S	
E1	
E2	
For Office Use Only	

Masters Project Final Report

(MCS)

2019

Project Title	Data Mining Approach for Studying the behavior of Weather Variations in Sri Lankan Context.
Student Name	H.W.V. Tissera
Registration No. & Index No.	2017/MCS/085 17440852
Supervisor's Name	Dr. M.G.N.A.S. Fernando

For Office Use ONLY

Data Mining Approach for Studying the behaviour of Weather Variations in Sri Lankan Context.

A dissertation submitted for the Degree of Master of Computer Science

H.W.V. Tissera
University of Colombo School of Computing
2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: H.W.V. Tissera

Registration Number: 2017/MCS/085

Index Number: 17440852

_____

Signature: Date: 19/11/2020

This is to certify that this thesis is based on the work of

Mr./Ms.

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. M.G.N.A.S. Fernando

_____

Signature: Date:

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATION

RNN	Recurrent Neural Networks
ANN	Artificial Neural Networks
ML	Machine Learning
CCSM3	Community Climate System Model version 3
SSTDM	Spatial and the Spatiotemporal Data Mining
MCRP	Markov-Chain Rainfall Prediction
ACF	Auto Correlation Function
ETL	Extract-Load-Transformation
ARIMA	Auto Regressive Integrated Moving Average
ADF	Augmented Dickey Fuller test
AR	Autoregressive
MA	Moving Average
ACF	(complete) auto-correlation function
PACF	partial autocorrelation function
AIC	Akaike’s information criterion
AICC	Akaike's bias-Corrected Information Criterion
BIC	Bayesian Information Criterion
ME	Mean Error
MSE	Mean Square Error
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percent Error

ABSTRACT

Weather forecasting is the task of determining the future conditions of the atmosphere for a given area. It is a one of the most challenging issues around the world for more than decades. Accurate weather prediction is more important because it has a direct impact on the social and the economic factors. For an example agricultural and industrial sectors are heavily depending on the weather predictions. Since Sri Lanka is a tropical country, it consists with mainly two seasons such as dry and wet season. These seasonal differences are highly impact on the economy of the country because agricultural products and rice production are act as the major role in Sri Lankan economy. Therefore, a reliable weather prediction is necessary to determine the best time to start planting and gaining maximum harvest. Further the meteorological elements such as rainfall, temperature, humidity and windspeed are immensely affect many aspects of human livelihood. They provide analytical support for the issues related to urban computing such as electric power generation planning, traffic flow prediction, air quality analysis and so on.

Therefore, in recent weather prediction has become a more important research area. Hence the researches are more concern with developing a reliable and accurate weather prediction model. So, the main goal of this research is to estimate the weather variations by utilizing the predictive analysis. During this analysis, various data mining and machine learning algorithms are used to develop a better weather prediction model. This research mainly introduces the Artificial Neural Network, Time series analysis, Regression analysis and Decision Tree approaches as the main data mining and machine learning techniques for predicating the weather conditions. Throughout this research it mainly concerns about the rainfall, temperature, windspeed, relative humidity and atmospheric pressure as the weather parameters for developing the predictive model.

The field of machine learning has received much interest from scientific community. Hence machine learning techniques like artificial neural networks are a good candidate for the prediction of weather conditions with large data sets. Also, weather forecasting with time series analysis has become an important mechanism in numerous meteorological applications as well as other environmental areas for determining the phenomena like temperature, relative humidity, pressure, rainfall etc. Therefore, the major concern of this research is to develop a weather prediction model using artificial neural networks and time series analysis while applying the other data mining and machine learning approaches such as and regression analysis and decision trees to achieve a reliable forecasting model.

ACKNOWLEDGEMENT

Foremost, I wish to express my sincere gratitude to my supervisor Dr. M.G.N.A.S. Fernando, senior lecturer of University of Colombo School of Computing - UCSC, for his continuous encouragement, guidance and support for my MCS project. Further his constant supervision with providing valuable resources and immense knowledge helped me in all the time of my research study as well as writing of the thesis.

Besides my supervisor, I also would like to thank MCS project coordinator Dr. Randil Pushpananda for giving all the support during the project work.

Also, I would like to thank all of my colleagues who gave me valuable ideas and support while always helping me to carry out the research successfully.

Finally, my appreciation goes for all of my family members for their valuable support and encouragement endlessly when the times are rough, to make my project success.

CHAPTER 1: INTRODUCTION

1.1. Introduction

Weather variation is a one of a major environmental concern for the livelihoods, food production, hydro-power operations, water availability and forest biodiversity in many countries around the world.[1] Moreover, these weather variations are widely impact on the developing countries located in the tropical regions of the world. Hence Sri Lanka has a high impact with weather variations comparing with the developed countries.

The impacts of weather variations cause an important environmental, economic and social challenges. Thousands of human lives are lost around the globe every year including significant damage on property and animal lives due to natural disasters such as flood, storms, landslides, heat waves, etc.[2] One of the main reasons for the above natural disasters is the unexpected weather conditions. Therefore, the ability to forecast the trends and the extreme events in the local weather variation has an important social and economic consequences.[3] It is essential to successfully address these weather variation challenges to develop the sustainability of modern living conditions, especially in domains such as human health, agriculture, global and regional economic systems and ecological management.

In this research it is being looked for developing a reliable weather predictive model using data mining and machine leaning techniques.

1.2 Problem Domain

This research is mainly correlated with some of the key areas in modern computer science such as data mining and machine learning domains. In this study, it describes how data mining and machine learning techniques can be applied for developing a reliable weather predictive model. Weather is one of a major key aspect of human life because it immensely affects the human activities. Hence it is very important to have reliability and the accurate weather predictions. Weather prediction is an essential application area in meteorology domain, and it is one of the most technologically and scientifically challenging issues around the world. Therefore, the main concern of this study is to develop a reliable weather predictive model using data mining and machine learning techniques.

1.3 Problem

There have been remarkable weather variations during the past few years and decades. Because of these unpredictable weather variations, human beings have faced numerous problems such as natural disasters like floods, earthquakes, tornadoes and tsunami. This will tremendously effects on economic, social and environmental stability.[3]

Therefore, there is a high requirement for a more reliable and accurate mechanism to detect the weather variations.[4] Although there is a necessity for a reliable and accurate weather prediction mechanism in Sri Lankan context, because of the issues with analyzing huge amount of data obtained from remote sensors like satellites, sensor networks, weather radars, etc... [3] and the lack of more precious domain knowledge along with the lack of new technologies compared with the foreign countries and uncertainty of the weather has make it difficult to come up with an accurate and reliable weather predictive model.

Though there is a mechanism for analyzing the weather variations in Sri Lanka, it is still in a poor level and no proper technique to analyze weather changes by applying the new technologies and predict it in advanced.[5] Hence it cannot facilitate the social and economic activities in a wide range.

Therefore, there is a high requirement for introducing a new and reliable weather predictive model by using new technologies like data mining and machine learning techniques to predict the weather variations in advanced for mitigating overall damage occurs due to weather variations.

1.4 Motivation

As mentioned in the above problem section, there is a lack of proper weather prediction mechanism to achieve reliable prediction results in Sri Lankan context. Although there are existing methods to predict the weather variations beforehand, the predictions are not quite accurate for making future decisions based on the predicted output. Hence the several unpleasant circumstances as natural disasters are occurred due to lack of weather forecasting methodologies, although there are precautions available for those issues. Therefore, there is a requirement for a reliable weather variation predictive model for Sri Lankan context. Hence it would be benefit for all the researches if this research gap is fulfilled. Therefore, the main motivation of this study is to take the initial step to reduce the above-mentioned research gap by analyzing the weather data using data mining and machine learning techniques.

1.5 Exact Computer Science Problem

There are considerable number of researches have been carried out under analyzing and forecasting the weather variations throughout several years. But somehow it still remains as ongoing researches due to some challenging factors associate with this research domain. Some of the factors are dependability of the weather variables on many human activities, advancement of various technologies that are directly related to this research domain such as evolutionary of the computation and the improvements in the measurement systems. Hence as mentioned in the problem section above, coming up a better and a reliable solution for weather prediction is still remain as a challenging task. Therefore, the actual computer science problem that is going to be addressed by this project is, applying data mining and machine learning methodologies for finding a better approach of predicting the weather variations. Predictive analytics is mainly driven by using predictive modelling and predictive models are generally include machine learning algorithms. These models can be trained over time for responding to new values or data by delivering the results as the requirements specified.

1.6 Research Contribution

1.6.1 Goal

The main goal of this study is to develop a reliable approach which is capable of predicting the weather variations using and data mining and machine learning techniques.

1.6.2 Objectives of the Study

- Identify the factors which are mainly influence on weather variations by doing the literature surveys and background studies.

  For achieving this objective, it needs to perform a critical study about the area of weather changes through literature surveys and background studies. When identifying the factors, it is necessary to give priority for the factors which causes high impact on the weather changes. Temperature, windspeed, relative humidity, windspeed and atmospheric pressure are some of the main factors which highly affect to the weather variations.

- Identify how the weather variations have been occurred in the Sri Lankan context according the identified factors.
- Identify the techniques that can be used to develop the weather predictive model.

Here it needs to perform a critical survey on techniques, for finding out the most appropriate mechanism to solve the problem. There are several techniques such as time series analysis, regression analysis and clustering analysis and artificial neural networks that can be useful when predicting the weather changes.

- Design and develop a reliable weather predictive model, as a solution for addressing the issues related to weather predictions in Sri Lanka.
  By considering the weather changing factors and the appropriate data mining and machine learning techniques, it needs to develop a weather predictive model which facilitate future predictions.

