ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Waleed Khalid Mahmood MAHMOOD

INVESTIGATION OF MACHINE LEARNING METHODS FOR PREDICTION OF OZONE CONCENTRATION TO DETERMINE OUTDOOR AIR QUALITY LEVEL

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

ADANA-2021

ABSTRACT

MSc THESIS

INVESTIGATION OF MACHINE LEARNING METHODS FOR PREDICTION OF OZONE CONCENTRATION TO DETERMINE OUTDOOR AIR QUALITY LEVEL

Waleed Khalid Mahmood MAHMOOD

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

Supervisor : Asst. Prof. Dr. Ercan AVŞAR
Year: 2021, Pages: 85
Juries : Asst. Prof. Dr. Ercan AVŞAR
: Asst. Prof. Dr. Ahmet AYDIN
: Asst. Prof. Dr. Gökay DİŞKEN

Due to an increase in the field of industrialization and urbanization that the world witnessed in recent years, the amount of air pollution is increased in various countries. This situation constitutes a risk for the humans especially those with respiratory diseases. Therefore, prediction of the air pollution level is an important task that may be useful for developing early warning systems. In this thesis study, performances of six different machine learning methods have been investigated for predicting outdoor ozone concentration in Istanbul, Turkey. These methods are support vector machines, multilayer perception, random forests, gradient boosted decision trees, k-nearest neighbor, and elastic net. The predictions are made according to one-step-ahead time-series forecasting scheme with a rolling walk forward validation. Six major atmosphere components that are based on air quality index readings, namely, SO_2, PM10, O_3, PM2.5, NO_2, and CO that are utilized to train the models. Firstly, prediction performance of these methods are calculated by training regression models. Next, the predicted values are converted into one of the five air quality index levels. This conversion makes it possible to calculate performance metrics regarding a multi-class classification problem. The lowest RMSE of 2246 was observed by using MLP and the highest accuracy score of 0.790 was observed by SVM.

Keywords: Machine Learning; Time-Series Forecasting; Regression Methods; Air Quality Index.

ÖZ

YÜKSEK LİSANS TEZİ

DIŞ ORTAM HAVA KALİTE SEVİYESİNİ BELİRLENMESİNDE OZON KONSANTRASYONU TAHMİNİ İÇİN MAKİNE ÖĞRENME YÖNTEMLERİNİN İNCELENMESİ

Waleed Khalid Mahmood MAHMOOD

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

Danışman : Dr. Öğr. Üyesi Ercan AVŞAR
Yıl: 2021, Sayfa: 85
Jüri : Dr. Öğr. Üyesi Ercan AVŞAR
: Dr. Öğr. Üyesi Ahmet AYDIN
: Dr. Öğr. Üyesi Gökay DİŞKEN

Son yıllarda dünyanın tanık olduğu sanayileşme ve kentleşme alanındaki artış nedeniyle çeşitli ülkelerde hava kirliliği miktarı artmaktadır. Bu durum insanlarda özellikle solunum yolu hastalığı olanlar için risk oluşturmaktadır. Bu nedenle hava kirliliği seviyesinin önceden tahmin edilmesi erken uyarı sistemlerinin geliştirilmesi konusunda faydalı olabilir. Bu tez çalışmasında, İstanbul, Türkiye'deki açık hava ozon konsantrasyonunu tahmin etmek için altı farklı makine öğrenme yönteminin performansları incelenmiştir. Bu yöntemler, destek vektör makineleri, çok katmanlı algılama, rastgele ormanlar, gradyan artırılmış karar ağaçları, k-en yakın komşu ve elastik ağdır. Tahminler, ileriye doğru yuvarlanan doğrulama metodu ile bir adım ileri zaman serisi tahmin şeklinde yapılır. Bu modellerin eğitimi için hava kalitesi indeks okumalarına dayanan altı ana atmosfer bileşeni olan SO_2, PM_{10}, O_3, $PM_{2.5}$, NO_2 ve CO ölçümleri kullanılmıştır. Öncelikle bu yöntemlerin tahmin performansı, regresyon modelleri eğitilerek hesaplanır. Daha sonra, tahmin edilen değerler beş hava kalitesi indeks seviyesinden birine dönüştürülür. Bu dönüştürme, çok sınıflı bir sınıflandırma problemine ilişkin performans ölçütlerinin hesaplanmasını mümkün kılar. En düşük RMSE 2246 ile MLP kullanılarak, en yüksek doğruluk puanı 0.790 ise SVM tarafından gözlenmiştir.

Anahtar Kelimeler:Makine Öğrenmesi; Zaman Serisi Tahmini; Regresyon Yöntemleri; Hava Kalitesi İndeksi.

EXTENDED SUMMARY

Recently, the world witnessed a remarkable development in the field of industrialization and urbanization, and the development is still taking place in clear steps forward. Usually, any developments are initially associated with risks that requires observation and control. One of these common problems associated with the development of industrialization and urbanization is ambient air pollution and its negative effects on life of citizens. Ambient air pollution (AAP) is an obvious mark of sustainable development because sources of AAP are responsible for producing climate pollutants such as Carbon dioxide (CO_2) or black carbon (BC). Recently, AAP has become a concern of most specialists and researchers, especially health organizations that claim to preserve the outdoor environment. Air pollution is considered the biggest environmental risk to health, as it occupies the fourth place on the list of the most common causes of death after tobacco, rising blood pressure and poor dietary habits. Air pollution is a major factor in the worsening of respiratory symptoms; in addition, it is one of the main causes of asthma attacks. Therefore, several of the researchers have sought to contribute to an investigation of air pollution dangerously and its relation with industrialization and urbanization. Therefore, prediction of the air pollution level is an important task that may be useful for developing early warning systems that are useful for the sensitive groups about the air pollution. This work evaluates the performances of six different machine learning methods for predicting the maximum outdoor ozone concentration for the next day.

In this study, an air quality dataset, which contains air pollution measurements for five years in Istanbul, is utilized to train the machine learning models. The utilized dataset consists of six major components of the atmosphere, namely, Sulfur Dioxide (SO_2), PM10, Ozone (O_3), PM2.5, Nitrogen Dioxide (NO_2), and Carbon Monoxide (CO) that was arranged on a sequence of time. Time series forecasting method is used to forecast the proposed time-series data. The

proposed data are preprocessed in many steps to be appropriate with time series forecasting method. The final obtained version of the dataset after preprocessing steps have consisted of 1013 row and 56 columns.

Six regression models based on machine learning methods are utilized to train the proposed TS dataset, namely, Support Vector Machine (SVM), Multilayer Perception (MLP), Random Forests (RF), Gradient Boosted Decision Trees (GBDT), k-Nearest Neighbor (K-NN), and Elastic Net (EN). To train and test the proposed models, the rolling walk forward validation method is used.

The first step, one-step-ahead are performed by using these models for measuring the regression performance. For the regression problem, the proposed models achieved approximate performance values. However, MLP model outperformed other proposed models, with the lowest RMSE score of 2246. EN model is the alternative model in the regression method, which has the second best performance after MLP model.

Next, the predicted values by these models with the observed data are converted to corresponding air quality levels by using US EPA formula. The values converted are labelled to the first five classes from the six classes of AQI standers of US EPA. The five classes are used to generate a multi-confusion matrix, which in turn allows to generate measures for the classification performance. For the classification problem, linear and non-linear SVM model is outperformed other proposed models, where, shown the highest accuracy score of 0.79.

ACKNOWLEDGMENTS

I want to express my deepest gratitude to Asst. Prof. Dr. Ercan AVŞAR for his supervision, orientation, motivation, constructive advice, and continued faith in me in this thesis. It was a huge honor for me to have Asst. Prof. Dr. Ercan AVŞAR as a supervisor.

I would like to thank to members of my thesis jury members, Asst. Prof. Dr. Ahmet AYDIN and Asst. Prof. Dr. Gökay DİŞKEN for all respectable knowledge and experience.

I want to share my sincere thanks and gratitude to the Turkish Government and people for giving me this opportunity and for their hospitality.

I would like to express big gratitude to my friend Mahmood ABED for helping me to get admission to Çukurova University and motivation me in all desperate moments.

I would like to express big gratitude to my best friend Hasan ALANTAKE for being beside me like a guardian angel during the exacerbation of my disease symptoms.

Finally, I wish to express my heartfelt thankfulness to my family for their patience, encouragement, and continuous moral support.

CONTENTS **PAGE**

LIST OF TABLE **PAGE**

LIST OF FIGURE **PAGE**

ABBREVIATIONS

AI	: Artificial Intelligence
ML	: Machine learning
AAP	: Ambient air pollution
CO	: Carbon dioxide
BC	: Black Carbon
WHO	: World Health Organization
OECD	: Organization for Economic Co-operation and Development
COPD	: Chronic Obstructive Pulmonary Disease
NO2	: Nitrogen Dioxide
PM	: Particulate matter
SO2	: Sulfur Dioxide
GAN	: Global Asthma Network
TSF	: Time-Series Forecasting
TS	: Time-Series
KNN	: K-nearest neighbor
SVM	: Support Vector Machines
GBDT	: Gradient Boosted Decision Tree
RF	: Random Forests
MLP	: Multilayer Perceptron
ANN	: Artificial Neural Networks
EN	: Elastic Net
US EPA	: United States Environmental Protection Agency
AQI	: Air Quality Index
CNB	: Classification Naive Bayes
LDA	: Linear Discriminant Analysis
StS	: Sequence-to-sequence
DL	: Deep learning

LSTM : Long Short-Term Memory
PCA : Principal Component Analysis
SARIMA : Autoregressive Integrated Moving Average
FTS : Fuzzy time series
RNN : Recurrent Neural Network
ARMA : Autoregressive Moving Average
MTL : Multi-Task Learning
BP-NN : Back-Propagation Neural Network
MLR : Multiple Linear Regression
ANFIS : Adaptive Neuro-Fuzzy Inference System
RMSE : Root-Mean-Square Deviation
BRTs : Boosted Regression Trees
FI : Feature Importance
EST : Echo State Network
MAE : Mean Absolute Error
FCMs : Fuzzy Cognitive Maps
EEG : Electroencephalogram
SWS : Scalar Wind Speed
WAQIP : World Air Quality Index Project
SD : Standard Deviation
WFV : Walk-Forward Validation
SVR : Support Vector Regression
RTs : Regression trees
LASSO : Least Absolute Shrinkage and Selection Operator
RBF : Radial Basis Function

1. INTRODUCTION

1.1. Ambient air pollution and sources

Air pollutants sources are classified into four main types according to the way they are configured at atmosphere. Air pollution sources in the atmosphere may be either formed by itself or emitted into it. The first type of source is primary air pollutants that are emitted into the atmosphere by exhaust pipe or chemical factory. The secondary air pollutants are formed by themselves within the atmosphere that arises as a result of chemical reactions of primary air pollutants. Gaseous air pollutants are the third source that presents as gases or vapor. The last source, particulate air pollutants include material in liquid or solid phase pending in the atmosphere (Lodgejr, 1996).

In recent years, the world has witnessed a massive increase in the field of industrialization and urbanization in addition to untapped population growth. As a consequence of this situation, there has been a negative impact on indoor and outdoor air quality. Ambient air pollution (AAP) is an obvious mark of sustainable development because sources of AAP are responsible for producing climate pollutants such as Carbon dioxide (CO_2) or black carbon (BC). Ambient air pollution is an international climate problem that is concerned with the researchers and specialist, which influences especially the health of the urban citizens. According to the World Health Organization (WHO), air pollution is considered to be the biggest environmental risk to health, as it occupies the fourth place of the list of the most common causes of death after tobacco, rising blood pressure and poor dietary habits (Health Organization, 2016). As stated by a recent survey, global mortality hazard factor of ambient air pollution is considered 5 million annually. Despite that, there are more than 90% of the world's population does not breath air in quality correspond of health-based guidelines (Vital Strategies, 2021). Air pollution monitoring capacity varies widely among from one city to another according to economy of the countries. Some high-income countries can be

establishing continuous high air pollution monitoring system of all their cities, but some low-income countries such as Africa and the Middle East are facing difficulty in establishing a system to monitor pollution (Vital Strategies, 2021). Thus, citizens in Africa or the Middle East suffered particularly much severe levels of pollutants than those high-income cities. According to WHO survey about the impact of ambient air pollution, in 2012 there is 3 million premature death worldwide equivalent to one death for every nine deaths due to ambient air pollution (Health Organization, 2016). Furthermore, an increase in the number of deaths was observed where the number of deaths rose to 4.2 million premature death worldwide, were about 87% of these deaths happen in Low- and Middle-income Countries (LMICs) (Health Organization, 2016). Hence, this rise in death confirms the survey that indicates an increase in the number of premature deaths 40% from ambient air pollution in the next two decades. Therefore, the goal of the United Nations Sustainable Development (UNSD) is developing a fortified strategy for the next two decades to limit air pollution hazards (Rafaj et al., 2018). Besides, the predictions of Organization for Economic Co-operation and Development (OECD) confirm the continuation of air pollution rising towards 2050, and it will be the prime reason for environmentally related deaths worldwide (OECD, 2012). Permanent exposure to air pollutants increases the risk of exposure to serious diseases, such as respiratory diseases (Asthma and Chronic Obstructive Pulmonary Disease (COPD)), lung cancer and the diseases relating to the heart and blood vessels (Health Organization, 2009).

In recent years, the researchers have focused on the dangers of indoor air pollution and their impact on the population. The reason for that concern is utmost for the citizens like elderly and children who spent most of the time indoor, in addition to young people turn to spend most of the time indoor cause of technological development and COVID-19. Despite the concern given towards indoor air pollution, outdoor air pollution was considered the major of many essential negative effects on universal public health. The importance of outdoor air

pollution is that it is the main source of indoor air pollution. Today, there is a lack of potential to prevent the outdoor particulate matter from seeping indoor environment yet (Blondeau et al., 2005). Therefore, outdoor air pollution is considered the major of many substantial negative effects on universal public health, in particular its negative impact on Asthma and COPD patients.

The six most hazardous pollutants of the atmospheric components are (Particulate matter ($PM_{2.5}$ and PM_{10}), Nitrogen Dioxide (NO_2), Sulfur Dioxide (SO_2), Ozone (O_3), and Carbon Monoxide (CO)) and they are reported to cause serious health problems (X. Q. Jiang et al., 2016). Each of the six components has different adverse effects. The concentrations of NO_2, SO_2, and PM have a direct impact on increasing mortality (Carey et al., 2013). While the concentration of O_3 directly impacts the increased risk of perforated appendicitis; the increase was estimated at 11-22% during 3-7 days of high O_3 exposure (Kaplan et al., 2013).

1.2. Impact air pollution on lung diseases

The concept of lung diseases (respiratory diseases) refers to the inability of the tissues and organs to receive enough amount of oxygen (like Asthma, allergic rhinitis and COPD). Asthma is a chronic respiratory disease. It causes airways to get inflamed and narrow eventually yielding difficulties in breathing. Global Asthma Network (GAN) studies indicate that the number of asthma patients in the world are about 0.334 billion with attributed death annually approximately 0.383 million (Ghoshal et al., 2020). COPD is a life-threatening lung disease distinguished by symptoms of wheeze and cough and difficulty breathing. Lung diseases may be a severe obstacle to the patient's ability to carry out daily functions, especially for the child, as it may be a sufficient reason for absenteeism from school and limitations of activities.

Previous studies have shown that the causes of asthma attacks are due to genetic or environmental factors (Ambient Air Pollution) (Toskala & Kennedy, 2015). Exposure to air pollution causes severe exacerbation in asthma and COPD

symptoms patients or can trigger a new case of an asthma patient (Seaton et al., 1995) and (Feldman, 2012). A survey conducted in China's Hubei province (2013-2018) about mortality due to asthma patients by ambient air pollution, they estimated that from over 7 thousand asthma death cases there are two of death were due to ambient air pollution daily (Beamer, 2019). Besides, it was statistically shown that more than 18% of air pollution deaths in 2016 were COPD cases (Health Organization, 2018). The researchers documented that the increase in the major components of air pollution is directly proportional to the increase in the mortality of asthma patients, such as increased O_3 proportionally was associated with increases a 9% of the deaths of asthma patients (Beamer, 2019). Ozone or as called trioxygen, is considered a dangerous component, to which exposure in a short time can lead to many health symptoms on the respiratory system; in addition to reduced lung function and airway inflammation (Matteelli & Saleri, 2008). According to an article, 1.1 million premature deaths worldwide annually were attributed to O3 exposure eventually causing respiratory diseases.

In addition, the potential health risks of other air pollution components, may be listed as follows:

- PM is associated with significant aggravated asthma problem; allergic symptoms; decrease in the growth of lung function.
- NO_2 is associated with nose, eyes, and throat; and can be reason respiratory illnesses and symptoms.
- SO_2 is associated with result in breathing difficulties of asthma patients and increased respiratory symptoms in children with asthma.
- CO is associated with reducing the ability of the blood to transfer oxygen.
- In addition to more than 187 hazardous air pollutants compounds believed that are associated with a significant aggravated respiratory problem (Matteelli & Saleri, 2008).

Conclusively, ambient air pollution is a serious global issue that requires the intervention of all authorities to take action. In high-income countries, AAP is considered a direct threat to the state's economy, also, the source of several serious diseases exposed to citizens. In low- and middle-income countries, air pollution is a fundamental factor in increased mortality, especially among those with respiratory diseases. The danger of air pollution is gradually increasing with increasing industrialization; preventing industrialization is not the right solution, however, the main goal is finding appropriate solutions according to this development rapidly. Therefore, the prediction of future values is a significant step that helps in developing warning systems for sensitive people.

1.3. Time series forecasting

Due to the urbanization and the increase in awareness, citizens have requirements commensurate with the current development, the first of which is with the daily forecast of the climate. The human is more curious about the only moving thing which never stops is time. Therefore, the curiosity is about things that change with the time change. Time-Series Forecasting (TSF) is a critical concept for method of real-world applications, to make machines predict the future values of a quantity, such as predict weather changes, traffic or health status of a person. Forecasting in an appropriate time will prevent damage from the natural calamities.

