JADAVPUR UNIVERSITY

MCA PROJECT REPORT

Predicting Missing Value from Multi-Site Time Series Atmospheric Ozone Data

A project report submitted in partial fulfilment of the requirements
for the degree of Master of Computer Application

in

Department of Computer Science and Engineering
Jadavpur University

By

Rajesh Sarkar
University Roll No.: 001910503038
University Registration No.: 149898 of 2019-20
Exam Roll No.: MCA226037

Under the Guidance of

Dr. Sarbani Roy
Professor
Department of Computer Science and Engineering
Faculty of Engineering and Technology,
Jadavpur University, Kolkata
June 2022

Declaration of Authorship

I, Rajesh Sarkar, declare that this project titled, “Predicting Missing Value from Multi-Site Time Series Atmospheric Ozone Data” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for degree at this University.
- Where any part of this project has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the project is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date:

To whom it may concern

This is to certify that the work on this project entitled “Predicting Missing Value from Multi-Site Time Series Atmospheric Ozone Data” has been satisfactorily completed by Rajesh Sarkar, Roll No: 001910503038, University Registration No.: 149898 of 2019-20. It is a bona-fide piece of work carried out under my supervision at Jadavpur University, Kolkata-700032, for partial fulfilment of the requirements for the degree of Master of Computer Application from the Department of Computer Science and Engineering, Jadavpur University for the academic session of 2019-2022.

_____

Prof. Sarbani Roy
(Supervisor)
Department of Computer Science and Engineering
Jadavpur University

_____

Prof. Anupam Sinha
Head of The Department
Department of Computer Science & Engineering
Jadavpur University

_____

Prof. Chandan Majumdar
Dean
Faculty of Engineering & Technology
Jadavpur University

Certificate of Approval

(Only in case the project is approved)

This is to certify that the project entitled "Predicting Missing Value from Multi-Site Time Series Atmospheric Ozone Data" is a bona-fide record of work carried out by Rajesh Sarkar in partial fulfilment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University during the period of January, 2022 to June, 2022. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Examiners:

_____
(Signature of Examiner)
Date:

_____
(Signature of Examiner)
Date:

Jadavpur University

Abstract

Faculty of Engineering and Technology, Jadavpur University
Department of Computer Science and Engineering

Master of Computer Application

By Rajesh Sarkar

Air quality of this days is a vital issue for human health. But not only for human but also for the earth. Increasing of harmful pollutants in air plays a major for global warming. Increase in daily life temperature causes the significant increase of sea level. Based on IPCC's 2021 report, city of India, Kolkata will be in underwater by 2030 unless a drastic change happens in climate change. Breathing in polluted air for years can cause of deadly disease, cancer. Many researchers actively started working on the air pollutant prediction model. Many of them worked on neural network-based models to predict the concentration of air pollutants. In this study, we will predict ozone concentration of a particular air quality monitoring station. We worked with five machine learning (Elastic Net, Decision Tree, Random Forest, SVR, LightGBM) and one deep learning (ANN) model and implementing five different feature selection methodologies (Radius based, Triangulation, Mutual Information, K-L divergence, Cluster based). We have used MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) to analyse our model's performance. LightGBM has given best result compared to all air pollutant prediction models.

Keywords: *missing value prediction, predictive analysis, ozone, feature selection*

Acknowledgements

On the submission of “Predicting Missing Atmospheric Ozone from Multi-Site Time Series Data”, I wish to express gratitude to the Department of Computer Science and Engineering for sanctioning a project work under Jadavpur University under which this work has been completed. I would like to express my sincere gratitude to my respected guide Dr. Sarbani Roy, Professor, Department of Computer Science and Engineering, Jadavpur University for her unfailing guidance, prolific encouragement, constructive suggestions and continuous involvement during each and every phase of this project. I feel deeply honoured that I got the opportunity to work under her guidance. I would like to express my heartfelt gratitude to Asif Iqbal Middya, Research Fellow, Jadavpur University, Kolkata, for his suggestions and unwavering support. I would also wish to thank Prof. Anupam Sinha, Head of Department of Computer Science and Engineering, Jadavpur University, Prof. Chandan Majumdar, Dean of Faculty of Engineering and Technology, Jadavpur University for providing me all the facilities and for their support to the activities of this project. I would like to express my gratitude and indebtedness to my parents and all my family members for their unbreakable belief, constant encouragement, moral support and guidance. Last, but not the least, I would like to thank all my classmates of Master of Computer Application batch of 2019-2022, for their co-operation and support. Their wealth of experience has been a source of strength for me throughout the duration of my work.

Regards,
Rajesh Sarkar
University Roll Number: 001910503038
University Registration No.: 149898 of 2019-20
Department of Computer Science and Engineering
Jadavpur University

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
MAE	**M**ean **A**bsolute **E**rror
RMSE	**R**oot **M**ean **S**core **E**rror
O_3	**O**zone
SVR	**S**upport **V**ector **R**egressor
GOSS	**G**radient-Based **O**ne-**S**ide **S**ampling
EFB	**E**xclusive **F**eature **B**undling
MI	**M**utual **I**nformation

In Dedication to my family for supporting me all the way!

Chapter 1

Introduction

An overview of this study is given in this chapter followed by the motivation, contribution and organization.

1.1 Overview

Air quality is one of the major issues of this century. Presence of harmful components in air is increasing day by day. Growing urbanization, demographical imbalance, high temperature difference, high emission of toxic gases from factories and vehicles are some the major reasons of polluting the air quality. Ozone (O_3), Particle matter ($PM_{2.5}$), Carbon monoxide (CO) etc. having high presence in air. This air pollutants are highly harmful for human health. In a response of this problem Air quality monitoring stations are set up in highly congested area to monitor the concentration of pollutants present in air. But regardless to say these stations are also faced some unavoidable circumstances such as technical glitch in machine, low maintenance etc. that prevents them to monitoring the air quality. To encounter this problem, prediction the concentration of air pollutant using various machine learning model is an interesting idea. Many studies have been done with different demographic region's data. Middya *et al.* [1] worked on deep learning and statistical model with six different air pollutants with a high accuracy rate. This study works on particularly on one air pollutant, Ozone. In this study, we implement five machine learning (Elastic Net, Decision Tree, Random Forest, SVR, LightGBM) and one deep learning model (ANN) using data of a highly congested area. It predicts the density of atmospheric ozone of a particular air monitoring station using remaining station's data which helps to aware of the air condition if the station is not working. We have used five feature selection methods and RMSE, MAE as accuracy metrics.

1.2 Motivation

Air quality of modern metro cities (e.g., Delhi) in India is falling day by day and it doesn't get the attention it deserves. Poor air quality can cause of some incurable disease like cancer. Today, when machine learning, artificial intelligence is used in almost every aspect of our daily

life to make it comfortable, then a study about pollutant prediction is a really important job. Many studies regarding this topic have been done in past. They mostly focus on some of the deep learning prediction models. Here, we have tried to focus on one pollutant specific dataset with a variety of machine learning models.