- Evaluate the proposed weather predictive model to ensure the accuracy and the reliability of the model.
  The proposed predictive model needs to be evaluated with the past data records of the weather changes and the evidences, to ensure the accuracy of the model. If the proposed model outputs are mapped correctly with past weather data evidences, then it can be determined as the solution is reliable and accurate.

1.7 Scope

This project involves developing a reliable weather predictive model for predicting the weather variations in Sri Lanka. The actual weather data which obtained from the Department of Meteorology in Sri Lanka are used for developing the proposed weather predictive model because they are highly related to the Sri Lankan domain. Further based on the availability of the services; data sets which obtained from the web APIs and the web sites such as weather underground[6] are used to collect the data sets. When analyzing the weather changes with the proposed weather predictive model, we specially focus on the weather variables which mainly affects to the weather variations such as temperature, wind speed, relative humidity, rainfall and atmospheric pressure as the inputs for this model. Based on the availability of the data related to the weather changes, other weather variables such as precipitation and the cloud formation also can be used as the inputs for this proposed weather predictive model.

Since the data we use for developing the predictive model are raw data, they need to be summarized using the suitable methods and should create an aggregated data set. Further Data mining and machine learning techniques are used for the analysis of the data when developing this new model. Finally, an evaluation is performed to check the accuracy and the reliability levels of the new weather predictive model.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In this chapter we give a critical review of the research in the area of weather variations for identifying the weather elements that lead to a high impact on weather variations and the available methodologies for implementing a reliable weather prediction model. For this purpose, this chapter has divided in to sub sections such as, main reasons for weather and climate variations and impacts associate with these variations, existing approaches for predicting the weather variations and the future trends in developing enhanced and reliable weather predictive models. This Chapter also define the research gaps with existing methodologies related to the weather variations by considering the literature review.

2.2 Main reasons for weather and climate variations and impacts associate with these variations.

There are several researches going on for developing more reliable and accurate, climate and weather predictive models. By using these predictive models it enables to evaluate the climate and weather variations happen in globally because it is one of the most important area for the existence and well-being of the humans.[3] [7] [2] Various researches have been conducted for identifying the main courses for the climate and weather changes.

Throughout those researches, researchers have identified that there are more common environmental factors which affect to the climate and weather changes immensely such as rainfall, temperature, relative humidity, air pressure, wind circulation, atmosphere, etc.[7] [8] Other than the above mentions major factors; distance from sea, ocean currents, relief and proximity to the equator also affect to the weather and climate changes.[3]

Other than the environmental factors, human related activities such as forest distortion, population increasing, emission of greenhouse gases, fuel burning, cities, farms and usage of lands also highly affect to the climate and weather changes.[3] There are mainly three types of climate related data which can be used for the knowledge discovery. They are temporal data, spatial data and the spatiotemporal data.[3] It is a rigorous task to discover knowledge from temporal, spatial and spatiotemporal data related to the climate changes.[3] There are several challenges associated with these data such as nonlinear dependence, long memory processes in time, analysis of extremes, regional change and nature of historical observations.[3] [7] Historical data may be incomplete and noisy. When data is noisy and incomplete it will

complicate the study of extremes and correlations.[7] On the other hand the weather and climate related data are generated from the remote sensors like satellite and weather radars, situ sensors and sensor networks and etc. The outputs get from the climate or earth system models produce terabytes of temporal, spatial and spatiotemporal data. Therefore, the rate of data generation and the storage is far exceeding.[3] Hence it lost the opportunity to analyze the data and get the insight knowledge about the data properly.

Because of the weather variations, there is a huge impact occurs on economic, social and environmental stability. It causes huge damage on human lives and living conditions of the humans. Weather variations will exacerbate the already worsening situations in industries such as agriculture, fisheries, financial sectors in developing countries while it will do economic damage to the developed countries.[3] [2] In the below section it describes how data mining and machine learning can be used to overcome the above-mentioned impacts caused by climate and weather variations.

2.3 Existing approaches for predicting the weather variations.

Among ongoing several researches for evaluating the weather variations happen in globally, one of the most reliable existing climate and weather modelling tools is Community Climate System Model version 3(CCSM3). CCSM3 can be identified as a framework which enables test and build different kind of climate models instead of acting as a single climate model.CCSM3 is consisting with four sub models such as land, sea-ice, ocean, atmosphere and connected with a coupler which exchanges information with the sub models.[3] Although there are accurate models like CCSM3 globally, it generates terabytes of data. Therefore, it is a challenge for mining or the analysis of the climate data. Climate refers to the average weather conditions over a long period of time, normally 30 years whereas, Weather refers to the atmospheric conditions in a specific place over short period of time. [9]. Hence climate models are used in globally and limited way because of the need of high processing power for terabytes of data. But the weather prediction models can be developed without high requirement of the super computers and more powerful hardware for processing the data. Although it is not necessarily super computers for processing the weather data, it is a rigorous task to mine and analysing the data.

Because of that spatial and the spatiotemporal data mining (SSTDM) methodologies need to be introduced. [3] [2] In the theoretical foundation of the SSTDM they imply that since leaning samples are not independent, the traditional data mining techniques are not enough for the

mining purpose. They emphasise that the correlations and the seasonal effects also need to consider when applying the data mining techniques.[3] [7] Some of the researches have conducted the researches to identify the various correlations between the factors which effects on the weather changes.[10] Also, they have identified that various data mining and machine learning techniques such as artificial neural networks, time series analysis [11], clustering analysis, regression analysis can be applied to explore the weather data accurately. [12] [2] when handling the nonlinear regressions support vector machines can be used.[2] With the rapid development of data mining and machine learning techniques, researches have tend to find more and more forecasting rules which can be applied for the meteorological and computer science area for determine the significance of data mining and machine learning in the area of weather forecasting. Weather forecasting is a one of the most sophisticated and challengeable task in machine learning domain because of the collection of massive volume of data, noise and the missing data, large search space and the complexity of the knowledge that needs to be discovered.[13]

One of the major factor which indicate the weather variations is rainfall. [7] [2] Therefore, there is need for an accurate mechanism for predicting the rainfall derivativities using data mining and machine learning techniques. Rainfall shows unique characteristics of high volatility and chaotic patterns that doesn't exists in other time series.[2] [14] Following are the machine learning algorithms that can be applied for rainfall prediction using rainfall derivatives. They are Markov chain analysis, Generic programming, Support vector machines, K-nearest neighbours, Radial basis neural networks, M5 rules and M5 model trees.[2] According to the literature, the most commonly and successful used rainfall prediction is the Markov-Chain extended with rainfall prediction(MCRP).[2] [15] MCRP is mainly have two stages. First stage is to produce an occurrence pathway using a Markov-chain (a sequence of rainy or dry days). The second stage is to generate a random rainfall amount (from a distribution) for every rainy day in the sequence. MCRP is having advantages such as it is a simplistic and lightweight algorithm for the problem of daily rainfall prediction.[2] But there are some disadvantages associate with MCRP such as heavily reliant on past information and that information are being reflective of the future. Because MCRP takes the past average value to be the major contribution for the future rainfall.[2] Therefore it produces weak predictive models that the annual derivations and short term behaviour in rainfall are not be captured. Hence it doesn't produce a general model that can be applied to the all cities.[7] [2] [15] Most of the machine learning applications are used for the short-term predictions such as for few hours and monthly amount of rain fall. Machine learning algorithms such as feed-forward back

propagation neural networks, multivariate time series analysis and Generic programming can apply to predict the rainfall derivatives.[2] [16] According to the literature, Artificial Neural Networks (ANN), Time Series Analysis and Regression Analysis [4] [17] [11] have been widely used for simulating and forecasting of meteorological variables such as rainfall, temperature, wind speed, relative humidity and atmospheric pressure. But those algorithms also support only when the specific conditions are satisfied. [18]

2.4 Future trends in developing enhanced and reliable weather predictive models.

According to the 2.1 section, developing a predictive model for identifying the weather variations is major concern in weather forecasting researches. For achieving this goal, data mining and machine learning approaches can be used to develop a reliable weather predictive model. But proposing an accurate and reliable mechanism and algorithm for predicting the weather variations is a research talent in data mining and machine learning research area.

Research	Achievements	Limitations
[3]	• Entire field of time series analysis and forecasting relies on auto correlation function (ACF) and Fourier transformation of ACF • Dependent among time series whether linear or non-linear remains important.	• Performance is low
[7]	• fingerprinting technique is used to predict extreme rainfall events • Consider rainfall data as well as the wind pattern data to develop the model • Performance is high.	• coarse-resolution data is used because of the unavailability of fine-resolution observed/reanalysis data
[2]	• Predicting accumulated rainfall amounts using a Sliding Window Algorithm • Predict the accumulated rainfall amounts (rather than predicting daily rainfall)	• Only consider about the rainfall data analysis
[13]	• Developed an approach called inexact field learning to produce high quality rules from the low-quality data, the FISH-NET algorithm.	• Only consider about the rainfall data analysis

	• FISH- NET algorithm achieves better prediction rate comparing with C4.5 algorithm, K- nearest neighbor method and discriminant analysis algorithm. • FISH- NET algorithm overcomes the LPA (Low prediction Accuracy) problem on large low-quality data sets.	
[19]	• Detecting climate changing patterns using multivariate time series. • Used clustering and cluster tracing technique	• Pattern may break in to periodically appearing substructures.

