnaghi. M (2018) [5] indicated an upgraded air superiority expectation strategy dependent based with the LightGBM type to anticipate the PM2.5 fixation on 35 air quality checking stations at Beijing during an assortment of 24 h. Light GBM is fast and may deal by way of the enormous magnitude of information and takes lower memory to run. But still, in our undertaking which nearly utilizes 1000 columns, it isn't fitting to utilize LGBM since they stand delicate to overfitting and will, without a lot of stretch overfit little information. It moreover inputs a great number of factors while coding.

Zhang.Y,Wang.Y, Gao.M, Ma.Q, Zhao.J, Zhang.R, Wang.Q and Zhang.L.H (2018) [6] utilized a spatio-fleeting framework, planned utilizing a LaSVM-considered online calculation. Execution of this framework is assessed by contrasting the expectation aftereffects of the AQI through that of a conventional SVM calculation. At this juncture, the preparing time altogether diminishes by expelling the non-bolster vector tests at the preparation step, and without diminishing the exactness. Be that as it will, the data records employed in this above-mentioned undertaking is imbalanced which doesn't

enable the design to exist appropriately.

Amado.T.M, Dela Cruz.J.C (2018) [7] utilized five prescient models, k-closest neighbours (KNN), support vector machine (SVM), Naïve-Bayesian classifier, irregular woodland, and neural system. The error metrics utilized were CV execution, Confusion Accuracy, and Logloss execution. Though they got good outcomes in lieu of Neural Networks, the utilization of neural system can prompt a more slow reaction taking place at a bigger scale, which we have exploited inside our information.

Zheng.Y, Yi.X, Li.M, Li.R, Shan.Z, Chang.E, Li.T (2015) [8] have used LR, Neural networks, a dynamic aggregator and inflection predictor. Indirectly, [2] derived from this. They assessed their design by means of information from 43 cities. Their usage of Neural Networks has made it much slower and further complicated related to other ML methods such as DTR models.

Hable-Khandeka.V and Srinath.P (2017) [9] have placed to light various Big Data perspectives by considering various heterogeneous data-sources and factors affecting the air condition. It has also explored recent tools meant for contemporary air quality monitoring. However, they had uncertainty and non-linearity within the system.

De Vito.S, Massera.E, Piga.M, Martinotto.L, Di Francia.G (2007) [10], has found a system which considers LR, SVR, Neural Networks, DTR and Lasso Regression. But since Lasso encourages shrinking of coefficients to 0, i.e. dropping those variates from model, Lasso Regression is measured as unreliable.

Sohn.S.H, Oh.S.C, Jo.B.W and Yeo.Y.K (2000) [11] have foreseen an Ozone development utilizing neural systems. In this work, they confirmed that by means of the assistance of various strategies for handling information, the custom of Machine Learning practices towards air quality expectations performs sensibly well. Similar to [1], the solid equipment reliance of neural systems attached to the processor plus the obscure time term of BPNNs have made this less proficient.

K.P.Singh, Shikha Gupta and Premanjali Rai Singh.K.P, Gupta.S, Rai.P (2013) [12] have used Principal Component Analysis within this work, using Single Decision Tree, Decision Tree Forest, Decision Tree Boost and Support Vector Machines. They established that both DTF as well as DTB have outperformed SVM. However, the designs

are very complex to generate besides the point that it could require a long amount of time.

Dragomir.E.G (2010)'s effort on air quality forecast utilizing k-closest neighbour calculations [13] is interrelated with [7], as they also have worked with a similar sort of algorithm, though on different scenarios. This catalogue is used to categorize the effluence quantity and on the road to inform the populace about some possible episodes of pollution. Nonetheless, the training data set of the prototype takes a huge amount of period to work.

Chaloulakou.A, Saisana.M, and Spyrellis.N (2003) [14] have worked on forecasting the subsequent twenty-four hour period's maximum every sixty-minute stretch ozone absorption in the Athens basin [14] using LR and neural networks, finding the neural networks as providing improved outcomes than linear regression. Neural networks may take ample time to process and could require a good deal of hardware dependence.

Vlachogianni.A, Kassomenos.P, Karppinen.A, Karakitsios.S, and Kukkonen.J (2011) [15] have used LR, ANN and DTR to compare the Air Quality Index in Greece. They established that ANN accomplishes improved performance to discover the day-to-day AQI and LR performs best to catch the hourly AQI.

E.Munoz, Mj Jimenez Come, M.L.Martin and F.J.TrujiloMunoz.E, Come.Mj.J, Martin.M.L and Trujilo.F.J (2013) [18] follow one-of-a-kind classification strategies that allows you to provide 24 h increase forecasts of the every day peaks of SO2 and PM10 concentrations within the Bay of Algeciras. K-nearest-neighbours, multilayer neural networks with backpropagation alongside support vector machines (SVMs) are the type techniques used. A resampling method with twofold cross-validation has been applied, the use of special excellent indexes to evaluate the overall performance of the prediction fashions. SVM fashions finished higher real fantastic charge and accuracy (ACC) first-rate indexes.

A.Kurt and A.B.OktayKurt.A and Oktay.A.B (2010) [19] used air toxin information from 10 distinctive air quality checking stations in Istanbul was utilized in anticipating Sulphur Dioxide (SO2), Carbon Monoxide (CO) and Particulate Matter (PM10) levels 3 days ahead of time for the Besiktas area. The discoveries are very agreeable. At the

point when the privilege neighboring areas are picked, the geographic models consistently yield lower mistake than the non-geographic models.

A.Kumar and P.Goyal'sKumar.A and Goyal.P (2011) [20] work was on creating estimating model for anticipating day by day AQI, which can be utilized as a premise of dynamic procedures. Right off the bat, the AQI has been assessed through a strategy utilized by US Environmental Protection Agency (USEPA) for various criteria contaminations as Respirable Suspended Particulate Matter (RSPM), Sulfur dioxide (SO2), Nitrogen dioxide (NO2) and Suspended Particulate Matter (SPM). Also, the day by day AQI for each season is estimated through three measurable models in particular time series auto regressive integrated moving average (ARIMA) (model 1), principal component regression (PCR) (model 2) and blend of both (model 3) in Delhi. The presentation of every one of the three models are assessed with the assistance of watched centralizations of pollutants, which mirrors that model 3 concurs well with watched esteems, when contrasted with the estimations of model 1 and model 2.

T.Bellander, N.Berglin, P.Gustavsson, T.Jonson, F.Nyberg and G.PershagenBellander.T, Berglind.N, Gustavsson.P, Jonson.T, Nyberg.F,Pershagen.G (2001) [21] developed an application on populace based case-control investigation of lung cancer in Stockholm, Sweden, was to utilize discharge information, scattering models, and Geographic Information Systems (GIS) to evaluate chronicled introduction to a few parts of surrounding air contamination. They additionally utilized straight intra-and extrapolation to get gauges for every single other year 1955-1990. At last, they connected yearly air contamination assessments to yearly organizes, yielding long haul private presentation files for every person. The outcomes demonstrate that GIS can be helpful for presentation appraisal in natural the study of disease transmission examines, given that definite geologically related introduction information are accessible for applicable timeframes.

M. Gao, L. Yin, and J. Ning'sGao.M, Yin.L, and Ning.J (2018) [22] study explored the attainability of utilizing ANN model with meteorological parameters as information factors to foresee ozone focus in the urban territory of Jinan, China. They right off the bat found that the engineering of system of neurons had little impact on the anticipating ability of ANN model. They found that the anticipating capacity of the closefisted ANN model was satisfactory. At last, vulnerability and affectability investigation were like-

wise performed dependent on Monte Carlo recreations (MCS). It was reasoned that the ANN could appropriately foresee the encompassing ozone level. Most extreme temperature, environmental weight, daylight length and greatest breeze speed were recognized as the prevail input factors fundamentally impacting the forecast of surrounding ozone fixations.

S. S. Roy, C. Pratyush, and C. Barna [23]RoyS.S, Pratyush.C, and Barna.C (2016) propose three prescient models for estimation of grouping of ozone gases noticeable all around which are Random Forest, Multivariate Adaptive Regression Splines and Classification and Regression Tree. Assessment of the forecast models demonstrates that the Multivariate Adaptive Regression Splines model portrays the dataset better and has accomplished altogether better expectation precision when contrasted with the Random Forest and Classification and Regression Tree. Additionally, Random Forest sets aside somewhat more effort for building the tree. Watching all the diagrams Multivariate Adaptive Regression Splines gives the nearest bend of both train and test set when analyzed. It very well may be presumed that Multivariate Adaptive Regression Splines can be an important device in foreseeing ozone for future.

A.Cotter, O.Shamir, N.Srebro and K.SridharanCotter.A, Shamir.O, Srebro.N, Sridharan.K (2011) [24] concentrate how mini-batch algorithmic group can be improved utilizing accelerated gradient methods. They give a novel investigation, which shows how standard gradient strategies may now and again be lacking to acquire a noteworthy spped-up. They propose a novel accelerated gradient calculation, which manages this inadequacy, and appreciates a consistently prevalent assurance. They finish up their paper with investigates real-world datasets, which approves their calculation and validates our theoretical insights.

2.2 Inference from the survey

From the survey, we were able to find that the usage of BPNN is not good as compared to the use of Decision Tree since the strong hardware dependence of neural networks on the processor and the unknown time duration of BPNNs make it less efficient. Decision trees are faster and more interpretable than BPNNs.

Use of CNN is also not very good since it consists of missing data, memory and unsupervised learning and overfitting. It is also much slower and more complicated compared to other machine learning models such as Decision Tree models, which do not cause any problems in case of missing data.

The problem with using a sparse dataset is that it has limited memory and limited processing power which makes the algorithm run very slow. DTR and RFR are more faster to train compared to the usage of sparse dataset.

It is not advisable to use LGBM since they remain delicate towards overfitting which could lead to simply overfitting trivial facts. It also takes in a lot of parameters while coding. We also cannot use imbalanced dataset as it doesn't permit the algorithm to be trained correctly.

The utilization of neural system can prompt a more slow reaction on a bigger scale, which is what is utilized in our information. We can also see that systems with uncertainty and non-linearity are not advisable.

Usage of Lasso Regression is cannot be considered as reliable since it can drop the variates from the model. We have also found that creation of the Decision Tree Boost and Decision Tree Forest are very complex and may require a lot of time.

CHAPTER 3

MODEL SELECTION

There are various types of machine learning models such as regression models, classification models, clustering models, etc. Each machine learning model solves a very specific type of problem. There are factors that should be looked upon while selecting the appropriate ML models such as, requirement or required output for the problem, type of data, availability of input and output data (as in supervised learning), and so on.