1.3 Contribution

The data is used in this study is collected and merged from the twenty-nine different air quality monitoring stations of Delhi. Here, we work on one pollutant specific data, Ozone. Our study is divided mainly in three parts- 1. Hypertunning 2. Feature selection, 3. Prediction model implementation.

Significant contributions of our work are-

- One pollutant specific optimal prediction model is identified by researching through various phases.
- Having a large amount of real-world data is advantageous when working in a real-world problem.
- Prediction model's output is analysed with decent amount of test size data which is more than 1800+ datapoints.
- Exploring some data analysis methodology helped to understand the underlying pattern of the dataset.

1.4 Organization

The following is how the rest of the work is organized, Related work with respect to our study and popular time series prediction models which are used in our study is discussed in chapter 2. Chapter 3 is starting with our problem statement and following with an overall workflow of our study process, data collection and a detail discussion on our approach towards missing value prediction. Describing experimental setup and a comparative analysis of our final output among all the prediction models is done in chapter 3. Chapter 5 consists of conclusion and future work.

Chapter 2

Preliminaries

Many research paper has been published regarding missing value prediction data in past. They used several types of variables e.g., meteorological, environmental for their work. Some of them used multiple pollutant dataset for air quality prediction. More number of air quality monitoring stations are setup in the congested area to make people aware about the air quality. In this chapter, we discussed about some of the previous work related to our topic along with the methods which are used to predict missing value.

2.1 Related Work

A number of studies have been done in order to estimate air quality in terms of ozone. Loya *et al.* [2] worked on atmospheric ozone time series data (hour by hour) of Mexico City between 2010-11 where they considered four chemical variables and four meteorological variables and they achieved a good accuracy. In the study of Kuwait's lower atmospheric ozone, Al-Alawi *et al.* [3] used seven environmental pollutants and five meteorological variables to develop their machine learning model in 2008 with a R2 score of 0.986. Aljanabi *et al.* [4] took an initiative to work on the ground-level ozone of Amman, Jordan in 2020 where they got a great improvement in result by applying various smoothing techniques e.g. Holt-Winters smoothing. Cardelino *et al.* [5] predicted the peak ozone concentration of the next day with 84% accuracy in a study in Atlanta metropolitan data. Barrero *et al.* [6] analysed the correlations between variables and O3 to predict ozone concentration of the next day. A deep learning study by Eslami *et al.* [7] on a dataset (2014-2016) showed some of the limitations of CNN model for prediction air quality. AlOmar *et al.* [8] worked on the W-ANN model and compared it with the classical ANN model where the W-ANN model produced more accuracy with lower errors. Neural network model (ANN) gives better accuracy compared with regressive models and ARIMA (auto regressive integrated moving average) is shown by Prybutok *et al.* [9]

Based on the three factors (nitrogen-di-oxide, temperature, relative humidity), Faris *et al.* [10] compared between two methodology of ANN- 1. MLP, 2. RBF. This study concludes MLP had the better accuracy compared to other. Another study by Kumar *et al.* [11] proved in his

study that MLP is one of the best prediction models for ozone (O_3) comparing with RBF and GRNN. This study result is based on the correlation between the observed and predicted value. Sheta *et al.* [12] used cycle reservoir with jumps (CRJ), a type of RNN to forecast ozone concentration in east of Croatia using factors like wind speed, relative humidity, wind direction, PM_{10}, temperature. Author has shown CRJ outperformed MLP and RBF network models.

Above studies shows that variety of research work have been done in the field of atmospheric ozone concentration in air. Multiple deep learning models like ANN, CNN, RNN and various forms this model has been used to analyse time series data. We also observed that different types of ANN model outperform most of the machine learning models for time series air pollutant dataset. These studies are heavily driven by meteorological and environmental pollutant data.

One pollutant-based dataset is used in our study using five types of times series analysis and six types of prediction models. Feature selection methodologies used in this study differentiate our work from the above-mentioned ones. Here, we have used radius-based approach, triangulation, mutual information, K-L divergence and cluster-based approach.

2.2 Methods

2.2.1 Elastic Net

Elastic Net is a regularization method. It uses the penalty of lasso and ridge regression which makes it more efficient. From a highly correlated group Lasso choose one variable from there whereas Elastic net tend to include ‘n’ number of variables till saturation. Elastic net incorporates a quadratic expression in the penalty to overcome the drawbacks of Lasso regression. This expression becomes ridge when used alone. This method works on two stages, first it determines the coefficient through ridge regression using square penalty and then scales down the coefficients through lasso regression using linear penalty.

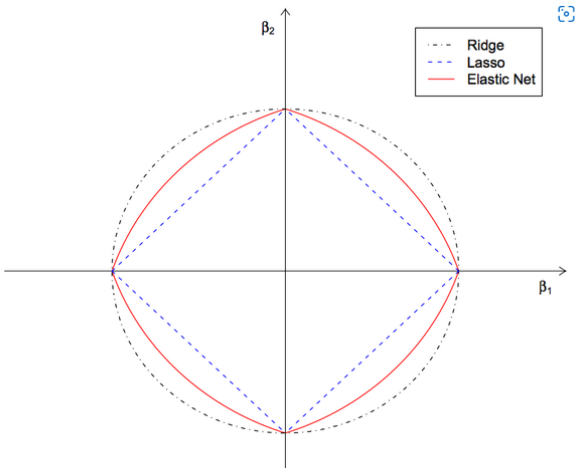

Fig 2.1: Penalty terms in the space of model parameters

2.2.2 Decision Tree

As the name suggests, this model works on a tree-like structure. Decision Tree is used for classification and regression, both the models. The dataset is broken down into smaller subsets and a tree is constructed simultaneously. Decision tree works on both numerical and categorical data. One node has one or more child nodes. Decision Tree regressor uses 'Standard Deviation Reduction'(SDR) to construct the tree. The attribute with highest SDR choose as root of the tree. Formula of SDR is-

$$\mathrm{SDR} = \mathrm{sd(T)} - \sum_i \frac{|T_i|}{|T|} \times \mathrm{sd(Ti)}$$

We need to calculate SDR for each attribute. Attributes with bigger SDR values are more likely to be chosen as a splitting attribute first in the Decision tree construction. Decision tree is also used as a classification method where the splitting criteria is Information Gain of the attributes.

2.2.3 Random Forest

Random Forest is a type of supervised machine learning model which uses ensemble learning methodology. Ensemble learning used the outputs from multiple machine learning models to forecast the final result. Ensemble learning can be of two types- 1. Boosting 2. Bagging. Random Forest is based on Bagging method. Bagging method use Bootstrap sampling technique. Bootstrap sampling is where the sample is selected randomly from dataset. For Fig 2.2, 1000 datasets are made with possibility of one datapoint can be present in multiple datasets.