Table 1: Summary of findings of the existing researches

2.5 Approach for filling the existing research gap by using literature review and novel methodologies.

Based on the literature review, there are large number of researches are conducting for solving the problems associate with weather forecasting. Researchers are very keen about finding the new methodologies for developing reliable and accurate weather forecasting models. Among those methodologies, developing weather predictive models using data mining and machine learning technologies has gained high consideration because data mining and machine learning methods are enabling to explore and find the hidden patterns in the data sets. [3] [2] [12] Though there are several data mining and machine leaning methods are available for exploring the weather related data such as Artificial Neural Networks (ANN), time series analysis, clustering analysis, regression analysis, support vector machines, K-nearest neighbours and Bayesian Networks [2] [19] there are several limitations associated with these analysing methods. Limitations addresses such as performance issues, handling noise and missing data, outlier detections, detecting co-relations between weather variables and etc. Therefore, there is a high requirement for introducing a new and reliable weather predictive model by using data mining and machine learning techniques while addressing the limitations of the existing weather forecasting models to overcoming the existing challenges. Further there is no proper mechanism to address the weather variations in Sri Lanka because of the less researches done related to the weather variations in Sri Lankan Context. Hence by developing a reliable weather forecasting model, it benefits for mitigating overall damage occurs due to weather variations in advance.

2.6 Summary

This chapter presented a critical review of the research in the area of data mining and machine learning approach for predicting the weather variations. As the major output of the literature review, we have identified the research problem as the unavailability of a reliable and accurate weather predictive model for predicting the weather variations in Sri Lankan context. Although there are several researches are going on for identifying the weather variations globally,[20] [1] [21] there is no proper mechanism to address the weather forecasting in Sri Lanka properly. Therefore, there is a necessity for predicting weather variations in Sri Lankan context for obtaining the economic, social and environmental stability.

CHAPTER 3: METHODOLOGY

3.1 Problem Analysis

In this section we investigate the research problem to gain more insight understanding about the problem domain for suggesting a better and practical solutions for solving it. According to the carry out study there is a requirement for a reliable weather predictive model for forecasting the weather variations. The aim of this study is to develop a weather predictive model using data mining and machine learning techniques. According to the literature, there are several factors that affects to the weather variations. Among those factors temperature, windspeed, relative humidity, atmospheric pressure and rain fall are mainly affecting to the weather changes. [1] [22]

Researchers have identified that there are so many environmental factors which affect to the weather changes apart from the main factors like temperature, windspeed, relative humidity, atmospheric pressure and rain fall. Other environmental factors can be categorized as distance from sea, ocean currents, relief and proximity to the equator.[3] Apart from the environmental factors, human related activities such as forest distortion, population increasing, emission of greenhouse gases, fuel burning, cities, farms and usage of lands also highly affect to the climate and weather changes.[3]

According to this study it only considers weather factors such as the temperature, wind speed, relative humidity, atmospheric pressure and rainfall for carrying out the research. During past years there are several researches have been conducted for identifying the relationships between these variables. But the problem is still remaining in the unsolvable condition because of the uncertainty of the weather variations and the dependability of the core weather variation factors with the other environmental factors. According to the literature researches have used the data mining and machine learning techniques to achieve the above-mentioned problem up to a considerable extent.[3] Data mining is defined as the process in databases that used to discover, extract and reveal previously unknown , hidden , meaningful and useful patterns. Data mining is the analysis step of the “knowledge discovery in databases” (KDD) process.[3] There are several data mining techniques that can be applied base on the requirement. Some of the data mining techniques are Regression Analysis, Classification, Association Rule Discovery, Clustering, decision trees and etc. Data mining techniques can be used for identifying the automatic pattern predictions based on the trend and behavior analysis and make predictions based on the trend and the behavior analysis.

Machine Learning (ML) is identified as the brain science of modern computer science. It can be categorized as a subbranch that comes under the tree of Artificial Intelligence (AI). By using machine learning techniques, it enables to study about the pattern description and predictions with the help of statistical based algorithms. These algorithms can act automatically without any human interference. Most of the scenarios these algorithms have three phases such as implementing, training and testing phases. After trained the algorithm with appropriate data, these algorithms can find some hidden information from newest data which the algorithm couldn't met before. Therefore, these algorithms act as filters that enables to separate and extract information from data. Hence machine learning techniques will replace the position of the newest computer brain science as it uses prior experiences which have met earlier or some basic instruction to determine and extract some unseen information from new data. Therefore, by using data mining and machine learning techniques it enables to develop a reliable weather prediction model.

3.2 Proposing Model / Design

The main focus of this project is to develop a solution model for predicting the weather conditions using data mining and machine learning approaches. As the first step it needs to identify the factors which causes the weather variations by doing the literature surveys and background studies. There are several factors that affects to the weather variations. Among those factors' temperature, wind speed, relative humidity, atmospheric pressure and rainfall are mainly considered as the Weather parameters in this study. Since these factors are specific data which belong to the meteorology domain, mainly the data sets are acquired from the Department of Meteorology in Sri Lanka. Other than the data obtained from the Department of Meteorology, the data sets are also acquiring from the web APIs and the web sites such as 'Weather Underground'[6] based on the availability and the applicability of the data to the Sri Lankan domain.

Data mining and Machine learning approaches are used to develop this weather predictive model. When applying the data mining techniques, it needs to have a comparably large data set for the analysis purpose. Since the data obtained from the various resources are raw data, we need to apply the preprocessing techniques and the other relevant techniques to summarize the data set with related variables. Then this aggregated data set can be used for the analysis of the weather variations.

The following sections describes how the data mining process can be carried out for the raw weather data that are collected from the various resources. After completing the data collection from various sources, preprocessing techniques need be applied. Data preprocessing consist with main stages such as data cleaning, data integration and data reduction. Under data cleaning step missing, duplicated and the faulty data will be removed from the database. During the data integration step, it converts the different data types which obtained from different databases into a single type. This can achieve by using data synchronization tools, data migration tools and the Extract-Load-Transformation (ETL) process. In the data reduction step, it will reduce the number of data and variables which will be suitable for a short-term analysis process if the analysis results are not changed. Then the data selection step is performed and set of samples that are appropriate for the query are selected. As the next step data transformation is performed and, in this step, it transforms data into relevant form required by the mining procedure. Finally, in the data mining step, relevant data mining techniques are applied for the final data set. When analyzing these summarized data, several data mining and machine learning techniques such as Artificial Neural Networks, Regression analysis, Time series analysis and Decision Trees are used to develop the weather prediction model. By applying the above techniques to the data set and observing output results we can identify the most appropriate techniques which gives the accurate data analysis related to the weather variations. And finally, the discovered knowledge is evaluated according to the novelty and the validity in the pattern evaluation process. Then based on the output results, it is possible to develop a weather predictive model which can predict the weather more reliably. After coming up with the predictive model we can divide the data set into training set and test set and evaluate the model for checking the accuracy with the future predictability of the proposed weather predictive model.

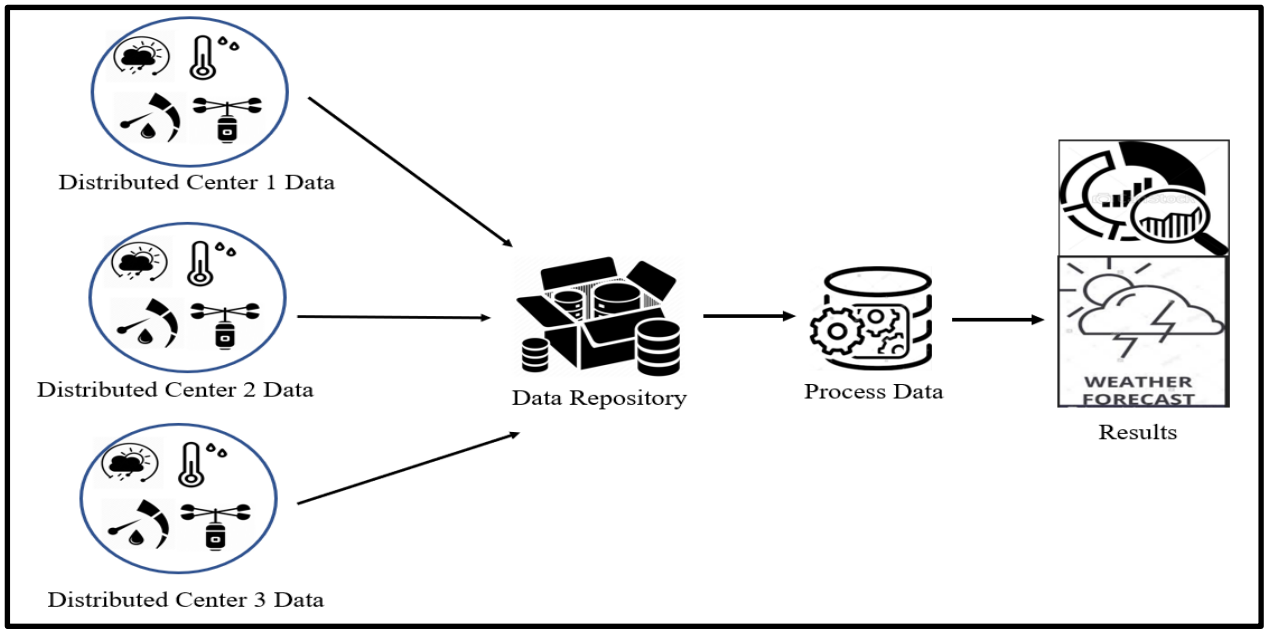


Figure 1: High Level Diagram

3.3 Dataset Creation

3.3.1 Weather Data Set

Generally, most of the machine learning techniques are involved with the data mining domain problems and associates with data sets. The data used in this work was collected from the 'Weather Underground'[6] website and the Department of Meteorology in Sri Lanka . As shown in Figure 2, data set consist of daily Temperature, Dew Point, Relative Humidity, Windspeed, Pressure, Rainfall and Precipitation measurements in the period **01/01/2013 to 30/09/2020** from weather station **Ratmalana Airport Station** at latitude **6.92** °N and longitude **79.83** °E in Colombo, Sri Lanka. The dataset consists with approximately 2820 data records.