TSF is a significant concept of machine learning (ML) that is neglected oftentimes. The reason it was neglected due to the hardness of handling with the time sequence. Dalrymple is the first who proved that importance of TSF in future prediction (Dalrymple, 1987). TSF is estimating actual unknown information regarding the future by fitting models on an orderly sequence of historical information. TSF difference is to add a time dimension, that means setting additional qualification and structure that providing extra information. Time series

data is a group of serious variable observations that include a series of observations ordered in time. Time series has four major components, may be listed as follows:

- Trend: is a long-term and medium movement that shows the general tendency of linear increasing and decreasing of the series data over prolonged time.
- Seasonal Variation: is a short-term movement having repeating seasonal pattern that works in an ordered and periodic manner, typically with a period of less than a year. It has a constant and recognized length.
- Cyclic Variations: are short-term variations in time series and has a cyclic pattern that operates over more than a year. It has a changeable and unknown length.
- Random or Irregular Movements: are irregular or purely random change in time series that mostly due to wars, flood or earthquakes. It has unexpected, unpredictable and uncontrollable variation (Bose & Mali, 2019).

Developing time-series forecasting systems using regression models is a primary undertaking in the field of engineering nowadays, and the study for enhancing the prediction models has never ended (Aijaz & Agarwal, 2019).

1.4. Problem statement

Awareness of ambient air pollution hazards and their impact on health of respiratory patients is increasing every day.

Therefore, several health organizations seek to provide availability of an appropriate and stable pollution-free environment for patients with respiratory disease. Besides, the responsible authorities and researchers seek to develop plans to address ambient air pollution, thus, addressing ambient air pollution occupies an

important place on the international agenda. The main purpose of treatment is not only to improve air quality, but there are many other benefits, the first of which is injury protection and enabling physical activity. Due to the advancements in technology, it became actionable to establish a system to reduce the risk of ambient air pollution more rapidly. Therefore, we aim to share that cooperation and exploit advancements to establish a lower cost integrated system of forecasting the hazards of pollution. As a result, the goal of the proposed system is applying various machine-learning methods with time-series data of air pollution to establish a future predication about the major pollutants.

In this thesis, time series forecasting (TSF) application was made on a sequence air quality dataset to predict the future values of O3 and classify these future values into one of the six air pollution levels. The TSF application is based on six different regression models, which are used for fitting time sequence dataset. These models are k-nearest neighbour (KNN), support vector machines (SVM), regression trees (gradient boosted decision tree (GBDT) and random forests (RF)), elastic net (EN) and artificial neural networks-multilayer perceptron (ANN-MLP). The proposed models are trained by the dataset that has six major pollution component collected over five years. Time sequence dataset is converted to a supervised dataset by posing O_3 as a target value of pollution to predict its next-day value.

After predicting the future value, this predicted value is classified into an air pollution level determined by The standard equation of United States Environmental Protection Agency (US EPA). The equation assigns an air quality index (AQI) to the pollutant, which is then used for determining the pollution level. The predicted concentration of pollution was labelled into five labels (good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy) to generate a confusion matrix, eventually. Although the air quality index is composed of six categories, ozone is restricted by the first five categories only.

The most important contributions of the thesis may be listed as follows:

- First, the thesis contributed to identifying and analyzing the most important steps necessary to establish an architecture of TSF.
- Thesis highlighted hazards of air pollution and it harmful to citizen with respiratory disease illustrated all the potential to build a protection model to control hazards of air pollution.
- In addition, the thesis compares the performance of many regression models and shows which one is most appropriate with TGF.
- Thesis highlighted the significance of using AQI for determining the degree of air pollution. Therefore, AQI was used to classify the observed data into five categories that were used to measure the performance of classification method.

1.5. Outline of the Thesis

The rest of this thesis is orderly as follows. Chapter 2 presents an overview of related works that are summarized for Time-Series forecasting method. Chapter 3 presents materials and methods that are used for establishing the architecture of TSF model. Chapter 4 the obtained results are presented and discussed with previous researches. Finally, Chapter 5 addressed the conclusion and future work.

2. LITERATURE REVIEW

In recent times, ambient air pollution problem has attracted the attention of many researchers. Therefore, several of these researches have focused on contributing to investigate the hazards and threats caused by air pollution and its relation with industrialization and urbanization. Most of the research activities aimed to show the potentials of technology for reducing the risk of air pollution. Thus, many of the methods were developed to solve and control an air pollution problem and most of the methods are utilized time series forecasting method.

In earlier years, the world has witnessed a remarkable development in the field of technology in addition to the reduction in the cost of computers. This development makes measuring and storing data easier and allows using ML. Bellinger et al. presented a detailed study of a wide variety of ML algorithms to mine and store dataset of air pollution and its effect on reducing pollution in the air (Bellinger et al., 2017). Kang et al. showed a future investigation on the threat of air pollution to citizens and the potential risks of gases spread in the smart city (Kang et al., 2018). They demonstrated many factors causing pollution together with corresponding solutions by ML approach based on time series forecasting. To this day, many researchers have continued to demonstrate the capabilities available in ML to eliminate the risks of air pollution. Zhu et al. illustrated the efficiency of ML approaches to predict air quality (D. Zhu et al., 2018). The proposed method refined models to forecast the concentration of air pollution based on data from previous days. The world still suffers from air pollution, especially in Tehran (the capital of Iran) where air pollution considered one of the major damage of the metropolitan such as the health issues. Therefore, Delavar et al presented studies to determine air pollutants by prediction models based on air pollution concentrations in Tehran (Delavar et al., 2019). The previous data of days, month, meteorology, nearest neighbor cities and topography were used as the input parameters. In addition, regression models were used to fit the dataset. They have

improved the error percentage to 94%. Willard et al. presented a structured and overall overview of ML framework in various fields (Willard et al., 2020). The survey presented an outline of several ML application that been performed. In addition, several ML application and method ideas were discussed for future research. All mentioned research has confirmed the efficiency of machine learning with ambient air pollution and the role was played to mitigate the pollution.

There has been increasing trend in the research efforts made on TSF algorithms particularly using statistical models. De Albuquerque Filho et al. presented an exhaustive evaluation of the time-series forecasting models (De Albuquerque Filho et al., 2013). The evaluated models utilized a time-series of air pollution concentration levels. The obtained results showed the suitability of the time-series forecasting methods for air pollution prediction. Constructing predictive models for the future requires statistical calculations for concluding the dependencies between a sequencing gap from the past and short-term future. Bontempi et al. presented an overview of one-step time-series forecasting methods for supervised learning and proved competitive of statistical models (Bontempi et al., 2013). The need for the availability of huge amounts of a time sequence of historical data was illustrated. In addition, machine-learning strategies for building the structure of time series forecasting were illustrated. Du et al. proposed a framework of deep learning (DL) with a sequence-to-sequence (StS) method to forecasting a time-series dataset (Du et al., 2018). Time-series dataset was addressed by long short-term memory (LSTM) model based on an encoder-decoder architecture. The experiment result showed a satisfying accuracy and competent to deal with the time sequence dataset.

TSF was considered as an appropriate method to predict the change in climate or air pollution. Several researchers proved its efficiency. For example, Gocheva-Ilieva et al. proposed a TSF model to indicate the higher component in the atmosphere of Blagoevgrad. Six pollutants of the atmosphere were analyzed by using the Principal Component Analysis (PCA) and Promax rotation methods

(Gocheva-Ilieva et al., 2014). The BoxJenkins stochastic autoregressive integrated moving average (SARIMA) models were applied to six pollutants to indicate the higher component in the atmosphere. present a ML model to predict ambient air pollution. The utilized dataset consisted of values collected from three different stations (suburban, residential, and industrial) (Rahman et al., 2015). Three methods were suggested to identify the most effective time-series methods. The suggested model was ARIMA, ANN, and three models of fuzzy time series (FTS). The result showed that ANN was more accurate for forecasting pollution. J. Zhu et al. proposed a hybrid-forecasting model to predict air pollution. The proposed model was established by integrating ensemble empirical mode decomposition (EEMD) and a combined forecasting model of sub-series selection (J. Zhu et al., 2018). For the reason of rising air pollution in the last several decades. Several environmental agencies were searched to forecast air pollution as solutions to save there citizen from future pollution. Reddy et al. proposed a TSF method to predict air pollution in Beijing and Delhi (Reddy et al., 2017). The proposed method was depended on an LSTM recurrent neural network (RNN) model to forecast the time-series dataset. The proposed model showed an equivalent accuracy compared with the baseline SVM. In addition, T. Liu et al. proposed the TSF model to enhance the prediction of air quality in Hong Kong (T. Liu et al., 2018). The proposed model was based on ARIMA with numerical forecasts. Time-series data was monitored from several stations in Hong Kong. The monitored dataset consisted of three essential components of air quality components were PM2.5, NO_2, and O_3. The deposition of harmful gases in the air has negative effects on the patient with respiratory dresses. Bhalgat et al. proposed an ML model to predict SO_2 concentration in the atmosphere (Bhalgat et al., 2019). Time-series model was used to forecast the SO_2 readings for a year or less.

Several cities suffer from the pollution that related to traffic especially who live near-road are most exposed. Traffic-related air pollution was presented by Carlsten et al. and was elucidated the risk of early exposure to traffic-related air

pollution associated with incident asthma a birth (Carlsten et al., 2011). One of the most effective solutions taken to mitigate traffic-related air pollution is planted tree barriers on roadsides. K. Hashad presented five ML methods to predict an appropriate location to plant trees (Hashad et al., 2020). The five methods that used to train and test dataset were a linear regression (LR), random forest, SVM, NN, and XGBoost (XGB).

Regression models in ML that are proven their efficiency to build the architecture of TSF. A large scale of the study was focused on regression problem while classification problem has no extensive study. Ahmed et al presented a comprehensive study that demonstrates a compare between the major regression models for time series forecasting (Ahmed et al., 2010). The purpose of the comparison is to clarify problems, faults and defects of the models in time series forecasting. The models ranked from best to worst were MLP, Gaussian processes, then Bayesian neural networks and SVR almost similar, then generalized regression NN and K-NN regression about tied, CART regression trees and last spot radial basis functions. In Beijing, air pollution became a health concern to citizens. Maheshwari & Lamba used supervised regression models to deal with the increasing pollution in Beijing (Maheshwari & Lamba, 2019). The six utilized regression models were decision tree (DT), ANN-MLP, K-NN, linear regression, Stochastic Gradient Descent and RF. Time-series dataset of hourly observation of air for 4 years was used for training the proposed model.

Support Vector Machine (SVM) is a supervised learning method in the ML algorithm. Primarily, SVM is used in classification tasks. SVM has been extended for regression and time series prediction due to its experimental error and a regularized code principle. Recently, several researchers rely on SVM and its role in the TSF method. Thissen et al. presented a model based on SVM to predict a time-series dataset (Thissen et al., 2003). SVM compared with the other two methods were autoregressive moving average (ARMA) and NN models. Two kinds of data set were used, the first real-world data was relatively easy and the second

data was a difficult nonlinear chaotic. For a difficult dataset, SVM has outperformed ARMA and NN models. For real-world data, the performances of all models were very close. One of the SVM studies was applied for is predicting financial based on TSF by Kim (Kim, 2003). The study was aimed to forecast the stock price index for over 9 years, and results were compared with the back-propagation NN model. The study proved that SVM is a promising alternative in financial forecasting. TSF was involved in many fields such as electric load, financial prediction and weather state. TSF cannot achieve accurately time-series data even by using familiar linear techniques. Therefore, time-series data required progressive algorithms for the forecasting process. Sapankevych and Sankar provided a comprehensive survey on TSF using ML (Sapankevych & Sankar, 2009). The survey was provided to the researcher with a brief tutorial, insight and advantages and challenges into the application of utilizing SVM. The survey was proved the potential of SVM accurate in time-series prediction, and demonstrated it outperform other non-linear forecast techniques including NN. Rubio et al. presented and evaluated a prediction model based on regression method for forecasting time-series information (Rubio et al., 2011). Least Squares SVM (LS-SVM) model with Kernel method was used to predict time-series dataset. Türkay and Demren present an application for forecasting electrical load based on SVM method. SVM was showed satisfactory results (Türkay & Demren, 2011).

For air pollution problems, Yeganeh et al. proposed a prediction model for forecasting the daily change in CO concentrations. SVM was utilized as the ideal candidate for this project. The dataset was used is an observation of CO concentrations for over 4 years in Tehran (Yeganeh et al., 2012). Due to rapid industrial and economic growth in China that companion high air pollution. Therefore, Liu et al. selected three cities in China to examine a forecasting model (B. C. Liu et al., 2017). Support Vector Regression (SVR) model was used to train the dataset. In Taiwan, Zhou et al. proposed a new framework that explores the combined Multi-output SVM with the Multi-Task Learning (MTL) method for

increasing the efficiency of TSF accuracy (Zhou et al., 2019). The dataset was selected from three stations in Taiwan, contained was an observation of PM2.5 concentrations for six years. The results demonstrate the potentate of the proposed model and its stability for handling a time-series dataset.

The most popular method in NN is MLP-NN that works on the principle of feedforward and training its data by error back-propagation (BP) algorithm. For advanced algorithms, many researchers are keen to find an appropriate tool to handle time-series dataset. One of the praised suggestion was MLP-NN, where they indicate is useful with large time-series dataset and it overcame large time-series dataset problems. Raudys and Mockus presented a comparison between MLP and ARMA method to forecast the time-series dataset (Raudys & Mockus, 1999). The dataset consisted of time-series data of real-world observation of seven economical countries. The results showed that for well-structured data MLP was better than ARMA, while for ill-structured data ARMA and MLP both of them was poor in prediction. Overall, the comparison showed preference MLP for handling the time series dataset. Shiblee et al. presented a novel approach to MLP with different error measures to forecast a time-series dataset (Shiblee et al., 2009). Deshpande proposed the ANN-MLP model for the TSF method (Deshpande, 2012). Time-series data was utilized for rainfall prediction in India where trained MLP predict rainfall successfully, and MSE reading was better than other NN. Another comparison between MLP and ARMA, where Voyant et al. presented a method based on AI to forecast a time series (TS) dataset (Voyant et al., 2014). ANN-MLP was utilized in this method; the results were compared with ARMA. The utilized dataset was TS of temperature change from 1999 to 2008 was utilized to forecast the global solar irradiation during 24 hours daily. The observed results were outperformed MLP than ARMA, where MLP of error prediction lower than ARMA by 1.3 points. In addition, Dudek presented an approach based on ANN-MLP to forecast the probabilistic electricity price (Dudek, 2016). TS dataset utilized to train the model was the price of electrical that cover three years.

In another study, MLP was utilized to forecast air pollution and the important role it plays in this area was explained. Akhtar et al. proposed a forecasting model for ambient air pollution in India (Akhtar et al., 2008). MLP was utilized to predict the concentration of PM10. The best result was observed by MLP compared to different algorithms, as it showed an overall accuracy of 98%. Baawain and Al-Serihi proposed a prediction model based on ANN to forecast air pollution (Baawain & Al-Serihi, 2014). The proposed model was trained by a dataset consist of major eight components of air pollution concentrations was monitored by stations in Oman. MLP method with BP algorithm was chosen from ANN methods to predict ambient air pollution. The results were shown well agreement between the concentrations of input and output data. In the principle of merge the characteristics of two models, He et al. present a hybrid model to overcome the limits of accuracy (He et al., 2015). MLP model combined with principal component analysis (PCA) were utilized to enhance the forecasting accuracy of the TS dataset. PM concentration during spring and winter was used as input data. The hybrid model showed a high-efficiency of the prediction accuracy of a PM concentration. Nowadays, statistics are being used frequently to predict climate change. Consequently, Choubin et al. presented a comprehensive study to compare MLP, multiple linear regression (MLR) and adaptive neuro-fuzzy inference system (ANFIS) models . (Choubin et al., 2016). The observed climate change in Iran was used as a dataset of the model. MLP showed outperformance than the other two models, where root-mean-square deviation (RMSE) was scored 0.86 and mean mean-square error was scored 0.74. Abderrahim et al. presented a prediction model based on an ANN method to forecast air quality in Algeria. MLP method was used to predict the concentration of PM10 (Abderrahim et al., 2016). The PM10 concentration was collected for over 4 years. Due to the hazards of airborne pollutants in urban, Rahimi proposed a model TSF model based on ANN for forecasting ambient air pollution (Rahimi, 2017). MLP and MLR models were used to predict the TS dataset. The dataset was used consist of NO_2 and NO_x

concentrations which collected during the two months of 2012. MLP model offer superiority over the MLR model was demonstrated.

K-nearest neighbors (KNN) is a lazy learning method used for both regression and classification methods. KNN was considered one of the classical methods has used for TSF, but KNN is not the most efficient if compared with SVM or ANNs. Using KNN with a time-series dataset was encouraged by several researchers. A novel forecasting technique was presented by Chaovalitwongse et al. (Chaovalitwongse et al., 2007). KNN model was utilized to predict the activities of the normal and abnormal brain by using an electroencephalogram observed dataset. The proposal has proven its efficiency to predict TS dataset. Ban et al. proposed a regression approach to forecast financial TS data (Ban et al., 2013). KNN model was utilized with the proposed approach. The S&P 500 stock data for nine years was used to train the proposed model. The proposed model showed well performance. In spite of the simplicity and intuitiveness of KNN in forecasting TS, only a few studies of this method subsist in TSF. Therefore, Al-Qahtani and Crone used the KNN regression model to predict electricity load (Al-Qahtani & Crone, 2013). TS dataset of the electricity load observed for over 7 years in the UK was utilized in the proposed method. The proposed model showed acceptable results. Due to the difficulty of managing and storing energy production, Wahid and Kim proposed an ML model to forecast the daily energy consumption in Korea (Wahid & Kim, 2016). KNN model was used in the proposed model to train TS of daily energy consumption. The observed accuracy results of the model were high, where showed 95.9% accurate results. In same year, Cai et al. improved the KNN model to predict short-term traffic (Cai et al., 2016). The improved KNN model was used in the improved model to forecast TS data collected from the real road segments in Beijing for 4 weeks. The improved model outperformed another four proposed models. Two years later, Priambodo and Jumaryadi presented another KNN model for forecasting short-time traffic (Priambodo & Jumaryadi, 2018). The utilized TS

data was collected from an IoT sensor in Denmark. KNN results were compared with NN results, where NN by using TS data showed a better result than KNN.

KNN was involved with researches that keen on reducing ambient air pollution risk. Dragomir used the KNN model for forecasting ambient air pollution (Dragomir, 2010). Many components of air pollution are used as input data of model, were recorded in June 2009. In the concept of integrating the characteristic of two methods to produce an enhanced hybrid model. Qin et al. presented a hybrid model based on a characteristic of AI techniques and statistical models (Qin et al., 2019). KNN and LSTM models were used to forecast air pollution TS data. The hybrid model was showed acceptable results for forecasting air pollution. Lubis et al. proposed a KNN model to forecast air pollution (Lubis et al., 2020). Principal Component Analysis (PCA) was used to enhance the forecasting accuracy of the KNN model. The utilized TS dataset contained PM10 and SO_2 attributes belonging to Pekanbaru city. The proposed model achieved accuracy results of 88%.

