National College of Ireland

SHORT TERM AIR QUALITY PREDICTION AND SUPERVISED MACHINE LEARNING ANALYSIS

MSc Research Project
Data Analytics

Nikhil Deshmukh
x16145003

School of Computing
National College of Ireland

Supervisor: Mr Sean McNally

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Nikhil Deshmukh
Student ID:	x16145003
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Mr Sean McNally
Submission Due Date:	11/12/2017
Project Title:	SHORT TERM AIR QUALITY PREDICTION AND SUPERVISED MACHINE LEARNING ANALYSIS
Word Count:	8151

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

SHORT TERM AIR QUALITY PREDICTION AND SUPERVISED MACHINE LEARNING ANALYSIS

Nikhil Deshmukh
x16145003
MSc Research Project in Data Analytics

11th December 2017

Abstract

Monitoring air quality pollutants form an important topic of atmospheric and environmental research due to the health effects caused by the pollutants present in the urban and suburban areas. The research evaluates forecasting models for predicting air pollution by exploiting various machine learning techniques. The goal is to determine the forecasting accuracy of the fine particulate matter(PM2.5) concentration in air by various time-series models and further evaluate classification models based on its capability to segregate the air pollution type. In this research data is sourced from Indian government website. The time series analysis is accomplished using time series models such as Auto Regression Integrated Moving Averages (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH), TBATS model, ARIMA with multivariate regressions (ARIMAX) and Dynamic Harmonic Regression (DHR). ARIMAX performs best among all with the lowest error. For classification K nearest neighbor (KNN), Artificial Neural Network (ANN) and Ensemble model are used. Through out the research, it was found that use of ensemble model improves the performance of classifier.

1 Introduction

1.1 Project Background and Motivation

In Delhi mean concentration of PM2.5 for past 3 years is around 150 µg/m^3 which is fifteen times higher than WHO guideline.PM2.5 is highly responsible pollutants for degrading the air quality, it has adverse health effects such as cardiovascular diseases, respiratory diseases, and premature deaths. Therefore, the development of effective and propitiative predictive model is of prime importance. Three methods have been used to predict PM2.5 statistical models, chemical transport, and machine learning. Statistical and chemical transport models have few drawbacks such as the negative correlation between different parameters and accuracy rely on an updated source. Due to the Complexity of air flow it is very difficult to get the updated sources.Machine learning technique such as KNN, ANN, Time series analysis and other hybrid models outperform the other two methods. That is why forecasting using machine learning techniques is extensively used not only to predict weather but air pollution as well. Deters et al. (2017)

Advance time series and classification methods such as ARIMA, ARIMAX, DHR, TBATS model, and hybrid model have been preferred over traditional mean method, decision trees due to temporal nature of advanced algorithms and volatility can be handled more efficiently Elman (1990)

Section 1 addresses the project specification which includes research question, purpose, research variables used in the study. Also, a brief overview about the Air Quality Index (AQI), PM2.5, machine learning, time series and classification models used concludes this section. Section 2 covers related work in air pollution prediction using machine learning. Literature covers both time series analysis and other prediction models used in this area. Section 3 covers design and methodology used for the project. In section 4 implementation of different predictive models, model comparison and validation is covered. Section 5 covers validation and section 6 concludes the paper with reference to future work in the area of air pollution prediction.

1.2 Project requirement specification

1.2.1 Research Question

Question1: *With what accuracy can the concentration of PM2.5 be predicted using time series machine learning methods?*
Question 2: *Can ensemble method better classify the air pollution as compared to traditional machine learning classification methods?*

1.2.2 Purpose

Purpose of this research is to find out with what accuracy the concentration of PM2.5 can be predicted for next 3 days and best method to classify air pollution. After statistical and numerical methods, machine learning methods are extensively used due to its robustness for volatility. Time series forecasting is not a new phenomenon for air pollution but use of sufficient factors considered for the analysis is lacking Khoshsima et al. (2014). Interestingly, PM2.5 is the prime influence for the visibility in the region as it affects the relative humidity, resulting in road accidents. Thus, to evade this situation and pre-warn the PM2.5 concentration, its prediction offers great potential which is the motivation behind this research Ni et al. (2017). As an evidence of the analysis of existing literature Ali and Tirumala (2016) ensembled or hybrid methods performs better than traditional models, hence compared and contrasted in this paper. Keeping in mind the work of existing researchers, this body has been developed which will be illustrated in section 2. Furthermore, the efficiency of the algorithm to predict air pollution offers the opportunity to provide health benefits. Hence, this paper will strictly focus on reduction of errors for time series prediction and on the other hand for classification it will focus on how accurately models classify.

1.3 Research Variable

Prediction of PM2.5 concentration is the first goal of this research, for which independent variable for time series study is PM 2.5 and fourteen dependent variables considered. This is multivariate time series analysis for which ARIMAX and DHR methods have been used. All the data has been taken from Indian government website [1].

[1] http://cpcb.nic.in/

For second data set, labels are created based on AQI, for classification analysis. AQI is calculated as per Indian government standards using break point concentrations of PM2.5, PM10, SO2 and NO2 concentration. Meteorological data temperature, humidity, pressure, rainfall and wind speed is accumulated from weather website [2] and concentration of pollutants is gathered from Indian government website[3].

The basis behind the selection of variables is discussed in the later chapter. Also, calculation of AQI will be discussed in this chapter.

1.4 AQI and PM2.5

"An AQI is defined as an overall scheme that transforms weighted values of individual air pollution related parameters (SO2, CO, visibility, etc.) into a single number or set of numbers. Sharma and Bhattacharya (2015) Different indexes have been proposed from long time for the calculation of air pollution, some examples are Green index, Fenstock Air Quality Index (AQI) and Ontario API. In most of the indexes either number of pollutants taken were less or aggregation function used had eclipse or ambiguity. That is why these indexes are not considered in this research. For example Ontario API uses only CO and SO2 concentration for calculation of air pollution index is not a good choice. Increased concentration of PM2.5 and PM10 are severely effecting the human health which can not be ignored. Goel et al. (2015)

For calculation of IND-AQI, concentration of SO2, NO2, PM2.5, PM10, O3, CO, SO2, NH3, Pb has been considered. Four pollutants SO2, NO2, PM2.5 and PM10 play major role in AQI calculation, where 3 out of these 4 pollutants are must for the calculation.

Ip= [(IHI - ILO)/ (BHI -BLO) * (Cp-BLO)] + ILO
where,
BHI= Breakpoint concentration greater or equal to given concentration
BLO= Breakpoint concentration smaller or equal to given concentration
IHI = AQI value corresponding to BHI
ILO = AQI value corresponding to BLO
Cp = Concentration of pollutant
AQI=Max (I1,I2,I3,...,In)

The maximum operator is selected for the calculation of AQI because it is free from eclipse and ambiguity.Other operator such as mean and exponential operators provides ambiguity, that is why AQI is selected over these indexes. Sharma and Bhattacharya (2015)
Below is the table which shows AQI category and AQI range:

AQI Category	AQI Range
Good	0-50
Satisfactory	51-100
Moderate	101-200
Poor	201-300
Very Poor	301-400
Severe	Above 400

Table 1: AQI Range and Category

[2]https://en.tutiempo.net/
[3]https://data.gov.in/

For short term time series prediction, PM2.5 has chosen as predictor variable because level of PM2.5 is very high in India Goel et al. (2015). Other pollutants and physical data are selected as predictor variables.

1.5 Machine Learning

Recently, statistical based Machine learning models are actively studied in place of traditional numerical methods because these methods are more influenced by the careful building of mathematical model Zhang et al. (2017). Broadly, machine learning is classified into 2 categories; supervised and unsupervised. In unsupervised learning, the dataset is modeled based on grouping and clustering to trace hidden knowledge from it where target variables are not clear Bougoudis et al. (2016). K-means clustering is an example of such models.