DATE	TEMP_MAX	TEMP_AVG	TEMP_MIN	DEWP_MAX	DEWP_AVG	DEWP_MIN	RH_MAX	RH_AVG	RH_MIN	WINDSP_MAX	WINDSP_AVG	WINDSP_MIN	PRESSURE_MAX	PRESSURE_AVG	PRESSURE_MIN	RAINFALL	PRECIPITATION
2013-01-01	88	82.5	77	74	72.3	69	91	72.8	54	8	6	3	29.8	29.8	29.7	22.3	0.54
2013-01-02	88	82.3	76	74	72.8	71	88	74.5	58	9	6.5	3	29.9	29.8	29.8	0	0
2013-01-03	84	79	77	75	74.3	74	91	85.5	71	12	5.3	1	29.9	29.8	29.8	0	0
2013-01-04	78	77	75	75	73.8	72	95	90.3	85	6	5	3	29.9	29.8	29.8	33.1	0.68
2013-01-05	83	83	83	70	70	70	64	64	64	12	10.5	9	29.8	29.7	29.7	8	0
2013-01-06	86	81.3	75	71	70.5	69	86	70	61	9	6.8	2	29.8	29.8	29.7	0	0
2013-01-07	83	77.5	74	76	72.5	68	95	85	76	9	5	0	29.8	29.8	29.7	0	0
2013-01-08	85	80	73	76	73.3	71	95	81.8	63	8	3.5	0	29.9	29.8	29.8	37.8	0.7
2013-01-09	83	79.3	77	75	73.7	72	95	82.7	73	13	8.7	6	29.8	29.8	29.8	32.2	0.68
2013-01-10	78	76.8	76	76	74.5	74	95	93.3	90	6	5.5	5	29.9	29.8	29.8	0	0

Figure 2: Weather Data Set

3.3.2 Data Preprocessing

In almost all the datasets that are used in the data mining applications are consisting with the raw data. Raw data is highly affected with the missing values, noise, and inconsistency. Therefore, the quality of the data has a high impact on the data mining results. In order to improve the quality of the data which use for the data mining, raw data needs to be pre-processed. By applying the pre-processing techniques, it will improve the efficiency and the ease of the mining process.

Therefore, in order to achieve a well-organized data set, following pre-processing techniques has applied to the Weather dataset.

- Estimating missing values

Finding the missing attribute values is a one of the most significant processes in data pre-processing phase. Also, it is a very important issue in data mining. Missing attribute values are more common in several real-world data sets. They possibly will come from the data collecting process or repeated diagnoses tests due to human error, machine failure or routine maintenance, any transformation in the experimental set up and indefinite data.[23] These incomplete datasets will cause the biasness due to difference between observed and unobserved data. Hence it is necessary to apply methods to estimate the missing values. There are two types of missing values. They are block missing (one- day data lost) and local missing (local non-continuous time series),which will differ in severity.[24] [18] By removing all data containing the missing values lead to a different interpretation for the characteristics of the real data. Therefore, understanding and handling of original circumstance with background knowledge to allocate the missing values seem to be a most favourable approach for handling missing attribute values.[23] But in actual scenarios, it is extremely complicated to know the unique meaning for the missing data or attributes. In practice there are several approaches to handle the missing information in an uncomplicated manner. For instance, the best way is substituting missing values with the global or class-conditional mean/mode. [23] Therefore, in this study the data associated with the block missing values are deleted from the dataset and local missing data is substituted with the mean value. Instead of the substitution of mean value method, linear interpolation method also can be used.

There are missing values in the Temperature, Relative Humidity, Rainfall and Precipitation attributes. So before applying the machine learning algorithms to the data set, those missing values need to be filled. The missing values of the Temperature, Relative Humidity, Rainfall and Precipitation attributes are replaced by the mean value of consecutive 7 days.

- Remove Null Values

One of the important steps in data wrangling is removing null values from the dataset. Because these null values adversely affect to the performance and accuracy of any machine learning algorithm. Hence it is important remove null values from the dataset before applying the machine learning algorithms to the dataset.

- Normalization of continuous variables

When there are continuous variables without normalization, it will sometimes cause in training failures in neural networks and the deep learning approaches. Therefore, it needs to use the min-max normalization to normalize every continuous variable in to [0,1]. And in the evaluation process it will re-normalize the predicted values in to the normal scale again. [18]

3.4 Solutions for Weather Prediction Approach

As described in the early chapters several data mining and machine learning algorithms and approaches can be applied for weather forecasting. These algorithms are different from each other with regards to their core concepts of analyzing the data and the approaches of applicability for the weather predictions. Therefore, the efficiency and the accuracy of the weather prediction model depend on the algorithm selected for the analyzing the data.

In this research basically we are using Artificial Neural Network, Time Series Analysis, Regression Analysis and Decision Tree Analysis for developing the weather prediction model. Python programming language is used to implement the machine learning algorithms.

3.4.1 Artificial Neural Network

Artificial neural network (ANN) is an effective machine learning technique for developing the computerized system that is capable of processing non- linear weather conditions inside a specific domain, and make predictions. Artificial neural network is inspired by biological neuron model and numbers of highly nonlinear neurons are interconnected for forming a network. [25] The neural network consists of three layers as input, hidden and output layers. These neurons are connected by links which comprises of weight. Weights represent the connection quality that exists between the neurons in the system.[17] [25] Basically Artificial neural network receives the input, then process the data and finally gives output with respect to input. When the system becomes more complex, the network is also becoming large.[25] A multilayer neural network consists of input layer, one or more hidden layer and output layer. As weather is data-intensive process and the dataset is highly non-linear therefore the predictions can be done accurately using artificial neural network. Also, an artificial neural

network is a powerful data-driven, self-adaptive, flexible computational tool that enables the capability of capturing nonlinear and complex underlying characteristics of any physical process with a high degree of accuracy.[25]

In this study, Long Short-Term Memory (LSTM) recurrent neural network architecture is used as one of a machine learning technique for make predictions of daily rainfall with use of other weather variables such as daily average temperature, windspeed, relative humidity and atmospheric pressure.

A Recurrent Neural Network (RNN) is a class of ANN where connections between units form a coordinated chart along a sequence. Recurrent neural network can utilize their internal memory to process sequences of inputs. [25] RNN can recall vital things about the information it receives. Hence it empowers extremely in forecasting what's coming next. This is the main motivation behind selecting the recurrent neural networks as the major machine learning technique for analyzing the data like weather, time series, speech, text and financial data. [4] [17] [25]

Long Short–Term Memory (LSTM) networks are an extension of recurrent neural network, which essentially broadens their memory. The units of Long-Short Term Memory networks are used as building units for the layers of a recurrent neural network, which is then often known as an LSTM network.[2] [25] Long Short–Term Memory allow recurrent neural networks to recall their inputs over a long period of time. The main reason is that the neural network can contain their data in memory which is much similar to the memory of a computer in the light of the fact LSTM can read, write and erase data from its memory. [25]

The proposed model for weather forecasting using recurrent neural network with LSTM algorithm predicts the rainfall as the output variable while taking average temperature, dew point, relative humidity, windspeed and atmospheric pressure as the input variables.

3.4.2 Time Series Analysis

Time series analysis consists with methods for analyzing the time series data in order to extract meaningful statistics and other characteristics of the data. Forecasting with time series analysis is an use of a model to predict future values based on previously observed values. [26] [11] This time series data can be defined as the data that is present in a series of particular time periods or intervals. Further the time series analysis and forecasting results has become a major methodology in numerous hydro-meteorological applications as it allows to study trends and variations in weather variables like temperature, wind speed, relative humidity, atmospheric pressure, and rainfall. [1] In this study mainly ARIMA (Auto Regressive Integrated Moving Average) models has been carried out to predict the daily average temperature, windspeed,

relative humidity and the atmospheric pressure. This ARIMA model also called as the Box-Jenkins models as well.[1] In the ARIMA model, autoregressive model of order p is conventionally classified as AR (p) and a moving average model with q terms is known as MA (q). A combined model that contains p AR-terms and q MA-terms is called an ARMA (p, q) model.[1] In order to make the nonstationary time series in to stationary time series, it is shifted by d lags, where in most cases d =1 and difference(d) is computed before further processing. This type of model is identified as ARIMA (p, d, q), where the symbol "I" denote as "integrated".[1] When identifying the most appropriate ARIMA model for a selected time series, Box and Jenkins [21] has proposed a methodology with four steps.

They are,

I. Identification of the model
II. Model parameter estimation
III. Perform diagnostic checking for measure appropriateness of the model
IV. Forecasting (Application of the finalized model) [1]

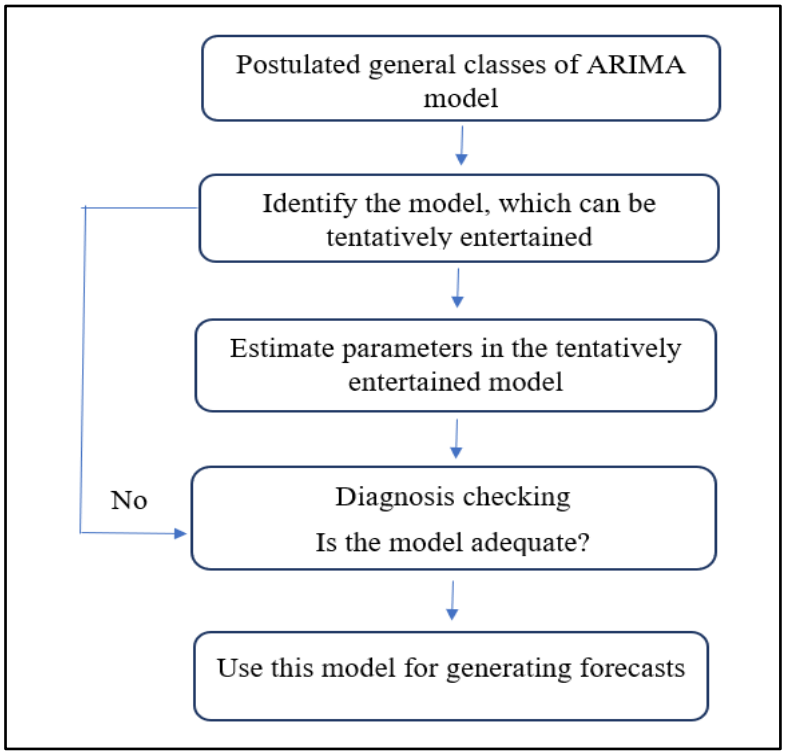


Figure 3: Box-Jenkins approach for selecting optimal ARIMA model

There are several characteristics associated with the time series data. The main characteristics are trend and the seasonality. Trend describes the increasing or decreasing value in the series and seasonality describes the possibility of having repeated short-term cycle in the series.