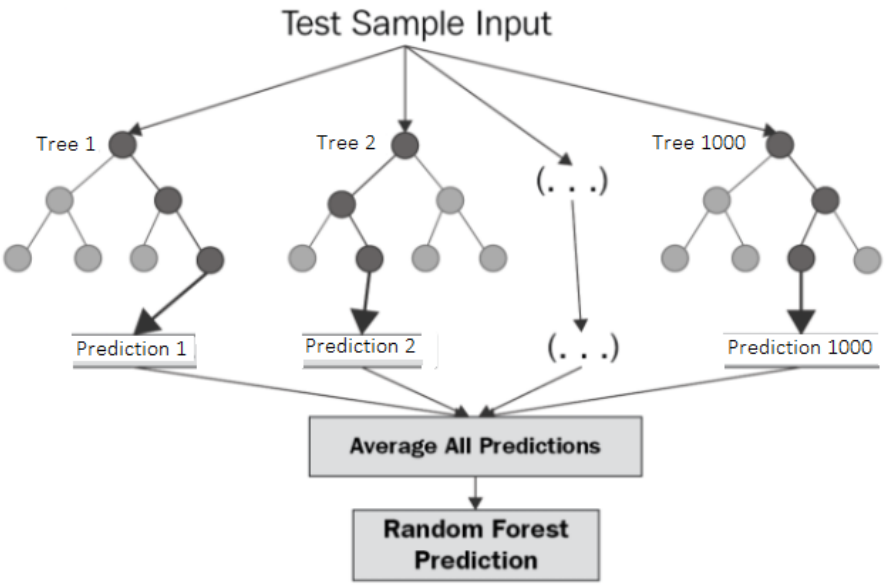

Fig 2.2: Structure of Random Forest

Random Forest use multiple Decision Tree model to predict its final output. In Fig 2.2, we see 1000 of Decision Tree is used to build Random Forest model. These trees are separated from each other and have separated output for each tree. In regression method, these outputs are continuous values. So, average of this values acts as the final result of Random Forest model.

2.2.4 Support Vector Regression

Support Vector Machine (SVM) is a classification model which predicts discrete values as output. SVM finds a hyperplane to separate the datapoints in a n-dimensional space where the shape of hyperplane is depending on the number of attributes. Now, the regression form of SVM is called Support Vector Regressor (SVR) [13]. SVR tries to find the best line within a threshold value. Distance from hyperplane to boundary line is called threshold value. SVR runs on a quadratic time complexity, that makes it inefficient for large datasets. Generally, Linear SVR is used for large dataset because of the time complexity where it uses only Linear Kernel.

2.2.5 Light GBM

A Gradient boosting framework, LightGBM is based on the Decision Tree concept to reduce the error term. It is mainly used to make Gradient Boosting technique faster and more distributed. This model works on two techniques-

1. Gradient-Based One-Side Sampling (GOSS)-
   This is a novel sampling technique which filters instances depending on the gradient. Instances with lower gradient is trained well than instances with higher gradient. Large gradient tends to be undertrained. Focusing on large gradient instances while sampling can affect the data distribution that's why GOSS keeps small gradient instances as well along with large gradient instances.

2. Exclusive Feature Bundling (EFB)-
   EFB reduces the time complexity of Light GBM by merging the feature. While working with large dataset, some features are there which are mutually exclusive i.e., they don't happen simultaneously. Light GBM is capable of figure out these features and merging them into one single feature that's reduces the run time of the model.

.

2.2.6 Artificial Neural Network

ANN is built on the concept of neurons present in our brain. These neurons accept billions of signals through Dendrites and pass the result through Axon to the other neuron. There are methods which are trying to replicate the working process of our brain. In particular here we are going to discuss about Artificial Neural Network (ANN). A Neural Network works like that where we give data points as input (Dendrites), then process in the hidden layer (Neuron body) and get output (Axon). Each input has an allocated weight.

A neuron has several input values. Each input value is multiplied with the assigned weight and all the multiplicated values are summed up. Now this number is passed to the hidden layer. In the Hidden Layer, we use different types of activation functions for better prediction such as 'Threshold function', 'Sigmoid function', 'Rectifier', 'Hyperbolic tangent'. Now, the number of hidden layers is varying depending on the size of training dataset.

ANN is based on Learning Algorithm. The best combination of weights of the input values is decided while the model is trained. After the training, model is ready to predict values for test dataset.

Chapter 3

Approach

In this chapter, we discussed about the problem statement followed by data collection, overall workflow of our missing value prediction and the approaches.

3.1 Problem Statement

The aim of this study is atmospheric ozone concentration prediction using five shallow learning models (e.g., Elastic Net, Decision Tree, Random Forest, SVR, LightGBM) and one deep learning model (ANN). Specifically, the objective is to predict the ozone concentration at air quality monitoring station. Prediction of accurate concentration of air pollutant is difficult because of rapidly changing weather and emission of air pollutant. It makes the task very difficult of prediction for a particular region at a particular time. This study primarily works on the time series data collected over a year from air quality monitoring stations of a highly congested region.

3.2 Data Collection

We have collected ground level ozone data from different Air Pollution Monitoring Stations situated in a highly congested city of India, Delhi. The data collected from the website of Central Pollution Control Board (https://www.cpcb.nic.in). We will predict the concentration of ozone for a particular station. Ozone is one of the highly harmful pollutants for human body. We have collected our data from twenty-nine different Air Pollution Monitoring Stations all around Delhi. The data was collected through out a year for everyday in an interval of one hour. We have gathered 9504 data points per station. Geometrical location of these air quality monitoring stations is given (see Table 3.1).

Table 3.1: Description of Air Pollution Monitoring Stations

Station Name	Latitude	Longitude
Anand-Vihar-DPCC	28.82284	77.10198
Ashok-Vihar-DPCC	28.7762	77.05107
Aya-Nagar-IMD	28.57103	77.0719
Bawana-DPCC	28.68468	77.07657
DTU-CPCB	28.63965	77.14626
Dwarka-S8-DPCC	28.75005	77.11126
East-Arjun-Nagar-CPCB	28.56789	77.25052
IHBAS-CPCB	28.47069	77.10994
ITO-CPCB	28.71051	77.24949
Jahangirpuri-DPCC	28.69538	77.18167
JNS-DPCC	28.57017	76.93376
KSSR-DPCC	28.63643	77.20107
Mandir-Marg-DPCC	28.67234	77.31526
MDCNS-DPCC	28.73282	77.17063
Mundka-DPCC	28.73253	77.11992

Station Name	Latitude	Longitude
Najafgarh-DPCC	28.68117	77.30252
Narela-CPCB	28.53135	77.19016
Nehru-Nagar-DPCC	28.67405	77.13102
Okhla-Phase-2-DPCC	28.58028	77.23383
Patparganj-DPCC	28.53079	77.27126
Punjabi-Bagh-DPCC	28.61128	77.23774
Pusa-DPCC	28.69979	77.16545
RK-Puram-DPCC	28.49857	77.26484
Rohini-DPCC	28.64762	77.31581
Sonia-Vihar-DPCC	28.56326	77.18694
Sri-Aurobindo-Marg-DPCC	28.65594	77.2949
Vivek-Vihar-DPCC	28.62862	77.24106
Wazirpur-DPCC	28.62376	77.28721
Shadipur-CPCB	28.65148	77.14731

3.3 Workflow

We have collected the air pollutant (Atmospheric Ozone) data from twenty-nine different air pollution monitoring stations (see Table 3.1). We will evaluate this data to predict the ozone concentration of our target station. The dataset had not contained any missing/garbage/NULL values.