In this experiment since we are using a data set that already consists of input data and output data, we are going to use supervised learning. The supervised learning is further classified into two types which are regression and classification. The regression method is used for predicting numerical values for a given input value (which is also a numerical value) and the classification method is used to predict the class or category for a given input value. Since the output of our data set is a numerical value (AQI value) we are going to use regression models for our predictions.

3.1 Regression Models

There are various types of regression model, however only a few types of regression models are used widely. In this project we have used some widely used regression models such as Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Support Vector Regression (SVR) along with Random Forest Regression (RFR).

3.1.1 Multiple Linear Regression

MLR, in like way, implied comparably as multiple regression, is a factual method that utilizes a couple of self-sufficient features to explain or anticipate the result of a variable (Fig 3.1). Multiple Linear Regression is an immediate method to manage showing the association between a response variable and explanatory factors. Multiple regression is

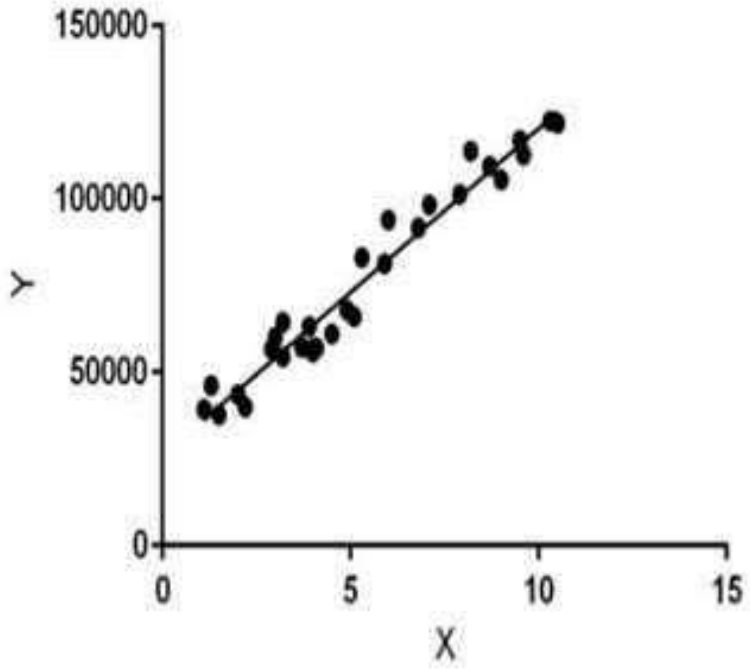


Figure 3.1: Multiple Linear Regression

an extension of linear (OLS) regression that utilizes only one explanatory variable. The aim of Multiple Linear Regression is to exhibit the direct association between the independent variables and response(dependent) variables. Mathematically, a MLR model can be represented by using the following equation:

$$y_i = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n \tag{3.1}$$

Where,

$y_i - dependent\ variable$

$b_0 - y\ intercept$

$x_1, x_2, ..., x_n - independent\ features$

$b_1, b_2, ..., b_n - weights\ of\ the\ respective\ independent\ features$

3.1.2 Support Vector Regression

In simple MLR, we attempt toward reducing the error estimate. However, in Support Vector Regression (SVR), we attempt to adapt the error inside a specific brink (Fig

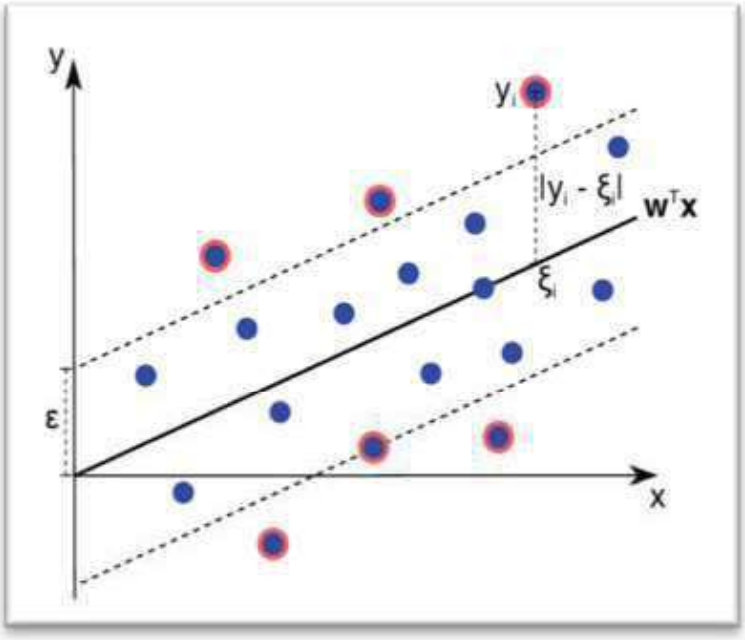


Figure 3.2: Support Vector Regression

3.2). We attempt to discover a regression plane, with the goal that all information of a set will be nearest to that plane. SVR uses unclear gauges from the Support Vector Machines (SVM) for game plan, with only several minor differentiations. Like SVM, SVR likewise chooses data points from the data sets and attempts to fit the rest of the data points dependent on these chose points. These are known as support vectors.

3.1.3 Decision Tree Regression

Decision Tree Regression Decision Tree Regression (DTR) builds regression or classification models as a tree construction. It isolates the given dataset towards more diminutive and humbler subsets whereas simultaneously a correlated decision tree is bit by bit made. (Fig 3.3).The conclusive event is a tree surrounded by decision nodes in addition to leaf nodes. Decision Tree remains a choice creation maneuver which employs a flowchart-like tree structure. A decision tree is made of nodes and edges connecting the nodes with each other. These internal nodes address a condition on the quality, every one of the branches discourse the after effect of the condition and every leaf node discourses a class name or results for a given subset of sources of data. The decision tree model is seen as genuinely extraordinary and generally, uses coordinated learning techniques. Tree-based frameworks enable sensible models with high accuracy, ampleness,

Figure 3.3: Decision Tree Regression

and straightforwardness of understanding.

3.1.4 Random Forest Regression

Random forests perform by developing a considerable count of decision trees at training time and bringing forth the mean expectation of the individual trees at testing time. (Fig 3.4).The essential thought behind this is to join various decision trees in deciding the final output as opposed to depending on singular decision trees. Random Forest Regression (RFR) is a further developed and hearty type of Decision Tree regression algorithm. In decision tree regression just a single choice tree is produced which is responsible for anticipating the result. In RFR arbitrary K data points are picked from the preparation set. Utilizing the picked K data sets a decision tree is created. This technique is rehashed N number of times, bringing about N particular Decision Trees. The normal expectation of the N decision trees is the Random Forest Prediction. Random Forest algorithm can foresee for both continous (genuine esteemed) and categorical information.

3.2 Error Metrics

3.2.1 Coefficient of R2 Determination

The Coefficient of R2 Determination is the extent of the fluctuation in the needy variable that can be anticipated from the explanatory variable. In regression, the R2 coefficient

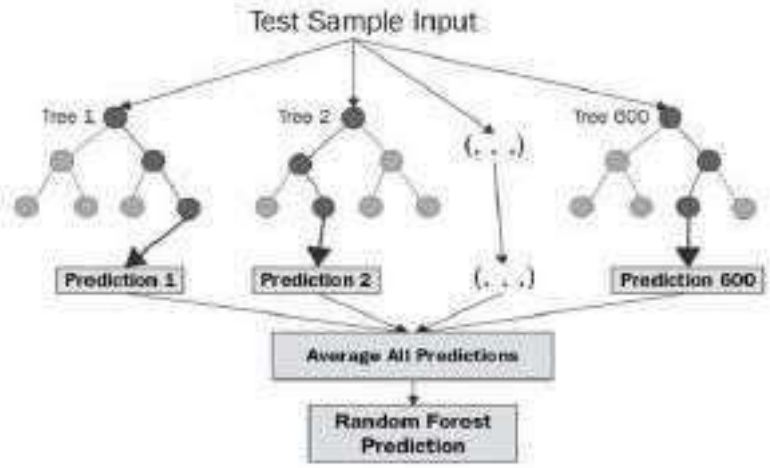


Figure 3.4: Random Forest Regression

of determination is a factual proportion of how well the regression forecasts surmised the genuine information points. It is a measurement utilized with regards to factual models whose principle design is either the forecast of eventual results or the examination concerning theories, based on other related data.It gives a proportion of how particularly watched results are imitated by the model, in view of the extent of absolute variety of results clarified by the model.

$$R^2 = \Sigma((y_p - y_m)^2)/\Sigma((y - y_m)^2) \tag{3.2}$$

Where,

$R^2 - Coefficient\ of\ R^2\ Determination$

$\Sigma -\ Mean$

$y_p - Predicted\ value$

$y_m - Mean\ value$

$y - Output\ value$

3.2.2 Mean Absolute Error

Mean Absolute Error MAE is a proportion of difference between two consistent factors. MAE isn't indistinguishable from RMSE; the utilization of squared separations hinders the translation of RMSE. In insights, MAE is a proportion of errors between matched perceptions communicating a similar phenomenon. The mean absolute error

utilizes a similar range as the information being estimated. This can be called scale-subordinate exactness measure and thusly it cannot be utilized in making correlations amidst courses using distinctive ranges. The mean absolute error is a typical proportion of gauge mistake in time series analysis, now and again utilized in confusion beside the further accepted content of mean absolute deviation.

$$MAE = (1/n)\Sigma|y_p - y_r| \tag{3.3}$$

Where,

$y_p - Predicted\ value$

$y_r - Observed\ value$

3.2.3 Root Mean Square Error

Root Mean Square Error RMSE is an algebraic recording method that additionally gauges the normal greatness of the error. It is a regularly utilized proportion of the alterations between values (test or masses regards) foreseen by means of a prototype or else a predictor and the characteristics seen. Root Mean Square Error is the root of the sample moment of the alterations amid real values in addition to predicted values or the quadratic average of these distinctions. The RMSE gives the amounts of the mistakes in assumptions for divergent occasions into a lone part of farsighted capability. RMSE is a proportion of exactness, to look at determining mistakes of dissimilar models for a precise dataset and not among datasets, since it's scale-subordinate.

$$RMSE = \sqrt{(\Sigma((y_p - y_r)^2))/n} \tag{3.4}$$

Where,

n - Total number of values

$y_p - Predicted\ value$

$y_r - Observed\ value$

3.2.4 Root Mean Square Logarithmic Error

In the event of Root Mean Square Logarithmic Error RMSLE, we attain the log regarding the expectations and predicted qualities. So fundamentally, what changes is the fluctuation that we are estimating. RMSLE metric just considers the relative mistake between the predicted and the actual worth and the size of the error isn't huge. RMSLE causes a bigger penalty for the underestimation of the original variable than the Overestimation. In straightforward words, more penalty is acquired when the anticipated Value is not exactly the Actual Value.On the other hand, less penalty is brought about when the anticipated worth is more than the real worth. The most remarkable property of the RMLSE is that it penalizes the underestimation of the genuine value more seriously than it accomplishes for the overestimation.