Supervised learning is the method where observed instances are labeled, or the target is known. Classification and regression are sub types of supervised machine learning. If target variable is continuous, regression is used, and for discrete variables classification techniques are used. As stated in the literature selection of classification methods are not trivial which offers number of comparison of classifiers for an explicit problem. Some examples of supervised learning models are KNN, ANN, Decision Trees(DT), Multiple regression, time series analysis using ARIMA, TBATS, ARIMAX and DHR.

1.5.1 Description of Time series models

Time series analysis is a key area in statistics that focuses on analyzing data set to study the features of the data and extract meaningful information and statistics from it. The main objective of time series study is to understand the time-dependent structure of single series (univariate time series) analysis and association among numerous series (multivariate time series) analysis . These time series are further subdivided which is not within the scope of this research. Both univariate and multivariate time series has been considered in this study. For ARIMA-GARCH and TBATS model univariate time series of PM2.5 taken while for ARIMAX and Dynamic harmonic regression, multivariate time series with other 14 variables were also opted. After conducting Box test, it has been analyzed that the time series data used in this study is non -stationary as p-value is below 0.05 which shows it is not a white noise. From the below ACF plot (Figure 1) of PM2.5 shows there is seasonality in the data as pattern is repeating after certain time period (95 lags).

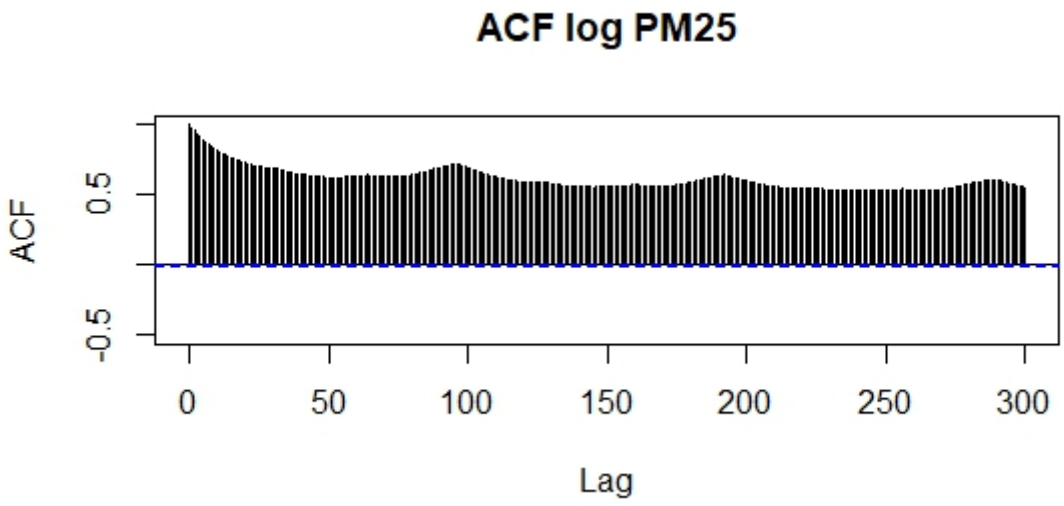

Figure 1: ACF plot for PM2.5

ARIMA-GARCH: Time series analysis is traditionally implemented by observing the auto-correlation of the time series. ARIMA models have widely used approach for time series modeling. ARIMA model is suitable for only linearly model because the model does not replicate recent changes and new information. To deal with the volatility and non-linearity ARIMA-GARCH model should be used. L-stern Group (2010) From the below QQ plot and residual plot there is non-linearity and cluster of volatility hence ARIMA-GARCH is used in this study.

TBATS Model: According to Livera et al. (2010) TBATS model is the combination of many models where the initial T is trigonometric. TBAT considers seasonal period, box-cox transformation parameters, ARMA errors, damping parameters and Fourier terms. All the choices made by model automatically, which makes model sometimes treacherous. Different parameter selected by TBAT model are as follows:

TBATS {W,{P,Q} ϕ, {M,K}}

Where,
W = Box-Cox transformation parameter
{P,Q} = ARMA errors
ϕ = Damping parameter
M,K = seasonal period and Fourier term

Some of the key advantages of the TBATS modeling are:

- Typical non-linear features can be handled efficiently which is commonly seen in real time series.
- It accepts a broad parameter space with the likelihood of better forecast.

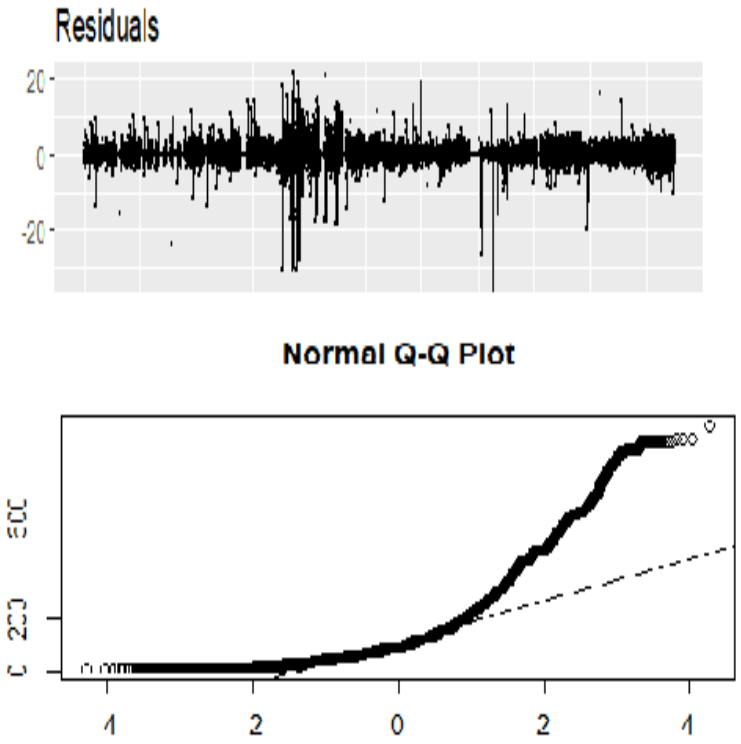

Figure 2: QQ plot and Residual

- For multiple seasonality, this method is very good.

In time series data used in this research has multiple seasonality and non-linear features are present, which is the reason behind selecting this model for the analysis.

ARIMAX and DHR: Both ARMA and ARIMA cannot include exogenous inputs in case of multivariate time series. Autoregressive integrated moving averages model with exogenous inputs (ARIMAX) model is applied, which is a good tool in time series prediction for multivariate time series Rob J (2011). Pankratz refers to the ARIMAX model as a dynamic regression model Livera et al. (2010). ARIMAX can be more flexible for the prediction by using exogenous input as other pollutants and weather variables. For dealing with the multivariate data and to get better forecast result ARIMAX is used in this study. The equation for the ARIMAX is same as the linear regression model only difference is the error term. The error term is white noise in case of linear regression but in this case, it will be ARIMA error. Zhang (2007)

Dynamic harmonic regression is also useful for multivariate time series, it uses Fourier term to handle seasonality.

sk (t) = sin (2πkt/ m)

ck (t) = cos (2πkt/ m)

yt = β0 + sum(k=1 to n) [αk sk$t$ + γk ck (t)] + et

It is the extension of the classical harmonic regression. In the above equation value of k defines seasonality where K cannot be more than half of the seasonal pattern. Value of K shows how many terms can be included and are coefficients and m is seasonal period.et is a non-seasonal error, Fourier terms assume that seasonality does not change all the time while in seasonal ARIMA model it allows it to evolve over the time. The benefit

of using Fourier term is that it can handle seasonality even though seasonal period m is very large. An example is a daily data where m is 365 that is the reason behind selecting this approach for this study as the frequency is 15 minutes and m is very high. Zavala et al. (2016)

1.5.2 Description of Classification Models

According to Rabeb et al. (2017) KNN is different from traditional machine learning models due to it's simple algorithm. This algorithm works on the idea of choosing nearest neighbors and the distance between elements which needs to be classify. Selecting the value of K is of prime importance. Among various measure, Euclidian distance is one of the commonly used measures. ANN is the intelligent system that has the capacity of learning and stabilizing the relationship between the variables, which mimics the human brain learning behavior Catalano et al. (2016). It maps any randomly selected input to the output node without any pre-determined mathematical assumption and creates a relationship by learning. As discussed in the literature, ANN outperforms many techniques for air quality predic that is why it has considered in this study.