In decision analysis and data mining, regression trees are considered one of the most familiar ML algorithms used widely in a regression problem because of its properties like being simple and straightforward. Regression trees showed optimistically result in the TSF method that encouraged the researchers of using it to analyze TS data. We touched on in our thesis a random forest (RF) and gradient boosting (GBM) regression method being the most famous and popular in TSF. Kane et al. presented the TSF model by using RF and ARIMA to clarify which is better in forecast TS data (Kane et al., 2014). Extremely pathogenic avian influenza in Egypt was used as a training dataset in the proposed model. The results showed RF outperforming ARIMA in the predictive ability of TS data. In 2017, Tyralis and Papacharalampous presented a study to evaluate RF model in TSF (Tyralis & Papacharalampous, 2017). RF model highest forecast performance when using a minimum number of TS data variables was observed. In the principle of merge the characteristics of two models,, Chen et al. proposed a hybrid model based on the GBDT model Augmented by the LSTM model (Chen et al., 2018). Electronic

health observed data was used as TS data to training the proposed model. The hybrid model considered a simplistic model that has two properties, generalizability and accessibility that make it easy to be performed on sophisticated TS problems. In the same year, Moore et al. presented an RF model to forecast future potential injuries to Alzheimer's disease (Moore et al., 2018). RF showed acceptable results were outperforming other models.

Regression trees also play a substantial role in building a forecasting model that aims to mitigate air pollution. Carslaw and Taylor presented the first work that explores the effect of the boosted regression trees (BRTs) model on forecast air pollution (Carslaw & Taylor, 2009). NO_x concentration data collected for four years in London was utilized in the proposed model. The BRTs showed the ability to use in air pollution forecasting. One of the common dangers of air pollution is its effect on solar radiation. Sun et al. proposed a developed RF model to forecast air pollution. The proposed model is developed by three kinds of data were air pollution index, meteorological data and solar radiation (Sun et al., 2016). RF model with air pollution index dataset was showed the best result compare with other input data. Feng et al. proposed a method to forecast air pollution in China (Feng et al., 2019). Recurrent Neural Network (RNN) and RF were used to forecast a TS dataset. The major six atmospheric components (CO, SO_2, NO_2, PM10, O_3 and PM2.5) was used to train the proposed model. the complex relationships among atmospheric components were revealed by feature importance (FI) that was generated by RF. where they observed that CO is an important component for shaping O_3, PM and NO_2; relative humidity is effected directly by Dew-point deficit. In Turkey, Altinçöp and Oktay presented a TSF method to predict air pollution (Altinçöp & Oktay, 2019). RF and ANN were used to forecast daily meteorological data in Istanbul. RF method showed outperforming in results accuracy than the ANN method.

The last utilized ML regression model in our thesis that we will touch on are the elastic net (EN). Elastic net is regulated linear regression that combined two

methods the ridge and lasso. Few researchers have dealt with EN, especially in the topics of TSF and forecasting air pollution. EN model was not used as much in TSF as its counterpart of ML regression models, therefore, the lack was motivated us to contribute to use EN in TSF. Most of the cited research in this section are published after 2016.

Xu et al. proposed a new approach called elastic ESN for adapting to the multivariate TS dataset (Xu et al., 2016). Echo state network (ESN) and EN were used to establish a TS forecasting model. EN used to solve all regression problems of ESN. The unknown weights were calculated by EN. The proposed model proved efficiency and characteristics with forecasting TS data. In the principle of preserving the environment from air pollution, Shahbazi et al. presented a study for imputing the missing air quality values (Shahbazi et al., 2018). The framework of the study is based on integrate two regression models were EN and neuro-fuzzy networks. Four air pollution component were used as input data in the proposed framework and they were collected from a monitoring network for air pollution in Tehran. Statistical criteria of regression methods were calculated, RMSE was 0.88, MAE was 0.73 and R-values was 0.81. Several researchers were attracted to predict the solar radiations; due to solar energy is the renewable and clean shape of energy. H. Jiang and Dong presented an investigation on using a square-root EN to forecast solar energy (H. Jiang & Dong, 2018). The proposed model outperformed the traditional methods. In the same year, W. Liu et al. proposed a TSF method for handling short-term load forecasting (W. Liu et al., 2018). Group method of data handling (GMDH) based on the EN method was used. GMDH-EN algorithm was showed a high-efficiency for load forecasting. In addition, the way to improve the combination prediction in TS dataset based on ML methods was discussed. Recently, spread of COVID-19 has become difficult to control; therefore, Johnsen and Gao presented an ML method for forecasting daily cases of COVID-19 (Johnsen & Gao, 2020). EN was utilized in the proposed method, as a conjectural

and simple model to forecast TS data. Elastic Net COVID-19 Forecaster (EN-CoF) showed high-accuracy with the feature of accessibility and generalization.

All of the mentioned studies underline the significance of ML and regression methods in the TSF domain and the possible ways of improving the related results. For the final part of related work, some of the TSF studies aiming to forecast tropospheric ozone (O_3) pollution and highlight its potential impact on the human health are covered.

Forecasting the ozone concentration is considered one of the vital matters due to its negative impact on the atmosphere and human's health. Wang and Lu presented a TF model that aimed to forecast the daily level of O_3 concentration in Hong Kong (D. Wang & Lu, 2006). MLP model was proposed to use for forecasting the O_3 TS dataset. MLP with a novel hybrid optimization method was showed a high-efficiency of O_3 prediction. Due to the arduous to handle ozone production by phenomenological models, Chelani proposed a linear regression model to forecast of daily maximum concentration of O_3 (Chelani, 2010). SVM was used to train daily O_3 concentration, which was collected for two years in India. In Turkey, (Özbay et al. proposed an ANN model to forecast the concentrations O_3 in the lower atmosphere (Özbay et al., 2011). MLP was applied to predict the TS dataset. The proposed model showed satisfying results in O_3 prediction. In Brazil, Luna et al. (Luna et al., 2014) proposed a model to forecast the concentration of ozone. PCA was utilized to analyze an air quality dataset, which was collected for two years in Rio de Janeiro City. SVM and ANN regression models were utilized to train a proposed TS dataset. Scalar wind speed (SWS), NO and NO_x have a high impact on the concentration of O_3 was observed by PCA. In addition, ANN and SVM models showed remarkably and acceptable results. The robustness of SVM, ANN and PCA were proved for forecasting the concentration of O_3. In Sydney, N. Jiang and Riley proposed a regression tree method to forecast ground-level ozone pollution (N. Jiang & Riley, 2015). RF was used to train the TS O_3 dataset, which was collected for 5 months. RF showed

outperforming CART algorithms, which demonstrated RF is a promising regression method for forecasting air pollution. Meng presented five regression models to forecast O_3 pollution (Meng, 2019). SVM, RF, AdaBoost, DT and Logistic Regression were used to train and forecast the concentration of O_3. The TS used dataset was derived from the UCI Website (https://www.uci.org/) of three cities. SVM showed high accuracy score of 0.949 compared with other proposed models. Recently, presented an overall review of O_3 concentrations, which was illustrated the hazards of air pollution and especially O3 (Yafouz et al., 2021). The technology development impact and benefits for preventing air pollution risks were highlighted during previous years. The previous works of ML, AI, O_3 prediction models were indicated, and their significant role in past and future plans for suppression of air pollution.

The main purpose of this study is a comprehensive and simplified illustration of the formation of TSF architecture and demonstrate the important features of the TSF method to encourage researchers to use and improve for forecasting future sequence data. The secondary purpose of this study is to help the universe from ambient air pollution especially people with respiratory disease. In addition, the side purpose is to demonstrate the appropriate ML models to predict TS data. Confidently, the proposed study will make an obvious renaissance of the TSF method in the literature.

3. MATERIALS AND METHODS

In this chapter, the work steps are demonstrated; where the used dataset is covered; dataset analyzing steps are mentioned sequentially; the architecture of the proposed system and the ML used models are illustrated for regression methods. In addition, the used methods for converting the results of regression to classification are demonstrated. All used methods are listed in a systematic sequentially.

3.1. Time Series Dataset

Time-series forecasting methods provide information about the near future based on a sequence of historical information called a time-series dataset. Time-series is a group of past readings that are arrayed according to a chronological sequence that was observed for univariate or multivariate. Univariate time-series is a sequence of data that depends only on a single variable. Multivariate time series (MTS) is consist of more than one variable was arranged in a sequence of time. MTS variables depend on each other, which are called dependency (Aishwarya Singh, 2018) ,

The air quality index (AQI) dataset is used as multivariate data to train the model, which consists of observation major atmospheric components (particulate matter (PM2.5 and PM10), Nitrogen Dioxide (NO_2), Sulfur Dioxide (SO_2), Ozone (O_3), Carbon Monoxide (CO) and other components like temperature and humidity). The used AQI dataset was published by the World Air Quality Index Project (WAQIP) (http://waqi.info/) at the time of spreading COVID-19 pandemic. The published AQI dataset involves readings from 380 major cities worldwide. These readings are being collected from several stations since 2016 and the dataset is still being updated daily. Four different arithmetic columns were provided by WAQIP for each component of the used air pollution dataset. First, the daily minimum value was calculated from the readings that were observed by several stations, referred to as "*min*". In the second column, the daily maximum value of

the group of data was calculated from the readings that were observed by several stations, referred to as "*max*". Then, the arithmetic average of the daily observed data was calculated as well, referred to as "*Mean*". In the final column, the sum of the squared distances of every expression of observed daily data was calculated, referred to as "*Variance*".

In this work, the data collected from Istanbul is chosen among 380 cities to be used in the experiments due to its vast area and high population density with high liveliness, in addition, to be considered a target of most tourists. Therefore, exposed to air pollution constantly (Yurtseven et al., 2018), which indicate to be the an appropriate city for analysis of air quality. The air quality measurements between the dates 26.12.2016 and 15.10.2020 belonging to Istanbul city are used as the dataset for this study. The final version of the dataset consists of 1013 row and 56 columns after the dataset preparation steps, which are explained in the upcoming subsections, are performed. The structure of the raw dataset is given in Table 3.1.

Table 3.1. An illustrative table of dataset contents.

Date	**Country: Turkey**	**City**	**Species of components**	**Air Pollutant Species**			
				Min	**Max**	**Mean**	**Variance**
12/26/2016 12:00:00 AM	TR	IST	SO2	0.6	20.8	3.6	150.56
12/26/2016 12:00:00 AM	TR	IST	PM10	1	119	32	2965.21
12/26/2016 12:00:00 AM	TR	IST	O3	0.5	18.3	2.5	136.67
12/26/2016 12:00:00 AM	TR	IST	NO2	6.9	76.1	28.8	2093
12/26/2016 12:00:00 AM	TR	IST	CO	0.1	59.3	7.6	1997.7
12/26/2016 12:00:00 AM	TR	IST	PM25	2	751	85	61407.9
…	…	…	…	…	…	…	…
10/15/2020 12:00:00 AM	TR	IST	SO2	0.2	44.3	2.9	1895.68
10/15/2020 12:00:00 AM	TR	IST	PM10	6	89	29	3163.9
10/15/2020 12:00:00 AM	TR	IST	O3	0.1	21.7	1.9	159.96
10/15/2020 12:00:00 AM	TR	IST	NO2	2.8	109.6	28.6	4147.72
10/15/2020 12:00:00 AM	TR	IST	CO	2.7	69.1	14	3263.06
10/15/2020 12:00:00 AM	TR	IST	PM25	37	71	56	743.54

3.2. Air Quality Index

All dataset published by WAQIP was AQI values that were converted before by WAQIP. AQI is an important index that indicates the level of pollution in the atmosphere based on concentrations of pollutant at a particular period. AQI is considered the ideal way to communicate between the government and its citizens regarding daily forecasted pollution and the level of pollution. Each government agency relies on a different AQI schedule with different health implications. The used dataset was accredited by the United States Environmental Protection Agency (US EPA) standards. The standards of US EPA were divided into 6 levels, each level indicated by a colour representing the state of pollution in the atmosphere. The green colour represents a good pollution level; this is followed by yellow, which means the level of pollution is moderate. Next, unhealthy for sensitivity group is indicated by the orange colour. The orange colour represents the hazardous threshold of all agency being a danger to people with respiratory diseases. The last three colours red, pink and maroon respectively indicate that they are unhealthy for all citizens, as illustrated in Table 3.2. All stations measure and monitor the concentrations of different air pollutants and save them. The measured concentrations are converted into AQI using the formula proposed by US EPA, which shown in equation (3.1).

$$\mathbf{I_p} = \frac{I_{HI} - I_{LO}}{BH_{HI} - BH_{LO}} \left(C_p - BH_{LO} \right) + I_{LO} \qquad (3.1)$$

Here I_p represents the AQI for pollutant p, C_p is the truncated concentration of pollutant p, BH_{HI} denotes the concentration breakpoint that is greater than or equal to C_p, BH_{LO} is the concentration breakpoint that is less than or equal to C_p, I_{HI} is the AQI value corresponding to BH_{HI}, I_{LO} represents the AQI value corresponding to BH_{LO}. Table 3.4 shown the breakpoint values to be used in equation (1) and the corresponding AQI values and levels (Mintz, 2013).

Table 3.2. AQI level by US EPA.

AQI	Air Pollution Levels	Health Implications
0-50	Good	Air quality is satisfactory, no risk.
50-100	Moderate	Air quality is acceptable.
100-150	Unhealthy for Sensitivity Group	Members of sensitive groups may experience health effects.
150-200	Unhealthy	Members of the public may experience health effects, for sensitive groups will be more serious health effects.
200-300	Very Unhealthy	The risk of health effects is increased for everyone; health alert.
300-500	Hazardous	Health warning for emergency conditions.

Table 3.3. Breakpoints for the AQI

The Breakpoints that equal to AQI standers

O3 (ppm) 8-hour	PM2.5 (mg/m3) 24-hour	PM10 (mg/m3) 24-hour	CO (ppm) 8-hour	SO2 (ppb) 1-hour	NO2 (ppb) 1-hour	AQI	AQI category
0.000 - 0.054	0.0 – 12.0	0 - 54	0.0 - 4.4	0 - 35	0 - 53	0 - 50	Good
0.055 - 0.070	12.1 – 35.4	55 - 154	4.5 - 9.4	36 - 75	54 - 100	51 - 100	Moderate
0.071 - 0.085	35.5 – 55.4	155 - 254	9.5 - 12.4	76 - 185	101 - 360	101 - 150	Unhealthy for Sensitive Groups
0.086 - 0.105	55.5 - 150.4	255 - 354	12.5 - 15.4	186 - 304	361 - 649	151 - 200	Unhealthy
0.106 - 0.200	150.5 - 250.4	355 - 424	15.5 - 30.4	305 - 604	650 - 1249	201 - 300	Very unhealthy
do not define	250.5 - 350.4	425 - 504	30.5 - 40.4	605 - 804	1250 - 1649	301 - 400	Hazardous
higher AQI values (≥ 301)	350.5 - 500.4	505 - 604	40.5 - 50.4	805 - 1004	1650 - 2049	401 - 500	Hazardous

3.3. Time-Series Dataset Pre-processing

As mentioned earlier, Istanbul city data was chosen among 380 cities as an initial step. Then the six major components for atmospheric pollution are chosen for forecasting, namely, Sulfur Dioxide (SO_2), PM_{10}, Ozone (O_3), $PM_{2.5}$, Nitrogen Dioxide (NO_2), and Carbon Monoxide (CO). After performing some preprocessing steps on this data, a dataset containing the min, max, mean, and variance of the

daily observation of the indicated six pollutants for 37 months of duration is obtained eventually. In the next subsection, the required preprocessing procedures for the TS dataset are explained.

3.3.1. One-Step Ahead Forecast

The main purpose of using TSF is to predict the unknown future data by using a sequence of historical information. Generally, TSF has two forecasting strategies. The first strategy is forecasting data at the next time step, which called the next day or one-step ahead forecasting strategy. The second strategy is forecasting data at more than one time step (An & Anh, 2016; Cheng & Wei, 2010). One-step ahead forecasting strategy is applied in the proposed model. The strategy based on the concept that reframing the main dataset to generate new data columns which was derived from the original data. Let t be the original columns that are already available. So, if the original columns are shifted by one-day back $(t-1)$ that will represent yesterday. Repeating the same method for seven days $(t-7)$ will represent the last week. Then, shifting one original column but by one-day forward $(t+1)$ that will represent tomorrow or the target set. Lastly, merge all of the gained data with the original TS data to obtain an entry TS dataset with a target set to train the proposed TSF model, which illustrated in Figure 3.1. In the used dataset, yesterday $(t-1)$ rows are added to the dataset, in addition, last-week $(t-7)$ rows are added as well, moreover, represent tomorrow or the target set $(t+1)$ is added to convert the TS dataset to supervised learning method.

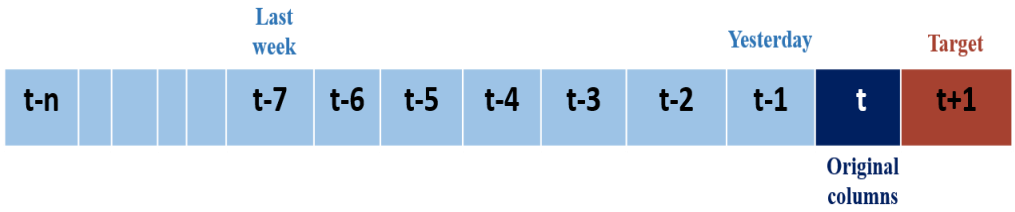

Figure 3.1. One-step Ahead strategy.

3.3.2. One-Hot Encoding

The coding method is a substantial step to preprocess the dataset to obtain satisfying results in ML. Due to most of the models do not work with categorical and float number dataset. The one-hot encoding method is the ideal solution to convert categorical and float number to numerical data that simple to handle it by ML models. One-hot encoding method uses N-bits to map the data to an N-dimensional vector, where each component of the vector is either one or zero. Here, "one" means existent while for “zero” means non-existent. In other words, only data that have the state of “one” is valid else that is non-valid (Seger, 2018). Table 3.4 illustrates a sample illustration of one-hot encoding with categorical data. As shown in Table 3.5, one-hot encoding was applied to encode the date of the proposed dataset. Therefore, an encoding vector with the length of 7-bits was used to determine the days of week. For example, "0001000" is a one-hot encoding vector indicating the dataset sample that corresponds to "Wednesday".

Table 3.4. One-Hot Encoding with categorical data.