Every model has a set of parameters. Default parameter setting doesn't give the best output every time. Our objective was to find the best parameter setting for each prediction model. We have used the processed dataset which have twenty-eight working attribute and one target attribute (Shadipur). Now, we used train_test_split() function of scikit-learn library to split the dataset into train and test data. Here, the test size is 20% and train size is 80% of the data.

Several methods are present to find the best parameter setting e.g., GridSearchCV, RandomSearchCV. We used mostly GridSearchCV for hyperparameter tunning. In some cases; we have done the process manually. We tried to use the mostly known and effective parameters for a model. For GridSearchCV, a dictionary is passed as parameter along with the model where

key of the dictionary is the parameters of the particular model and value is a list of possible combination of those parameters. Now train the returned model. Then, we got our best parameter setting for the particular prediction model. GridSearchCV is not used for deep learning model tuning. Keras tuner is used for ANN tuning. For ANN, number of hidden layers, number of neurons and learning rate is considered primarily for tuning. After getting these values, we varied epochs and batch size using a for loop for better accuracy. Beside these techniques, we also tuned some model manually as the execution time is very high. For manual tuning, we consider one parameter at a time while others at their best value possible. The parameter is not tuned already but is in use then we made sure it is in its default value. Once we achieved our best setting with 20% test size for a particular model then we varied the test size [10%-40%] in an interval of 5% and analyse the accuracy. Here, training and testing time is also calculated for different test size.

Mean absolute error (MAE) and Root mean square error (RMSE) were used to analyse the more suitable value of the parameters. And we used this parameter setting for training our dataset in future. R2 score and correlation between predicted and actual value is also considered. Low value of RMSE and MAE indicates less error in predicted data.

Importance of each feature in the final output varies and here feature selection plays a very important role. Selecting the more important feature and dropping the less important feature can improve our accuracy by a good margin. Computing large data is costly in terms of time. To reduce the time complexity while maintaining the accuracy, feature selection is one of the best options. We have done several feature selection methods e.g., Radius based, Triangulation, Mutual Information, K-L Divergence. We also used an unsupervised leaning technique (K-means clustering) to split our dataset into different number of clusters based on their real time location (longitude, latitude). In radius-based feature selection method, we tried to differentiate the data based on their radius from the target station. Mutual Information between each and every station to target station helps us to detect which attributes are more related in nature with the target attribute. In working with this relation, we tried another methodology is called K-L divergence which works on the similarity of probability distribution between two attributes. These approaches gave us a decent number of datasets to work on.
Here, train_test_split() is done with the datasets which we get from feature selection. Here, the ratio of train and test data is 80:20. The parameters used here is the best setting we get from hypertuning. As an example, in Random Forest, best parameter setting we got is- n_estimators

is 500, bootstarp is True and warm_start is True. So, we use this setting in our feature selection dataset and the test size is 20%. Now we trained the model with training data using these setting and then predict the output for testing data. RMSE and MAE is used to analyse the results.

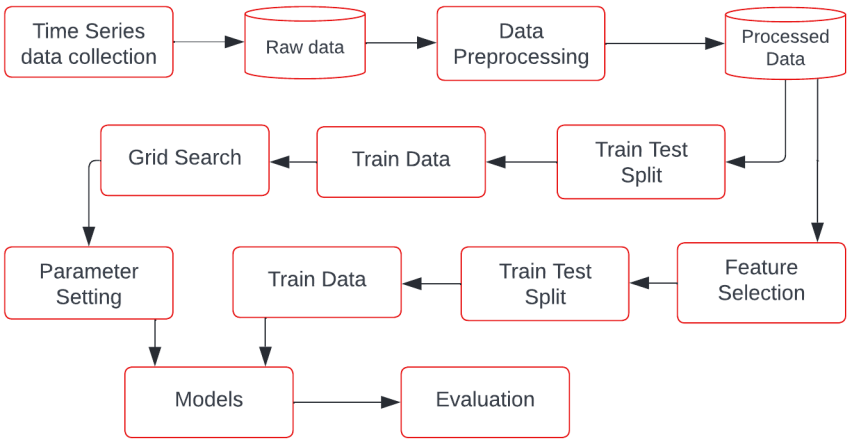

Fig 3.1: Flowchart for the missing value prediction

3.4 Proposed Approach for missing value Prediction

The density of atmospheric ozone in air can be depend on many factors such as industry, traffic, temperature difference, demography etc. Hence, this prediction model may not feasible to another region. We tried to explore some of the feature selection techniques and prediction models for this particular dataset to predict our output as much accurate as possible. Here, we are going to discuss about feature selection and prediction models.

A detail classification of methods used in this study is show in Fig 3.2. Several studies have been done in air pollutant using various techniques and models. In this study we have worked on five machine learning (Elastic Net, Decision Tree, Random Forest, SVR, Light GBM) and one deep learning model (ANN). A generalize description of prediction models are given in chapter 2. Methodologies used for feature selection are based on the distance (Radius-based, Cluster-based, Triangulation) or based on the relation between a feature and target attribute (Mutual Information, K-L Divergence).

An overall workflow of the proposed approach is shown in Fig. 3.1. One pollutant (atmospheric ozone) specific dataset was given of which the data was merged from twenty-nine different air quality monitoring stations. The raw dataset didn't have any missing value so that we didn't need to work on any kind of missing value imputation techniques.

A machine learning model have a bunch of parameters as it's arguments and its quite difficult to find out the best value for it. We carried out a hyper-tunning technique to find out our best parameter setting for each model mentioned in Fig 3.1 with a test size of 20%. We used GridSearchCV as of our parameter tunning technique for shallow learning and keras-tuner for ANN (artificial neural network) model tunning. We have also done manual tunning in some cases where tuning functions took much time to execute. We tried to consider the parameters which has a good impact on the final output. In case of ANN tunning, first we considered a range of hidden layers [2 to 20], number of neurons [32 to 512], learning rate. In addition to that, later vary batch size, epoch, learning rate to get the best result of ANN. Manual tunning (e.g., Random Forest, SVR) is done in order to optimize the execution time of some model and here we consider to tune the parameters one by one by fixing the values of other parameters to their best value.