Recently, the Hybrid models have been widely used for air pollution prediction. In most of the literature, ensemble models perform better than the traditional one and provide an unbiased result. Ensemble models are built to overcome the problem of weak predictors. It has advantage to alleviate the small sample size to reduce the over fitting of the data Singh et al. (2013). Use of any algorithm with boosting increases the accuracy of the model and extensively used in many applications.

2 Related Work

In the rapid development of economy, air pollution is the cause of many health problems. That is why it is extremely necessary to predict air pollution by the government and local health agencies Kong (2017) and Anenberg et al. (2016). AQI is a tool, introduced by Environmental Protection Agency (EPA) to measure pollutants level in the air Sharma and Bhattacharya (2015).

Ali and Tirumala (2016) developed the classification and regression model to discriminate AQI level during different seasons using single decision tree (SDT), decision tree forest (DTF), decision tree boost (DTB), Support Vector Machines(SVM) and ensemble methods. Principle component analysis (PCA) is performed to get the correlation among pollutants. Bagging and boosting ensembled models outperform the traditional SVM model. In a similar research Singh et al. (2013) has used set of SVM for very purpose of Data mining. CPU execution time(minutes) and accuracy is compared for single SVM and ensembled models, where 11 % more accuracy is obtained by ensemble model. However, there is diversity dilemma in boosting algorithms due to which there is a chance that model is very specific. Hence, dealing with diversity is necessary to get better generalize performance. Xi et al. (2015) investigated that to get better performance, more features should be included. The author opted to meteorological parameters along with the AQI and pollutants concentration. SVM, DT, Random Forest(RF), linear regression, boosting and combination of all is used, where the performance of ensemble model is best among all. There is a limitation to this study that even after considering sufficient features, samples included in the study are less which can result in misinformation of the prediction results. PM 2.5 plays a major role in the formation of AQI.

Deters et al. (2017) analyzed the PM2.5 and weather data to predict air pollution. Abundant data has taken for the analysis, but no evidence of feature engineering due to which question of trust arises on the accuracy of the result. The author used binary classification with the use of Randomise Binary Trees(RBTs) which provides general rules for classification of PM2.5. Also, the concentration of PM2.5 is predicted greater than 20 micrograms per meter cube, which seems to be more influenced by other factors rather than only meteorological parameters. Similarly, Zhan et al. (2017) also focused on PM2.5 and meteorological data focusing on Geographically-Weighted Gradient Boosting Machine (GW-GBM) learning method which is improved by GBM through building spatial smoothing kernels to weigh the loss function. This method can handle missing data by using surrogate splits. Other pollutants such as NO2, SO2, PM10 are not considered for the research which is a major cause of air pollution. Also, this model overcomes the estimation bias problem and uses 10-fold cross validation for the evaluation which can reduce the most common problem of over fitting.

In another study by Ip et al. (2010) LeastSquare -Support Vector Machines(LS-SVM) and multi-layer perceptrons (MLP) was used to overcome the problem of overfitting of previous study. Pollution concentration and meteorological data have taken for the analysis and corresponding analysis shows the seven out of sixteen directions were associated with pollutant levels. The author concluded that LS-SVM predicts with better accuracy and lower error rate as compared to MLP. To further improve China's PM2.5 prediction Multi-Channel Ensemble Learning via Supervised Assignment (MELSA) has been proposed by Zhang et al. (2017) considering all the supporting pollutants but weather affecting cause are ignored in this study. Data from 35 stations has been gathered so that analysis will not be biased towards only one area of the city. Since the data has been accumulated from 35 different stations of Beijing, bright chances are there that all the gathered data has very different value and it is difficult to generalize it as one data set. For example, wind speed can differ for a different part of the cities. The ensemble approach of voting using bagging boosting is opted by the author. In the literature, there is no evidence of validation. Considering result without validation and drawing a conclusion about the accuracy and error can result in misclassification. Tikhe et al. (2014) performed validation with different number of epochs on ANN and analyzed that it gives the best performance by using least epochs. Author implemented two methods ANN and Genetic Programming (GP) for PM2.5 prediction and based on Root Mean Squared Error(RMSE) concluded that GP performs better than ANN.For ANN validation 36 models have been tested by trial and errors method ranges from 40 % - 85 % which assures the validation of model and accurate result. As GP only used here for the short-term forecast, for long-term forecasting reliability of this method needs to be analyzed.

Some researchers focused mainly on other pollutant concentration like Ozone, PM10. Raimondo et al. (2011) performed ANN and SVM with backward selection algorithm (the notion of relativity entropy) inferred by information theory. ANN with 8 hidden nodes performs best compared to different kernels of SVM. One limitation of this study is that it does not consider enough weather data. According to Sudarsan et al. (2010) meteorological data is of prime importance in the process. Rabeb et al. (2017) also focused on ozone concentration and applied 3 different SVM kernels where linear performs the best among all the 3. The author achieved 97 % accuracy by using KNN algorithm keeping the value of K to 1 (Euclidian distance). Straight away keeping value ok K =1 is not good as there is a chance of overfitting the module, keeping K value not very large

or small is always a good practice.Chiwewe et al. (2016) and Shaban et al. (2016) have conducted similar research focusing on ozone concentration while considering weather elements. In both studies, ANN is commonly used but in the study conducted by Chiwewe et al. (2016), M5P trees give the best result among SVM, ANN, and M5P. Interestingly, all the data collected by Shaban from the same station rather than relying on the data from different supercomputers which gives good accuracy using ANN up to 80%.

The idea of using the Neural network in time series analysis is not a new notion. One way to build IT2FNN is to fuzzify a neural network is used by Castro (2008). They focused on prediction of ozone using different architectures of the fuzzy neural network. The author concluded that results improved using IT2FNN with least RMSE of 5.6. The only prediction of ozone does not give the exact idea of air pollution. Use of univariate time series is a downside of this because other predictor variables can change in the entire result. A similar study has been done by Benvenuto and Marani (2000) where short-term prediction has been done using ANN and ARIMAX model. Author has focused on PM10 prediction using multiple predictors and weather elements. Also, the author suggests that if univariate time series is used with a single layer in ANN then it works as multiple regression. The result of ANN has not been contrasted with ARMAX models because these last are a specific instance of ANNs based on similar sources of info. one limitation is that Partial Auto Correlation Function (PACF) is not analyzed and correlation among all predictors has not determined which may result in the wrong prediction. In slightly different research by Lee et al. (2012), different order of Box-Jenkins ARIMA conducted on air pollution index (API) which is made of pollutants. Ljung-Box test has been done on data to check the P-value for white noise. Ni et al. (2017) has conducted in-depth correlation analysis using both Pearson and Spearman coefficient to get a linear and non-linear relationship and decided to target PM2.5 for ARIMA model with least RMSE of 6.76 micrograms per mete cube. 1-year hourly data has taken with the frequency of 1 hour to predict next days prediction which is the base of this paper. Even though it predicts with good accuracy the overall prediction shows one-day lag because of abruptly changing the daily value of PM 2.5.

To overcome this problem every 15-minute frequency data has considered for the analysis to see the daily change in PM2.5 concentration. Advanced statistical model like ARIMAX, TBATS are also implemented which is discussed in the implementation section.

3 Methodology

This research is based on the 'Cross Industry Standard Process for Data Mining' (CRISP-DM) methodology. CRISP is hierarchical process model which has four level breakdowns Ncr et al. (2000). The method is modified as per the need of this research as the sequence of phases is not rigid. CRISP is most preferred methodology by professionals, in a survey CRISP won by most number votes for the best model Piatetsky-Shapiro (2014).

Most of the phases of this methodology successfully collaborate along with the phases of the project.Understanding the business and identifying the problem is one of the most crucial parts of the process. Short term air pollution forecasting is the goal of this project.