(3.5)

RMSLE = $\sqrt{(1/n)\Sigma(\log(y_p + 1) - \log(y_r + 1))}$

Where,

$y_p - Predicted\ value$

$y_r - Observed\ value$

CHAPTER 4

IMPLEMENTATION AND DESIGN

For this experiment we are using scikit-learn library. Scikit-learn remains a library in Python that provides many built-in unsupervised and supervised learning algorithms. The following section does a deeper discussion on the methods, and procedure followed in this experiment.

4.1 Data Collection

India has 28 states and a population of roughly around 1.37 billion. The hourly monitored values of the gases, obtained from the Central Pollution Control Board, is present in Fig 4.1. The data set we are using for this experiment has AQI data collected from 19 different states. This data set is taken from Open Government Data (OGD) Platform India (Fig 4.2). This site gives Current National AQI esteems from various checking places through India. The impurities checked are SO2 NO2, PM10 and PM2.5, CO, O3, etc. This site provides the data set in the form of xml file. Thus after converting the xml file into csv file (Fig 4.3) we had the following columns in our data set: state, city, station, date, time, concentration of contaminants like PM2.5, PM10, NO2, NH3, SO2, CO, O3, AQI value, and Predominant Matter (Fig 4.4).

4.2 Missing Value Processing

Because of different reasons, missing qualities in the informational index are very common in meteorological informational collections. The missing values might diminish the nature of the model produced by ML algorithms. There are various techniques for dealing with missing values. Deleting all the rows which have missing value is one of the techniques however this technique is not suitable if the size of the data set is small. Doing so in a small data set could result in the loss of essential data and the model

Table 3.11 Breakpoints for AQI Scale 0-500 (units: $\mu g/m^3$ unless mentioned otherwise)

AQI Category (Range)	PM_{10} 24-hr	$PM_{2.5}$ 24-hr	NO_2 24-hr	O_3 8-hr	CO 8-hr (mg/m^3)	SO_2 24-hr	NH_3 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.6 –1.0
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1- 10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200	2.1-3.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Severe (401-500)	430+	250+	400+	748+*	34+	1600+	1800+	3.5+

*One hourly monitoring (for mathematical calculation only)

Image credit: National Air Quality Index Report by Central Pollution Control Board

Figure 4.1: National Air Quality Index Report

Figure 4.2: Open Government Data Platform

Figure 4.3: Transformation from XML to CS

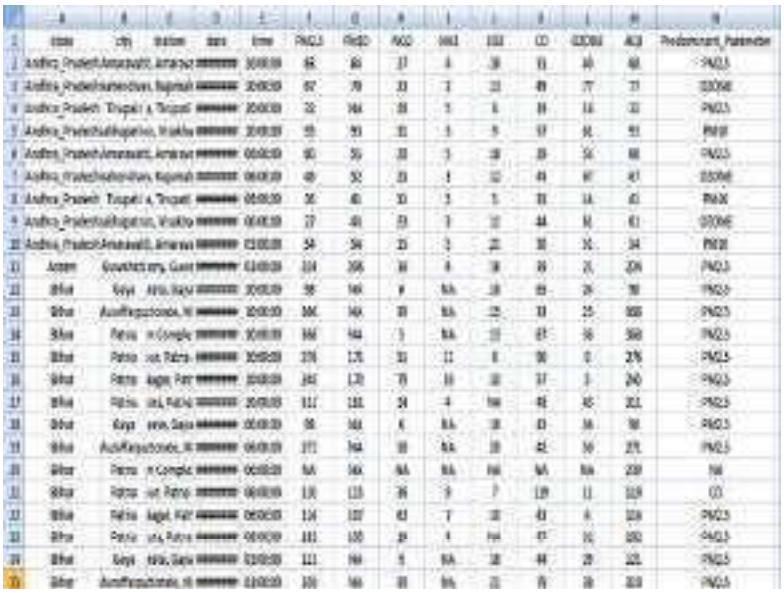


Figure 4.4: Transformed CSV File

generated by the ML algorithms would not be accurate. We also cannot delete rows of a data set that consists of sensitive data.

Another way for handling the missing values can be done by replacing the missing points by the mean, median or mode of the column. We have used a similar technique for this experiment. In this experiment, we replaced the missing values with the mean of that column for a particular state. We have used the following formula for calculating the mean and processing missing values.

$$(mean)_s = (\Sigma(Z_i)_s)/(n_s) \tag{4.1}$$

Where,

$Z_i - pollutant\ concentration\ for\ a\ state\ s$

$n_s - total\ number\ of\ rows\ for\ a\ state\ s$

$s - a\ unique\ state$

4.3 Feature Selection

As mentioned earlier our data set consists of the features such as state, city, station, date, time, PM2.5, PM10, NO2, NH3, SO2, CO, O3, AQI value, Predominant Matter. Feature selection is important for generating accurate models. In AI and Statistics, feature selection, or else variable selection stands as the method towards selecting a subset of applicable important outputs (factors, indicators) for usage in model progress. In feature selection, a feature or an independent variable is selected if it has a statistically significant effect on the outcome variable. This statistical significance is determined by the p-value. In general, if the p-value for a feature is less than 0.05, then we can say that the feature has a significant influence on the outcome variable, and thus we say that the feature is statistically significant. (Table 4.1)

Table 4.1: Estimation of p-values for concentration of pollutant

Feature	**p-value**	**Outcome**
PM2.5	0.00	Significant
PM10	0.00	Significant
NO2	3.91e - 84	Significant
NH3	4.09e - 109	Significant
SO2	6.61e - 0	Significant
CO	1.70e - 53	Significant
O3	4.90e - 08	Significant

Hence, we see that all the pollutants have a significant effect on the outcome variable (AQI). Thus all the pollutant features can be considered in our feature set. Additionally to these features we have also added state, city and station columns in our feature set. The reason behind this is that the average AQI value of a given state or place stays almost constant with few minor fluctuations. Even with these fluctuations the AQI value almost stays pretty close to its average value. For example, consider the states Delhi and Chennai. Their average AQI index is around 164 and 93 respectively. Thus on any given day, their AQI indexes would be closer to their average values such as 166 and 91 for Delhi and Chennai respectively. Therefore our final feature set consist columns of state, city, station and all pollutant matters.

4.4 Data Transformation and Feature Scaling

Our feature set consists of both numeric data (concentration of pollutants) and non-numeric data (state, city, station). Since regression algorithms are built on a statistical model, it uses numeric data for its computation and model generation. Thus we cannot directly use our feature set because of non-numeric data. Therefore we need to transform the data set by converting the string values (non-numeric data) into integer values (numeric data). This can be done in a few ways. The first way is to assign a unique numeric value to each non-numeric values. For example, assign value 1 to state Andhra Pradesh, value 2 to Assam and so till the last state is assigned. Follow the same procedure for the city and station column. Look at the following table for a better understanding.

Table 4.2: Value assigning to each state

State	**Numeric Equivalent**
Andhra Pradesh	1
Assam	2
....	....
....	....
West Bengal	19

This methodology may look right but there is one major issue with this technique. This method will create a regression model which will give higher priority to the state having larger value (West Bengal with assigned value 19) and lower priority to the state having a lower value (Andhra Pradesh with assigned value 1). Thus it will lead to inaccurate predictions.

Another method for changing the non-numerical information into numerical information can be accomplished by making a different new section for each non-numeric values. Create a new column for each non-numeric value and assign the value 1 if the row contains that non-numeric value else assign 0 if otherwise. Consider Table 4.3 for better understanding.

The data set in Table 4.3 is converted into the data set in Table 4.4 (new feature set)

Table 4.3: Data set sample before transformation

State	**.....**	**PM2.5**
Andhra Pradesh	.....	68
Assam	....	141
.....	.....	.....
.....	.....	.....
West Bengal	.....	133

Table 4.4: Data set sample after transformation

Andhra Pradesh	**Assam**	**.....**	**West Bengal**	**.....**	**PM2.5**
1	0	.....	0	.....	68
0	1	.....	0	.....	141
.	.	.....	.	.....	.
.	.	.....	.	.....	.
0	0	.....	1	.....	133

Thus the regression model generated using the new feature set will give equal priority to every state and can now make more accurate predictions.

Once we have achieved our new transformed data set it is time for feature scaling. Feature Scaling is a strategy that is to institutionalize the free features present in the data set in a settled space. It is applied when the independent variables vary in order of magnitudes from other independent variables. It chiefly contributes through normalizing the data kept in between a precise array, and it similarly aids in hastening the calculations in an algorithm. It essentially supports in normalizing the data kept in between a precise array and, it similarly aids in hastening the calculations in an algori. To perform feature scaling we used StandardScaler class present in sklearn.preprocessing module.

Thus after performing all the data transformation and feature scaling on our data set, the new data set obtained contains about 334 columns and over 1500 rows. We then split the feature set to two components creating training set and testing set. The size of the test set came to be 0.25 times that of the feature set.

4.5 Architecture Diagram and Activity Diagram

The architecture diagram for this project is given in Fig 4.5. The raw data is taken from the OGD platform and is converted into XML file database. From there, we take the data and we converth the XML file to a CSV file (Data Store). From there, we classify the data into Features and Labels. From there, using the four regression methods, namely MLR, SVR, DTR and RFR, we test and train the data using the four Error Metrics, which are Coefficient of R2 Determination, MAE, RMSE and RMSLE. Then, we compare the values and find the method which has performed best on the training data as well as the testing data.

The activity diagram for this project is given in Fig 4.6. The raw data is taken from the OGD platform and is converted into XML file database. From there, we take the data and we convert the From there, using the four regression methods, namely MLR, SVR, DTR and RFR, we test and train the data using the four Error Metrics, which are Coefficient of R2 Determination, MAE, RMSE and RMSLE. Then, we compare the values and find the method which has performed best on the training data as well as the testing data.