After tunning the models using the complete dataset, feature selection is carried out to understand the characteristics of the dataset. A total of five approaches (Radius based, Triangulation, Mutual Information, K-L Divergence, Cluster Based) are implemented to eliminate the least contributed attributes.

- **Radius based approach**

  Atmospheric ozone is likely to travel in neighbouring region from the origin region. Hence, it could also affect the regions which are close to the origin. We consider the stations present in a particular radius from the target station. Different radius [5 to 25] in an interval of 5 are taken into consideration.

- **Triangulation**

  Triangulation [14] is a method to define the boundaries based on the geographical subdivision of a region. Quadtree data structure is used to implement this method. This method captures surface feature properly as it is aligned with the feature's boundary.

- **Mutual Information**

  Mutual information (MI) [15] defines the uncertainties between two attributes. High MI means a large reduction in uncertainties and low MI means small reduction. MI is zero means the attributes are independent.

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y}.$$

  X, Y = Discrete variables

  $Pxy\ (x, y)$ = Joint probability distribution

  $P_X(x) = \sum_y P_{XY}(x,y)$, $P_Y(y) = \sum_x P_{XY}(x,y)$

- **Kullback-Leibler Divergence**

  K-L Divergence [16] score is a measurement of how much one probability distribution varies from another. $D_{KL}(P||Q)$ can be think of as a statistical distance that measures how far the distribution Q is from P. $D_{KL}(P||Q)$ close to zero means the distance is small. Considering attributes with small distances with respect to target attribute have more chance to perform better comparing with other attributes.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right)$$

$$D_{\mathrm{KL}}(Q \parallel P) = \sum_{x \in \mathcal{X}} Q(x) \ln\left(\frac{Q(x)}{P(x)}\right)$$

  K-L divergence is also called relative entropy where $D_{KL}(P||Q)$ is called relative entropy of P with respect to Q. It means, the information gain if P were chosen instead of Q.

- **Cluster based approach**

  An unsupervised learning model (K-means clustering) is used to split the stations into different clusters. This machine learning model used the geographical location (longitude, latitude) of the air quality monitoring stations to calculate the distance between them. We varied the number of clusters [2 to 20] in an interval of 2 and took the stations that present in the same cluster with the target station.

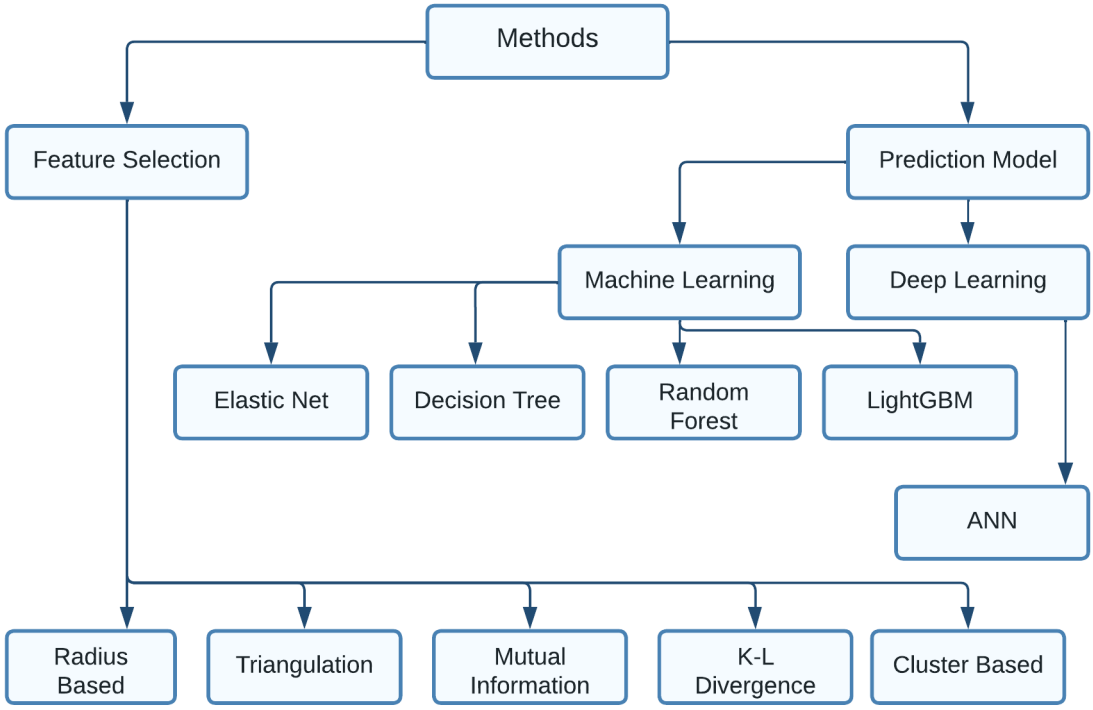

Fig 3.2: A taxonomy of the approaches evaluated in this study for feature selection and prediction models.

Chapter 4

Result and Analysis

This chapter has a brief discussion on our experimental result and analysis. Firstly, we have noted down all the configuration of the machine where the experiment is carried out followed by a detail discussion of our result.

4.1 Experimental Setup

This study is done on a windows PC with Intel(R) Core (TM) i5-4460 CPU@ 3.20GHz with 8 GB RAM (DDR3). We worked on this experiment in python language. Various python libraries were also used e.g., NumPy, Pandas, Matplotlib, scikit learn. NumPy (https://numpy.org) is used in working with array. Pandas (https://pandas.pydata.org) is used for analysis of data. Matplotlib (https://matplotlib.org) is used for visualization of data. Scikit learn (https://scikit-learn.org) is used for implementing models and various metrics (MAE, RMSE) to judge the error of the output. Keras (https://keras.io) to build the layers of ANN model and keras tuner (https://keras.io/keras_tuner/) for tunning of ANN.

4.2 Results

In this study, we worked with one pollutant specific time series data which is collected from twenty-nine different air quality monitoring stations situated in Delhi. The ozone concentration is collected through out a year for everyday in an interval of one hour. We set our target station at Shadipur. In raw data, there had not any missing value. A time series data visualization is given in Fig 4.1 using a whisker plot.