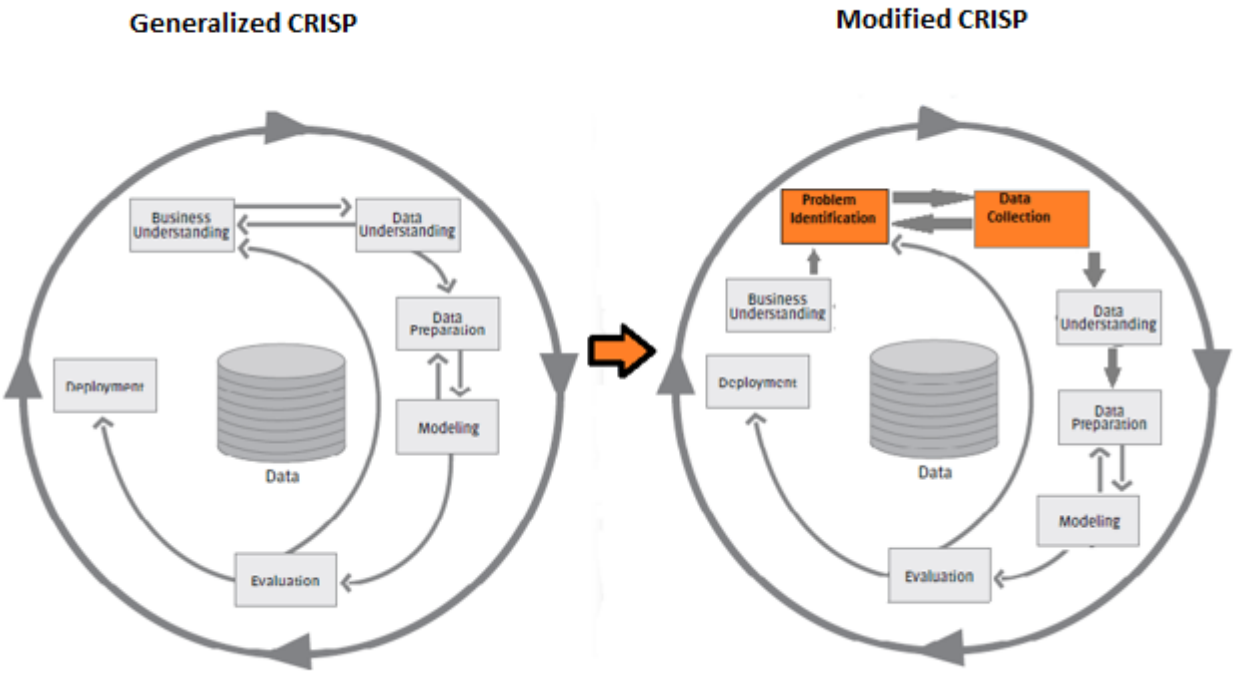

Figure 3: Modified CHRISP DM

3.1 Methodology for Data collection

Data set for Delhi has been finalized for the investigation because Delhi is worlds most polluted metropolitan city in 2014. By forecasting the pollution some actions can be taken to control it in the near future (Anenberg et al.; 2016)

3.2 Methodology For Cleaning Data

Data preparation is one of the important parts of any data mining method to improve the accuracy of the result. "Data cleaning deals with data problems once they have occurred." (Broeck et al.; 2005)

Firstly, among all the variables only essential variables for the time series analysis have been selected, and all variables were standardized in Python using pandas library. Data is interpolated using forward fill (ffill) function and lambda x mean function with window = 96. ffill functions fills the missing values in the data with previous values and, lambda x mean function fills missing value with mean of selected window. ffill function was used to fill missing values in the data set because there was not high fluctuation of PM2.5 concentration within a day, and there was a high probability that missing values were same as previous available value. back word fill (bfill) function was not used because values were closed to the previous values not with the future values. Lambda x function with window 96 was used for some attributes which had high fluctuating values. Every 15 minute data has 96 values each day so, this function fill missing value with average value of 96 rows. This selection of window was appropriate rather than selecting the mean

for complete data because mean value for the day was closer than selecting all values. After cleaning all the variables different variables are merged into a single file using the pd.concate function of python. Concatenation defines a structure to the data which is necessary for further analysis (Jair et al.; 2017).

To clean classification data set, libraries such as tidyr, dplyr used. Instead of cleaning data in excel it is easy to use these libraries, which also provide join functions to merge data sets. Using inner join function on date key weather data set and pollution data set were merged.Initially, this data set was not labeled but labels were created based on the AQI value as per the government standards.

Outlier detection is one of the important steps in cleaning the data to get the accurate result Buzzi-ferraris and Manenti (2011). With the help of box plot, outlier is checked and handled. Data with missing values are omitted and some impractical values are scaled by a mean function of python.

3.3 Methodology For Handling Class Imbalance

In any labelled data set if there is majority of one class then it has the problem of class imbalance. Class imbalance problem can be found in any real case. Imbalance class is associated with misclassification. There are two ways to handle the problem; solution at data level and at algorithm level. Oversampling technique is popularly used technique at data level in which data is added to balance the class. The undersampling technique is opposite of this in which random samples are deleted from majority class (Santoso1 et al.; 2017). The hybrid technique is used in this study where some amount of data was deleted from the majority class and data was added to a minority class. The benefit of hybrid sampling is that it increases the performance time and chances of over fitting is less. Also, information loss was less compared to random under sampling. Before handling class imbalance, all the 3 models were misclassifying with high rate of error. Hence, overcoming class imbalance is necessary before moving forward for modeling.

3.4 Feature Selection And Evaluation

Feature selection and evaluation is an important part of the data analysis process to increase the performance of the models. According to Ni et al. (2017) correlation analysis shows how closely two variables of the data set are related. By observing Pearson's coefficient author says there is high correlation of Wind speed, NO2, SO2, PM10 and CO with PM2.5. Selecting appropriate variables can result in a dramatic increase in the performance of the model. Feature engineering and relevance analysis have been considered from the literature but the implementation of this will be the part of future work.

3.5 Time Series Analysis

In the time series research, basic models such as mean method, Naive method and ETS used for forecasting before implementing ARIMA but forecasting range was very wide and residual results were not good. Due to non-stationary nature of the data, ARMA method was also not in the scope. Box-Cox test is the test to check if the data is white nois or not by observing the value of p. Once Box Test assured that data is not white noise and after decomposing (Figure 4) the trend, seasonality, and randomness, ARIMA modeling was

done. Different order of ARIMA was tried on the data but auto ARIMA with the order (1,1,2) was best fit for the data. Due to non-linearity and volatility, GARCH model is also implemented on it. Because ARIMA model can not handle non linearity efficiently, it was necessary to use GARCH model. TBATS method has been implemented as multiple seasonality was there in the data which could not be recognized easily. As mentioned in the section 1, TBATS model selects the combination of several different parameters which are best fit for the data and build the model.Parameters selected by tbats() model were lambda values, arma errors, Fourier terms and damping parameters. While applying dynamic harmonic regression, trail and error method was used for K value starting from 1 to 40. When trying different values of K, it was increased in such a way that, Akaike Information Criterion(AICc) value should decrease. Once, AICc value stops decreasing that is the perfect value of K for the model. (Rob J; 2011)