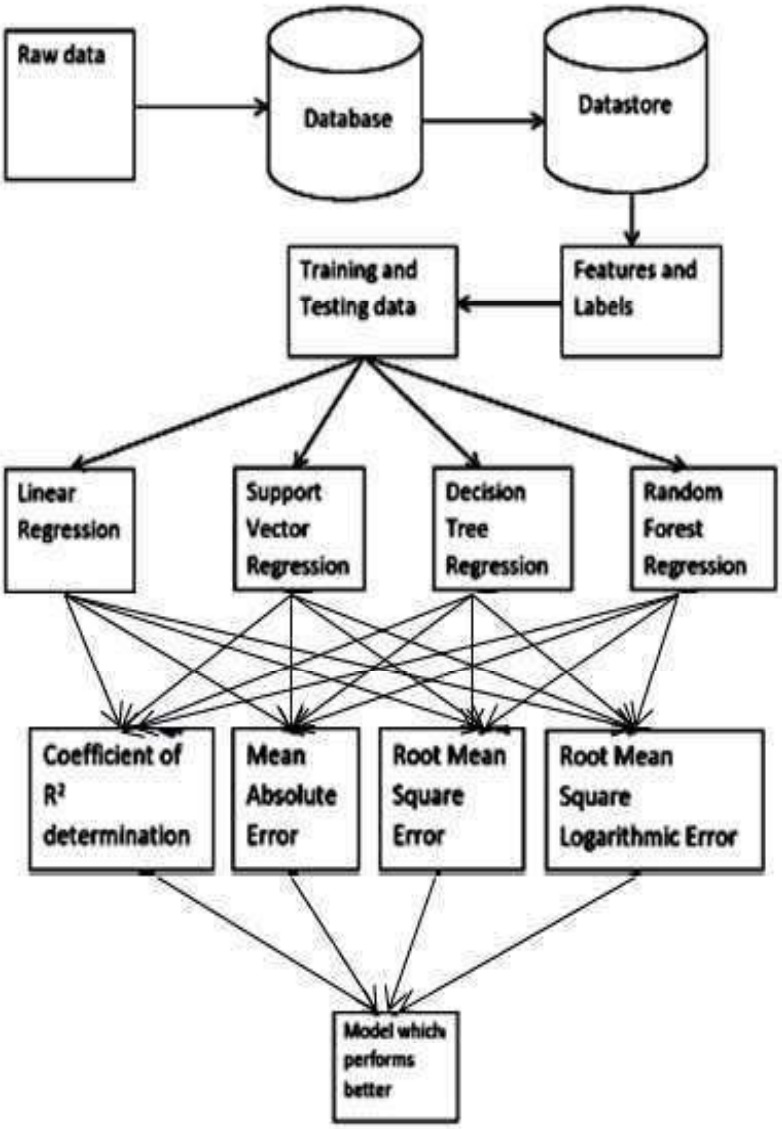


Figure 4.5: Architecture Diagram

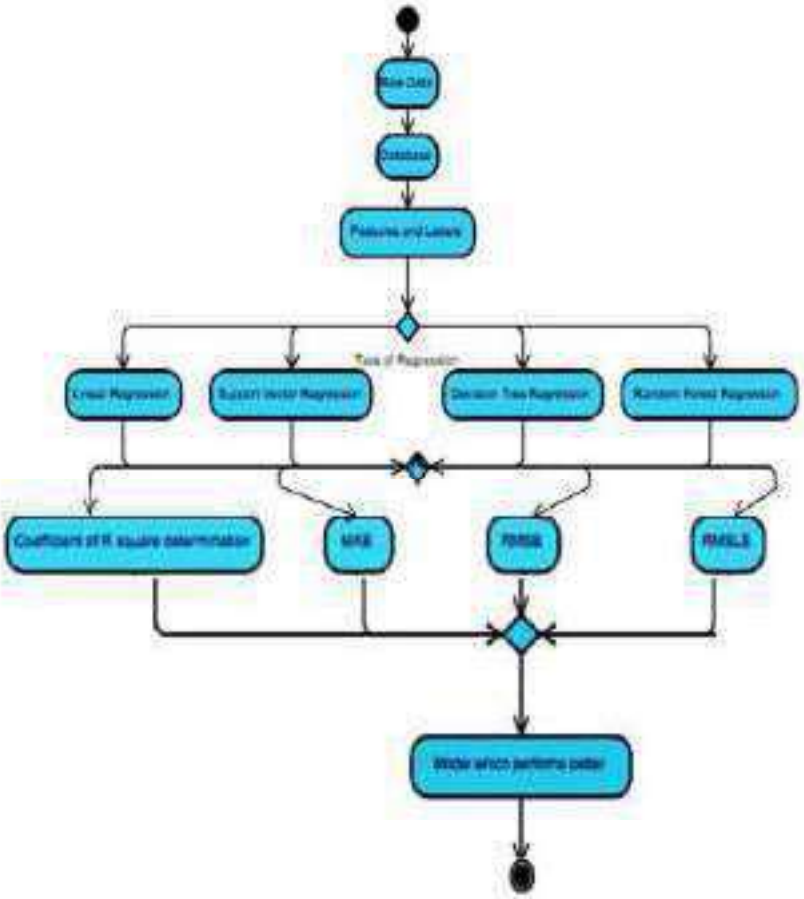


Figure 4.6: Activity Diagram

CHAPTER 5

RESULTS

5.1 Output

Table 5.1: Model performance on the training set

Models	R2	MAE	RMSE	RMSLE
MLR	0.9965	3.2952	5.9334	0.0595
DTR	1.0000	0.0000	0.0000	0.0000
RFR	0.9996	0.7106	2.0237	0.0195
SVR	0.9494	16.076	22.628	0.1423

Table 5.2: Model performance on the testing set

Models	R2	MAE	RMSE	RMSLE
MLR	0.9965	3.4796	5.4973	0.0517
DTR	0.9955	2.354	6.2370	0.0563
RFR	0.9982	1.7016	3.8577	0.0422
SVR	0.9164	19.0722	27.0025	0.1686

5.2 Inference

The outcomes were obtained by contrasting the working of various machine learning algorithms (mentioned earlier) and by judging their performance on various error metrics.

The two tables (Fig 5.1 and 5.2) show the performance of the ML models on both the training and testing set. We see that the decision tree regression performed extremely well on the training set but did not perform that good on testing data. The reason behind this could be that the DTR model was mugging up the values and its results

- Results on training set:

models	R^2	RMSE	MAE	RMSLE
MLR	0.9965	5.9334	3.2952	0.0595
Decision Tree	1.0000	0.0000	0.0000	0.0000
Random Forest	0.9996	2.0237	0.7106	0.0195
SVR	0.9494	22.628	16.076	0.1423

Figure 5.1: Training Data Result

- Results on testing set:

Models	R^2	RMSE	MAE	RMSLE
MLR	0.9965	5.4973	3.4796	0.0517
Decision Tree	0.9955	6.2370	2.354	0.0563
Random Forest	0.9982	3.8577	1.7016	0.0422
SVR	0.9164	27.0025	19.0722	0.1686

Figure 5.2: Testing Data Result

(i.e. overfitting it) from the training set. Thus when shown the same training set it predicted very accurately, but when fed with new data (testing data) it could not perform that well. On the other hand, Random Forest Regression performed well on both the training set and testing set. This can be clarified because of the more robust nature of RFR. As mentioned earlier RFR generates N no. of Decision Trees from K selected data points and the average prediction of these N decision trees is the Random Forest Prediction. Thus while generating a model the bad predictions done by few decision trees (overfitted trees) are out-numbered by the good predicting decision trees due to the average prediction. We also see that MLR performed quite well on both training and testing sets, but not as good as random forest regression. Since the MLR generated a linear model based on training points, it produced less error while predicting the training set, but overall it had the same result on training set and testing set. At last, we see that the support vector regression performed the worst out of all other regression models. This could be due to the complex nature of SVR. Unlike other regression models where the error rate is minimized, SVR tries to fit the error within a certain threshold by fitting the hyper plane that has the maximum no. of data points between the boundary lines.

The figures 5.3 - 5.10 show the training and testing values of MLR, SVR, DTR and RFR.

The figures 5.11 - 5.14 show the comparison of various machine learning models on the training set and the testing set.

The scatter plots (Fig 5.15-5.18) shows the performance of various regressions for predicting the test data. We have taken AQI values on the y-axis and the pollutant concentration on the x-axis. The test data is described by the blue dots and the predicted values are represented by the orange dots.

For the MLR model(Fig 5.15), since the R2 value is 0.9965 we see that most of the predicted values overlap the real values. Thus the model gives a very good fit.

Fig 5.16 represents the performance of the SVR model. We see that the model did not perform well as the predicted points are far off from their actual point (test data). Thus it has a low R2 score and high error values.

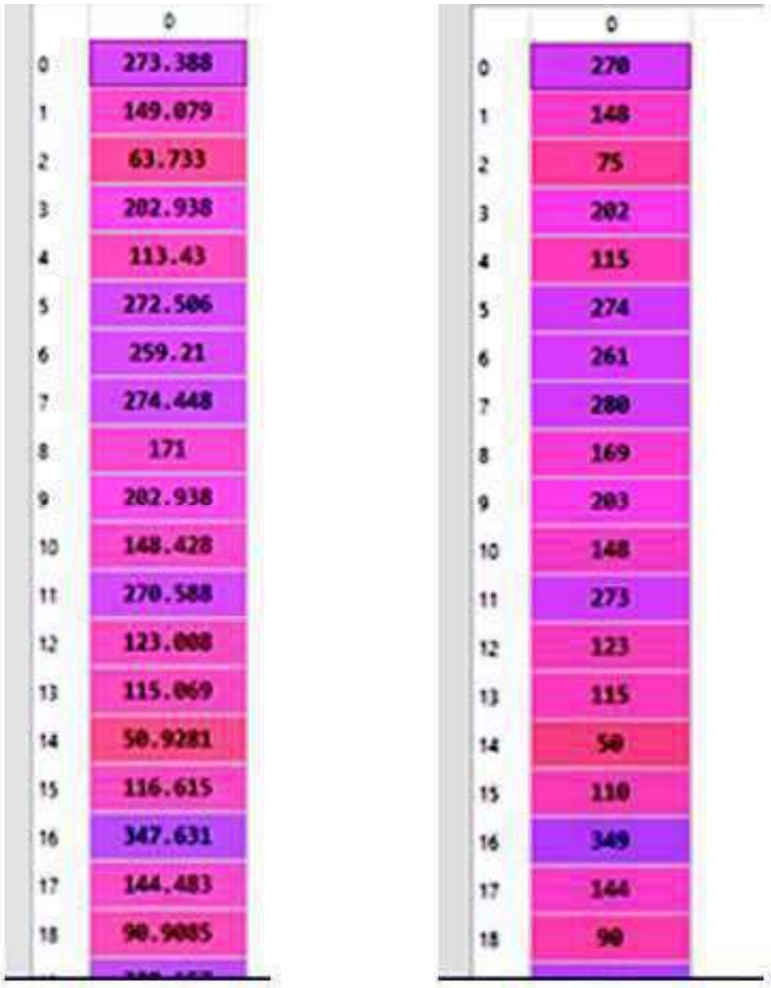
Figures 5.17 and 5.18 show the performance of DTR and RFR respectively. From