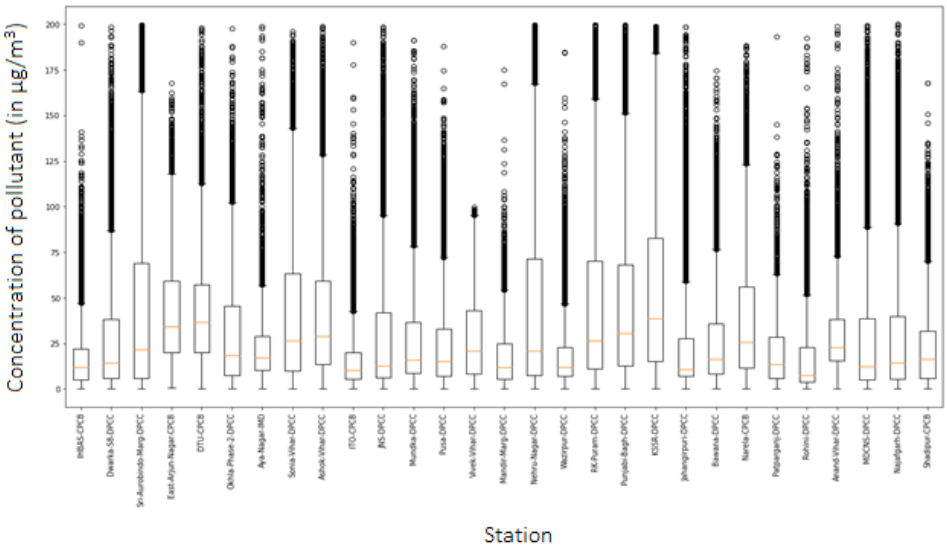

Fig 4.1: Data visualization using whisker Box plot of ozone for twenty-nine different stations

Evaluation metrics are important to determine the accuracy of the models. We used Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) and beside this, we have also evaluated our model with R2 score and correlation between the actual and predicted output.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

$$\text{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=0}^{n}(y_{test} - y_{pred}),\ TSS = \sum_{i=0}^{n}(y_{test} - y_{mean})$$

A brief discussion about the experimental result is provided in this section. As discussed in chapter 3, hyper tuning is done with whole dataset. Default Train size is 80% and test size is 20%. GridSearchCV is used for hyper tuning the machine learning models and keras tuner for deep learning model. Fig. 4.2 describes the training and testing time comparison among the models with best parameter setting with default train and test size. Random Forest took a significant increase among all of them in training. LightGBM and SVR took almost equal time to train. ANN is a deep learning model and consisting of five hidden layers so it took a while to execute. Though, it depends on factors like the number of layers, epochs, batch size etc.

Random Forest executed fast in testing compared to other models where SVR took the highest time to execute in testing.

Training and testing time is also machine dependent and may miscalculate the times because of the buffer access time.

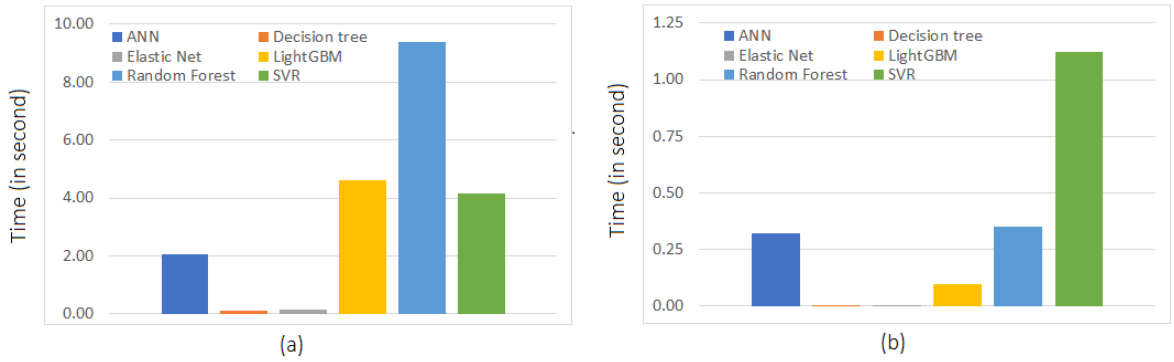

Fig 4.2: Execution time of different models (a) Training time (b) Testing time

RMSE and MAE is considered to compare the prediction models. Fig 4.3 indicates that six prediction model's RMSE and MAE value comparison which is implemented using best parameter setting. Here, we can see that LightGBM have the lowest value and Decision tree has the highest value in both MAE and RMSE. Interesting to see that Random Forest and SVR both the model's accuracy is very close to LightGBM. ANN and Elastic Net could not perform well with respect to LightGBM. Exclusive feature bundling of LightGBM gives it an edge over other prediction model.

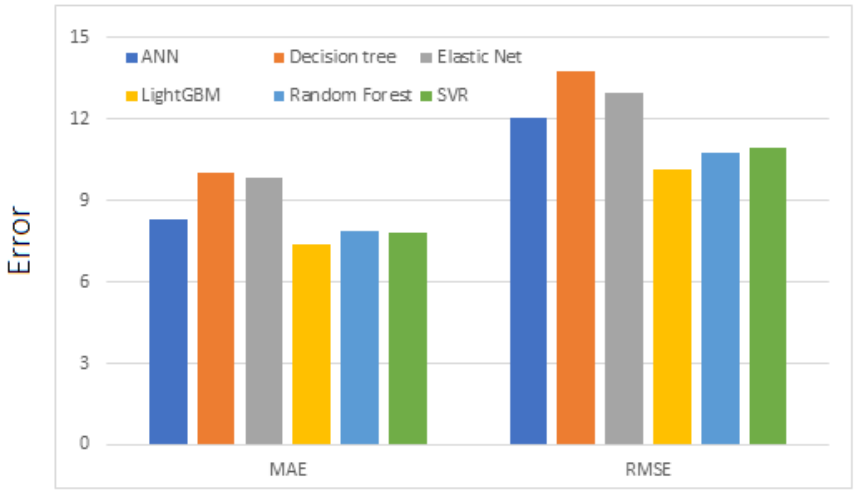

Fig 4.3: MAE and RMSE comparison for different prediction models.

LightGBM took a decent time and performed best amongst the models. The actual data vs predicted data of LightGBM model is shown in Fig 4.4 (a). x-axes represent the actual values and y-axes represent the predicted values. We can compare this with our worst performed model in term of accuracy, Decision Tree in Fig 4.4 (b). Overlapping points in Fig 4.4 (a) tends to form a strong, positive and linear relation between actual and predicted value. In a comparison, if we look at the Fig 4.4 (b), datapoints are not dense and spread over the graph. Fig 4.4 (b) indicates a weak and nonlinear relation between actual and predicted value. A comparison of correlation and R2 score amongst prediction in Fig 4.5 is given that shows LightGBM has a better correlation between predicted and actual value. Random forest and SVR has a very small difference with LightGBM in Fig 4.6.