Therefore, as the first step of developing the Box-Jenkins model, it is necessary to determine the stationarity of the time series and check whether there is any notable seasonality that needs to be addressed. The stationary data is data that doesn't change the mean, variance and autocorrelation factors over the time. In order to test the stationarity of the time series, an Augmented Dickey Fuller test (ADF) is needed to perform on the time series dataset. An Augmented Dickey Fuller test (ADF) is a testing strategy for performing a unit root in time series data samples.[20] The result value of the Augmented Dickey-Fuller (ADF) statistic is a negative number.[20] [27] The more negative the number is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.[27] Unit root test tests whether a time series variable is non-stationary or not. [28] Detecting the seasonality of the data set is the major concern during the model identification stage. if there is a seasonality exists in the data set,

it is necessary to identify the order for the seasonal autoregressive (AR) term and order for the seasonal moving average (MA) term. And also, it needs to find the stationarity of the non-seasonality parameter (d) by performing the Augmented Dickey-Fuller test. If it determined that the time series is stationary, then the value for the non- seasonality parameter (d) becomes zero. (d =0).

The next step of developing the Box-Jenkins model is to determine the orders of the seasonal autoregressive (AR) term, p and the seasonal moving average (MA) term q in the ARMA (p, q) model. This can be achieved by the inspection results of the ACF, PACF autocorrelation plots. Then based on the inspection results of the ACF, PACF autocorrelation plots, the most appropriated orders of the ARIMA models can be determined and evaluated using the AIC criterion. When evaluating the model quality, popular criteria are Akaike's information criterion (AIC) Akaike's bias-corrected information criterion (AICC) and Bayesian information criterion (BIC). AIC is mainly used to select between different models where the lowest score is the most preferable. During this study we have used the AIC criterion to select the most appropriate model.

The third step of developing the Box-Jenkins model which is diagnostic checking to measure appropriateness of the identified ARIMA model can be achieved through the model residuals of the ACF and PACF plots. The final step as well as the end goal of ARIMA model which is forecasting the one or more future time steps of the time series can be achieved through applying the measurements obtained in the previous steps to the ARIMA model. ARIMA model can be defined as follows.[29]

ARIMA (p, d, q)

p - Number of autoregressive terms

d - Number of nonseasonal differences needed for stationarity

q - Number of lagged forecast errors in the prediction equation

3.4.3 Regression Analysis

Linear Regression is a machine learning algorithm based on supervised learning.[4] It is a direct technique of demonstrating the connection between a scalar reaction, also known as the dependent variable (Y) and one or more explanatory variables or independent factors (X).[2] [4] . The generalized formula for a Linear Regression model can be present as follows.[4]

$$\hat{y} = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + ... + \beta(p\text{-}n)\, x(p\text{-}n) + E$$

$\hat{y}$ - predicted outcome variable (dependent variable)

x_j - predictor variables (independent variables) for j = 1, 2,., p-1 parameters

β_0 - Intercept or the value of $\hat{y}$ when each xj equals zero

β_j - change in $\hat{y}$ based on a one-unit change in one of the corresponding xj

E - random error term associated with the difference between the predicted $\hat{y}_i$ value and the actual y_i value

The most basic form of building a Linear Regression model relies on an algorithm known as Ordinary Least Squares which finds the combination of βj's values which minimize the E term.[2] The key assumption of the linear regression technique is the linear relationship between the dependent variable and each independent variable. Pearson correlation coefficient is a one way of assessing the linearity between dependent variable and independent variables.[11] The Pearson correlation coefficient (r) is a measurement of the amount of linear correlation between equal length arrays which outputs a value ranging -1 to 1. Correlation values ranging from 0 to 1 represent increasingly strong positive correlation while correlation values from 0 to -1 represent negative correlation.[30] Following Table 2 represent the clear-cut boundaries for the levels with strength of a correlation coefficient.

Correlation Value	Interpretation
0.8 - 1.0	Very Strong
0.6 - 0.8	Strong

0.4 - 0.6	Moderate
0.2 - 0.4	Weak
0.0 - 0.2	Very Weak

Table 2: Correlation Coefficient Interpretation

In linear regression models a technique known as step-wise regression is applied to add or remove variables from the model and assess the statistical significance of each variable on the resultant model.[30]

The proposed model for weather forecasting using regression analysis predicts the Average temperature considering the previous three days weather parameter values of average, minimum and maximum values of Dew Point, Relative Humidity, Windspeed, Atmospheric Pressure and Precipitation.

3.4.4 Decision Tree Analysis

Decision tree analysis is one of the predictive modelling approach use in data mining and machine learning techniques.[31] It is a graphical representation of decisions and their corresponding effects in both qualitatively and quantitatively. The structure of the methodology is in the form of a tree; hence it is named as decision tree analysis. Decision tree analysis is used to scale the decision after it has been made and as viewing its consequences under imagined conditions can give an insight understanding of what to expect in similar conditions in near future.[32]

In this study, the Decision tree classifier is used to determine the conditions for day being a Rainy day or a Not Rainy day.

3.5 Implementation

This section discusses the implementation details of the proposed weather prediction model. As mentioned in the section 3.2, mainly there are four machine learning techniques have been used for analyzing the weather data set. They are Artificial Neural Networks, Time series analysis, Regression analysis and Decision tree analysis. Section 3.5.1 provides the implementation details of the weather forecasting model using recurrent neural network while section 3.5.2 addresses the implementation details using time series analysis, section 3.5.3 includes the regression analysis and section 3.5.4 includes the decision tree analysis implementation details.

3.5.1 Implementation details related to Recurrent Neural Network

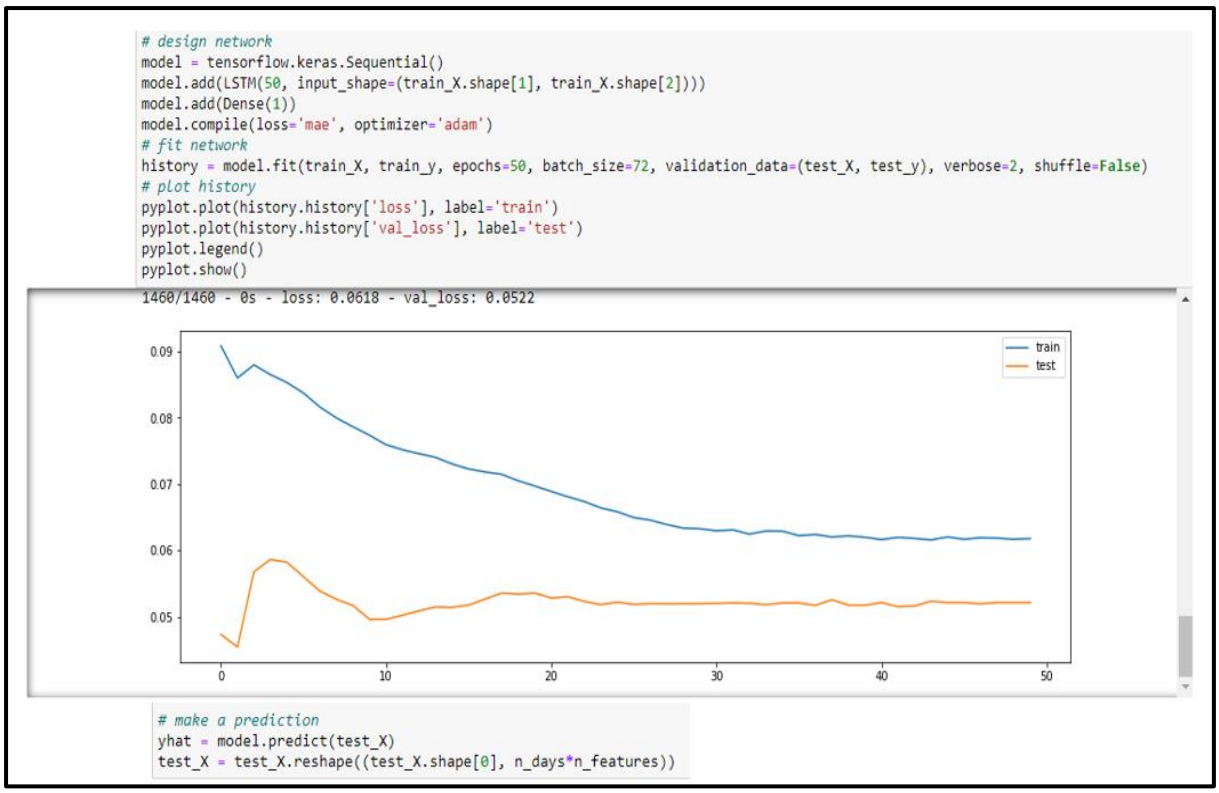
This study is carried out using recurrent neural network with LSTM algorithm to predicts the rainfall as the output variable while taking average temperature, dew point, relative humidity, windspeed and atmospheric pressure as the input variables.

```
# Prepare the Weather dataset for the LSTM
# 1.Framing the dataset as a supervised learning problem
# 2.Normalizing the input variables.

# convert series to supervised learning
def series_to_supervised(data, n_in=1, n_out=1, dropnan=True):
    n_vars = 1 if type(data) is list else data.shape[1]
    df = DataFrame(data)
    cols, names = list(), list()
# input sequence (t-n, ... t-1)
    for i in range(n_in, 0, -1):
        cols.append(df.shift(i))
        names += [('var%d(t-%d)' % (j+1, i)) for j in range(n_vars)]
# forecast sequence (t, t+1, ... t+n)
    for i in range(0, n_out):
        cols.append(df.shift(-i))
        if i == 0:
            names += [('var%d(t)' % (j+1)) for j in range(n_vars)]
        else:
            names += [('var%d(t+%d)' % (j+1, i)) for j in range(n_vars)]
# put it all together
    agg = concat(cols, axis=1)
    agg.columns = names
# drop rows with NaN values
    if dropnan:
        agg.dropna(inplace=True)
    return agg

# normalize features
scaler = MinMaxScaler(feature_range=(0, 1))
scaled = scaler.fit_transform(values)
scaled
```