Label Encoding

Names of people	Categorical	Age
Ahmed	1	22
Ali	2	30
Khalid	3	45

One-Hot Encoding

Ahmed	Ali	Khalid	Age
1	0	0	22
0	1	0	30
0	0	1	45

Table 3.5. One-Hot Encoding scheme for coding the day information in the dataset.

days	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
12/25/2016	1	0	0	0	0	0	0
12/26/2016	0	1	0	0	0	0	0
12/27/2016	0	0	1	0	0	0	0
12/28/2016	0	0	0	1	0	0	0
12/29/2016	0	0	0	0	1	0	0
12/30/2016	0	0	0	0	0	1	0
12/31/2016	0	0	0	0	0	0	1

3.3.3. The columns of the dataset

As mentioned earlier, the data uploaded from WAQIP was the content of 380 cities, and then the Istanbul city dataset extracted and cleaned from missed data. The final extracted data contains 24 columns, where six major components of the atmosphere (SO_2, O_3, PM_{10}, PM_{25}, NO_2, CO), and there are four readings (Max, Min, Mean, Variance) for each of these components. As a result, the extracted data 24 columns that are arranged horizontally. To let the dataset to cover the information about the date, some extra columns are added by means of one-hot encoding. Yesterday columns are added to the dataset, which are extracted from the maximum of six components by using the One-Step Ahead method, and then by using the same method, the last-week columns are added to the dataset. In other words, nineteen columns are added to the dataset that contains the encoded dataset by using the One-Hot Encoding method, where the first 7 columns are represented the day of the week (day0-day6), and the next columns are represented the months of the year (month0-month11). The final column is the target column is added by

using the One-Step Ahead method, which contains the tomorrow information (next day). Finally, the dataset contains 56 columns representing the features, where the first 24 columns represent the observed data by WAQIP. The next 12 columns represent the time dimensional information for yesterday and last week that were added by using the One-Step Ahead method. The following 19 columns represent the encoding data that were added by using the One-Hot Encoding, the final column is represented the target column is added by using the One-Step Ahead method, as shown in the Table (3.6).

Table 3.6. The explanations of the columns in the dataset.

The total content of 56 columns								
6 atmospheric components (O3, SO2, PM10, PM25, CO, NO2) × 4 measure for each one (Max, Mine, Mean, Variance)= (24 elements)				**6 elements for time dimensional of yesterday + 6 elements for time dimensional of last-week= (12 elements) (max readings)**		**7 elements for days in week encoding plus 12 elements for months encodings = 19 elements**		**Target**
O3-max	O3-min	O3-mean	O3-variance	O3-max for yesterday	O3-max for last-week	Day0 to day6 (**Wednesday**)	Month 0 to month 11 (**April**)	O3_max (next -day)
28.5	0.5	6.5	405.8	27.3	18.3	0001000 (7 elements)	000100000000 (12 elements)	27.7

3.3.4. Standardization

In general, most of the dataset contents are on different readings and scale. This means that the same data may be represented by two significantly different values collected on different scales such as 2 meters and 2,000 millimeters. In ML that is very unacceptable, where the models require all input columns of data be measured on a notionally common scale to improve the results. Two popular methods significantly were used to adjust dataset values on a common scale. The

first method is the process of scaling the data linearly in a specific range which is typically selected as [0,1]. The formula for applying this type of scaling is given in equation (3.2)

$$\boldsymbol{X_{scaled}} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3.2)$$

Where, X is represented the input feature, X_{min} the minimum value of the feature and X_{max} the maximum value of the feature.

The other scaling method, z-score normalization, rescale the input feature values such that the scaled feature vector has zero mean and unit standard deviation (SD). The formula employing the z-score normalization is given in equation (3.3).

$$\boldsymbol{z_{score}} = \frac{\mathrm{X} - \mu}{\sigma} \quad (3.3)$$

Here, X represents the input feature, μ and σ are the mean and the standard deviation values of the feature vector, respectively.

Z-score was used to rescale the proposed air pollution dataset, where the function for calculating the mean, SD and Z-score were generated. The generated function is applied to continuous feature values. In other words, one-hot encoded features were not modified by this standardization process.

3.3.5. Walk-Forward Validation

Generally, the basic steps that must be performed in the process of building ML algorithms split the train-test sets in a random way and use the K-fold cross-validation method. Each of these steps have important characteristic and benefits to improve ML model performance. These typical methods are avoided in TSF problems because the sequence of the data should not be distorted during training of a TSF model. In other words, the TSF method depends on analyzing the dataset

according to a time sequence, which means consideration of time sequence when using split and validation is a requirement. Train-test split method that esteem the time sequence was used on the TS dataset. For validation, the Walk-Forward validation method was utilized to avoid common errors. Walk-Forward Validation (WFV) method based on generating a loop with many iterations for forecasting one value (next-day value) for each iteration of the test set. The forecasted value will combine with the train set for the next forecasting in the next iteration. The prediction process continues until the end of iterations, which are equal to the number of the test set. Walk-Forward validation has two types are Anchored Walk Forward and Rolling Walk-Forward.

Anchored WFV is based on the approach of increasing the number of samples in the training set (or called optimization window) in each iteration. For every single iteration, the test value that was predicted is added to the optimization window and neglected from the test set, which means an increase in the length of optimization window and decrease in the test set (Benhamou et al., 2020). Figure 3.2 illustrates the mechanism of Anchored Walk Forward validation.

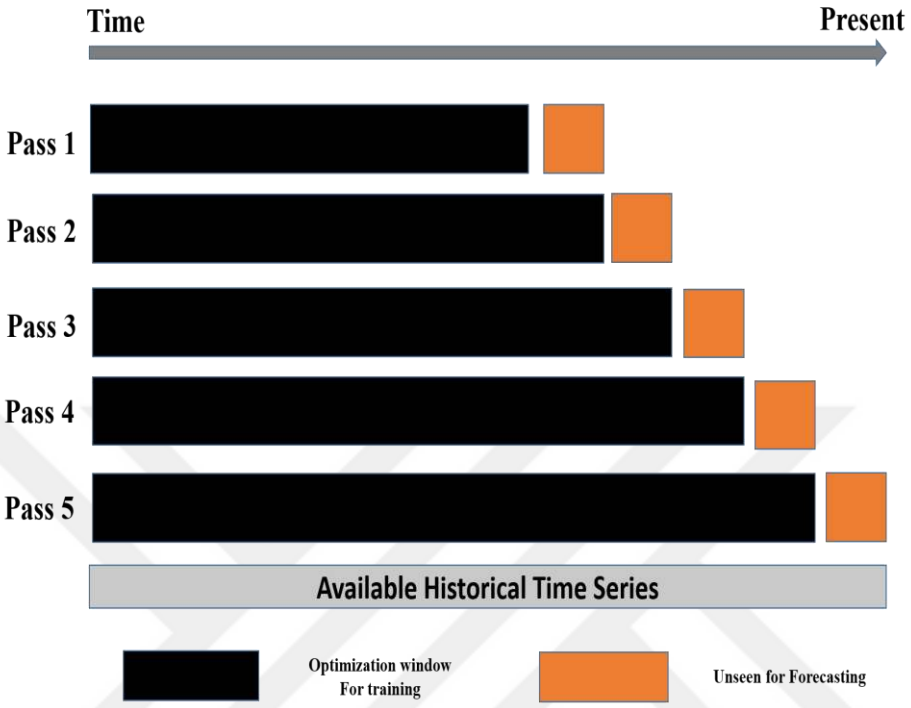

Figure 3.2. Anchored Walk Forward validation.

Rolling WFV is based on the approach of stability of the optimization window during iterations steps with a decrease in the test set. For each iteration, the test value that was predicted is added to the optimization window with the stability of optimization window size and neglected from the test set. Which means when the predicted value will be added the optimization window size will be shifted forward by one value instead of the value that was added (Podsiadlo & Rybinski, 2015). Figure 3.3 illustrates the mechanism of Rolling Walk Forward validation which is the method followed in this thesis for training and validation of the TSF models. When WFV is employed, the number of training samples is selected as 607 and the amount of samples in test set is 406. In other word, a window of size 607 samples is shifted on time axis 406 times and the models are trained and tested at every step.

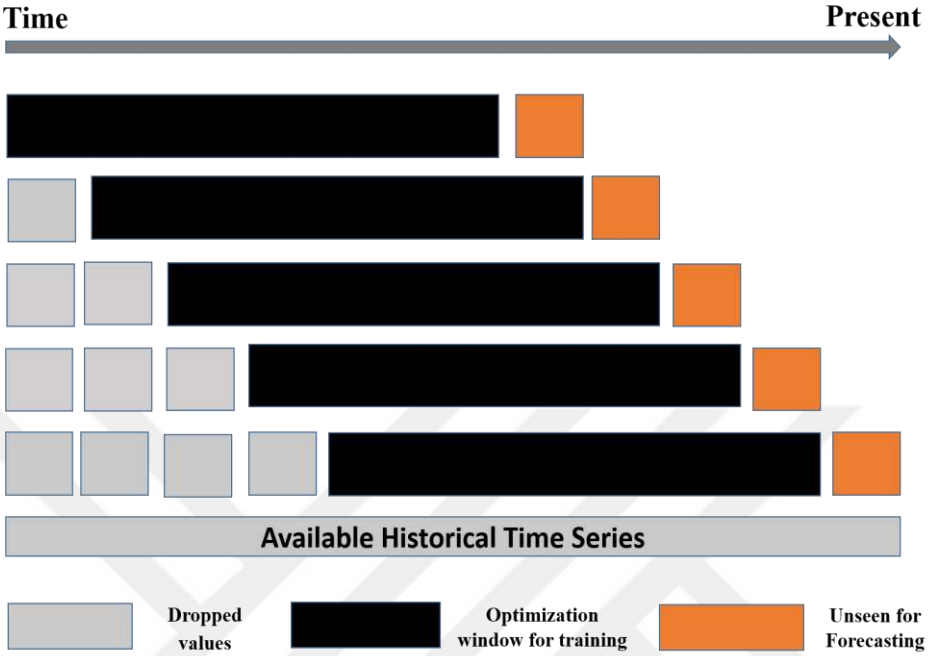

Figure 3.3. Rolling Walk Forward validation.

3.4. Regression Methods

Regression is a significant prediction technique in ML, which is used to find the connection between the target variables and predicted variables. Usually, the regression methods are used in TSF to determine the causal effect connection between input and output values. Another benefit of the regression methods is clarifying the strength of multiple output variables that impact the input variable. In addition, the significance is in permitting us to compare the impacts of dataset values that measured on different scales (Fahrmeir et al., 2013). Due to the benefits of regression methods, it is an ideal candidate for forecasting TS data. Six models based on regression methods for training the TS dataset were utilized in this thesis study. The purpose of using several models was to find the appropriate model to forecast the next day. The predicted next-day measurements are then assigned an

AQI level with which classification performance is evaluated. In the following subsections, the six regression models that were used are elaborated.

3.4.1. Support Vector Regression

SVM is a method originally developed for classification problems. Several decades earlier, Vapnik-Chervonenkis theory proved that SVR is a satisfactory model to apply to be regression problems (Vapnik, 1999). SVR model was developed to perform predictions on unseen data and it is a supervised machine learning algorithm that achieves high performance in TSF applications (Sapankevych & Sankar, 2009). SVM is a linear learning machine defined in equation (3.4), which the formula constantly utilized to solve regression problem.

$$\boldsymbol{f(x)} = (w.x) + b \tag{3.4}$$

Here, w represents the weight vector, b is the bias term, and x is the input dataset. The primary step for learning the weights with high generalization is regularization of the weight by penalizing the largest ones. The previous part was explained the linear formula of SVR. If the function, $f(x)$, is optimized in a non-linear way, the mapping function ,$\boldsymbol{\phi(x)}$, is used as shown in equation (3.5). $\boldsymbol{\phi(x)}$ will be used for mapping the input data into higher dimensional "feature" space or kernel space (Awad & Khanna, 2015). Three types of $\boldsymbol{\phi(x)}$ are accredited, were polynomial, hyperbolic tangent and Gaussian function "Radial Basis Function" (RBF). RBF is used in the proposed TSF method, besides linear SVR is used as well. Equation (3.6) represent the Kernel function formula for RBF, μ represents the mean of the feature vector and γ is the gamma is a measure of the amount of the spread in the Gaussian function that is utilized to measure the similarity between two points. If a small gamma is used that will determine the large variance

of Gaussian function that means each two points are similar even if are not close to each other and vice versa.

$$\boldsymbol{f}(\boldsymbol{x}) = (w.\phi(x)) + b \tag{3.5}$$

$$\boldsymbol{\phi}(\boldsymbol{x}, \mu) = exp(-\gamma\|x - \mu\|^2) \tag{3.6}$$

The idea behind SVR is illustrated in figure 3.4 involving two dashed lines (decision boundary). If L represents the distance from the green line (i.e. the hyperplane) to the red lines, then these distance values can be denoted by $\pm\varepsilon$. Therefore (Alakh, 2020), the formula representing the decision boundary is:

$$(w.x) + b = \pm\varepsilon \tag{3.7}$$

The main goal is to determine the decision boundary at 'L' space from the hyperplane such that data points closest to the hyperplane or the support vectors are inside the boundary line (Alakh, 2020). The parameter "C" or regularization parameter is tuned by several readings to reveal the best available value for the proposed model. Its function is tuning the margin for decision function. If the small margin is encouraged, the parameter "C" value should be large, thus will obtain a smaller misclassification rate during training and vice versa. However, large values of the regularization parameter may cause overfitting. Therefore, it should be tuned carefully.

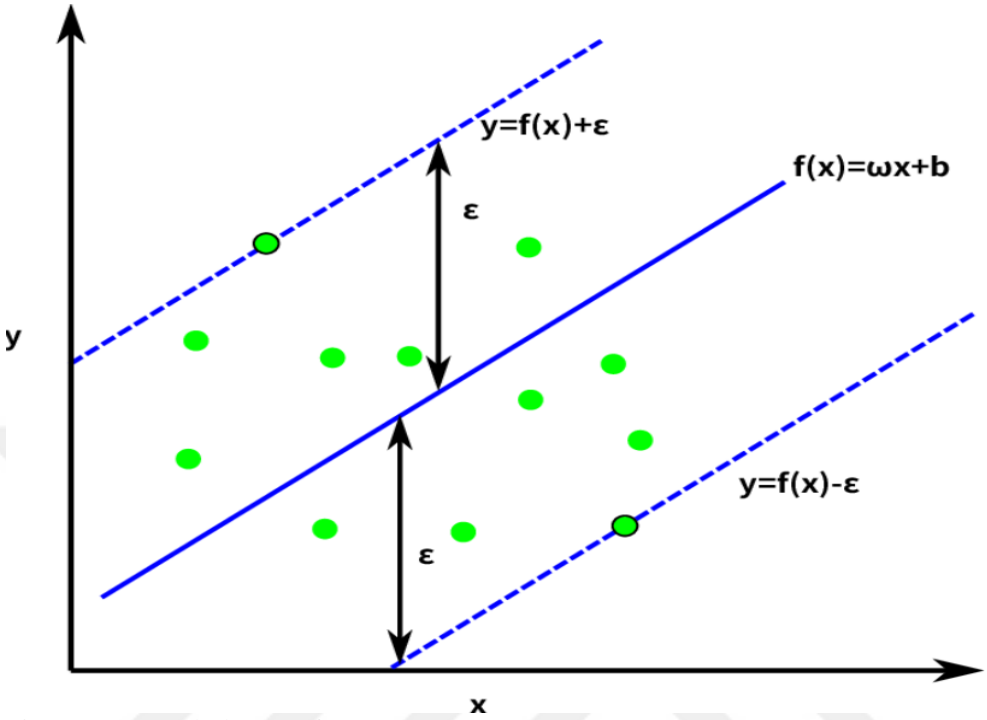

Figure 3.4. Work theory of SVR.

3.4.2. Multilayer Perceptron

The Artificial Neural Networks (ANN) is a connection of nodes or computation units similar to human neurons, so they are called a neural network. The architecture ANN consists of three layers as shown in Figure 3.5. The first layer is called the input layer containing several units or neurons which represent the seen values of the dataset for analysis. The hidden layer is the second layer between the input and output layer, which is considered as a nonlinear transformation of the input data. In other words, the hidden layer functions apply weight to the input values, then point them as output by an activation function. The weight is a numeric number that represents the connection line of the units between each other. The final layer is the output of neurons (S.-C. Wang, 2003). The ANN output formula is represented in equation (3.8).

$$\boldsymbol{h_i} = \sigma(\sum_{j=1}^{N} V_{ij} X_j + T_i^{hid}) \quad (3.8)$$

Here, h_i represents the ANN output of nodes i, σ is the activation function that imposes nonlinearities to the calculations made in the hidden layer. X_j are the input units, V_{ij} are the weights, and T_i^{hid} hidden layer threshold.

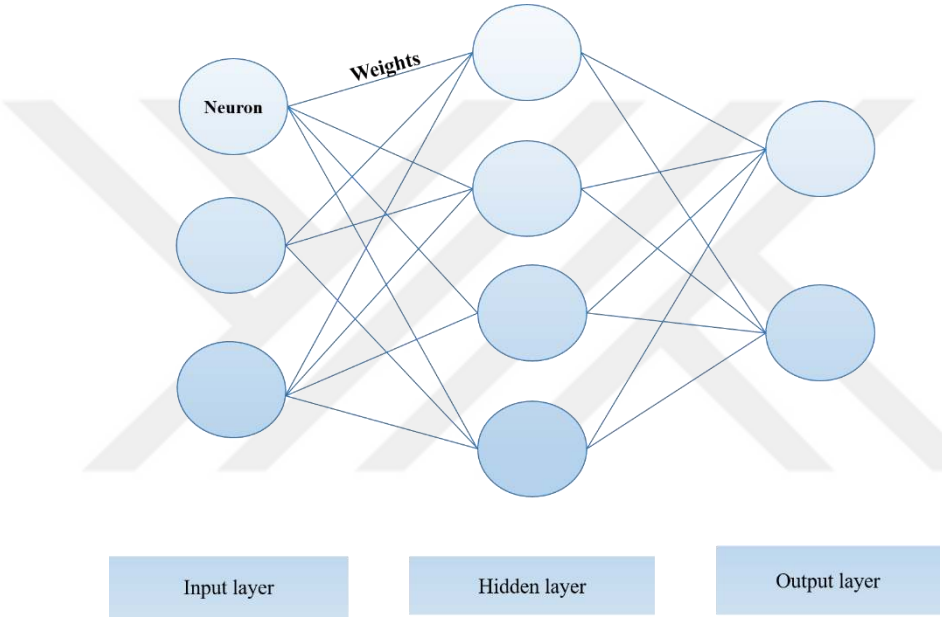

Figure 3.5. ANN architecture.