	0
0	305.932
1	198.821
2	71.9051
3	73.9967
4	281.457
5	166.24
6	133.229
7	345.046
8	73.8804
9	166.61
10	175.213
11	143.33
12	188.114
13	90.4582
14	244.138
15	179.812
16	269.786
17	335.947
18	122.541

	0
0	306
1	196
2	69
3	71
4	281
5	167
6	148
7	345
8	79
9	169
10	172
11	141
12	189
13	95
14	239
15	177
16	269
17	334
18	123

Figure 5.3: MLR y-test data - Predicted data on the left, Observed data on the right

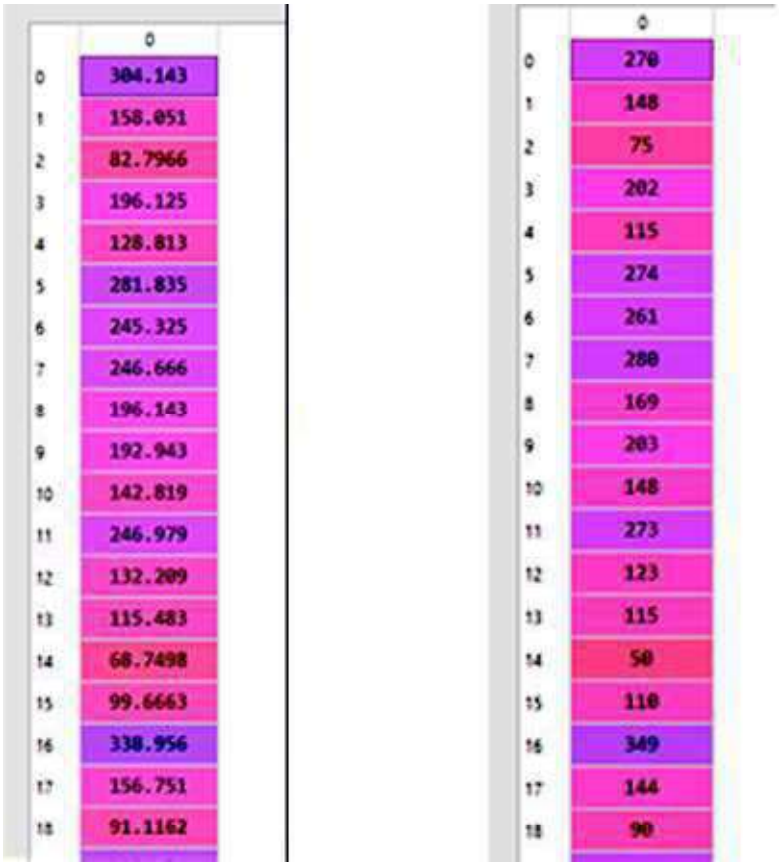
the figures, we can see that both the models have predicted accurately. Almost all the predict points overlaps with the test data points. Thus they have high R2 value and fewer errors.



	0
0	273.388
1	149.079
2	63.733
3	202.938
4	113.43
5	272.506
6	259.21
7	274.448
8	171
9	202.938
10	148.428
11	270.588
12	123.008
13	115.069
14	50.9281
15	116.615
16	347.631
17	144.483
18	90.9085

	0
0	270
1	148
2	75
3	202
4	115
5	274
6	261
7	280
8	169
9	203
10	148
11	273
12	123
13	115
14	50
15	110
16	349
17	144
18	90

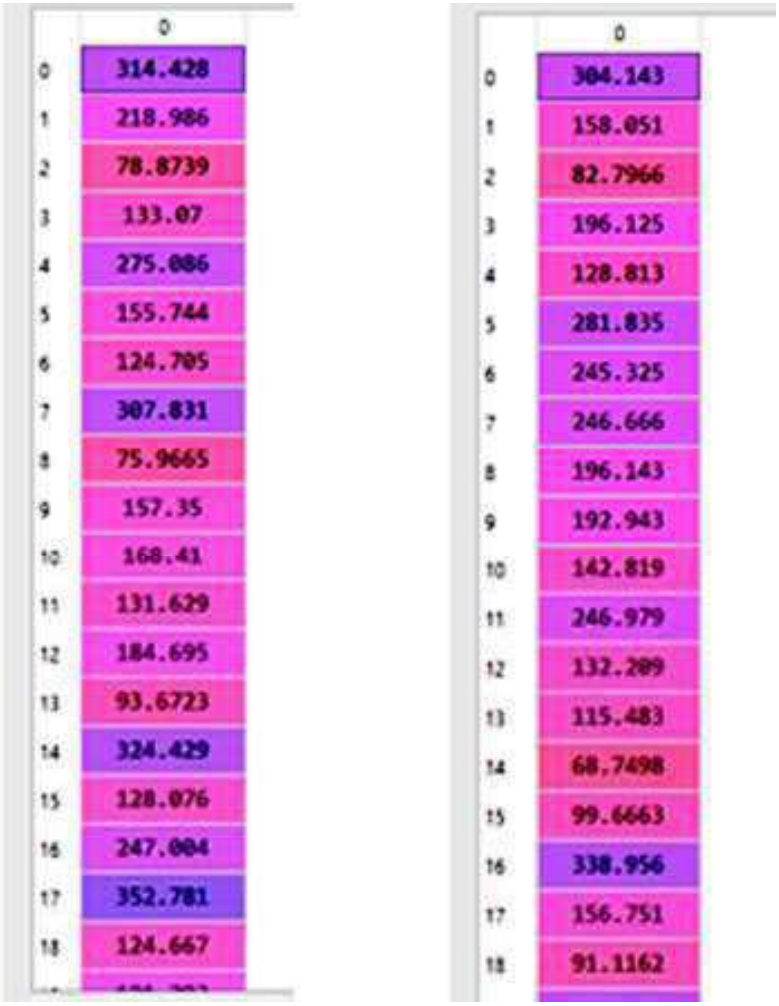
Figure 5.4: MLR y-train data - Predicted data on the left, Observed data on the right



	0
0	304.143
1	158.051
2	82.7966
3	196.125
4	128.813
5	281.835
6	245.325
7	246.666
8	196.143
9	192.943
10	142.819
11	246.979
12	132.209
13	115.483
14	68.7498
15	99.6663
16	338.956
17	156.751
18	91.1162

	0
0	270
1	148
2	75
3	202
4	115
5	274
6	261
7	280
8	169
9	203
10	148
11	273
12	123
13	115
14	50
15	110
16	349
17	144
18	90

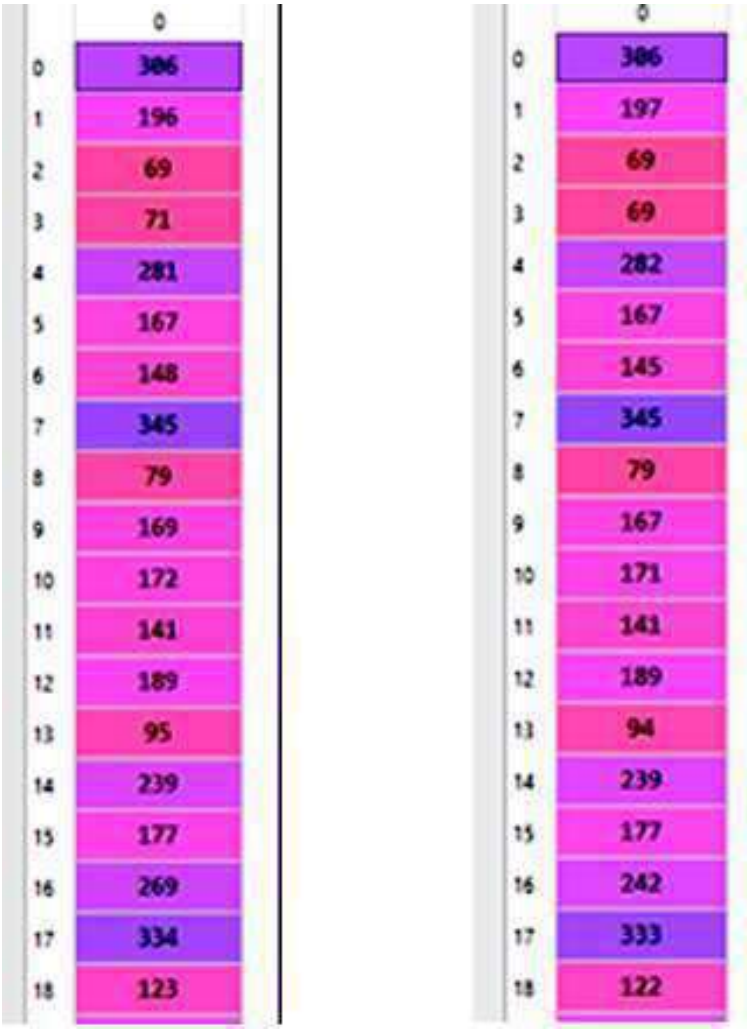
Figure 5.5: SVR y-test data - Predicted data on the left, Observed data on the right



	0
0	314.428
1	218.986
2	78.8739
3	133.07
4	275.086
5	155.744
6	124.705
7	307.831
8	75.9665
9	157.35
10	168.41
11	131.629
12	184.695
13	93.6723
14	324.429
15	128.076
16	247.004
17	352.781
18	124.667

	0
0	304.143
1	158.051
2	82.7966
3	196.125
4	128.813
5	281.835
6	245.325
7	246.666
8	196.143
9	192.943
10	142.819
11	246.979
12	132.289
13	115.483
14	68.7498
15	99.6663
16	338.956
17	156.751
18	91.1162

Figure 5.6: SVR y-train data - Predicted data on the left, Observed data on the right



	0
0	386
1	196
2	69
3	71
4	281
5	167
6	148
7	345
8	79
9	169
10	172
11	141
12	189
13	95
14	239
15	177
16	269
17	334
18	123