Figure 4: Code segment for preparing weather dataset for the LSTM




```
# design network
model = tensorflow.keras.Sequential()
model.add(LSTM(50, input_shape=(train_X.shape[1], train_X.shape[2])))
model.add(Dense(1))
model.compile(loss='mae', optimizer='adam')
# fit network
history = model.fit(train_X, train_y, epochs=50, batch_size=72, validation_data=(test_X, test_y), verbose=2, shuffle=False)
# plot history
pyplot.plot(history.history['loss'], label='train')
pyplot.plot(history.history['val_loss'], label='test')
pyplot.legend()
pyplot.show()
```

```
1460/1460 - 0s - loss: 0.0618 - val_loss: 0.0522
```

```
# make a prediction
yhat = model.predict(test_X)
test_X = test_X.reshape((test_X.shape[0], n_days*n_features))
```

Figure 5: Code segment for designing neural network and make predictions

3.5.2 Implementation details related to Time Series Analysis

This study is carried out using ARIMA model for predicting the weather variations under time series analysis. This model has implemented for forecasting the values relates to the weather variables such as temperature, windspeed, relative humidity, rainfall and atmospheric pressure.

One of a main requirements of the time series analysis is to check the stationarity of the time series and determine whether there is any notable seasonality that needs to be addressed. To accomplish the above requirement, it is necessary to perform the Augmented Dickey-Fuller test. The following python code segment can be used to check the stationarity of the time series data by performing an Augmented Dickey Fuller test.

```python
# check stationarity in time series data

def check_stationarity(df):
    # method1: plot the time series to check for trend and seasonality
    df.plot(figsize=(10, 10))

    # method 2: check if histogram fits a Gaussian Curve, then split data into two parts, calculate means and variances
    #and see if they vary
    df.hist(figsize=(10, 10))
    plt.show()

    X = df["TEMP_AVG"].values
    split = int(len(X) / 2)
    X1, X2 = X[0:split], X[split:]
    mean1, mean2 = X1.mean(), X2.mean()
    var1, var2 = X1.var(), X2.var()
    print('mean1=%f, mean2=%f' % (mean1, mean2))
    print('variance1=%f, variance2=%f' % (var1, var2))

    # if corresponding means and variances differ slightly (by less than 10), we consider that the time series
    # might be stationary
    if (abs(mean1-mean2) <= 10 and abs(var1-var2) <= 10):
        print("Time Series may be Stationary, since means and variances vary only slightly.\n")
    else:
        print("Time Series may NOT be Stationary, since means and variances vary significantly.\n")

    # method3: statistical test (Augmented Dickey-Fuller statistic)
    print("Performing Augmented Dickey-Fuller Test to confirm stationarity...")

    result = adfuller(X)
    print('ADF Statistic: %f' % result[0])
    print('p-value: %f' % result[1])

    p = result[1]
    if (p > 0.05):
        print("Time Series is NOT Stationary, since p-value > 0.05")
        df = df.diff()  # differencing to make data stationary
        return False
    else:
        print("Time Series is Stationary, since p-value <= 0.05")
        return True
```

```python
print ("Stationarity Check for %s" % 'Colombo')
is_stationary = check_stationarity(df)
```

Figure 6: Code segment for checking stationarity

After checking the stationarity of the weather data set, it needs to check the most appropriate values for the ARIMA model parameters such as p, d and q. The following code segment describes the method to find the ARIMA model parameters.

```python
p_range = q_range = list(range(0,3))  # taking values from 0 to 2

aic_values = []
bic_values = []
pq_values = []

for p in p_range:
    for q in q_range:
        try:
            model = ARIMA(df, order=(p, d, q))
            results = model.fit(disp=-1)
            aic_values.append(ARMAResults.aic(results))
            bic_values.append(ARMAResults.bic(results))
            pq_values.append((p, q))
        except:
            pass

best_pq = pq_values[aic_values.index(min(aic_values))]  # (p,q) corresponding to lowest AIC score
print("(p,q) corresponding to lowest AIC score: ", best_pq)
```

Figure 7: Code segment for identify ARIMA model parameters

After identifying the ARIMA model parameters, diagnostic checking is conducting to measure the appropriateness of the identified model. This is done using model residual values. Model residuals can be calculated using the below code segment.

```
df1 = pd.Series(df['TEMP_AVG'].values,index = df.index)

# model residuals = Original series - fitted values
(df1 - arima_model.fittedvalues)
```

Figure 8: Code segment for calculating the residuals

Finally, it needs to predict the one or more future values related to the selected weather parameter depending on the model developed using time series analysis. Following code segment describes the way of forecasting the future values.

```
# drop-down menu to select number of Days for which predictions are required
days_drop_down_menu = widgets.Dropdown(
    options=list(range(1,151)),
    value=10,
    description='No. of Days:',
    disabled=False,
)

days_drop_down_menu
```

```
num_days = days_drop_down_menu.value
sample_forecast = arima_model.forecast(steps=num_days)[0]
sample_forecast
```

```
from datetime import date, timedelta
i = 1
start_date = last_date_dataset
print("Last date of data set =", start_date)
delta = timedelta(days=1)
for x in sample_forecast:
    da= (start_date + timedelta(days=i)).strftime("%Y-%m-%d")
    print(da ,": ", x, 'F')
    i += 1
```

Figure 9: Code segment for forecasting the values

3.5.3 Implementation details related to Regression Analysis

```
# (1) select a significance value
alpha = 0.05

# (2) Fit the model(use Ordinary Least Squares algorithm)
model = sm.OLS(y, X).fit()

# (3) evaluate the coefficients' p-values
model.summary()
```

```
X = X.drop('TEMP_MIN_3', axis=1)
model = sm.OLS(y, X).fit()
model.summary()
```

```
# instantiate the regressor class
regressor = LinearRegression()
# fit the build the model by fitting the regressor to the training data
regressor.fit(X_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

```
# make a prediction set using the test set
prediction = regressor.predict(X_test)
```

Figure 10: Code segment for applying the Ordinary Least Squares algorithm and make predictions

3.5.4 Implementation details related to Decision Tree Analysis

```
# Convert to a Classification Task
# Binarize the RainFall to 0 or 1. Rain=1 & No Rain = 0
clean_data = data.copy()
clean_data['RainFall_label'] = (clean_data['RAINFALL'] == 'Rain')*1
print(clean_data['RainFall_label'])
```

```
# Target is stored in 'y'.
y=clean_data[['RainFall_label']].copy()
```

```
# Use Sensor Signals as Features to Predict Rainfall
features = ['TEMP_AVG','DEWP_AVG','RH_AVG','WINDSP_AVG','PRESSURE_AVG']
```

```
# Perform Test and Train split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=324)
```

```
# Fit on Train Set
rainfall_classifier = DecisionTreeClassifier(max_leaf_nodes=10, random_state=0)
decision_tree = rainfall_classifier.fit(X_train, y_train)
```