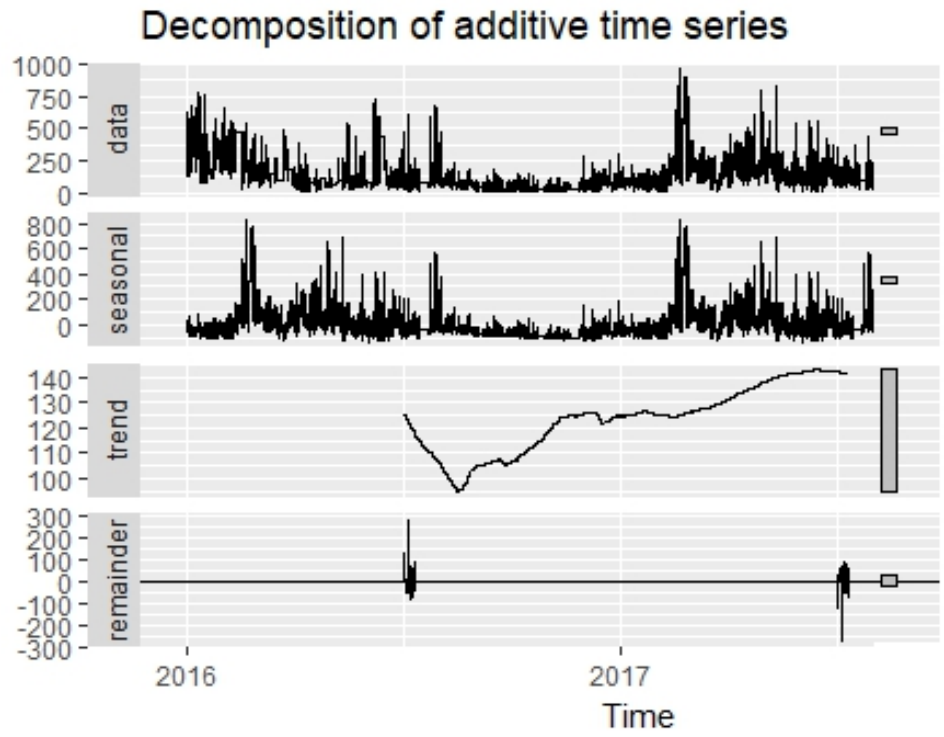

Figure 4: Decomposition of time series

3.6 Classification

Decision tree and Naive base were also considered before opting ANN, KNN and Hybrid models. Top 3 models with good classification performance have been finalized for this research. KNN with different values of K has been implemented, for K value 1 accuracy was good but the model was too specific. Detailed analysis regarding selection of K value and distant function will be covered in the section 4.

ANN is also implemented using a different combinations of learning rate. Training cycles from 1000 to 5000 were tried by increasing 500 epochs at a time. Different training cycles were tried on the data to check the effect of this on accuracy and error. Before modeling, data was shuffled and normalized to get more appropriate results.

Different combinations like KNN- RF, ANN-Naive, and ensemble models is opted

in the research as recently it is popularly used. Hybrid model of Knn and boosting is finalized for evaluation as it was giving the highest accuracy.

4 Implementation

Implementation of the project is divided into two parts, in the first part implementation of time series model and in the second part implementation for classification models is discussed.

4.1 Implementation Of Time Series Models

In this section, all the implementation parameters related to the time series models is discussed. Rstudio is used for implementing all the models.

4.1.1 ARIMA-GARCH model

Before implementing the model, it is converted to time series using ts() function available in R, with the sub-daily frequency of fifteen minutes. The first check before implementing was to decompose the seasonality, trend, and randomness (Figure 4). By using decompose function it is checked that there is some seasonality, as well as the trend. In the starting there is downward trend and later continuous upward trend can be observed from the figure. After decomposing the time series, Dickey-Fuller test has been conducted to see if the time series is stationary or not. p-value for this test is below 0.05 that tells that it is a non-stationary time series, which rejects the alternate hypothesis. A p-value below 0.5 for Box Test assures that the data is not a white noise and there are some trend and seasonality. From the above Figur 1, ACF plot shows some seasonal patterns. Log and differencing has been taken to make time series stationary as shown in the Figure 6, which is an important part of the ARIMA modeling. Differencing of 1 has been taken to make time series stationary as over differencing can cause an increase in standard deviation. According to L-stern Group (2010), AICc is another way to check the best order fit for the ARIMA model. order of p,d, and q should be changed in such a way that AICc should be minimum. Instead of trying different orders auto.arima selects value of p,d and q automatically, which is best fit for the data. In this research values are selected as (1,1,2). Box-Cox test gives value of lambda as 1(no transformation) which is taken as input while implementing auto arima. Box-Cox value represents the value of transformation required to stabilize the variance of the time series Anikender and Goyal (2011). If it gives value of -1 which means it inversely transforms the time series.

Anikender and Goyal (2011)

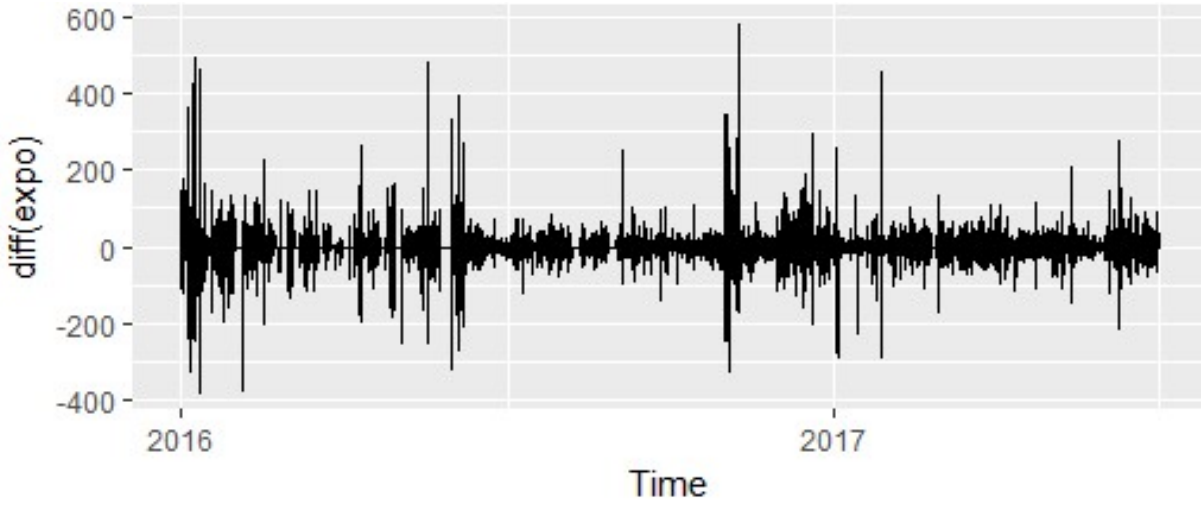

Figure 5: Differencing of time series

For fitting the ARIMA model, 80% and 20% split for training and testing the data was selected, which is not an accurate measure for testing the performance of a model. Cv function in forecast library available in R gives a handy tool for cross-validation. rolling window time series cross-validation is used with n = 10. After applying cross validation value of MSE was increased as time increases. It means for forecasting later instances accuracy decreases vice verse. After fitting the model PM2.5 concentration was predicted for 3 days. Residual is checked which is not a white noise in this case as p-value is less than 0.05. Residuals shows non-linearity and volatility in the PM2.5 data. One more way to confirm is to check QQ plot of the data, if it is linear or not (Figure 2). The p-value for arch residual is above 0.05 which shows that residual is white noise and model is perfect for this scenario. Arima-arch residual QQ plots confirm the validity of the GARCH model. Anikender and Goyal (2011)