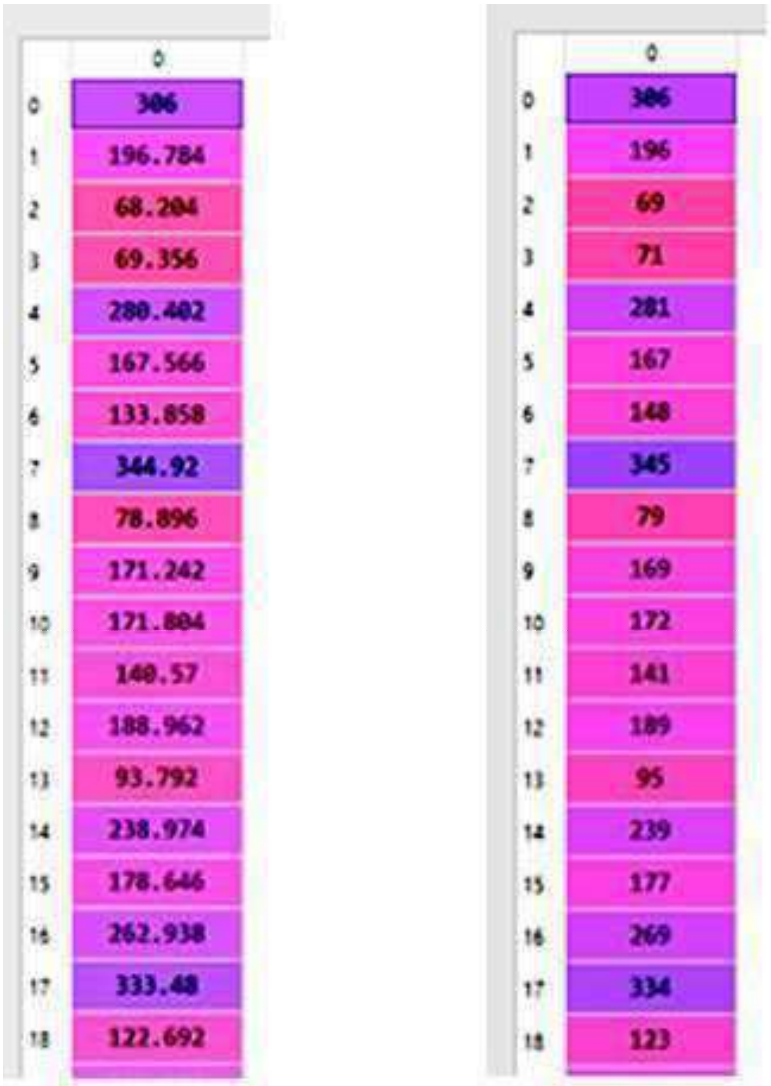
	0
0	386
1	197
2	69
3	69
4	282
5	167
6	145
7	345
8	79
9	167
10	171
11	141
12	189
13	94
14	239
15	177
16	242
17	333
18	122

Figure 5.7: DTR y-test data - Predicted data on the left, Observed data on the right

	0
0	270
1	148
2	75
3	202
4	115
5	274
6	261
7	280
8	169
9	203
10	148
11	273
12	123
13	115
14	50
15	110
16	349
17	144
18	90

	0
0	270
1	148
2	75
3	202
4	115
5	274
6	261
7	280
8	169
9	203
10	148
11	273
12	123
13	115
14	50
15	110
16	349
17	144
18	90

Figure 5.8: DTR y-train data - Predicted data on the left, Observed data on the right



	0
0	306
1	196.784
2	68.284
3	69.356
4	280.402
5	167.566
6	133.858
7	344.92
8	78.896
9	171.242
10	171.804
11	140.57
12	188.962
13	93.792
14	238.974
15	178.646
16	262.938
17	333.48
18	122.692

	0
0	306
1	196
2	69
3	71
4	281
5	167
6	148
7	345
8	79
9	169
10	172
11	141
12	189
13	95
14	239
15	177
16	269
17	334
18	123

Figure 5.9: RFR y-test data - Predicted data on the left, Observed data on the right

	0
0	269.724
1	147.878
2	75.226
3	201.954
4	115.558
5	273.836
6	260.962
7	279.966
8	166.282
9	282.896
10	147.912
11	272.982
12	122.934
13	115.236
14	50.762
15	109.876
16	349.004
17	144.212
18	90.282

	0
0	270
1	148
2	75
3	202
4	115
5	274
6	261
7	280
8	169
9	283
10	148
11	273
12	123
13	115
14	50
15	110
16	349
17	144
18	90

Figure 5.10: RFR y-train data - Predicted data on the left, Observed data on the right