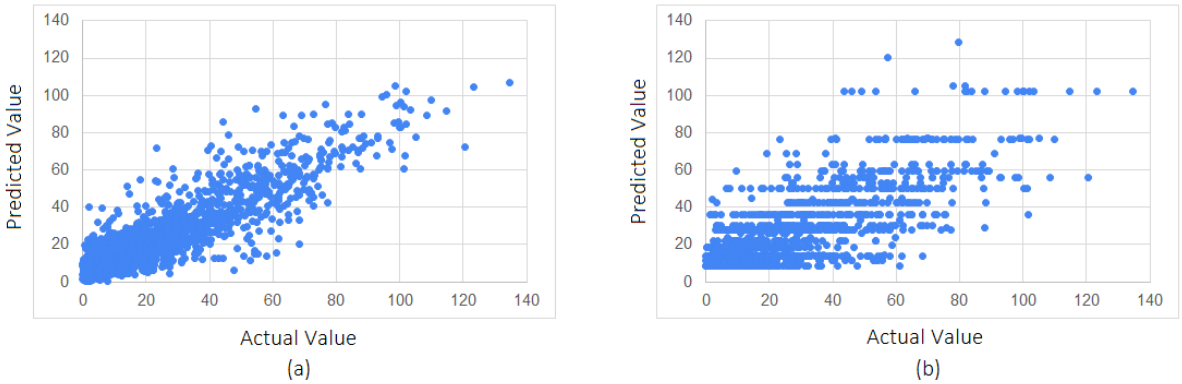

Fig 4.4: Comparison between (a) LightGBM (b) Decision Tree of actual vs prediction datapoints

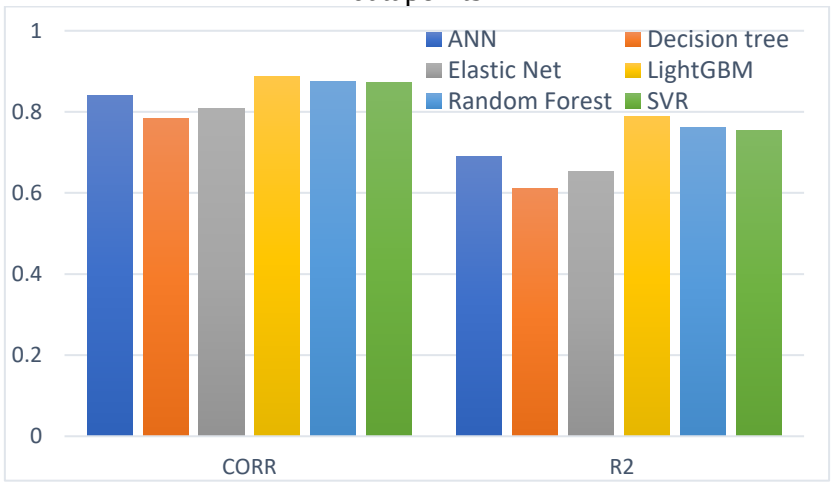

Fig 4.5: Correlation and R2 score comparison for different prediction models.

Five types of feature selection are done in this study- Radius based, Triangulation, Mutual Information, K-L divergence, Cluster based. We are going to analyse the results of these

methods implemented in LightGBM. In Fig 4.6, we can observe that MAE and RMSE is decreasing when the number of attributes is increasing in dataset. That means, including a greater number of stations are effective to low down the errors in predicted values. Fig 4.7 strengthens the fact when MAE and RMSE from hyper tuned model is compared with triangulation model, the error increased. In triangulation, for our target station we have only three neighbouring stations as attribute. Like this, for a lower radius in Fig 4.6 (a), higher mutual information in Fig 4.6 (b), lower K-L divergence in Fig 4.6 (c) and greater number of clusters in Fig 4.6 (d) contains lower number of attributes and the increase in MAE and RMSE is reflected in Fig 4.6.

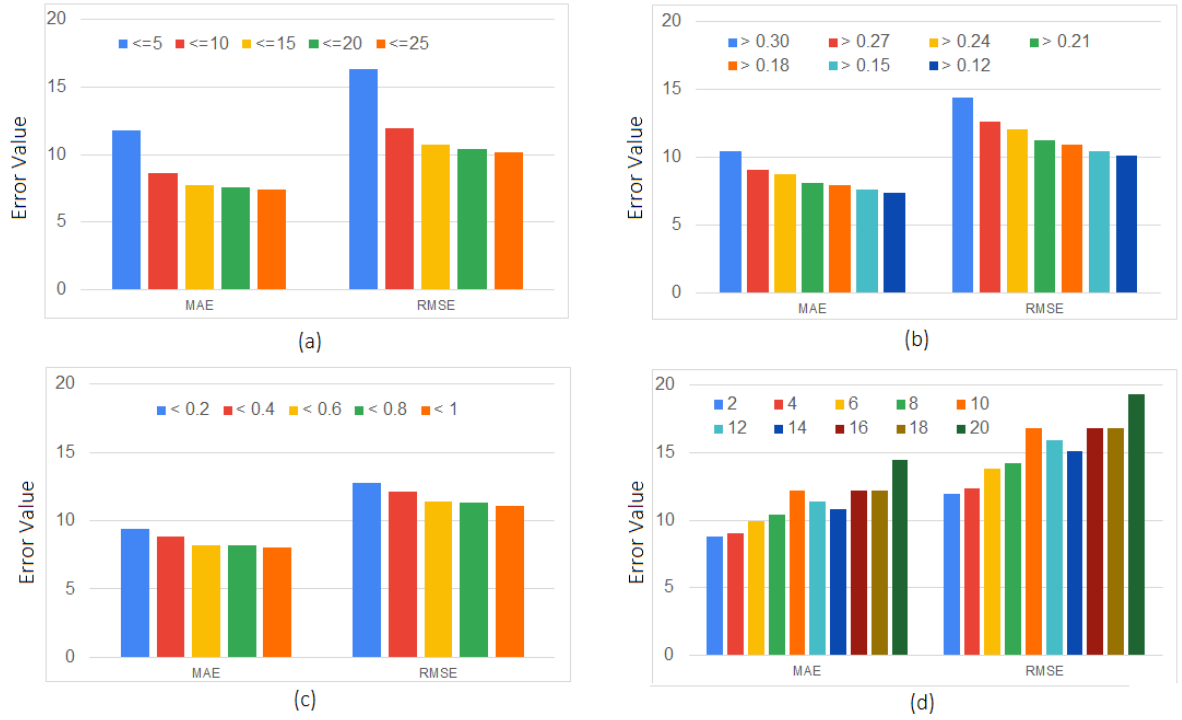

Fig 4.6: RMSE and MAE comparison for different feature selection methods using LightGBM (a) Radius based (b) Mutual Information (c) K-L divergence (d) Cluster based