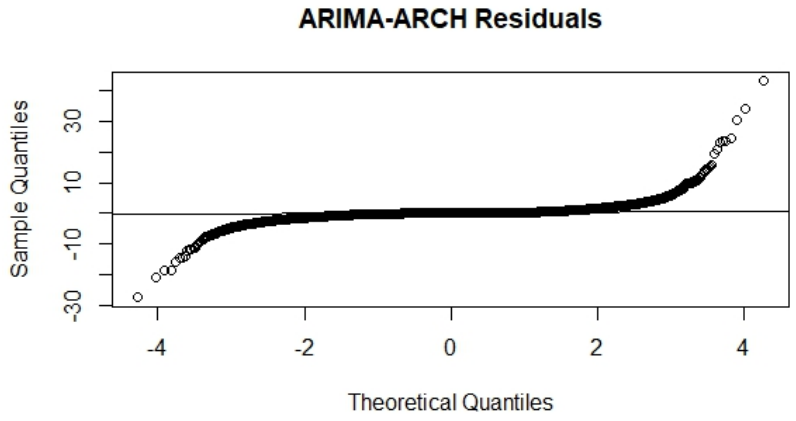

Figure 6: Arima-Arch residual

4.1.2 TBATS model

TBATS models are good where multiple seasonal periods are there or seasonality is hard to observe. In the data set used in this research, there are multiple seasonal patterns observed. TBATS takes very less time to build the model, but it takes a long time to train, as TBATS checks several combination of parameters while implementing. It is the combination of many other models and it must check several values; lambda values, arma errors, Fourier term, damping parameter. After converting data into time series, it needs to be passed into the tbats functions with h argumen to get the forecast. Time series data of PM2.5 is passed to the function and results were observed. From the below Figure 7, arma errors are 0, no Fourier terms are taken, and Box-Cox parameter is 0.424 with a seasonal period of 0.954. It can be clearly seen that predicted value is first increasing and then decreasing over the time for next 3 days. Residual is checked after implementing the model. Even though value of p is very less for the residual it cannot be concluded that model is not at all fit for the given time series, sometimes when the forecasting values are too narrow or wide this problem occurs. Rob J (2011)

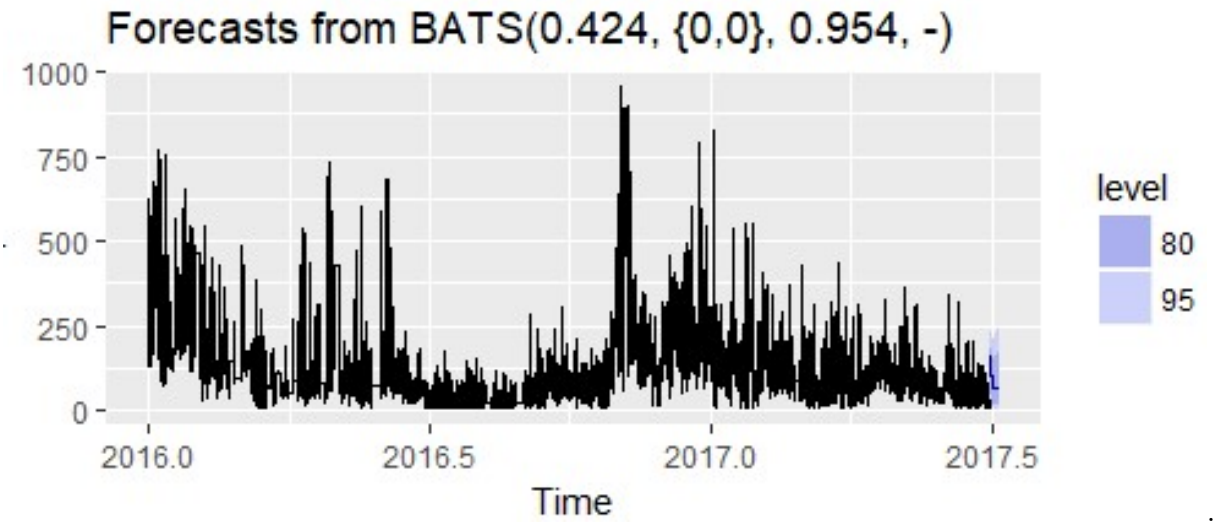

Figure 7: Forecast From TBAT model

4.1.3 ARIMAX And DHR

For multivariate time series, ARIMAX and DHR are implemented in the research. Fitting ARIMAX model is not very different than fitting ARIMA model. Same as ARIMA, auto arima function is used and additionally, Xreg argument is used for the predictor variables. In this case, 14 Xreg arguments have been used which are pollutants concentration as well as weather data such as temperature, pressure etc. For this data set, it fits regression model with ARIMA errors (3,1,5). Generally, arima coefficients are not easy to interpret but regression coefficients can be interpreted. Changes in all the variables with respect to PM2.5 can be observed from the value of the coefficient. For example, change in PM10 increases by 4 % when PM2.5 concentration increases by 1%. All the arima and regression coefficients can be seen from the summary of the fit. In this model regression takes care of predictors while ARIMA takes care of short-term dynamicsZavala et al. (2016). Future values have passed in the xreg arguments for all the 14 predictors and forecasted value can

be checked for PM 2.5. Xreg argument passed in the auto arima function for predictor variable is Xreg (12.43,121.47,56.4,816.67,10.7,580,26.3,381.33,1.43,128.2,65,0.3,3.53, 5.3) These values from 12.43 to 5.3 are the value of predictor variables in the time series data set. Residual check gives p-value above 0.05 which shows residual is white noise and model is good for the data set.

DHR uses Fourier term in which sin and cosine functions are used. Sin and Cosine can approximate any frequency into the periodic frequency. Zavala et al. (2016). These frequencies are called harmonic frequency . Arima error in the model is non-seasonal so seasonal component will be set to False. After the box-cox test lambda value achieved is 0.42 which is used here in the argument while fitting the model. DHR works same as ARIMAX model except that it takes xreg argument as Fourier term. All the predictors will be considered as Flourier frequencies. K value is selected to minimize AICc and with K value 6, minimum AICc value observed. In the forecast function, k value 6 with h value as 288 is passed, which instruct the model that value for future instances are predicted, not for the past instances. This analysis is done on sub-daily data so the use of DHR is very efficient here. Residual check gives P-value of 0.74 which is quite above the 0.05 and residual is a white noise and model is good for the predictions.

4.2 Implementation Of Classification Models

Rapid miner is used for implementing all the classification models for this research. KNN is one of the simplest algorithm among all the machine learning models. Euclidian distance has taken for building the model and K value is taken as 7.

Choosing k value is a critical part of fitting the model. Several trial and error methods were performed where k =1, 3, 7, 9, 11, 13. In many cases, by selecting K value very small(for example '1'), over fitting occurred. Irrespective of the good accuracy model is of no use. One of the good practice for selecting K should be near to square root of training sample. Straight away this values cant be selected some nearer values should also be tried. There are approximately 200 trainig samples used for classfication analysis so K valu is selected as 7. Squared Euclidian, Chi-Square, correlation, Euclidian are some examples of distances Sam and Sp (2011). Among which Euclidian is most popular, because of its simplicity it takes very less time to train the data set.That is the reason behind selecting Euclidian distance in this research. Using cross-validation performance increased by 10%.

$$\text{Euclidean Distance (x,y)} = \text{squareroot}((\text{x2} - \text{x1})^2 + (\text{y2} - \text{y1})^2 - - - - - - -)$$

ANN is also very popular black box algorithm. While implementing ANN labels were changed from string to numeric values and the missing operator is selected. Due to different ranges in the columns, complete data has normalized, and shuffling has implemented. Shuffling is necessary to bring all variable at comparable state. Different training cycles have tried on the data set from 2000 to 10000 with a different learning rate of 0.3 to 0.6. Even after increasing the learning rate performance was not improved. Training cycle had a direct impact on the performance of the neural network, but the time taken increased rapidly. While applying 10-fold cross-validation. The automatic sampling process is selected, rapid miner selects the best sampling suitable.

Several combinations of a hybrid model like KNN- RF, ANN-Naive, KNN-SVM, decision tree-Knn have been tried but the combination of Grdient boosting and KNN improves performance drastically. For random forest, minimal gain taken was 0.1, leaf size

taken is 2 and size of split selected as 4. Three level pre-pruning was selected over post pruning as the time taken was more while doing post pruning and performance was not improved. In our model different values of learning rate and sample rate is selected and auto distribution has chosen. Model is implemented using a different number of trees from 10-30. Even after increasing number of trees, model performance was not increased and this method was not included in the evaluation. Boosting method dramatically improves model accuracy, based on the concept of averaging many rough rules of thumbs Elith et al. (2008) . Boosting diminishes bias of many small models with low variance. Maximum depth is taken as 5 with a learning rate of 0.1, along with it KNN is selected. Again value of k is selected as 7 as mentioned above.