```
type(rainfall_classifier)
```

```
sklearn.tree._classes.DecisionTreeClassifier
```

```
# Predict on Test Set
predictions = rainfall_classifier.predict(X_test)
```

Figure 11: Code segment for analyzing the weather dataset with decision tree

CHAPTER 4: RESULTS AND EVALUATION

4.1 Introduction

This project is focused on developing a reliable weather prediction model for forecasting weather variations using data mining and machine learning techniques. The trustworthiness of a particular model is depending on the testing, evaluation and the validation of the model. Evaluation process can be defined as understanding a model and identifying how effectively it works for a particular purpose. It is the process of obtaining the value of a model, as the model is a representation of an object. Validation is the process of obtaining or testing the "truth" of the model. Since all the predictive models are incomplete representation of the actual reality, it is not seeking the perfect representation of the truth, but instead it finds the value in imperfect representation provided by the model. Generally, it can predict how good or bad the model is relative to observations based on the values obtained through evaluation process.

Since the purpose of this project is to develop a reliable weather prediction model, it requires real measured data for the correct forecast run. The data set is obtained from the website https://www.wunderground.com/ . And this website gives weather forecast for the next five days. These weather data are collected from automatic weather stations by using intelligent sensors. This data collection needs to be processed for the estimation of the optimal weather conditions depending on the predictive model. The aim of this predictive model is to use machine learning techniques for predicting the temperature of the next day at any particular city in Sri Lanka based on the weather data of the current day of the city. Meteorological parameters such as Temperature, Wind Speed, Humidity and Pressure are used to predict the weather variations. After finalizing the data set, the data needs to be processed using machine learning techniques such as Time Series analysis, Linear Regression, Neural Networks and Deep Learning etc. Then the outputs we get from above mentioned machine learning models need to be evaluated for developing a better weather predictive model.

When we are evaluating the weather prediction models like forecasting models, the most appropriate evaluation approach is the mathematical proof because all the observations are represented using numbers. In the evaluation process the models that created using different machine learning approaches needs to compare with each other and the model which gives the reliable values needs to be selected as the most optimal weather prediction model. The forecast accuracy delivered by the weather prediction model through evaluation can be measured using following methodologies.

- Percentage of accuracy expressed as a ratio between the predicted and measured value.

  This method allows to compare a value of particular weather parameter predicted by the numerical weather prediction model with the value of that actual weather parameter measured on the ground meteorological stations.

  The accuracy of the numerical weather model can be obtained through the following formula.

$$\text{Accuracy} = \frac{Value\ (Predicted)}{Value\ (Measured)} * 100\ (\%)$$

  The main objective of this method is to show which weather model can predict the local phenomenon with the greater accuracy.

- Pivot table providing data regarding the number of cases when the phenomenon was predicted either correctly or falsely.

  This method helps to identify the frequency of cases where the phenomenon was predicted and where it actually occurred, and in all possible combinations. This method allows detailed analysis than the relative accuracy method of the forecast because if the data set, we collected is uneven, then it can lead to wrong interpretation of measured data. (when the number of observations in different classes differ greatly)

  In this case pivot table consist with four fields.

A: Intervention: Number of cases when the phenomenon was predicted and really happened. Good forecast.

B: Error: The phenomenon was not predicted and occurred. Wrong forecast.

C: False: The phenomenon was predicted and did not arise. Wrong forecast.

D: Correct: The phenomenon was not predicted and did not happen. Good forecast

		Forecast	
		+	-
Measurement	+	A Intervention	B Error
	-	C False	D Correct

Figure 12: Pivot Table

Based on the output of the above forecasting we can analyze the forecasts with verification criteria such as Probability of Detection (POD), False Alarm Ratio (FAR) and etc.
When evaluating a weather prediction model there are always errors associated with it. The main reason is the observations we do have are generally not perfect and there is lack of information. Therefore, there is a necessity for testing and evaluating the observations which enables to understand the errors and ensure the accuracy.[2] Hence the model evaluation can be performed by calculating the Mean Error (ME) and the Root Mean Square Error (RMSE) between the forecast values and the observed values of a particular weather parameter.[4] [22]

Mean Error (ME) explains the difference between the forecast and the observed values while indicating the systematic error. If the forecast is perfect, the ME value is equal to zero while positive error values shows the overestimation and negative values shows the underestimation.

$$ME = \frac{\sum_{n=1}^{n=N}(P_n - O_n)}{N-1}$$ where

P_n: Forecast value

O_n: Observed value

N: Number of observations

Root Mean Square Error (RMSE) explains the difference between values predicted by the weather prediction model and the values actually measured on the ground meteorological stations. The individual differences are called as residuals and RMSE will aggregate them into a single measure of predictive power. RMSE value provide information regarding the total amplitude of the error by disregarding the signal of positive or negative unlike Mean Error.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{n=N}(P_n - O_n)^2}{N-1}}$$ where

P_n: Forecast value

O_n: Observed value

N: Number of observations

After identifying the optimal weather prediction model, evaluation is necessary to perform for checking the accuracy and the validity of the proposed weather prediction model. For this we simulate the proposed weather prediction model at a particular point in the system and then compare that simulation output against the set of actual weather observation results which

happened in the past. Then based on the accuracy of the mapping of the past incidents with the outputs of the proposed prediction model, we can validate the performance and the accuracy of the proposed weather prediction model.

If the model can predict the weather variations accurately which happened in the past based on the past weather input data, then the accuracy and the reliability of the proposed predictive model becomes high.

4.2 Results of Work

4.2.1 Results related to the Artificial Neural Network

In this section we are going to check the efficiency and the accuracy level of the developed weather predictive model using the Artificial Neural Network. Following are the Test results achieved from neural network for predicting the rainfall as the output variable while taking average temperature, dew point, relative humidity, windspeed and atmospheric pressure as the input variables.

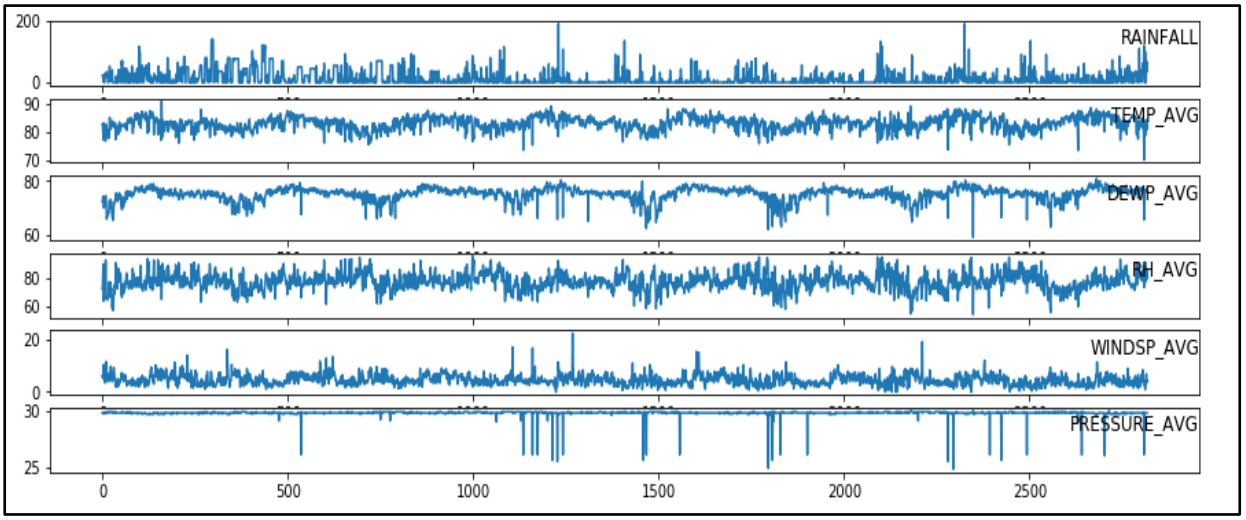


Figure 13: Variations of Weather Data

The following plot visualizes the predicted values for Rainfall.

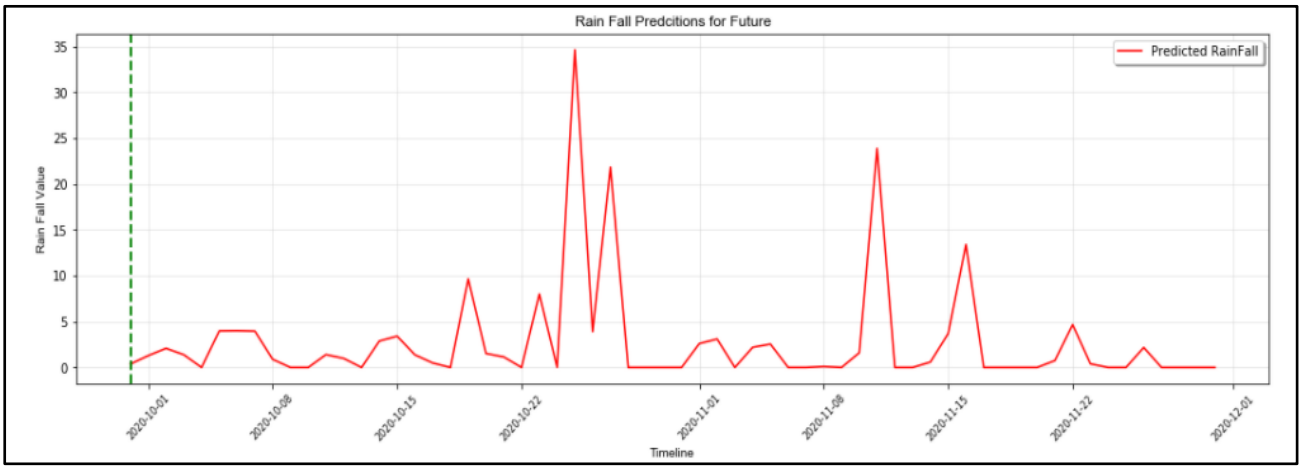


Figure 14: Graph of Rainfall Prediction for Future

The accuracy of the model needs to be tested after finalizing the model. In this study Mean Squared Error, Mean Absolute Error and Median Absolute Error are used to test the accuracy of the model.