Multilayer Perceptron (MLP) model is a supervised model of ANN algorithms that contains multiple hidden layers and work on the principle of connecting node between each other. It does not involve any computation blocks with loops. In other words, the information flows only in the forward direction, which means from the input layer through the hidden layer to output and this is called a feed-forward method. In MLP model, the output is connected to entire nodes of both hidden and input layers, which is called fully connected (Gupta, 2013). MLP model was indicated as an appropriate candidate for the TSF model,

due to its potential for resolving the regression problems by a non-linear tool, its potential for approximating the smooth function even if not use a prior presumption of data distribution (Kukkonen et al., 2003).

MLP model is applied in the proposed TSF model. The hidden layer parameter is an MLP parameter used to adjust the number of nodes and the hidden layer. The hidden layer parameter is tuned to an appropriate one by selecting among three variables for the TSF model.

3.4.3. K-Nearest Neighbors

K-Nearest Neighbors (KNN) model is one of the simplest non-parametric ML algorithms was used in both classification and regression methods. KNN model is based on the lazy learning method that means deferred the generalization of the training set, which is called instance-based learning. KNN is referred to as highly effective when was used for analyzing the TS dataset (Martínez et al., 2019).

KNN works on the principle of feature similarity for forecasting the next-day value. Which means by depending on the values of features in the training set, the value that wants to be predicted was assigned as a new point. Then, the distances between the assigned new point and the points of the training set were calculated. Based on the calculated distance, the closest "K" points were selected. In the final step, the new data is forecasted by calculating the average of "K" values of the closest assigned points (Aishwarya Singh, 2018).

Equation (3.9) represents the formula used to calculate the distance between the new and input points.

$$\boldsymbol{D_{ij}} = \sqrt{\sum_{m-1}^{n}(X_{im} - X_{jm})^2} \tag{3.9}$$

Here, D_{ij} represents the distance measure, m is the specific time, X_{im} denote the input dataset, X_{jm} is the seen test set with unseen target. Equation (3.10) represents the formula used to calculate the next-day value,

$$\boldsymbol{N} = \frac{\sum_i^K X_i}{K} \tag{3.10}$$

where N denotes the next-day value, X_i are the closest "K" values, K being the number of point closest to the target. Figure 3.6 illustrates the idea behind the K-nearest neighbors model with "K=3" was chosen.

The effective and only parameter in KNN is the "K parameter". K parameter is used to determine the number of neighbor's points that are chosen to calculate the next-day value. K parameter is tuned to an appropriate three variables for the TSF model.

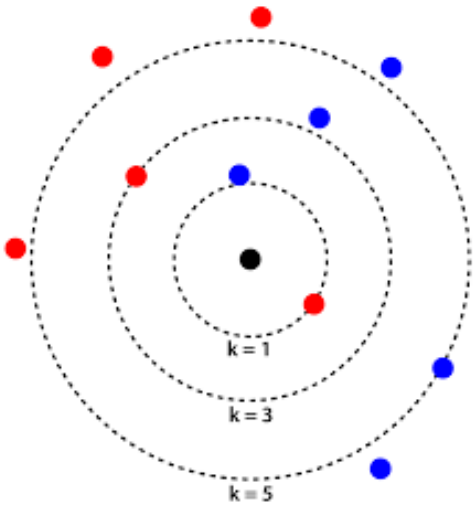

Figure 3.6. KNN model for different values of K

3.4.4. Regression Trees

Regression trees (RTs) is a popular modelling tool in ML, which works on the idea of partition the input dataset into small sets, in which the sets are more homogeneous with esteem to the response. For realizing the homogeneity, RTs determine predictor and value of sets, depth of the tree, and finally the prediction equation for the leaf nodes. Even though RTs have particular weaknesses like the instability of the model and the performance of predictive model not being optimal, they are still important and powerful technique in the ML methods due to a number of characteristics that are desirable for modelling problems that may be listed as follows:

- First, the generation of RTs contains straightforward steps that are easy to implement.
- The logic behind RT architecture efficiently addresses several types of predictors even though when there is no pre-processing step.
- Furthermore, it is extremely efficient in handling the missing data.

In the regression trees, there are two types of nodes namely, "Decision" node and "Leaf" node. "Leaf" node is used as the output of decisions that were made by the "Decision" node. Figure 3.7 shows a sample architecture for RTs. Regression tree method forms the fundamental basis of many alternative complex trees that work on a similar forecasting principle. Gradient Boosted Decision Trees (GBDT) and Random Forest (RF) models are two examples of such variants. Both of these two complex algorithms are used in the proposed TSF model.

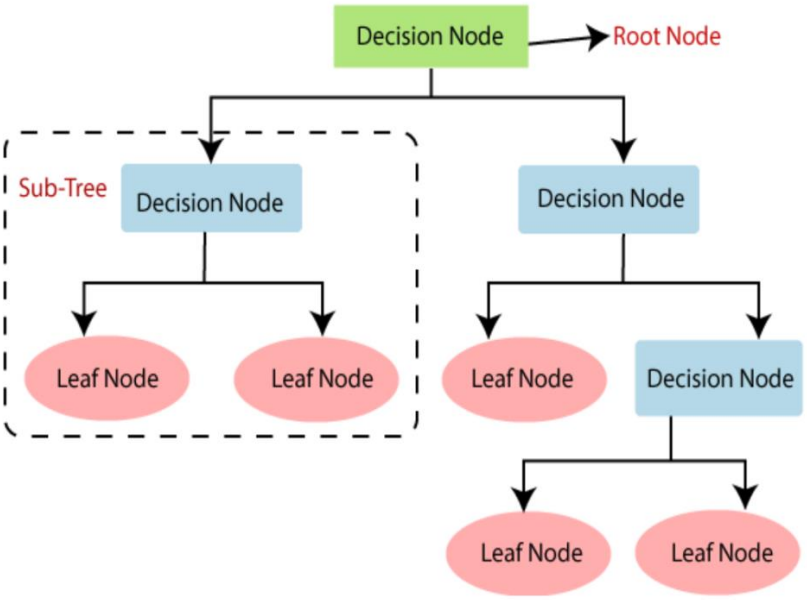

Figure 3.7. RTs architecture.

RF is an algorithm constructed by evaluation of several previous generalizations of the authentic bagging method carefully (Breiman, 2001). The previous generalizations started when Breiman unveiled his theory for improving the bagging, which indicated to the tree construction process will adding randomness for reducing the correlation between predictors, which was Breiman theory (Kuhn & Johnson, 2013). Several researchers were started adjusting the model by adding the randomness to the algorithm during the learning process may be listed as follows:

- (Ho, 1998) presented an idea to build full trees founded on random subsets.
- (Dietterich, 2000), presented an improved approach called "random split selection". This idea was indicted to use a random subset of the superior k foreteller at each split in the tree for building the trees. Which means choosing a randomized number of split “K“ within a tree that ensure the

"best" tree would be included in the ensemble. The number of K will be determined by validation.

- In addition, (Breiman, 2000) sought to disorganize the tree framework by adding noise to the response.

RFs reduces the variance of the overall dataset to any learner in the set. This eventually results in a low bias by selecting powerful, complex learners. Although, additional trees are included in the RF model, it still may be considered as computationally efficient. RFs are tuned by several parameters and the parameter that is responsible for the quantity of trees is more effective. Performances of three different values for the number of sequential trees are observed for tuning the appropriate tree number in the proposed TSF model.

The second algorithm is gradient boosting machines that is a statistical algorithm that was founded by (Friedman, 2001) and works for both regression and classification methods. Friedman showed that GBDT outperforms other boosting algorithms and it is less prone to overfitting. GBDT is considered as an elegant and simple algorithm with highly adaptable to various problems. Usage of a number of weak learners and minimizing the loss function in the model are the main principles of GBDT. Reducing the loss function in the model is achieved by initializing decision trees until the ideal forecast of the response is obtained (Touzani et al., 2018). A typical GBDT model is tuned by several parameters; the parameter that responsible for learning rate, tree depth, boosting stages to perform and the minimum samples in leaf nodes are tuned. The number of iteration parameter (boosting stages) is tuned by observing the model performance of the model for three different values in the proposed TSF model and other parameters fixed on an appropriate one variable.

3.4.5. Elastic Net

Tikhonov Regularization (Ridge regression) (Groetsch, 1986; Hoerl, 2020) and LASSO (least absolute shrinkage and selection operator) (Tibshirani, 1996) are statistical methods in machine learning that are used in regression analysis. Decrements in the complexity of the model can be considered to have a negative impact on the model response. Thus, instead of extracting the complex variables, the penalization method on the complex variables proposed by LASSO and Ridge models can be used. The LASSO algorithm uses the taxicab metric to penalize the model variables simultaneously fitting the input dataset. This is also called L1-norm penalty. The Ridge algorithm uses the squared coefficients method to penalize the model variables for fitting the input dataset. This is also called L2-norm penalty.

A linear regression model that uses both L1 and L2 penalty called Elastic Net (EN). Elastic Net (EN) is considered as an improved method that avoids the shortcomings of the Ridge and LASSO models. The minimized loss function associated with EN model is given below (Zou & Hastie, 2005):

$$\boldsymbol{L_{en}}(\widehat{\boldsymbol{\beta}}) = \frac{\sum_{i=1}^{n}\left(y_i - x_i^J \widehat{\beta}\right)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j-1}^{m}\widehat{\boldsymbol{\beta}}_j^2 + \alpha\sum_{j-1}^{m}\left|\widehat{\boldsymbol{\beta}}_j\right|\right) \quad (3.11)$$

Where λ represents the model parameter, x is the input variables matrix, y denotes the unseen set, $\boldsymbol{\beta}$ is the weight vector. In addition, alpha (α) presents the confusion parameter, which combines the LASSO and Ridge models. When $\alpha = 1$, the elastic net is matched to lasso regression, while $\alpha = 0$ to ridge regression. In the proposed TSF model the Ridge, LASSO and EN algorithms are applied. Only the EN model is mentioned in the result table because it outperforms the others. Alpha parameter is tuned to an appropriate three variables for the TSF model.

3.5. Performance metrics

As mentioned before, utilized dataset contains six components of air pollution in the atmosphere that was converted to the AQI standards of Europe. First of all, the concentration of O_3 is predicted after pre-processing the dataset. Then the predicted concentration is converted into corresponding AQI by using the standard formula of AQI in equation (3.1). Hence, for each of the steps, various performance measures are calculated depending on the used method. Since the regression models are used in the first step, root-mean-square error (RMSE) and R-squared (R^2) are calculated.

RMSE is used to measure the difference between the input data observed and the predicted values by the regression models. Equation (3.12) represents the used formula for calculating RMSE,

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_{Ti}-Y_{pi}\right)^2}{n}} \tag{3.12}$$

where n is denotes the number of data points, Y_{Ti} and Y_{pi} represent the actual and predicted values, respectively.

R^2 is a statistical measure used for calculating the extent variance between the observed and predicted values. The formula used for calculating R^2 is given in equation (3.13),

$$\boldsymbol{R^2} = 1 - \frac{RSS}{TSS} \tag{3.13}$$

where RSS denotes the residual sum of squares as clarified by equation (3.14).

$$\boldsymbol{RSS} = \sum_i \left(Y_{Ti} - Y_{pi}\right)^2 \tag{3.14}$$

In addition, TSS denotes the total sum of square calculated as shown in equation (3.15).

$$\boldsymbol{TSS} = \sum_i (Y_{Ti} - Y_{mi})^2 \tag{3.15}$$

Then, the classification method is accomplished by comparing the concentration of the observed components with predicted values. First, the actual level and assigned level of concentration are calculated by equation (3.1). In the second step, the actual level and assigned level are labelled according to the AQI of the US EPA standard in Table 3.2. Then, a confusion matrix is generated depending on labelled data. Confusion matrix or as called error matrix is a specified table, which was used for visualization of the efficiency of a classification algorithm (Visa et al., 2011). Figure 3.8 shows the confusion matrix for a binary classification problem.

		Predicted / classified	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Figure 3.8. A confusion matrix for two classes.

In order to visualize all level of pollution, a multi-class confusion matrix is generated due to air quality is defined according to five categories. Figure 3.9

shows the multi-class confusion matrix that is used in the proposed TSF model. After a multi-class confusion matrix generated, precision, accuracy, f1-score, and recall tests are measured in order to determine the efficiency of the proposed model in the classification method. When calculating these metrics, the "good" class of air pollution level is selected as the majority class and the entries for TP, FP, TN, and FN are generated accordingly.

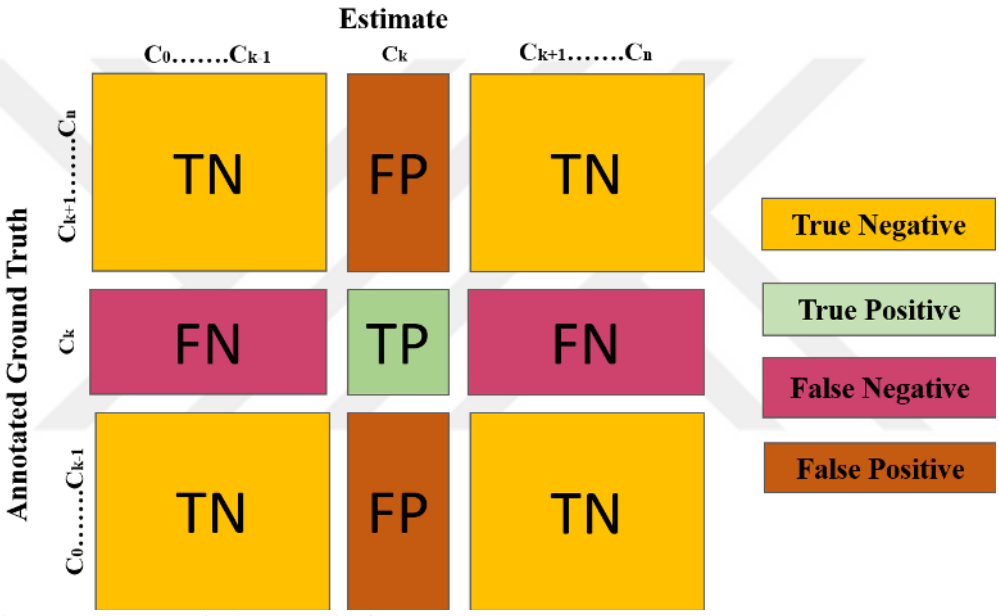

Figure 3.9. A confusion matrix for multi-class.

Accuracy is the major performance measure of the true predictions calculated by assigning the ratio of correctly predicted observations divided by the total observations. Equation (3.16) shows the utilized formula to calculate accuracy value.

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{3.16}$$

Precision is a useful measure to determine the rate of correct predictions indicating model success. It is calculated by dividing the value of correctly positive predicted observation (TP) by the total predicted positive observations (TP+FP). Equation (3.17) shows the formula for calculating the precision value.

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \tag{3.17}$$

Recall or Sensitivity denotes the rate of the actual positives predicted by the model output of that are actually positive. It is calculated by dividing the value of correctly positive predicted observation (TP) to total observations in the actual class (TP+FN). The expression for calculating the recall value is given in equation (3.18).

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} \tag{3.18}$$

F1 Score is a weighted average function of precision and recall. It takes into account both false positives and false negatives, considered a good measurement to seek an equilibrium between precision and recall. Equation (3.19) shows the utilized formula to calculate F1 Score.

$$\boldsymbol{F1\ Score} = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \tag{3.19}$$

4. EXPERIMENTAL RESULTS

The TS dataset of air pollution in Istanbul was prepared to contain six major pollution components. In addition, in order to apply the TSF model to the TS dataset, the dataset was pre-processed, and eventually, the previous measurements for one day and one week ago are included as features in the dataset. Ozone (O_3) was considered as the target pollutant to be forecasted for the next-day because O_3 has a direct impact on increasing the mortality related with air pollution. The selected algorithms for training the chosen dataset are mentioned and explained.

As indicated earlier, the predictions are performed by the mentioned regression algorithms and their corresponding performance values are calculated. Then, the predicted values by the regression algorithms and the actual values are used for converting the concentrations into an AQI. Hence, the converted data by equation (1) are categorized into five classes to generate the confusion matrix, which allows for calculating the performance of the air quality level assignments. In addition, the regression models' performance of classification problems and regression problems results are compared. Moreover, performances of the regression models with different parameters are observed and appropriate parameters for this prediction task are determined.

In this section, the performance of the used models based on the ML techniques are be mentioned and illustrated as well as the used parameters for tuning the models are presented. Table 4.1 shows the used parameters for tuning the model. For each of the regression models, the parameters yielding the lowest RMSE are used for calculating the AQI for air quality level assignment.

Table 4.1. The examined methods and the corresponding parameters.

Method	Parameters	Tried values for the parameters
KNN	K	3,5,7,9,11,13,15, 18, 20
SVM	Kernel type	Linear, Gaussian (RBF)
	C	1,3,4,7,9, 10-200, 550, 700
	gamma	0.1, 0.01, 0.001, 0.0008, 0.0003, 0.0001
RF	Number of estimators	100, 300, 400, 500, 700
EN	Alpha	0.1- 0.9
MLP	Number of maximum iterations	200, 300 , 400, 500 , 700, 1000
	Hidden layer size	(10,5), (14,7), (20,10), (50,25), (25,), (50,), (70,), (100,), (150,), (200,), (250,), (300,)
GBDT	Number of estimators	300, 400, 500, 700, 900

For KNN method, three different values for K were used in testing as shown in Table 4.1. The lowest results of regression analysis and highest results of classification analysis were achieved when K=3. Table 4.2 shows the used three neighbors in testing for both regression and classification.

Table 4.2. Regression and classification results for KNN model with different parameters

KNN	K=3	K=5	K=7
R^2	**0.45**	0.39	0.36
RMSE	**2594**	2886	3013
Accuracy	**0.74**	0.74	0.74
Recall	**0.74**	0.74	0.74
F1-score	**0.76**	0.75	0.75
Precision	**0.76**	0.75	0.75

By implementing KNN with K=3, three different figures plotted to clarify the action of the proposed model. Figure 4.1 shows the comparison plot of input and predicted values and Figure 4.2 presents the scatter plot of these values. Figure 4.3 shows the corresponding confusion matrix.