Figure 5.11: Contrast between R2 on various ML Models

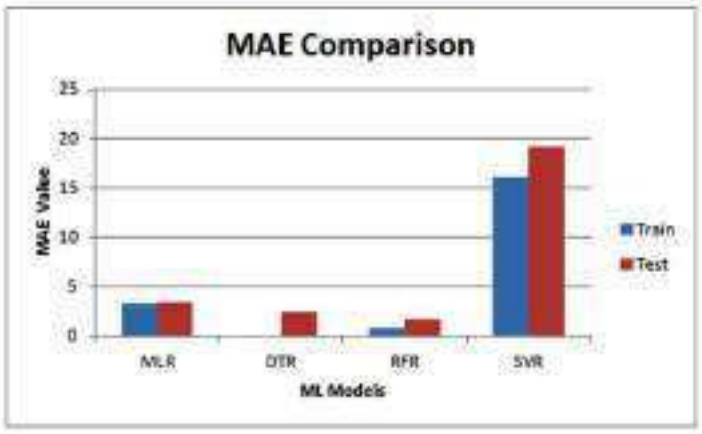


Figure 5.12: Contrast between MAE on various ML Models

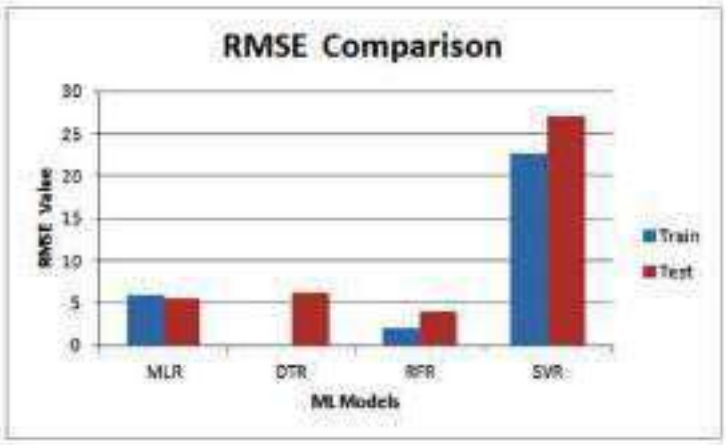


Figure 5.13: Contrast between RMSE on various ML Models

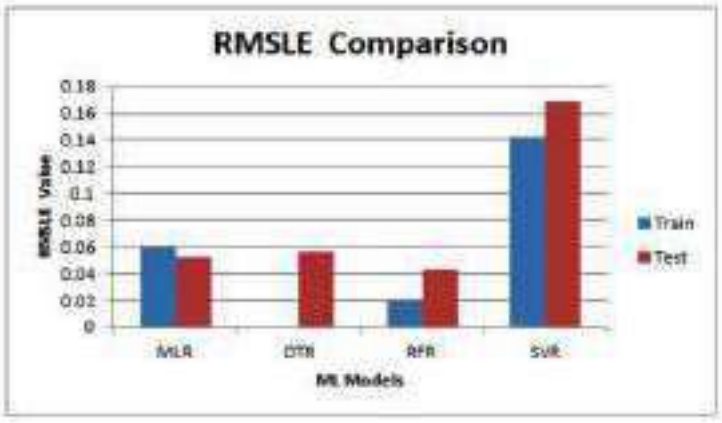


Figure 5.14: Contrast between RMSLE on various ML Models

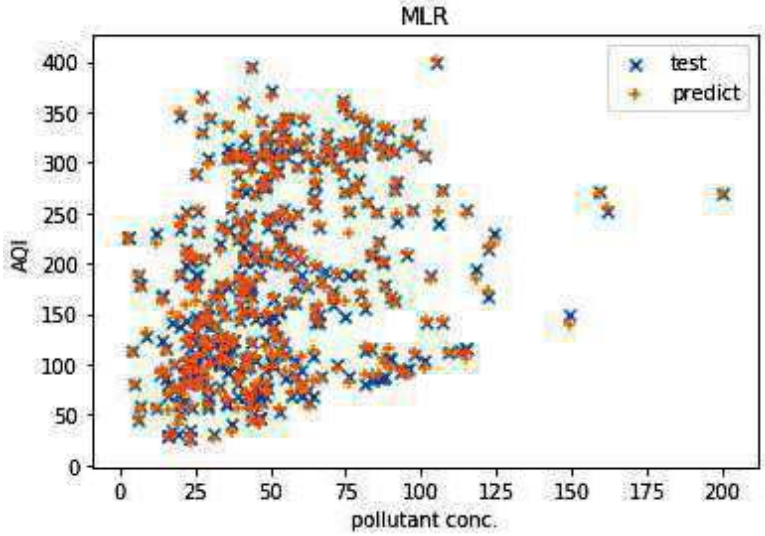


Figure 5.15: Performance of MLR

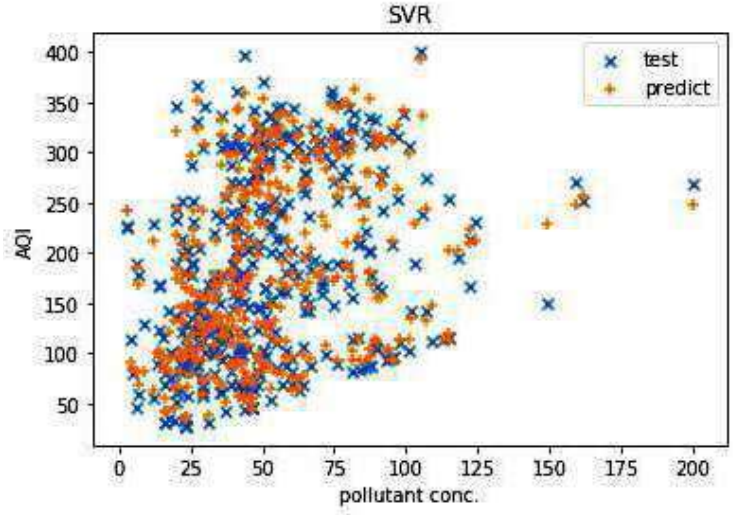


Figure 5.16: Performance of SVR

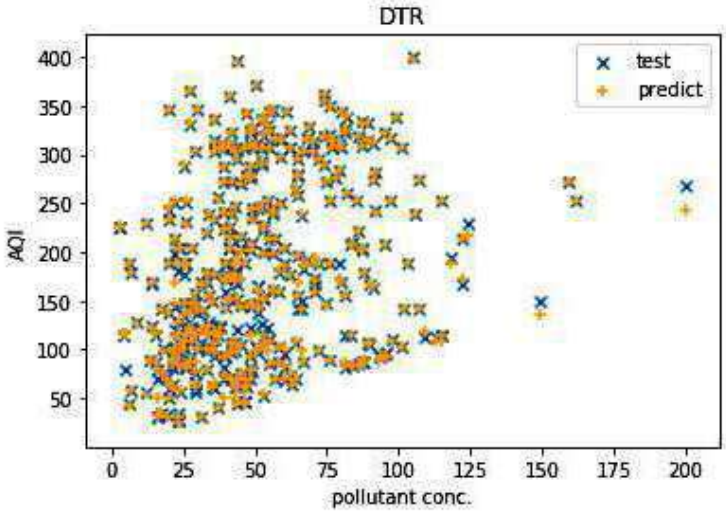


Figure 5.17: Performance of DTR

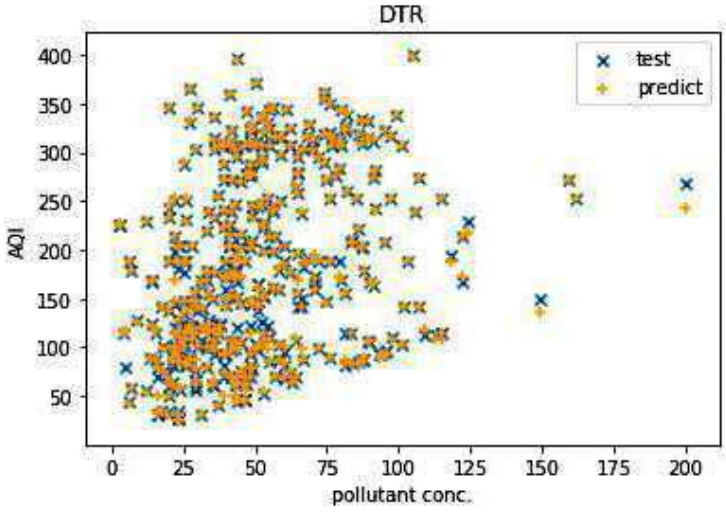


Figure 5.18: Performance of RFR

CHAPTER 6

CONCLUSION

So far we have compared 4 ML models namely MLR, SVR, DTR and RFR and their performance were discussed in details. From the results obtained we can conclude that Random Forest algorithm played out the best in comparison with other models, while SVR performed the poorest. Multiple Linear Regression model performed equally good on both training set and testing set, thus making it the second-best ML model. The performance of the Decision Tree Regression model remained satisfactory.

The following reasons could be the possible explanation for better performance by Random Forest Regression:

• RFR creates countless small decision trees and therefore, it is possible that the number of decision trees providing correct results outdoes the number of decision tree providing incorrect results.

• Random forest algorithm performs better even if the dataset have outliers in them.

• Random forest can generalize data very well and does not require feature scaling.

The goal of this project is to measure the AQI in a particular area by by means of ML methods such as MLR, SVR, DTR and RFR. We have used error metrics such as Coefficient of R2 determination, MAE, RMSE and RMSLE. From the results obtained we can wrap up saying that Random Forest algorithm outdid the other models, and SVR performed the poorest. Multiple Linear Regression model performed equally good on both training set and testing set, thus making it the second-best ML model. The performance of the Decision Tree Regression model was adequate. The future work for this project can consist of using other machine learning models to reduce the dimensionality, such as Principal Component Analysis, Ensemble trees or Variance filters. Another enhancement on or work can be done if we can get the data of all the States and Union Territories of India. Since the Open Government Data Platform of India gives data for only 19 states, currently it is not possible. Another work worth pursuing is the use median or mode for missing-value processing, instead of using mean.

REFERENCES

1. Liu. B, Shi. C, Li. J, Li. Y, Lang. J, Gu. R (2019) "COMPARISON OF DIFFERENT MACHINE LEARNING METHODS TO FORECAST AIR QUALITY INDEX". Frontier Computing. FC 2018. Lecture Notes in Electrical Engineering **542** Springer, Singapore
2. Keller.C.A, Evans.M.J, Kutz.J. N Pawson.S (2017) "MACHINE LEARNING AND AIR QUALITY MODELING". *2017 IEEE International Conference on Big Data (Big Data).* IEEE Publication. Dec 2017. pp. 4570-4576
3. Ganesh.S, Modali.S.H, Palreddy.S.R and Arulmozhivarman.P (2017) "FORECASTING AIR QUALITY INDEX USING REGRESSION MODELS: A CASE STUDY ON DELHI AND HOUSTON". *2017 International Conference on Trends in Electronics and Informatics (ICEI)*. IEEE Publication. May 2017. pp. 248-254
4. Soh.P, Chang.J and Huang.J (2017) "ADAPTIVE DEEP LEARNING-BASED AIR QUALITY PREDICTION MODEL USING THE MOST RELEVANT SPATIAL-TEMPORAL RELATIONS". IEEE Access **6.** pp. 38186-38199
5. Ghaemi. Z, Alimohammadi. A Farnaghi. M (2018) "LASVM-BASED BIG DATA LEARNING SYSTEM FOR DYNAMICPREDICTION OF AIR POLLUTION IN TEHRAN". Environmental Monitoring and Assessment 190. Springer International Publishing
6. Zhang.Y,Wang.Y, Gao.M, Ma.Q, Zhao.J, Zhang.R, Wang.Q and Zhang.L.H (2018) "A PREDICTIVE DATA FEATURE EXPLORATION-BASED AIR QUALITY PREDICTION APPROACH". IEEE Access **7**. pp. 30732-30743
7. Amado.T.M, Dela Cruz.J.C (2018) "DEVELOPMENT OF MACHINE LEARNING-BASED PREDICTIVE MODELS FOR AIR QUALITY MONITORING AND CHARACTERIZATION". *TENCON 2018 - 2018 IEEE Region 10 Conference*. IEEE Publication. Oct 2018. pp. 0668-0672
8. Zheng.Y, Yi.X, Li.M, Li.R, Shan.Z, Chang.E, Li.T (2015) "FORECASTING FINE-GRAINED AIR QUALITY BASED ON BIG DATA". *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Aug 2015. pp. 2267-2276
9. Hable-Khandeka.V and Srinath (2017) "MACHINE LEARNING TECHNIQUES FOR AIR QUALITY FORECASTING AND STUDY ON REAL-TIME AIR QUALITY MONITORING". *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. IEEE Publication. Aug 2017. pp. 1-6.
10. De Vito.S, Massera.E, Piga.M, Martinotto.L, Di Francia.G (2007) "ON FIELD CALIBRATION OF AN ELECTRONIC NOSE FOR BENZENE ESTIMATION IN AN URBAN POLLUTION MONITORING SCENARIO". Sensors and Actuators B: Chemical **129** (2) .pp.750-757
11. Sohn.S.H, Oh.S.C, Jo.B.W and Yeo.Y.K (2000) "PREDICTION OF OZONE FORMATION BASED ON NEURAL NETWORK". Journal of Environmental Engineering ASCE **126** (8)

12. Singh.K.P, Gupta.S, Rai.P (2013) "IDENTIFYING POLLUTION SOURCES AND PREDICTING URBAN AIR QUALITY USING ENSEMBLE LEARNING METHODS". Atmospheric Environment ASCE **80**. pp. 426-437
13. Dragomir.E.G (2010) "AIR QUALITY INDEX PREDICTION USING K-NEAREST NEIGHBOR TECHNIQUE". Bulletin of PG University of Ploiesti, Series Mathematics, Informatics, Physics **LXII** (1). pp.103-108
14. Chaloulakou.A, Saisana.M, and Spyrellis.N (2003) "COMPARATIVE ASSESSMENT OF NEURAL NETWORKS AND REGRESSION MODELS FOR FORECASTING SUMMER TIME OZONE IN ATHENS". Science of the Total Environment **313** (1-3). pp. 1-13
15. Vlachogianni.A, Kassomenos.P, Karppinen.A, Karakitsios.S, and Kukkonen.J (2011) "EVALUATION OF A MULTIPLE REGRESSION MODEL FOR THE FORECASTING OF THE CONCENTRATIONS OF NOX AND PM10 IN ATHENS AND HELSINKI". Science of The Total Environment **409** (8). pp. 1559-1571
16. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/ *WHO* (2014)
17. Breiman.L (2001) *"RANDOM FORESTS"*. Machine Learning. Springer Publication. **45**. pp. 5-32
18. Munoz.E, Come.Mj.J, Martin.M.L and Trujilo.F.J (2013) "PREDICTION OF PM10 AND SO2 EXCEEDANCES TO CONTROL AIR POLLUTION IN THE BAY OF ALGECIRAS, SPAIN". Stochastic Environmental Research and Risk Assessment **28**. pp.1409-1420
19. Kurt.A and Oktay.A.B (2010). "FORECASTING AIR POLLUTANT INDICATOR LEVELS WITH GEOGRAPHIC MODELS 3 DAYS IN ADVANCE USING NEURAL NETWORKS". Expert Systems with Applicatons **37** (12) pp. 7986-7992
20. Kumar.A and Goyal.P (2011) "FORECASTING OF DAILY AIR QUALITY INDEX IN DELHI". Science of The Total Environment **409** (24). pp. 5517-5523
21. Bellander.T, Berglind.N, Gustavsson.P, Jonson.T, Nyberg.F, Pershagen.G (2001) "USING GEOGRAPHIC INFORMATION SYSTEMS TO ASSESS INDIVIDUAL HISTORICAL EXPOSURE TO AIR POLLUTION FROM TRAFFIC AND HOUSE HEATING IN STOCKHOLM". Environmental Health Perspectives. **109** (6). pp. 633–639.
22. Gao.M, Yin.L, and Ning.J (2018) "ARTICIAL NEURAL NETWORK MODEL FOR OZONE CONCENTRATION ESTIMATION AND MONTE CARLO ANALYSIS". Atmospheric Environment **184.** pp. 129-139
23. RoyS.S, Pratyush.C, and Barna.C (2016) "PREDICTING OZONE LAYER CONCENTRATION USING MULTIVARIATE ADAPTIVE REGRESSION SPLINES, RANDOM FOREST AND CLASSIFICATION AND REGRESSION TREE". *Proceedings of the 5th International Workshop Soft Computing Applications*. **634**. Springer, Cham Publications. pp. 140-152
24. Cotter.A, Shamir.O, Srebro.N, Sridharan.K (2011) "BETTER MINI-BATCH ALGORITHMS VIA ACCELERATED GRADIENT METHODS". *Conference on Advances in Neural Information Processing Systems 24*. NIPS 2011 Publication. pp. 1647–1655