4.3 Model Comparison

In the time series analysis, the base error measures were Root Mean Squared Error(RMSE) and Mean Absolute Percentage Errors (MAPE). Some other measures, Mean Squared error (MSE) and Mean Absolute Errors(MAE) have not been considered. MAE and MSE depends on the scale of the data. MAE is not good for 0 or small values. MAPE and RMSE of the models have been compared for the accuracy of time series model. For comparison among all the four models RMSE and for comparison between same type of models (univariate and multivariate) MAPE is selected. MAPE is good measure for the data set with same scales. Rob J (2011)

Residual values, ACF plots are also checked for the performance with BOX test and AICC value where needed. All measure can be misleading if time series value is not of constant frequency over the time. Equations for RMSE and MAPE are :

Let y_i denote the ith observation and $\hat{y}_i$ denote a forecast of y_i
forecast error $e_i = y_i - \hat{y}_i$
percentage error $p_i = 100e_i/y_i$

MAPE =mean($|p_i|$)

RMSE=$\sqrt{\text{mean}(e_i^2)}$

For classification models, confusion matrix is the base to calculate accuracy. An accuracy is a number of correctly classified instances. These metrics can be misleading in case of the imbalanced data set that is why, error, precision, and recall are also analyzed. Accuracy = (TP + TN) / (TP + TN + FP + FN)

4.4 Validation

For the validation of time series models 80-20 split was used and rolling forecasting origin (time series) cross-validation has been selected. Suitable plots and demographic statistics have been selected where needed.

For classification models Cross-validation is performed using 10-fold cross-validation. All models without cross-validation and with cross-validation are implemented. After implementation of cross-validation there is significant increase in the performance of all the models. This shows the importance of cross-validation, also it performs testing on different parts of the data set which gives a very general result.

5 Evaluation

5.1 Time Series Model Evaluation

Models	RMSE(%)	MAPE(%)	Training Time(Sec)
ARIMA-GARCH	17.503	6.87	11
TBATS	17.41	7.35	72
ARIMAX	17.23	8.46	85
DHR	17.67	6.92	90

Performance Of Time Series Models

Forecast errors are the difference between observed value and point forecast. From the above table, it can be clearly seen that ARIMAX performs best among all the models as the RMSE value is the lowest (17.23). One reason for this, ARIMAX considered the relationship among all the variable and selects p and q value to reduce AICc. Also, upon analysis of regression coefficients, CO shows the best relationship with respect to PM2.5. On the other hand, DHR performance is poorest among all models. It considered 14 factors as Fourier frequencies for prediction, this makes model more complex, result in poor performance. Also, seasonal patterns are not clear and selecting the Fourier terms by increasing or decreasing the value of K, is difficult to see the AICc value is reducing or not. In terms of performance time also, ARIMAX outperforms TBATS and DHR. TBATS considers many combinations of models and DHR converts all frequencies in sine and cosine which is the reason for the increase in training time. ARIMA is the best model in terms of training the model with 11seconds.

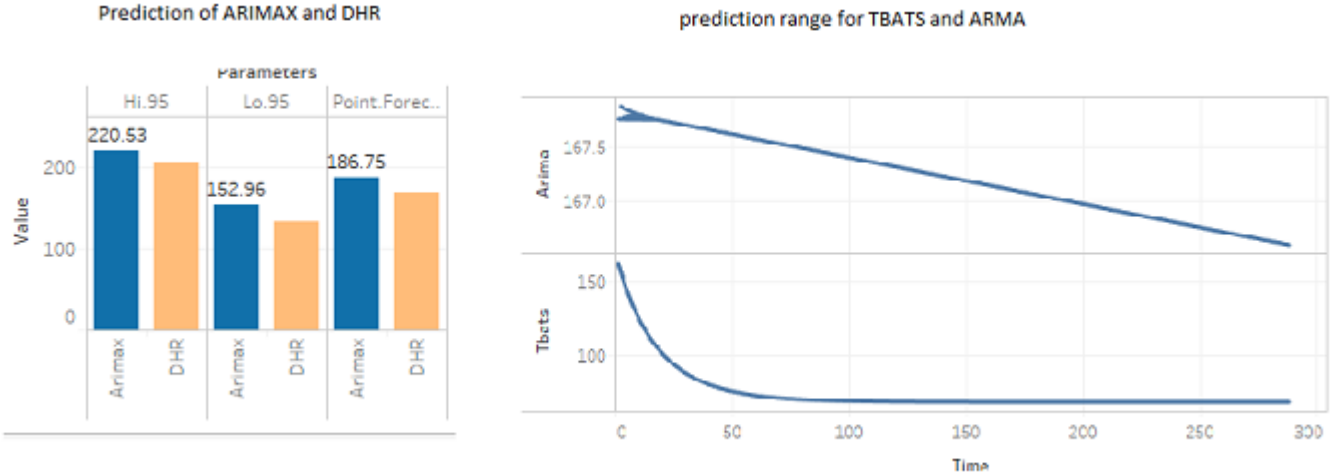

Figure 8: Performance Comparison

For univariate time series, ARIMA performs better than TBATS model which has lower MAPE of 6.87%. TBATS selects all the combination of different parameters with automatic choices, which sometimes are not very appropriate, result in lower performance. TBATS forecast graph of next 3 days shows at first value of PM2.5 decrease from 180 to 60 for the first day and then gets constant for next 2 days as shown in the Figure 8. Forecasting value for ARIMA shows a constant decrease over the 3 days of time. From the analysis of forecasting result it can be observed that prediction width of High 80% and low 80% value is approximately 45 for ARIMA but for TBATS it is comparatively higher, varying from 60 -110. One reason for this can be considering many terms and for prediction makes window width broader for TBATS.

For Multivariate time series analysis, DHR performs better than ARIMAX in terms of error with lower MAPE. One reason behind this can be an appropriate selection of terms to be included which is K. Second reson could be all variables approximated as periodic frequencies. There is not huge performance time gap between both models, high value and low-value prediction has not a big difference for both models with an approximate window length of 70(figure 8). In both models, same number of predictor variable has chosen. The only difference is that DHR consider these frequencies as Fourier frequencies and it applies multiple regression with ARIMA error. Training time is almost same for both the models. So, the overall performance of DHR is better than ARIMAX in multivariate time series analysis.

5.2 Classification Model Evaluation

Models	Accuracy(%)	Error(%)	Execution Time(Sec)
KNN	76	24	2
ANN	86.5	13.5	146
Ensemble	94.7	4.2	5

Performance Of Classification Model

Ensemble classifier outperformed among all with the highest accuracy and lowest error. KNN performs with the lowest accuracy and highest error. One reason behind that is KNN does not perform very good when dimensions are high, as in this research 14 variables are selected for the analysis. To get better performance, other distances then Euclidian can be considered which is a part of future work. One benefit of this algorithm is that it has very fast training phase but takes more memory. ANN takes the highest time for execution which is 146 seconds but provides quite a good accuracy of 86.5% with a low error rate of 13.5%. It takes more time as training cycle increases, it learns from the data and makes no assumption about it. Weighted recall which is a number of positive cases correctly predicted is quite good for ANN. Increased execution time and lower performance than ensemble model raise the question of the value of using an ANN model over ensemble classifier. As mentioned by (Ali and Tirumala; 2016) in the literature hybrid models performs better than traditional classifiers which has been confirmed by the above result with the highest accuracy of 94.7 % and lowest error measure of 4.2%. Boosting are the algorithms which change weak learners to strong and prediction are combined the weighted majority of both KNN and boosting, that is why it outperformed traditional classification models. In any classifier accuracy and error, measures should be satisfactory as a result use of hybrid models is justified.

6 Conclusion and Future Work

The main objective of this study is to implement several time series models and classification models for air pollution prediction and compare the performance of models. PM2.5 has been predicted with the highest accuracy by ARIMAX model (lowest RMSE). PM2.5 concentration for future three days has been successfully predicted using all the 4 time series models. For univariate time series ARIMA-GARCH and for multivariate time series DHR performs better with lower MAPE values. Despite, the fact that DHR performs worst among all the models(if RMSE is taken the only measure for the decision), it has better performance in its own category with lower MAPE. Use of TBATS model was relatively new approach in air pollution prediction. Among all the classification models, Ensemble model performs the best with the highest accuracy which confirms the finding from many researchers as discussed in the literature.