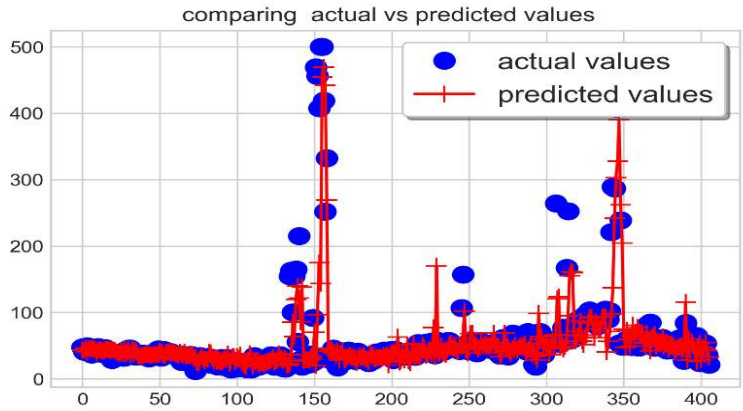

Figure 4.1. The comparison of input and predicted values by KNN method.

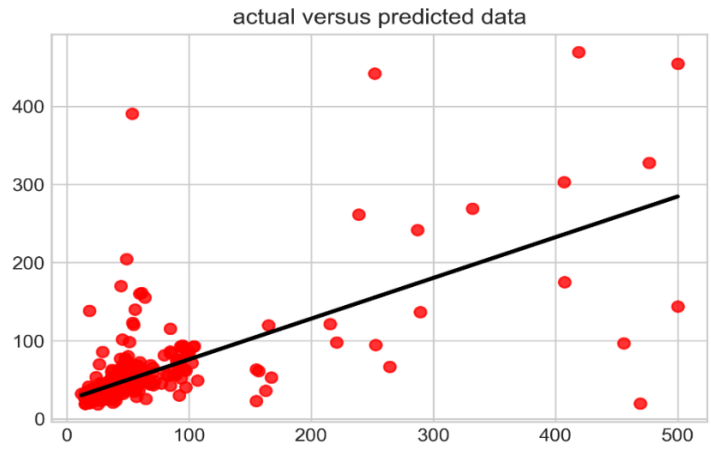

Figure 4.2. The scatter plot of input and predicted values of implement KNN (x-axis: actual values, y-axis: predicted values).

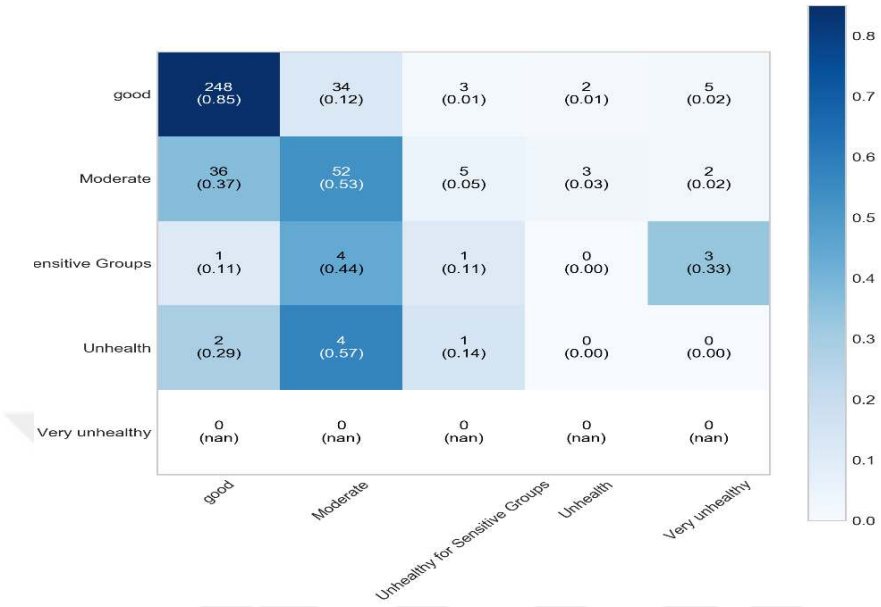

Figure 4.3. The confusion matrix of implement KNN (columns denote actual values and the rows denote predicted values)

For SVM method, the linear and nonlinear methods are used. For the linear SVM, four different values (1, 3, 5, and 7) are used for the regularization parameter, C. For the non-linear method, RBF kernel is used with several C and gamma variables. Table 4.3 shown the best three variables of using linear and non-linear. The best regression result was by the linear method when $C = 1$. For classification step, a non-linear SVM achieves the highest metrics.

Table 4.3. The Regression and classification results for SVM models with different parameters

SVM	**C=1**	**C=3**	**C=550 Gamma=0.001 Kernel= RBF**
R^2	**0.52**	0.49	0.42
RMSE	**2284**	2426	2737
Accuracy	0.77	0.76	**0.79**
Recall	0.77	0.76	**0.79**
F1-score	0.77	0.77	**0.79**
Precision	0.77	0.77	**0.785**

For linear SVM with $C = 1$, three different figures plotted to visualize the predictions of the model. Figure 4.4 shows the comparison plot of input and predicted values of linear SVM implementation. Figure 4.5 presents the corresponding scatter plot of the predictions and the actual values. Figure 4.6 shows the confusion matrix of non-linear SVM implementation that achieves the best classification results.

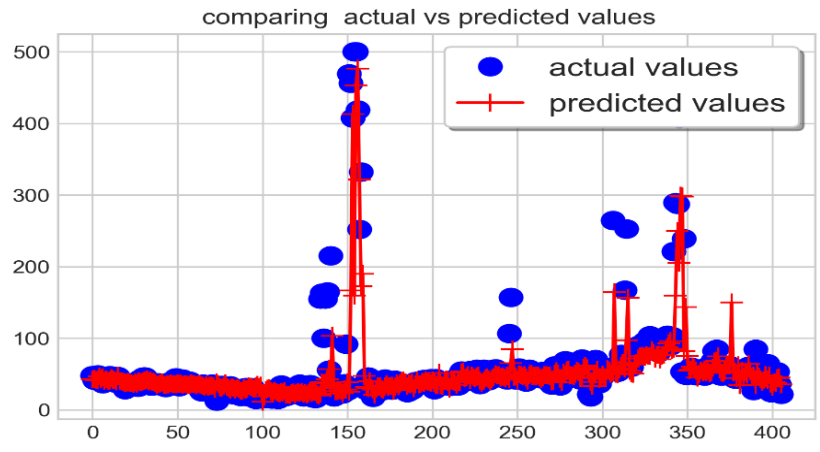

Figure 4.4. The comparison of observed and predicted values by SVM.

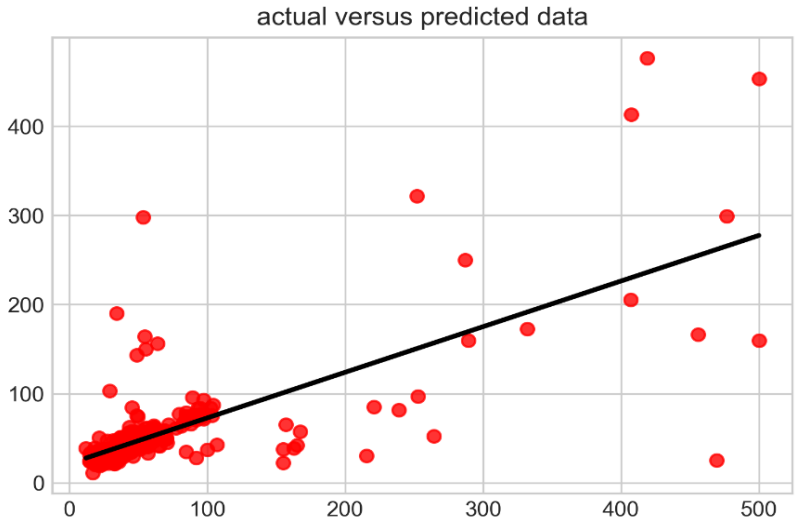

Figure 4.5. The scatter plot of observed and predicted values of implement SVM (x-axis: actual values, y-axis: predicted values)

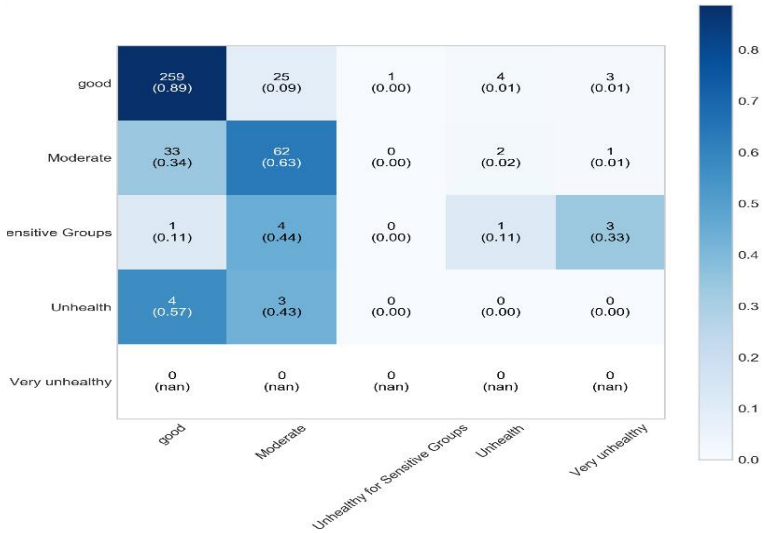

Figure 4.6. The confusion matrix of non-linear SVM implementation (columns denote actual values and the rows denote predicted values)

As for ANN-MLP method, the number of maximum iterations was set to 300. Hidden layer parameter is tuned by trying several values and the best three appropriate hidden layer variables are presented in Table 4.4 for the proposed TSF model. The lowest RMSE result was obtained by MLP.

Table 4.4. The Regression and classification results for MLP models with different parameters.

MLP with max_iter=300	**Hidden layer sizes= (14,7)**	**Hidden layer sizes = (70,)**	**Hidden layer sizes= (20,10)**
R^2	**0.52**	0.51	0.51
RMSE	**2246.7**	2301.5	2305
Accuracy	0.68	**0.70**	0.67
Recall	0.68	**0.70**	0.67
F1-score	0.70	**0.72**	0.70
Precision	0.76	**0.76**	0.74

For ANN-MLP with the first hidden layer equal to 14 units and the second hidden layer equal to 7 units, three different figures are plotted to allow for visual evaluation of the proposed model as well. Figure 4.7 shows the comparison plot of input and predicted values of MLP model and Figure 4.8 contains the scatter plot of these values. Figure 4.9 shows the corresponding confusion matrix of the values output by the MLP model**.**

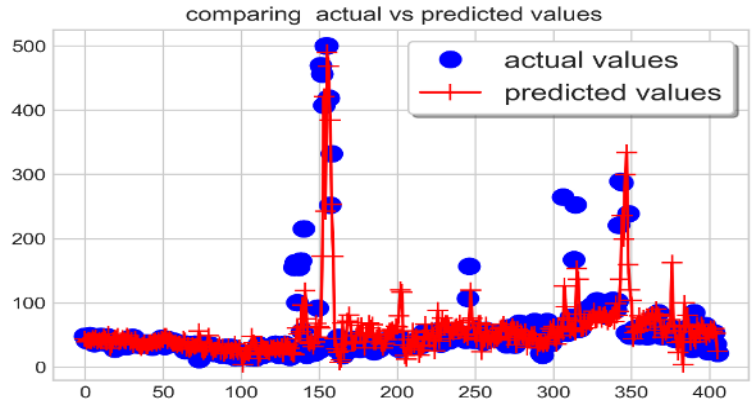

Figure 4.7. The comparison of observed and predicted values by MLP.

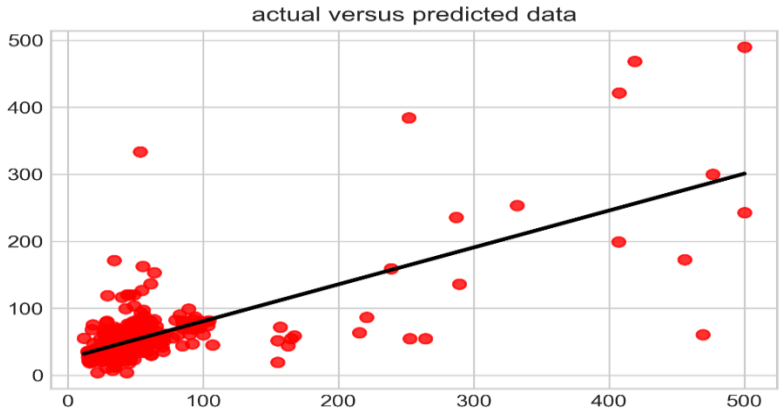

Figure 4.8. The scatter plot of observed and predicted values of implement MLP (x-axis: actual values, y-axis: predicted values)

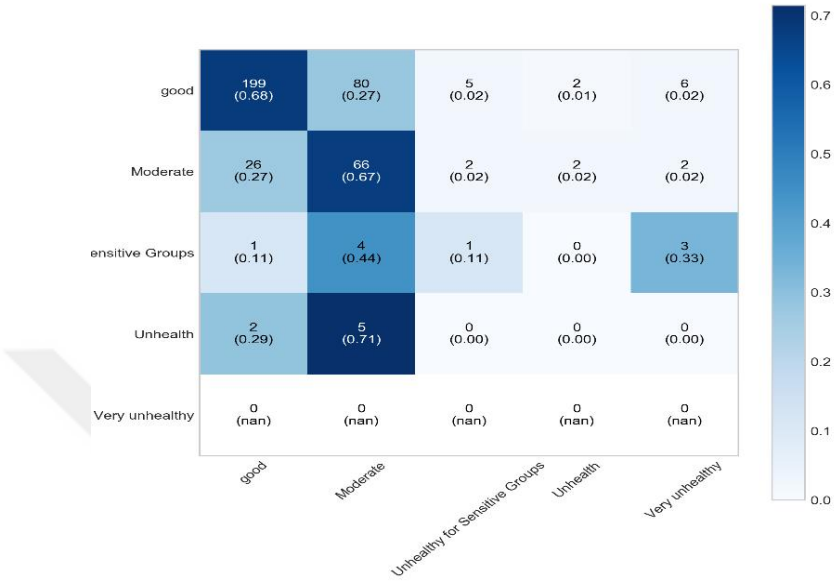

Figure 4.9. The confusion matrix of ANN-MLP implementation (columns denote actual values and the rows denote predicted values)

The EN model was tuned by only the alpha parameter, which is the most influencing parameter. As shown in Table 4.5, the best value for alpha was found to be 0.9 both for regression and classification steps.

Table 4.5. The Regression and classification results for EN models with different parameters

EN	Alpha =0.9	Alpha =0.5	Alpha =0.1
R^2	**0.52**	0.52	0.51
RMSE	**2250**	2284	2592
Accuracy	**0.72**	0.70	0.68
recall	**0.72**	0.70	0.68
F1-score	**0.74**	0.71	0.70
Precision	**0.79**	0.76	0.74

For EN with alpha equal to 0.9, three different figures are plotted to clarify the action of the proposed model. Figure 4.10 shows the comparison plot of observed and predicted values of EN implementation, Figure 4.11 presents the scatter plot of these values. Figure 4.12 illustrates the corresponding confusion matrix of the EN model.

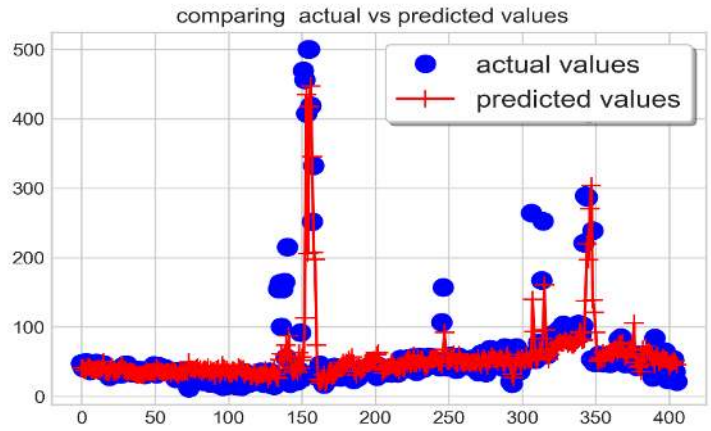

Figure 4.10. The comparison of observed and predicted values by EN method.

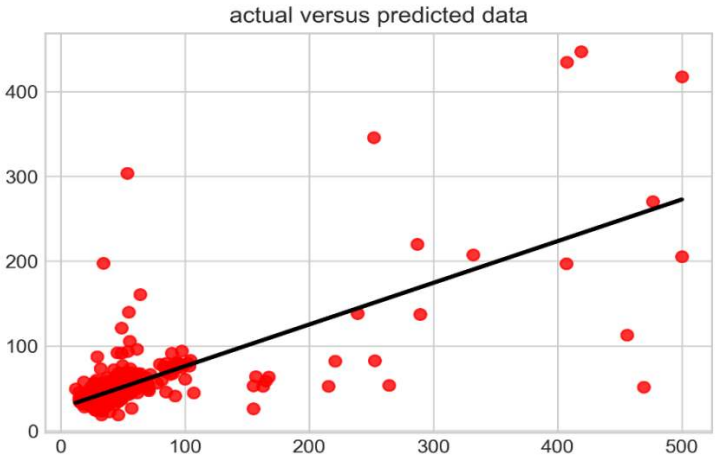

Figure 4.11. The scatter plot of observed and predicted values of EN method (x-axis: actual values, y-axis: predicted values)

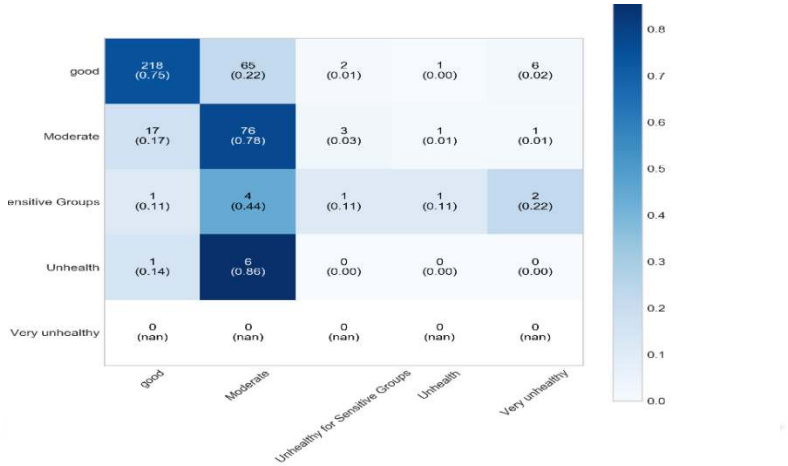

Figure 4.12. The confusion matrix of EN implementation (columns denote actual values and the rows denote predicted values)

As for regression trees, RF and GBDT were tuned by several parameters. For GBDT algorithm, learning rate, tree depth, and leaf nodes were set on an appropriate variable, and the number of iterations is tuned by three different variables. For RF, the number of sequential trees was tuned by three different variables, as well. Table 4.6 presents the best two readings was obtained by tuned variables on RF and GBDT. Where, RF outperforms GBDT in an analysis of both regression and classification method.