APPENDIX A

SAMPLE CODE FOR CONVERSION OF XML TO CSV

```
import xml.etree.ElementTree as et
import csv
tree = et.parse("weather data xml
data_aqi_cpcb(14).xml")
root = tree.getroot()

    arr = []

    pm2_array = []
pm10_array = []
no2_array = []
nh3_array = []
so2_array = []
co_array = []
o3_array = []
aqi_val_array = []
predominant_para_array = []
date_array = []
time_array = []
state_array = []
city_array = []
station_array = []

    for country in root.findall("Country"):
for state in country.findall("State"):
```

dt = station.get("lastupdate")

dt = dt.split()

date$_array.append(dt[0])$

$time_array.append(dt[1])$

$pol_arr = [”PM2.5”, ”PM10”, ”NO2”, ”NH3”, ”SO2”, ”CO”, ”OZONE”]$

$aqi_presen t = ”yes”$

$forpindexinstation.findall(”Pollutant_Index”) :$

$print(pindex.get('Avg'))$

$if(pindex.get('id') == ”PM2.5”) :$

$pm2_array.append(pindex.get('Avg'))$

$pol_arr.remove('PM2.5')$

$elif(pindex.get('id') == ”CO”) :$

$co_array.append(pindex.get('Avg'))$

$pol_arr.remove(”CO”)$

if(len(pol$_arr)! = 0) :$

$while(len(pol_arr)! = 0) :$

$ele = pol_arr.pop(0)$

$if(ele == ”PM2.5”) :$

if(station.find('Air$_Quality_Index')isnotNone) :$

$aqi_val_array.append(station.find('Air_QualityIndex').get('Value'))$

$predominant_para_array.append(station.find('Air_QualityIndex').get$

$('Predominant_Parameter'))$

$else :$

$aqi_val_array.append('NA')$

$predominant_para_array.append('NA')$

data$_row = []$

$fora, b, c, d, e, i, j, k, l, m, n, p, q, rin$

$zip(state_array,$

$city_array, station_array, date_array, time_array, pm2_array,$
$pm10_array, no2_array, nh3_array, so2_array, co_array, o3_array, aqi_val_array,$
$predominant_para_array):$
$data_row.append([a, b, c, d, e, i, j, k, l, m, n, p, q, r])$

data$_r ow.insert(0, [$"$state$", "$city$", "$station$", "$date$", "$time$", "$PM2.5$", "$PM10$",
"$NO2$", "$NH3$", "$SO2$", "CO", "$OZONE$", "AQI", "$Predominant_Parameter$"])
$foriindata_row:$
$print(i)$
$print("********")$

with open("state$_w eather_a qi_d ata_t wo.csv", 'a', newline = '')as file:$
$writer = csv.writer(file)$
$writer.writerows(data_row)$
$print("filewritten")$

APPENDIX B

SAMPLE CODE FOR MACHINE LEARNING METHODS

import numpy as np

import pandas as pd

df = pd.read$_csv(''dataset$

$state_weather_aqi_data_mf2.csv'')$

x1 = df.iloc[:,:12].values

z1 = pd.DataFrame(x1)

y1 = df.iloc[:,12:13].values

z2 = pd.DataFrame(y1)

from sklearn.preprocessing import OneHotEncoder

ohe = OneHotEncoder()

x$_new1 = pd.DataFrame(ohe.fit_transform(x1[:, [0]]).toarray())state$

$x_new2 = pd.DataFrame(ohe.fit_transform(x1[:, [1]]).toarray())city$

$x_new3 = pd.DataFrame(ohe.fit_transform(x1[:, [2]]).toarray())station$

$feature_set = pd.concat([x_new1, x_new2, x_new3, pd.DataFrame(x1[:, 5 : 12])],$

$axis = 1, sort = False)$

importing ml libraries

from sklearn.model$_selectionimporttrain_test_split$

$fromsklearn.linear_modelimportLinearRegression$

$fromsklearn.preprocessingimportPolynomialFeatures$

$fromsklearn.treeimportDecisionTreeRegressor$

$fromsklearn.ensembleimportRandomForestRegressor$

$fromsklearn.svmimportSVR$

x$_train, x_test, y_train, y_test = train_test_split(feature_set, y1, test_size = 0.25, random_state = 0)$

———————————————————

—- test data prediction ————————

———————————————————

multiple linear regression model

mreg = LinearRegression()

mreg.fit(x$_train, y_train)$

mlr$_{yp}redict = mreg.predict(x_test)$

———————————————————

decision tree regression model

dec$_tree = DecisionTreeRegressor(random_state = 0)$

$dec_tree.fit(x_train, y_train)$

dt$_{yp}redict = dec_tree.predict(x_test)$

———————————————————

random forest regression model

random forest with 500 trees

$\text{rt}_reg = RandomForestRegressor(n_estimators = 500, random_state = 0)$

$rt_reg.fit(x_train, y_train)$

$\text{rt}_ypredict = rt_reg.predict(x_test)$

---

support vector regression model

— feature scaling the paramenters for better results —

from sklearn.preprocessing import StandardScaler

$\text{sc}_x = StandardScaler()$

$sc_y = StandardScaler()$

$x_train_svr = sc_x.fit_transform(x_train)$

$y_train_svr = sc_y.fit_transform(y_train)$

$\text{svr}_reg = SVR()$

$svr_reg.fit(x_train_svr, y_train_svr)$

$\text{svr}_ypredict = sc_y.inverse_transform(svr_reg.predict(sc_x.$
$transform(x_test)))$

—- training data prediction ——

—- MLR ——

$\text{mlr}_ytp_rmse = sqrt(metrics.mean_squared_error(y_train, mreg.predict$
$(x_train)))$

$mlr_ytp_mae = metrics.mean_absolute_error(y_train, mreg.predict$
$(x_train))$

$mlr_ytp_r2 = metrics.r2_score(y_train, mreg.predict(x_train))$

$m1 = mreg.predict(x_train)$

$mlr_ytp_rmsle = rmsle(y_train, m1)$

—— decision tree reg ——

$\text{dt}_ytp_rmse = sqrt(metrics.mean_squared_error(y_train, dec_tree.predict(x_train)))$

$dt_ytp_mae = metrics.mean_absolute_error(y_train, dec_tree.predict(x_train))$

$dt_ytp_r2 = metrics.r2_score(y_train, dec_tree.predict(x_train))$

$dt_ytp_rmsle = rmsle(y_train, dec_tree.predict(x_train))$

—— random forest reg ——

$\text{rf}_ytp_rmse = sqrt(metrics.mean_squared_error(y_train, rt_reg.predict(x_train)))$

$rf_ytp_mae = metrics.mean_absolute_error(y_train, rt_reg.predict(x_train))$

$rf_ytp_r2 = metrics.r2_score(y_train, rt_reg.predict(x_train))$

$rf_ytp_rmsle = rmsle(y_train, rt_reg.predict(x_train))$

—— svr ——

$\text{svr}_ytp_rmse = sqrt(metrics.mean_squared_error(y_train, sc_y.inverse_transform(svr_reg.predict(sc_x.transform(x_train)))))$

$svr_ytp_mae = metrics.mean_absolute_error(y_train, sc_y.inverse_transform(svr_reg.predict(sc_x.transform(x_train))))$

$svr_ytp_r2 = metrics.r2_score(y_train, sc_y.inverse_transform(svr_reg.predict(sc_x.transform(x_train))))$

$svr_ytp_rmsle = rmsle(y_train, sc_y.inverse_transform(svr_reg.predict(sc_x.transform(x_train))))$

==========

========== RESULT ==========

==========

print("evaluating on training data:")

print("models^{2}")

print("MLR0$\widehat{}$.4f1$\widehat{}$.4f2$\widehat{}$.4f3$\widehat{}$.4f".format(mlr$_{y}tp_{r}2, mlr_{y}tp_{r}mse, mlr_{y}tp_{m}ae, mlr_{y}tp_{r}msle$))

print("DTR0$\widehat{}$.4f1$\widehat{}$.4f2$\widehat{}$.4f3$\widehat{}$.4f".format(dt$_{y}tp_{r}2, dt_{y}tp_{r}mse, dt_{y}tp_{m}ae, dt_{y}tp_{r}msle$))

print("evaluating on testing data:")

print("models^{2}")

print("RFR0$\widehat{}$.4f1$\widehat{}$.4f2$\widehat{}$.4f3$\widehat{}$.4f".format(r2$_{r}t, rmse_{r}t, mae_{r}t, rmsle_{r}t$))

print("SVR0$\widehat{}$.4f1$\widehat{}$.4f2$\widehat{}$.4f3$\widehat{}$.4f".format(r2$_{s}vr, rmse_{s}vr, mae_{s}vr, rmsle_{s}vr$))

Format - I

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University u/s 3 of UGC Act, 1956)

Office of Controller of Examinations

REPORT FOR PLAGIARISM CHECK ON THE DISSERTATION/PROJECT REPORTS FOR UG/PG PROGRAMMES

(To be attached in the dissertation/ project report)

1	Name of the Candidate **(IN BLOCK LETTERS)**	
2	Address of the Candidate	**Mobile Number :**
3	Registration Number	
4	Date of Birth	
5	Department	
6	Faculty	
7	Title of the Dissertation/Project	
8	Whether the above project/dissertation is done by	Individual or group : (Strike whichever is not applicable) a) If the project/ dissertation is done in group, then how many students together completed the project : b) Mention the Name & Register number of other candidates :
9	Name and address of the Supervisor / Guide	**Mail ID : Mobile Number :**
10	Name and address of the Co-Supervisor / Co- Guide (if any)	**Mail ID : Mobile Number :**

11	Software Used			
12	Date of Verification			
13	**Plagiarism Details: (to attach the final report from the software)**			
Chapter	**Title of the Chapter**	**Percentage of similarity index (including self citation)**	**Percentage of similarity index (Excluding self citation)**	**% of plagiarism after excluding Quotes, Bibliography, etc.,**
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
Appendices				

I / We declare that the above information have been verified and found true to the best of my / our knowledge.

Signature of the Candidate	**Name & Signature of the Staff (Who uses the plagiarism check software)**
Name & Signature of the Supervisor/Guide	**Name & Signature of the Co-Supervisor/Co-Guide**
Name & Signature of the HOD	