Limitations of this study is that, models have not implemented in the real-time for predicting into future. Feature selection is considered from the literature, but feature engineering algorithms have not been implemented in this research, which may be explored for future work. Hybrid time series models and deep learning models can be implemented in future to predict all the pollutants concentration (SO2, NO2, PM10 and PM2.5). From the concentration of pollutants, AQI value will be calculated. Based upon AQI future values can be classified that, how polluted air will be based on point forecast. Further deep researches can embrace more explanatory factors in the model. This would elongate to improve models predictive accuracy.

Acknowledgment

Firstly, I would like to express my sincere and faithful gratitude to my supervisor Sean McNally for support he has given me during my Masters thesis. His guidance helped me learn immensely during this short time. I would like to thank all the data analytics faculty. Finally, I would like to thank my parents and friends who supported me during this journey.

References

Ali, S. and Tirumala, S. S. (2016). Performance Analysis of SVM ensemble methods for Air Pollution Data, pp. 211–215.

Anenberg, S. C., Belova, A., Brandt, J., Fann, N., Greco, S., Guttikunda, S., Heroux, M.-e., Hurley, F., Krzyzanowski, M., Medina, S., Miller, B., Pandey, K., Roos, J. and Dingenen, R. V. (2016). Survey of Ambient Air Pollution Health Risk Assessment To, **36**(9): 1718–1737.

Anikender and Goyal, P. (2011). Forecasting of air quality in del using principal component.

Benvenuto, F. and Marani, A. (2000). Neural networks for environmental problems : data quality control and air pollution nowcasting , **2**(3): 281–292.

Bougoudis, I., Demertzis, K. and Iliadis, L. (2016). EANN HISYCOL a hybrid compu-

tational intelligence system for combined machine learning : the case of air pollution modeling in Athens, pp. 1191–1206.

Broeck, J. V. D., Cunningham, S. A., Eeckels, R. and Herbst, K. (2005). Data Cleaning : Detecting , Diagnosing , and Editing Data Abnormalities, **2**(10): 966–971.

Buzzi-ferraris, G. and Manenti, F. (2011). Outlier detection in large data sets, *Computers and Chemical Engineering* **35**(2): 388–390.
URL: *http://dx.doi.org/10.1016/j.compchemeng.2010.11.004*

Castro, J. R. (2008). A Hybrid Learning Algorithm for Interval Type-2 Fuzzy Neural Networks in Time Series Prediction for the Case of Air Pollution *, pp. 14–19.

Catalano, M., Galatioto, F., Bell, M., Namdeo, A. and Bergantino, A. S. (2016). Environmental Science & Policy Improving the prediction of air pollution peak episodes generated by urban transport networks, *Environmental Science and Policy* **60**: 69–83.
URL: *http://dx.doi.org/10.1016/j.envsci.2016.03.008*

Chiwewe, M, T., Ditsela and Jeofrey (2016). Machine Learning Based Estimation of Ozone Using Spatio-Temporal Data from Air Quality Monitoring Stations, pp. 58–63.

Deters, J. K., Zalakeviciute, R., Gonzalez, M. and Rybarczyk, Y. (2017). Modeling PM 2 . 5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters, **2017**.

Elith, J., Leathwick, J. R. and Hastie, T. (2008). A working guide to boosted regression trees, (Ml): 802–813.

Elman, J. (1990). Finding Structure in Time, **211**: 179–211.

Goel, R., Gani, S., Guttikunda, S. K., Wilson, D. and Tiwari, G. (2015). On-road PM 2 . 5 pollution exposure in multiple transport microenvironments in Delhi, *Atmospheric Environment* **123**: 129–138.
URL: *http://dx.doi.org/10.1016/j.atmosenv.2015.10.037*

Ip, W. F., Vong, C. M., Yang, J. Y. and Wong, P. K. (2010). Least Squares Support Vector Prediction for Daily Atmospheric Pollutant Level, pp. 1–6.

Jair, H., Palacios, G., Andrés, R., Toledo, J., Albeiro, G., Pantoja, H. and Martínez, Á. A. (2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change, **2**(3): 598–604.

Khoshsima, M., Ahmadi-givi, F., Bidokhti, A. A. and Sabetghadam, S. (2014). Impact of meteorological parameters on relation between aerosol optical indices and air pollution in a sub-urban area, **68**: 46–57.

Kong, H. (2017). Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine : The Case of, pp. 1–19.

L-stern Group, L.-S. (2010). Time Series Analysis with ARIMA ARCH / GARCH model in R, pp. 1–19.

Lee, M. H., Rahman, N. H. A., Latif, M. T., Nor, M. E., Kamisan, N. A. B. et al. (2012). Seasonal arima for forecasting air pollution index: A case study, *American Journal of Applied Sciences* **9**(4): 570–578.

Livera, A. M. D., Hyndman, R. J. and Snyder, R. D. (2010). Forecasting time series with complex seasonal patterns using exponential smoothing Forecasting time series with complex seasonal patterns using exponential smoothing, (October).

Ncr, P. C., Spss, J. C., Ncr, R. K., Spss, T. K., Daimlerchrysler, T. R., Spss, C. S. and Daimlerchrysler, R. W. (2000). Crisp-dm 1.0.

Ni, X. Y., Huang, H. and Du, W. P. (2017). Relevance analysis and short-term prediction of PM 2 . 5 concentrations in Beijing based on multi-source data, **150**: 146–161.

Piatetsky-Shapiro, G. (2014). Kdnuggets methodology poll.

Rabeb, F., Souhir, B. and Abdennaceur, K. (2017). Ozone monitoring using SVM and KNN.

Raimondo, G., Montuori, A., Moniaci, W., Pasero, E. and Almkvist, E. (2011). A machine learning tool to forecast pm 10 level.

Rob J, H. (2011). Forecasting: principles and practice.
URL: *https://www.otexts.org/fpp/8*

Sam, S. and Sp, S. U. A. (2011). Distance concepts.

Santoso1, B., Wijayanto1, H., Notodiputro1, K. A. and Sartono, B. (2017). Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review.

Shaban, K. B., Kadri, A. and Rezk, E. (2016). Urban air pollution monitoring system with forecasting models, *IEEE Sensors Journal* **16**(8): 2598–2606.

Sharma, M. and Bhattacharya, A. (2015). National Air Aqulitu Index Central Pollution Control Boar.

Singh, K. P., Gupta, S. and Rai, P. (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods, **80**: 426–437.

Sudarsan, J., Maurya, D., Singh, R. and Feroz, O. M. (2010). Role of weather data in validating air quality models, *Recent Advances in Space Technology Services and Climate Change (RSTSCC), 2010*, IEEE, pp. 50–54.

Tikhe, S. S., Khare, K. C. and Londhe, S. N. (2014). Short Term Air Quality Forecast Using Data Driven Approaches, **4**(1): 224–236.

Xi, X., Wei, Z., Xiaoguang, R., Yijie, W., Xinxin, B., Wenjun, Y. and Jin, D. (2015). A Comprehensive Evaluation of Air Pollution Prediction Improvement by a Machine Learning Method, pp. 176–181.

Zavala, A. J., Messina, A. R. and Framework, A. T. M. (2016). Dynamic Harmonic Regression Approach to Wind Power Generation Forecasting.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Zhu, L. and Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM 2 . 5 concentrations across China using a spatially explicit machine learning algorithm, **155**: 129–139.

Zhang, C., Yan, J., Li, Y., Sun, F., Yan, J., Zhang, D., Rui, X. and Bie, R. (2017). Early Air Pollution Forecasting as a Service : an Ensemble Learning Approach, pp. 636–643.

Zhang, R. (2007). SAS Global Forum 2007 Data Mining and Predictive Modeling Regression, Auto-Regression, Dynamic Regression —– A Practical Modeling Example in Financial Industry, pp. 1–12.