Table 4.6. The Regression and classification results for RF and GBDT models with different parameters

RT	**RF** number of estimators=500	**GBDT** number of estimators =300 tree depth=8 leaf nodes =8 learning rate= 0.01
R^2	**0.46**	0.42
RMSE	**2548**	2745
Accuracy	**0.714**	0.71
recall	**0.714**	0.71
F1-score	**0.74**	0.73
Precision	**0.78**	0.76

Hence, RF is used to plot the three different figures to visualize the performance of the model. Figure 4.13 presents the comparison plot of observed and predicted values of RF implementation, Figure 4.14 involves the scatter plot of these values. Figure 4.15 shows the corresponding confusion matrix of the RF model.

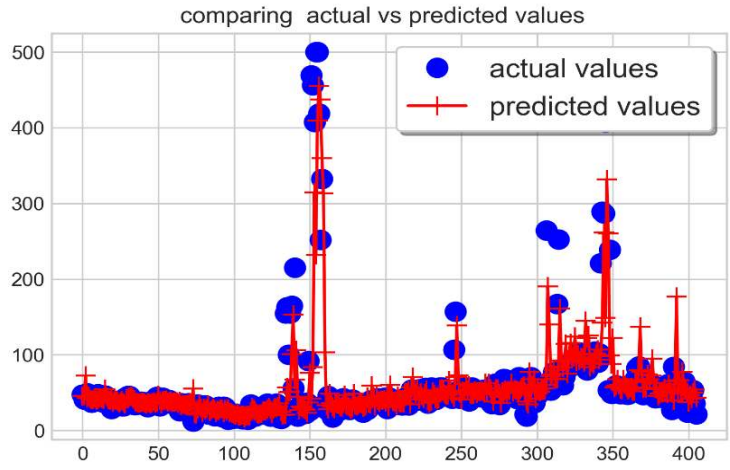

Figure 4.13. The comparison of observed and predicted values by RF method.

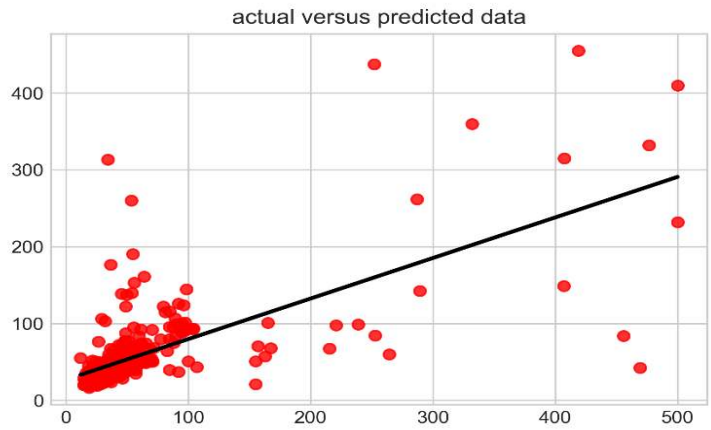

Figure 4.14. The scatter plot of observed and predicted values of RF method (x-axis: actual values, y-axis: predicted values)

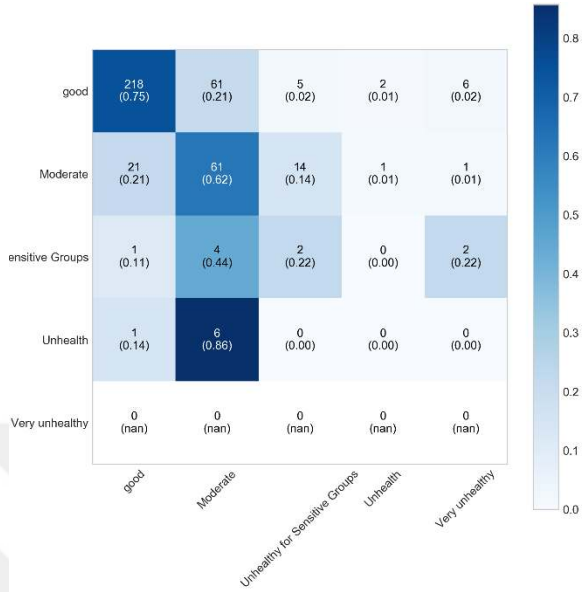

Figure 4.15. The confusion matrix of RF implementation (columns denote actual values and the rows denote predicted values)

In the previous part, the results obtained by the regression models with the best parameters were indicated separately. For the comparison stage of regression models performance, Table 4.7 is generated to collectively present the achieved best results for the models.

Table 4.7. Regression results for ozone concentration prediction

	KNN	SVM	RF	EN	MLP	GBDT
RMSE	2594	2284	2548	2250	**2246**	2745
R^2	0.45	**0.52**	0.46	**0.52**	**0.52**	0.42

As can be seen in Table 4.7**,** the measured R^2 and RMSE scores have a negative correlation in general. However, the lowest R^2 score is recorded by MLP model obviously with another two methods were SVM and EN. In addition, the lowest RMSE score is recorded by MLP model as well. Thus, the lowest error of

the analyzed methods is achieved by MLP model. In other words, the MLP method outperforms other models in the regression part and it is the most appropriate regression model for analyzing the TS air pollution dataset. The alternative regression method for analyzing TS dataset is the EN method as shown in Table 4.7, which is very close to results obtained by MLP regression method.

As earlier mentioned, the AQI levels of the observed and predicted values are categorized into five classes for measuring the classification performance. Table 4.8 shows the best results obtained by the classification method of the proposed models. The highest performance of the classification method is obtained by non-linear SVM model with C=550 and gamma= 0.001. In addition, linear SVM method appeared to provide satisfactory results as well. This means that the SVM method outperforms other models used in the experiments for the classification method and it is the most appropriate classification model for analyzing the TS air pollution dataset.

Table 4.8. Classification results for ozone concentration prediction

	KNN	**SVR**	**RF**	**EN**	**MLP**	**GBDT**
Accuracy	0.756	**0.790**	0.714	0.726	0.697	0.709
recall	0.756	**0.790**	0.714	0.73	0.697	0.709
F1-score	0.757	**0.788**	0.740	0.743	0.716	0.732
Precision	0.76	**0.786**	**0.780**	**0.786**	0.765	0.760

As shown in Figures 4.7 and 4.8, conclusively, the majority of the errors in the predictions occur when the O_3 concentration is very high. This means that the regression models are not able to learn and recognizer the patterns related with the high O_3 concentration times. This problem can be originating from the fact that the

samples of high O_3 concentration are very rare in the dataset. It is also notable that there are no samples belonging to “Hazardous” air pollution level corresponding to highest AQI values. Therefore, the next-day prediction scheme by the analyzed TSF models is difficult to handle with the samples having such a distribution. In the related literature, the same problem was encountered by Burrows et al. for analzing the ozone concentration (Burrows et al., 1995). They mentioned that the majority of the errors made by the ML model was due to the high concentration of the ozone. Their obtained results were close to the results of our proposed method.

Furthermore, in light of the obtained results by the proposed model, Zheng et al. also used ML models to forecast the air quality (Zheng et al., 2018). RF, MLP, SVM, LSTM, and ARIMA models were used for training TS dataset. The maximum obtained accuracy was obtained by the parametric models was 0.75 and the maximum obtained accuracy was obtained by ARIMA and LSTM was 0.72, which means less than the results are obtained by our proposed model. However, a multiple kernel learning (MKL) hybrid model was suggested which overcomes all of the mentioned results. This implies most of the hybrid model are more efficacy than common models.

Finally, Table 4.8 shows the accuracy and recall values that were used to measure the classification performance are identical of the proposed models. Which means the ability that is classified the positive samples correctly is the same as the ability that is used to classify negative samples correctly, thus the used models are "balanced". In other words, a balance may have occurred as a result of the observed dataset are converted to AQI according to equation (3.1), then, are returned back to real concentrations to measure the classification performance.

Finally, the correlation between the air pollutant features are calculated if they seem to be redundant. For this purpose, Pearson correlation coefficient has been calculated between two features. This measure is limited between -1 and 1, where if the correlation results of those features are 1 which indicates a full positive correlation, vice versa. While the results 0 means there is no correlation.

Moreover, if the result is between -0.5 and 0.5, this means that those input features have less prominent correlation. Table 4.9 shows the Pearson's Correlation between the used dataset features. Three results crossed the -0.5 and 0.5, were "SO_2 and CO" with a correlation score of 0.56, "CO and NO_2" with a correlation score of 0.614, and the highest obtained correlation score of 0.636 by "SO_2 and NO_2", which is reasonably good. All other results obtained indicate that a less notable correlation.

Table 4.9. The Pearson's Correlation of the used dataset features.

Dataset Features	SO_2	PM10	PM25	CO	O_3	NO_2
SO_2	1.00	0.333	-0.014	**0.561**	0.061	**0.636**
PM10	0.333	1.00	0.279	0.346	-0.032	0.437
PM25	-0.014	0.279	1.00	-0.010	-0.108	0.147
CO	**0.561**	0.346	-0.010	1.00	0.022	**0.614**
O_3	0.061	-0.032	-0.108	0.022	1.00	0.001
NO_2	**0.636**	0.437	0.147	**0.614**	0.001	1.00

5. CONCLUSION AND FUTURE WORK

In this study, ML methods used to analyze the air pollution time-series dataset for forecasting the concentration of O_3 of the next-day. One-step ahead of time series forecasting method was performed for forecasting the test set. Regarding the train set was generated by the rolling walk forward validation method. Therefore, the used data processing was in successive and distinct steps regarding time series forecasting. Many models were tested on time series data, however, six specific machine learning methods, namely, SVM, MLP, GBDT, RF, EN, and KNN models were chosen as the basic models for time-series forecasting experiments. A dataset consisting of 1013 row and 56 columns was generated from the publicly available AQI data published by WAQIP.

In the first step, the proposed regression methods were performed on the proposed dataset based on the time series forecasting method, where the lowest errors were observed by MLP model. Therefore, MLP can be considered an appropriate algorithm to establish TSF architecture. In addition, EN is the alternative algorithm that was achieved results that are very close to MLP algorithm.

In the next step, the actual values and the predicted AQI values were assigned a concentration level of air quality by using the formula proposed by US EPA. As a result, the assignment data was labelled into a multi-class for measuring the classification performance, where linear and non-linear SVR algorithms observed the highest result that outperformed others.

Therefore, it may be concluded that MLP is suitable when prediction of ozone concentration is required. On the other hand, SVR can be a better choice if the aim is to predict the level of AQI instead of the concentration amount.

Furthermore, the training and testing of these models have been performed with different combinations of their significant parameters. The parameters with

which the best performances obtained are considered to be the appropriate and utilized for final comparison between the ML models.

According to the results, majority of the errors are caused by the samples corresponding to high ozone concentration. Therefore, increasing the number of samples for high pollution measurements, may help to improve the efficiency of the methods. In the future, some hybrid algorithms may be used to select parameters and optimize models. In that case, better results may be achieved with less amount of manual computation.

REFERENCES

Abderrahim, H., Chellali, M. R., & Hamou, A. (2016). Forecasting PM10 in Algiers: efficacy of multilayer perceptron networks. *Environmental Science and Pollution Research*, *23*(2), 1634–1641. https://doi.org/10.1007/s11356-015-5406-6

Ahmed, N. K., Atiya, A. F., El Gayar, N., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, *29*(5), 594–621. https://doi.org/10.1080/07474938.2010.481556

Aijaz, I., & Agarwal, P. (2019). A Study on Time Series Forecasting using Hybridization of Time Series Models and Neural Networks. *Recent Advances in Computer Science and Communications*, *13*(5), 827–832. https://doi.org/10.2174/1573401315666190619112842

Aishwarya Singh. (2018). A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code). *A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python Code)*, 1. https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

AISHWARYA SINGH. (2018). WEB-A Multivariate Time Series Guide to Forecasting and Modeling (with Python codes). *Analytics Vidhya*, *5000*, 1–27. https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/

Akhtar, A., Masood, S., Gupta, C., & Masood, A. (2008). Prediction and analysis of pollution levels in delhi using multilayer perceptron. *Advances in Intelligent Systems and Computing*, *542*, 563–572. https://doi.org/10.1007/978-981-10-3223-3_54

Al-Qahtani, F. H., & Crone, S. F. (2013). Multivariate k-nearest neighbour regression for time series data - A novel algorithm for forecasting UK electricity demand. *Proceedings of the International Joint Conference on Neural Networks*, *February 2018*. https://doi.org/10.1109/IJCNN.2013.6706742

Alakh, S. (2020). Support Vector Regression Tutorial for Machine Learning. *Analythicsvidhya.Com*. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

Altinçöp, H., & Oktay, A. B. (2019). Air Pollution Forecasting with Random Forest Time Series Analysis. *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*, 8–12. https://doi.org/10.1109/IDAP.2018.8620768

An, N. H., & Anh, D. T. (2016). Comparison of Strategies for Multi-step-Ahead Prediction of Time Series Using Neural Network. *Proceedings - 2015 International Conference on Advanced Computing and Applications, ACOMP 2015*, 142–149. https://doi.org/10.1109/ACOMP.2015.24

Awad, M., & Khanna, R. (2015). Efficient learning machines: Theories, concepts, and applications for engineers and system designers. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, 1–248. https://doi.org/10.1007/978-1-4302-5990-9

Baawain, M. S., & Al-Serihi, A. S. (2014). Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network. *Aerosol and Air Quality Research*, *14*(1), 124–134. https://doi.org/10.4209/aaqr.2013.06.0191

Ban, T., Zhang, R., Pang, S., Sarrafzadeh, A., & Inoue, D. (2013). Referential kNN regression for financial time series forecasting. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8226 LNCS*(PART 1), 601–608. https://doi.org/10.1007/978-3-642-42054-2_75

Beamer, P. I. (2019). Editorials: Air pollution contributes to asthma deaths. *American Journal of Respiratory and Critical Care Medicine*, *200*(1), 1–2. https://doi.org/10.1164/rccm.201903-0579ED

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, *17*(1), 1–19. https://doi.org/10.1186/s12889-017-4914-3

Benhamou, E., Saltiel, D., Ungari, S., & Mukhopadhyay, A. (2020). Time your hedge with Deep Reinforcement Learning. *ArXiv*. https://doi.org/10.2139/ssrn.3693614

Bhalgat, P., Pitale, S., & Bhoite, S. (2019). Air Quality Prediction using Machine Learning Algorithms. *International Journal of Computer Applications Technology and Research*, *8*(9), 367–370. https://doi.org/10.7753/ijcatr0809.1006

Blondeau, P., Iordache, V., Poupard, O., Genin, D., & Allard, F. (2005). Relationship between outdoor and indoor air quality in eight French schools. *Indoor Air*, *15*(1), 2–12. https://doi.org/10.1111/j.1600-0668.2004.00263.x

Bontempi, G., Ben Taieb, S., & Le Borgne, Y.-A. (2013). Machine Learning Strategies for Time Series Forecasting. *European Business Intelligence Summer School*, *138*, 62–77. https://doi.org/https://doi.org/10.1007/978-3-642-36318-4_3

Bose, M., & Mali, K. (2019). Designing fuzzy time series forecasting models: A survey. *International Journal of Approximate Reasoning*, *111*, 78–99. https://doi.org/10.1016/j.ijar.2019.05.002

Breiman, L. (2000). Randomizing outputs to increase prediction accuracy. *Machine Learning*, *40*(3), 229–242. https://doi.org/10.1023/A:1007682208299

BREIMAN, L. (2001). Random forests. *Machine Learning*, 1–122. https://doi.org/10.1201/9780429469275-8

Burrows, W. R., Benjamin, M., Beauchamp, S. ., Lord, E. R., & McCollor, D. (1995). CART Decision-Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone for the Vancouver, Montreal, and Atlantic Regions of Canada. *Journal of Applied Meteorology*, *34*(8), 1848–1862. https://doi.org/10.1175/1520-0450(1995)034<1848:CDTSAA>2.0.CO;2

Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., & Sun, J. (2016). A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. *Transportation Research Part C: Emerging Technologies*, *62*(June 2020), 21–34. https://doi.org/10.1016/j.trc.2015.11.002

Carey, I. M., Atkinson, R. W., Kent, A. J., Van Staa, T., Cook, D. G., & Anderson, H. R. (2013). Mortality associations with long-term exposure to outdoor air pollution in a national English cohort. *American Journal of Respiratory and Critical Care Medicine*, *187*(11), 1226–1233. https://doi.org/10.1164/rccm.201210-1758OC

Carlsten, C., Dybuncio, A., Becker, A., Chan-Yeung, M., & Brauer, M. (2011). Traffic-related air pollution and incident asthma in a high-risk birth cohort. *Occupational and Environmental Medicine*, *68*(4), 291–295. https://doi.org/10.1136/oem.2010.055152

Carslaw, D. C., & Taylor, P. J. (2009). Analysis of air pollution data at a mixed source location using boosted regression trees. *Atmospheric Environment*, *43*(22–23), 3563–3570. https://doi.org/10.1016/j.atmosenv.2009.04.001

Chaovalitwongse, W. A., Fan, Y. J., & Sachdeo, R. C. (2007). On the time series K-nearest neighbor classification of abnormal brain activity. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *37*(6), 1005–1016. https://doi.org/10.1109/TSMCA.2007.897589

Chelani, A. B. (2010). Prediction of daily maximum ground ozone concentration using support vector machine. *Environmental Monitoring and Assessment*, *162*(1–4), 169–176. https://doi.org/10.1007/s10661-009-0785-0

Chen, H., Lundberg, S., & Lee, S. I. (2018). Hybrid gradient boosting trees and neural networks for forecasting operating room data. *ArXiv*, *Nips*.