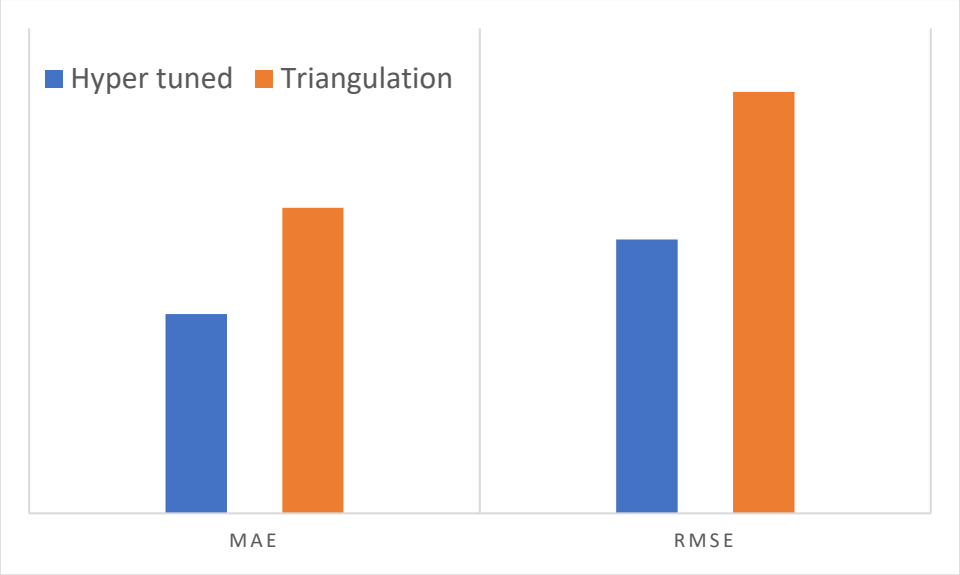

Fig 4.7: MAE and RMSE comparison of Triangulation and Hyper tuned value of LightGBM

Now, if we analyse the Fig 4.6 carefully, we can observe that the decrease of MAE and RMSE is negligible after a point. In Fig 4.6 (a) MAE of radius 20 is 7.59 and MAE of radius 25 is 7.39. This difference cannot impact our prediction value but the exclusion of feature decreases the time complexity. This negligible difference in MAE while the number of feature decreases in the dataset can be seen mostly in Radius based method in Fig 4.6 (a), Mutual Information in Fig 4.6 (b) and K-L Divergence in Fig 4.6 (c). This shows that lesser number of features can be used to forecast the air pollutant data while maintaining the accuracy.

In comparison of six different prediction models, LightGBM performs most efficiently with a value of MAE = 7.39262009 and RMSE = 10.1694019. Random forest and SVR seems promising as it's best score of MAE = 7.8998512 and RMSE = 10.75692, MAE = 7.78872969 and RMSE = 10.92472716 respectively.

Chapter 5

Conclusion and Future Work

In this chapter we discussed about the overall work at a glance. A further discussion about the future possibilities and improvisation of this experiment is there.

5.1 Conclusion

We have worked on five machine learning and one deep learning prediction model in this study. Finding most accurate concentration of air pollutant for an air pollution monitoring station. We had a yearlong data which is taken every day in a span of one hour. Training models with 80% data and testing it with remaining 20% of data using the Hyper-tuned parameters. We have also worked on feature selection of the actual dataset. We have used MAE, RMSE to evaluate the model performance. LightGBM have performed best in terms of RMSE and MAE but Random Forest and SVR also had a very close result to LightGBM. Feature selection methods (Radius Based, Mutual Information, K-L Divergence) are suitable to reduce some fewer effective features from dataset to decrease the time complexity of the model.

5.2 Future work

In future, deep learning models (e.g., CNN, RNN, LSTM, Bi-LSTM etc.) can be used to predict pollutant concentration which may outperform the machine learning model. We have worked on five different logics for analysis of actual data, so we can work on other popular analysis methodologies (e.g., ADF test, Autocorrelation etc.). More number of air quality monitoring station data will help this method to perform better. Most importantly, this study will raise awareness amongst the people suffered the most.

Reference

1. Asif Iqbal Middya, Sarbani Roy, "Pollutant specific optimal deep learning and statistical model building for air quality prediction", *Environmental Pollution* 2022, https://doi.org/10.1016/j.envpol.2022.118972.

2. Nahun Loya, Ivan Olmos Pineda, David Pinto, Helena Gomez-Adorno, Yuridiana Aleman, "Forecast of Air Quality Based on Ozone by Decision Tree and Neural Networks", 2012

3. Saleh M. Al-Alawi, Sabah A. Abdul-Wahab, Charles S. Bakheit, "Combining principal component regression and artificial neural networks for more accurate predictions of ground level ozone".

4. Maryam Aljanabi, Mohammad Shkoukani, Mohammad Hijjawi, "Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan", DOI: 10.1007/s11633-020-1233-4

5. Carlos Cardelino , Michael Chang , Jim St. John , Bill Murphey , Jeff Cordle, Rafael Ballagas, Lynda Patterson , Ken Powell , Jim Stogner & Susan Zimmer-Dauphinee, "Ozone Predictions in Atlanta, Gorgia: Analysis of the 1999 Ozone Season", https://doi.org/10.1080/10473289.2001.10464342

6. M.A. Barrero, J.O. Grimalt, L.Canton, "Prediction of daily ozone concentration maxima in the urban atmosphere", doi:10.1016/j.chemolab.2005.07.003

7. Ebrahim Eslami, Yunsoo Choi, Yannic Lops, Alqamah Sayeed, "A real time hourly ozone prediction system using deep convolutional neural network".

8. Mohamed Khalid AlOmar, Mohammed Majeed Hameed, Mohammed Abdulhakim AlSaadi, "Multi hours ahead prediction of surface ozone gas concentration: Robust artificial intelligence approach", 2020, DOI:https://doi.org/10.1016/j.apr.2020.06.024

9. Victor R. Prybutok , Junsub Yi , David Mitchell, "Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations", 1999

10. Hossam Faris, Mouhammd Alkasassbeh, Ali Rodan, "Artificial Neural Networks for Surface Ozone Prediction: Models and Analysis", 2013

11. Navneet Kumar, Anirban Middey, Padma S. Rao, "Prediction and examination of seasonal variation of ozone with meteorological parameter through artificial neural network at NEERI, Nagpur, India", http://dx.doi.org/10.1016/j.uclim.2017.04.003

12. Alaa Sheta, Hossam Faris, Ali Rodan, Elvira Kovac-Andric, Ala M.Al-Zoubi, "Cycle reservoir with regular jumps for prediction ozone concentrations: two real cases from the east of Croatia",

13. Efficient Learning Machines by Mariette Awad and Rahul Khanna. [67-70]

14. Josann Duane, Amgad Saleh, “The Feature Point Triangulation Method For Spatial Subdivision”.

15. www.scholarpedia.org/article/Mutual_information#Definition

16. https://machinelearningmastery.com/divergence-between-probability-distributions/