Cheng, C. H., & Wei, L. Y. (2010). One step-ahead ANFIS time series model for forecasting electricity loads. *Optimization and Engineering*, *11*(2), 303–317. https://doi.org/10.1007/s11081-009-9091-5

Choubin, B., Khalighi-Sigaroodi, S., Malekian, A., & Kişi, Ö. (2016). Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. *Hydrological Sciences Journal*, *61*(6), 1001–1009. https://doi.org/10.1080/02626667.2014.966721

Dalrymple, D. J. (1987). Sales forecasting practices. Results from a United States survey. *International Journal of Forecasting*, *3*(3–4), 379–391. https://doi.org/10.1016/0169-2070(87)90031-8

De Albuquerque Filho, F. S., Madeiro, F., Fernandes, S. M. M., De Mattos Neto, P. S. G., & Ferreira, T. A. E. (2013). Time-series forecasting of pollutant concentration levels using particle swarm optimization and artificial neural networks. *Quimica Nova*, *36*(6), 783–789. https://doi.org/10.1590/S0100-40422013000600007

Delavar, M. R., Gholami, A., Shiran, G. R., Rashidi, Y., Nakhaeizadeh, G. R., Fedra, K., & Afshar, S. H. (2019). A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran. *ISPRS International Journal of Geo-Information*, *8*(2). https://doi.org/10.3390/ijgi8020099

Deshpande, R. R. (2012). On The Rainfall Time Series Prediction Using Multilayer Perceptron Artificial Neural Network. *International Journal of Emerging Technology and Advanced Engineering*, *2*(1), 148–153. http://www.ijetae.com/files/Volume2Issue1/IJETAE_0112_28.pdf

DIETTERICH, T. G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, *40*, 139–157.

Dragomir, E. G. (2010). Air Quality Index Prediction using K-Nearest Neighbor Technique. *Seria Matematică - Informatică - Fizică*, *LXII*(1), 103–108. http://bmif.unde.ro/docs/20101/pdf_final_12 EDragomir.pdf

Du, S., Li, T., & Horng, S. J. (2018). Time Series Forecasting Using Sequence-to-Sequence Deep Learning Framework. *Proceedings - International Symposium on Parallel Architectures, Algorithms and Programming, PAAP*, *2018-Decem*, 171–176. https://doi.org/10.1109/PAAP.2018.00037

Dudek, G. (2016). Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting. *International Journal of Forecasting*, *32*(3), 1057–1060. https://doi.org/10.1016/j.ijforecast.2015.11.009

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). Regression: Models, methods and applications. In *Regression: Models, Methods and Applications* (Vol. 9783642343). https://doi.org/10.1007/978-3-642-34333-9

Feldman, C. (2012). Respiratory System Diseases, Ruminants. *SpringerReference*. https://doi.org/10.1007/springerreference_142787

Feng, R., Zheng, H. jun, Gao, H., Zhang, A. ran, Huang, C., Zhang, J. xi, Luo, K., & Fan, J. ren. (2019). Recurrent Neural Network and random forest for analysis and accurate forecast of atmospheric pollutants: A case study in Hangzhou, China. *Journal of Cleaner Production*, *231*, 1005–1015. https://doi.org/10.1016/j.jclepro.2019.05.319

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Institute of Mathematical Statistics Stable*, *29*(5), 1189–1232. https://doi.org/10.1146/annurev-cellbio-092910-154008

Ghoshal, A., Waghray, P., Dsouza, G., Saluja, M., Agarwal, M., Goyal, A., Limaye, S., Balki, A., Bhatnagar, S., Jain, M., Tikkiwal, S., Vaidya, A., Lopez, M., Hegde, R., & Gogtay, J. (2020). Real-world evaluation of the clinical safety and efficacy of fluticasone/formoterol FDC via the Revolizer® in patients with persistent asthma in India. In *Pulmonary Pharmacology and Therapeutics* (Vol. 60). Elsevier Ltd. https://doi.org/10.1016/j.pupt.2019.101869

Gocheva-Ilieva, S. G., Ivanov, A. V., Voynikova, D. S., & Boyadzhiev, D. T. (2014). Time series analysis and forecasting for air pollution in small urban area: An SARIMA and factor analysis approach. *Stochastic Environmental Research and Risk Assessment*, *28*(4), 1045–1060. https://doi.org/10.1007/s00477-013-0800-4

Groetsch, C. (1986). The theory of Tikhonov regularization for Fredholm equations of the first kind. *SIAM Review*, *28*(1), 116–118. https://doi.org/10.1137/1028033

Gupta, N. (2013). Artificial Neural Network. *Institute of Engineerinf and Technology*, *3*(1), 24–28. https://www.iiste.org/

He, H. Di, Lu, W. Z., & Xue, Y. (2015). Prediction of particulate matters at urban intersection by using multilayer perceptron model based on principal components. *Stochastic Environmental Research and Risk Assessment*, *29*(8), 2107–2114. https://doi.org/10.1007/s00477-014-0989-x

Health Organization, W. (2009). Global health risks: mortality and burden of disease attributable to selected major risks. In *GLOBAL HEALTH RISKS*. http://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf

Health Organization, W. (2016). Ambient air pollution: a global assessment of exposure and burden of disease. *Clean Air Journal*, *26*(2), 6. https://doi.org/10.17159/2410-972x/2016/v26n2a4

Health Organization, W. (2018). *Burden of disease from ambient air pollution for 2016*. 1–4.

Ho, T. K. (1998). *00709601.Pdf*. *20*(8), 832–844.

Hoerl, R. W. (2020). Ridge Regression: A Historical Context. *Technometrics*, *62*(4), 420–425. https://doi.org/10.1080/00401706.2020.1742207

Jiang, H., & Dong, Y. (2018). A novel model based on square root elastic net and artificial neural network for forecasting global solar radiation. *Complexity*, *2018*. https://doi.org/10.1155/2018/8135193

Jiang, N., & Riley, M. L. (2015). Exploring the Utility of the Random Forest Method for Forecasting Ozone Pollution in SYDNEY. *Journal of Environment Protection and Sustainable Development*, *1*(5), 245–254.

Jiang, X. Q., Mei, X. D., & Feng, D. (2016). Air pollution and chronic airway diseases: What should people know and do? *Journal of Thoracic Disease*, *8*(1), E31–E40. https://doi.org/10.3978/j.issn.2072-1439.2015.11.50

Johnsen, T. K., & Gao, J. Z. (2020). Elastic Net to Forecast COVID-19 Cases. *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies, 3ICT 2020*. https://doi.org/10.1109/3ICT51146.2020.9311968

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/1471-2105-15-276

Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., & Xie, G. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. *International Journal of Environmental Science and Development*, *9*(1), 8–16. https://doi.org/10.18178/ijesd.2018.9.1.1066

Kaplan, G. G., Tanyingoh, D., Dixon, E., Johnson, M., Wheeler, A. J., Myers, R. P., Bertazzon, S., Saini, V., Madsen, K., Ghosh, S., & Villeneuve, P. J. (2013). Ambient ozone concentrations and the risk of perforated and nonperforated appendicitis: A multicity case-crossover study. *Environmental Health Perspectives*, *121*(8), 939–943. https://doi.org/10.1289/ehp.1206085

Khaled Hashad, Jiajun Gu, Bo Yang, Moren Rong, Edric Chen, Xiaoxin, M. and K. M. (2020). Designing roadside green infrastructure to mitigate traffic-related air pollution using machine learning. *Science of the Total Environment*, *773*. https://doi.org/10.1016/j.scitotenv.2020.144760

Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1–2), 307–319. https://doi.org/10.1016/S0925-2312(03)00372-2

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., & Cawley, G. (2003). Extensive evaluation of neural network models for the prediction of NO 2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment*, *37*(32), 4539–4550. https://doi.org/10.1016/S1352-2310(03)00583-1

Liu, B. C., Binaykia, A., Chang, P. C., Tiwari, M. K., & Tsao, C. C. (2017). Urban air quality forecasting based on multidimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *PLoS ONE*, *12*(7), 1–17. https://doi.org/10.1371/journal.pone.0179763

Liu, T., Lau, A. K. H., Sandbrink, K., & Fung, J. C. H. (2018). Time Series Forecasting of Air Quality Based On Regional Numerical Modeling in Hong Kong. *Journal of Geophysical Research: Atmospheres*, *123*(8), 4175–4196. https://doi.org/10.1002/2017JD028052

Liu, W., Dou, Z., Wang, W., Liu, Y., Zou, H., Zhang, B., & Hou, S. (2018). Short-term load forecasting based on elastic net improved GMDH and difference degree weighting optimizations. *Applied Sciences (Switzerland)*, *8*(9). https://doi.org/10.3390/app8091603

Lodgejr, J. (1996). Air quality guidelines. Global update 2005. Particulate matter, ozone, nitrogen dioxide and sulfur dioxide. *Environmental Science and Pollution Research*, *3*(91), 23–23. http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/pre2009/air-quality-guidelines.-global-update-2005.-particulate-matter,-ozone,-nitrogen-dioxide-and-sulfur-dioxide

Lubis, A. H., Sihombing, P., & Nababan, E. B. (2020). Analysis of Accuracy Improvement in K-Nearest Neighbor using Principal Component Analysis (PCA). *Journal of Physics: Conference Series*, *1566*(1), 2–10. https://doi.org/10.1088/1742-6596/1566/1/012062

Luna, A. S., Paredes, M. L. L., de Oliveira, G. C. G., & Corrêa, S. M. (2014). Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. *Atmospheric Environment*, *98*, 98–104. https://doi.org/10.1016/j.atmosenv.2014.08.060

Maheshwari, K., & Lamba, S. (2019). Air Quality Prediction using Supervised Regression Model. *IEEE International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2019*, *Ml*. https://doi.org/10.1109/ICICT46931.2019.8977694

Martínez, F., Frías, M. P., Pérez, M. D., & Rivera, A. J. (2019). A methodology for applying k-nearest neighbor to time series forecasting. *Artificial Intelligence Review*, *52*(3), 2019–2037. https://doi.org/10.1007/s10462-017-9593-z

Matteelli, A., & Saleri, N. (2008). Respiratory Diseases. *Travel Medicine*, *2*, 561–572. https://doi.org/10.1016/B978-0-323-03453-1.10057-4

Meng, Z. (2019). Ground Ozone Level Prediction Using Machine Learning. *Journal of Software Engineering and Applications*, *12*(10), 423–431. https://doi.org/10.4236/jsea.2019.1210026

Mintz, D. (2013). Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI). *Environmental Protection*, *May*, 1–28. https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf

Moore, P. J., Gallacher, J., & Lyons, T. J. (2018). Random forest prediction of Alzheimer's disease using pairwise selection from time series data. *ArXiv*, *Mci*, 1–14.

OECD. (2012). *OECD Environmental Outlook to 2050*. http://www.oecd-ilibrary.org/environment/oecd-environmental-outlook-to-2050_9789264122246-en

Özbay, B., Keskin, G. A., Doğruparmak, Ş. Ç., & Ayberk, S. (2011). Predicting tropospheric ozone concentrations in different temporal scales by using multilayer perceptron models. *Ecological Informatics*, *6*(3–4), 242–247. https://doi.org/10.1016/j.ecoinf.2011.03.003

Podsiadlo, M., & Rybinski, H. (2015). *Application of Fuzzy Rough Sets to Financial Time Series Forecasting*. https://doi.org/10.1007/978-3-319-19941-2 38

Priambodo, B., & Jumaryadi, Y. (2018). Time Series Traffic Speed Prediction Using k-Nearest Neighbour Based on Similar Traffic Data. *MATEC Web of Conferences*, *218*. https://doi.org/10.1051/matecconf/201821803021

Qin, Z., Cen, C., & Guo, X. (2019). Prediction of Air Quality Based on KNN-LSTM. *Journal of Physics: Conference Series*, *1237*(4). https://doi.org/10.1088/1742-6596/1237/4/042030

Rafaj, P., Kiesewetter, G., Gül, T., Schöpp, W., Cofala, J., Klimont, Z., Purohit, P., Heyes, C., Amann, M., Borken-Kleefeld, J., & Cozzi, L. (2018). Outlook for clean air in the context of sustainable development goals. *Global Environmental Change*, *53*(September), 1–11. https://doi.org/10.1016/j.gloenvcha.2018.08.008

Rahimi, A. (2017). Short-term prediction of NO2 and NOx concentrations using multilayer perceptron neural network: a case study of Tabriz, Iran. *Ecological Processes*, *6*(1). https://doi.org/10.1186/s13717-016-0069-x

Rahman, N. H. A., Lee, M. H., Suhartono, & Latif, M. T. (2015). Artificial neural networks and fuzzy time series forecasting: an application to air quality. *Quality and Quantity*, *49*(6), 2633–2647. https://doi.org/10.1007/s11135-014-0132-6

Raudys, A., & Mockus, J. (1999). Comparison of ARMA and Multilayer Perceptron Based Methods for Economic Time Series Forecasting. *Informatica*, *10*(2), 231–244. https://doi.org/10.3233/INF-1999-10207

Reddy, V., Yedavalli, P., Mohanty, S., & Nakhat, U. (2017). *Deep Air: Forecasting Air Pollution in Beijing, China*. https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/deep-air-forecasting_final.pdf

Rubio, G., Pomares, H., Rojas, I., & Herrera, L. J. (2011). A heuristic method for parameter selection in LS-SVM: Application to time series prediction. *International Journal of Forecasting*, *27*(3), 725–739. https://doi.org/10.1016/j.ijforecast.2010.02.007

Sapankevych, N., & Sankar, R. (2009). Time series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, *4*(2), 24–38. https://doi.org/10.1109/MCI.2009.932254

Seaton, A., Godden, D., MacNee, W., & Donaldson, K. (1995). Particulate air pollution and acute health effects. *The Lancet*, *345*(8943), 176–178. https://doi.org/10.1016/S0140-6736(95)90173-6

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. *Degree Project Technology*, 41. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-237426%0Ahttp://www.diva-portal.org/smash/get/diva2:1259073/FULLTEXT01.pdf

Shahbazi, H., Karimi, S., Hosseini, V., Yazgi, D., & Torbatian, S. (2018). A novel regression imputation framework for Tehran air pollution monitoring network using outputs from WRF and CAMx models. *Atmospheric Environment*, *187*, 24–33. https://doi.org/10.1016/j.atmosenv.2018.05.055

Shiblee, M., Kalra, P. K., & Chandra, B. (2009). Time series prediction with multilayer perceptron (MLP): A new generalized error based approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5507 LNCS*(PART 2), 37–44. https://doi.org/10.1007/978-3-642-03040-6_5

Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., Lu, C., & Zhao, N. (2016). Assessing the potential of random forest method for estimating solar radiation using air pollution index. *Energy Conversion and Management*, *119*, 121–129. https://doi.org/10.1016/j.enconman.2016.04.051

Thissen, U., Van Brakel, R., De Weijer, A. P., Melssen, W. J., & Buydens, L. M. C. (2003). Using support vector machines for time series prediction. *Chemometrics and Intelligent Laboratory Systems*, *69*(1–2), 35–49. https://doi.org/10.1016/S0169-7439(03)00111-4

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Toskala, E., & Kennedy, D. W. (2015). Asthma risk factors. *International Forum of Allergy and Rhinology*, *5*(September), S11–S16. https://doi.org/10.1002/alr.21557

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533–1543. https://doi.org/10.1016/j.enbuild.2017.11.039

Türkay, B. E., & Demren, D. (2011). Electrical load forecasting using support vector machines: A case study. *International Review of Electrical Engineering*, *6*(5), 2411–2418.

Tyralis, H., & Papacharalampous, G. (2017). Variable selection in time series forecasting using random forests. *Algorithms*, *10*(4). https://doi.org/10.3390/a10040114

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Visa, S., Inoue, A., & Ralescu, A. (2011). Proceedings of the Twenty-second Midwest Artificial Intelligence and Cognitive Science Conference. *Maics*, *710*, 120–127.

Vital Strategies. (2021). *Accelerating City Progress on Clean Air: Innovation and Action Guide*. www.vitalstrategies.org/cleanairguide%0Avitalstrategies.org

Voyant, C., Randimbivololona, P., Nivet, M. L., Paoli, C., & Muselli, M. (2014). Twenty four hours ahead global irradiation forecasting using multi-layer perceptron. In *Meteorological Applications* (Vol. 21, Issue 3, pp. 644–655). https://doi.org/10.1002/met.1387

Wahid, F., & Kim, D. H. (2016). A prediction approach for demand analysis of energy consumption using K-nearest neighbor in residential buildings. *International Journal of Smart Home*, *10*(2), 97–108. https://doi.org/10.14257/ijsh.2016.10.2.10

Wang, D., & Lu, W. Z. (2006). Forecasting of ozone level in time series using MLP model with a novel hybrid training algorithm. *Atmospheric Environment*, *40*(5), 913–924. https://doi.org/10.1016/j.atmosenv.2005.10.042

Wang, S.-C. (2003). Artificial neural network. In *Interdisciplinary Computing in Java Programming* (Vol. 743). https://doi.org/https://doi.org/10.1007/978-1-4615-0377-4_5

Willard, J. D., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating physics-based modeling with machine learning: A survey. *ArXiv*, *1*(1), 1–34.

Xu, M., Han, M., & Member, S. (2016). *Multivariate Time Series Prediction*. *46*(10), 2173–2183.

Yafouz, A., Ahmed, A. N., Zaini, N., & El-Shafie, A. (2021). Ozone Concentration Forecasting Based on Artificial Intelligence Techniques: A Systematic Review. *Water, Air, & Soil Pollution*, *232*(2), 79. https://doi.org/10.1007/s11270-021-04989-5

Yeganeh, B., Motlagh, M. S. P., Rashidi, Y., & Kamalan, H. (2012). Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. *Atmospheric Environment*, *55*, 357–365. https://doi.org/10.1016/j.atmosenv.2012.02.092

Yurtseven, E., Vehid, S., Bosat, M., Köksal, S., & Yurtseven, C. N. (2018). Assessment of ambient air pollution in Istanbul during 2003-2013. *Iranian Journal of Public Health*, *47*(8), 1137–1144.

Zheng, H., Li, H., Lu, X., & Ruan, T. (2018). A multiple kernel learning approach for air quality prediction. *Advances in Meteorology*, *2018*. https://doi.org/10.1155/2018/3506394

Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., Wang, Y. S., & Kang, C. C. (2019). Multi-output support vector machine for regional multi-step-ahead PM2.5 forecasting. *Science of the Total Environment*, *651*, 230–240. https://doi.org/10.1016/j.scitotenv.2018.09.111

Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. *Big Data and Cognitive Computing*, *2*(1), 1–15. https://doi.org/10.3390/bdcc2010005

Zhu, J., Wu, P., Chen, H., Zhou, L., & Tao, Z. (2018). A hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model. *International Journal of Environmental Research and Public Health*, *15*(9), 1–19. https://doi.org/10.3390/ijerph15091941

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

BIOGRAPHY

He received his B.Sc. degree in Communications Engineering Department from Al-MA'MON College in 2015. His research areas are IoT, Machine Learning, and Computer Vision.