```
The Mean Squared Error: 18.834 mm
The Mean Absolute Error: 9.95  mm
The Median Absolute Error: 3.32 mm
```

Figure 15: Test Results of Accuracy

Since the above results comparably small values we can proceed with the proposed model and predict the rainfall values for the future.

	RAINFALL
2020-09-30	0.372852
2020-10-01	1.286049
2020-10-02	2.085554
2020-10-03	1.386329
2020-10-04	0.000000
2020-10-05	3.987707
2020-10-06	4.005827
2020-10-07	3.960228
2020-10-08	0.886315
2020-10-09	0.000000
2020-10-10	0.000000
2020-10-11	1.402495
2020-10-12	0.979780
2020-10-13	0.000000
2020-10-14	2.879857
2020-10-15	3.412855
2020-10-16	1.378173
2020-10-17	0.489618

Figure 16: Predicted Values for Rainfall

4.2.2 Results related to the Time Series Analysis

In this section we are going to check the efficiency and the accuracy level of the developed weather predictive model using the Time Series Analysis. Following are the Test results achieved from the Augmented Dickey Fuller test which is conducted for checking the stationarity of the weather parameters such as Average Temperature, Average windspeed and Average humidity in the dataset.

4.2.2.1 Test results for Average Temperature

Following visual plot indicate the average temperature for the Colombo weather data set. By examining the plot, we can notice that there haven't any prominent trend associated with it. Therefore, there is no need for transforming the data to achieve the variance stability. Further by calculating the mean and the variance we can confirm the above conclusion. Mean and the variance is calculated in the below Figure 19.

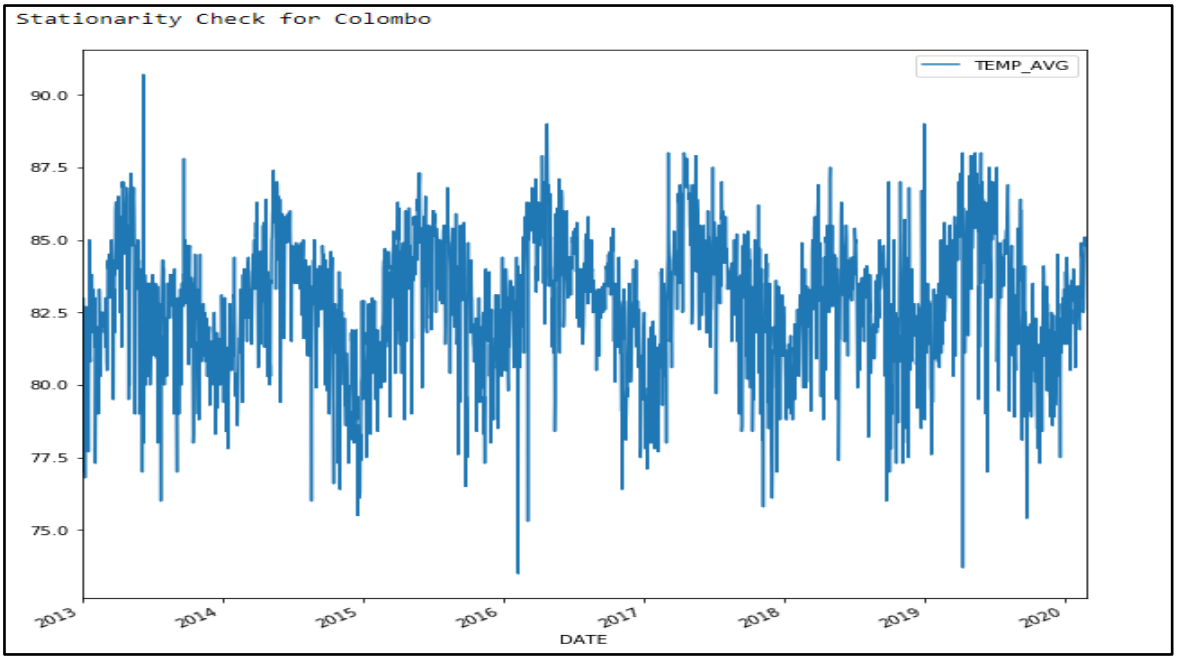


Figure 17: Variations of Average Temperature Data

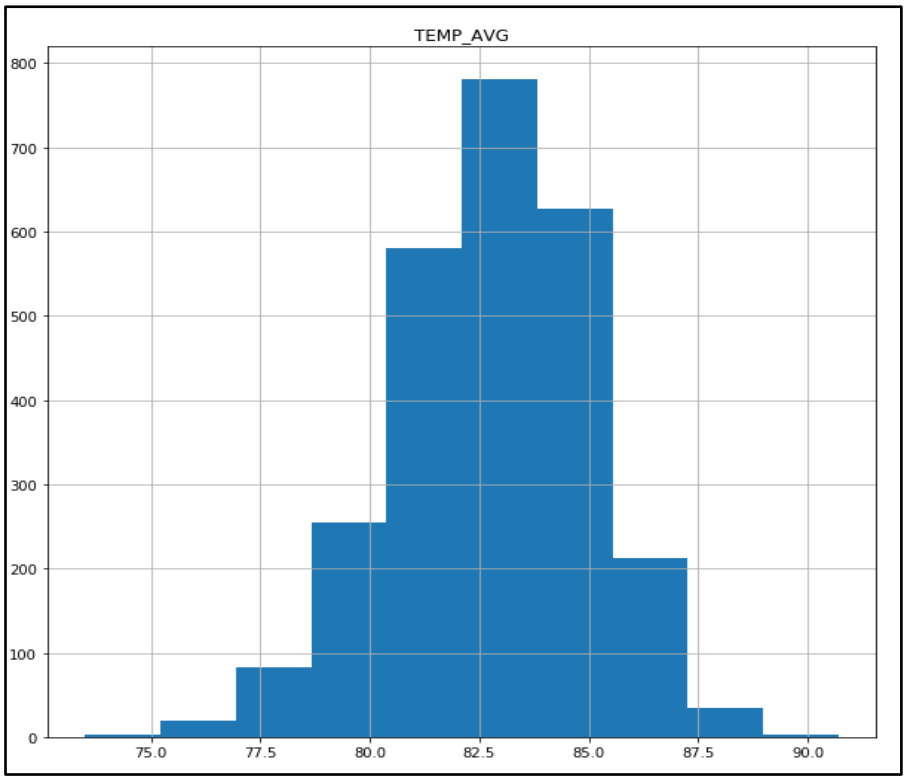


Figure 18: Histogram for Variations of Average Temperature

Since Augmented Dickey-Fuller test is performed to check the stationarity of the time series data, by analyzing the output results we can determine the stationarity of the Average Temperature weather variable.

```
mean1=82.786143, mean2=82.793846
variance1=5.017822, variance2=5.035639
Time Series may be Stationary, since means and variances vary only slightly.

Performing Augmented Dickey-Fuller Test to confirm stationarity...
ADF Statistic: -3.987886
p-value: 0.001475
Time Series is Stationary, since p-value <= 0.05
```

Figure 19: Output of the Augmented Dickey Fuller Test for Average Temperature

According to the output results of the Augmented Dickey Fuller (ADF) Test, we can notice that the ADF test statistic is -3.987886 and the calculated p-value is 0.001475. Since p-value is <= 0.05, it indicates that the null hypothesis which means H_0: Time series is non-stationary, may be rejected.[33] Hence as a result we can conclude that data for the Average Temperature weather variable is stationary and there is no need for differencing or transformation to achieve

the stationarity of the data. Therefore, the differencing parameter (d) in the ARIMA model is zero (d=0).

After identifying that the dataset is stationary, then the ACF, PACF autocorrelation plots need to draw for further analysis. Then based on the output results of the ACF and PACF autocorrelation plots, we can determine the most appropriate orders of the ARIMA models by evaluating the AIC criterion.

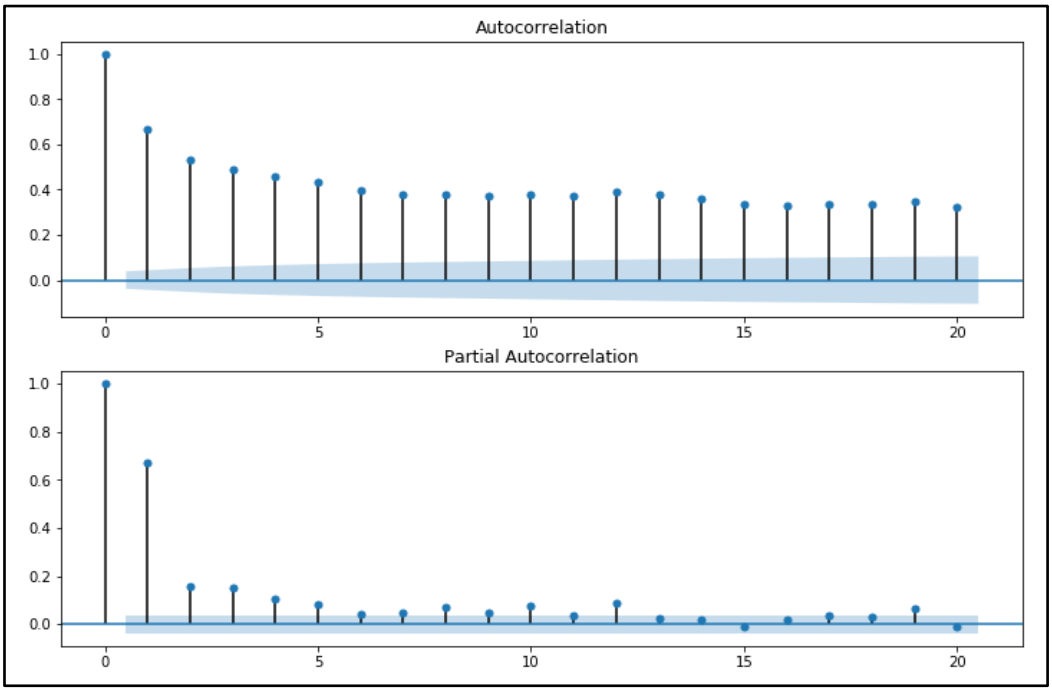


Figure 20: Output of the ACF, PACF Autocorrelation Plots for Average Temperature

Then based on the outputs of the ACF and PACF autocorrelation plots, the ARIMA model parameters need to be obtained.

```
Lowest AIC value: 9802.007104277538
(p,q) corresponding to lowest AIC score:  (2, 3)
```

Figure 21: ARIMA Model Parameters for Average Temperature

Based on the performance, the best fitted values of the ARIMA model p and q value are p = 2 and q=3. Since the data set is stationary, d = 0.

After estimating the model parameters, it needs to perform the diagnostic checking to measure appropriateness of the identified ARIMA model. This can be obtained through the model residuals of the ACF and PACF plots. According to the Figure.22 the spikes at the different lags are within the statistical confidence area of the ACF and PACF plots of residuals. Hence,

we can determine that the ARIMA (2,0,3) model is appropriate for forecasting average temperature data.

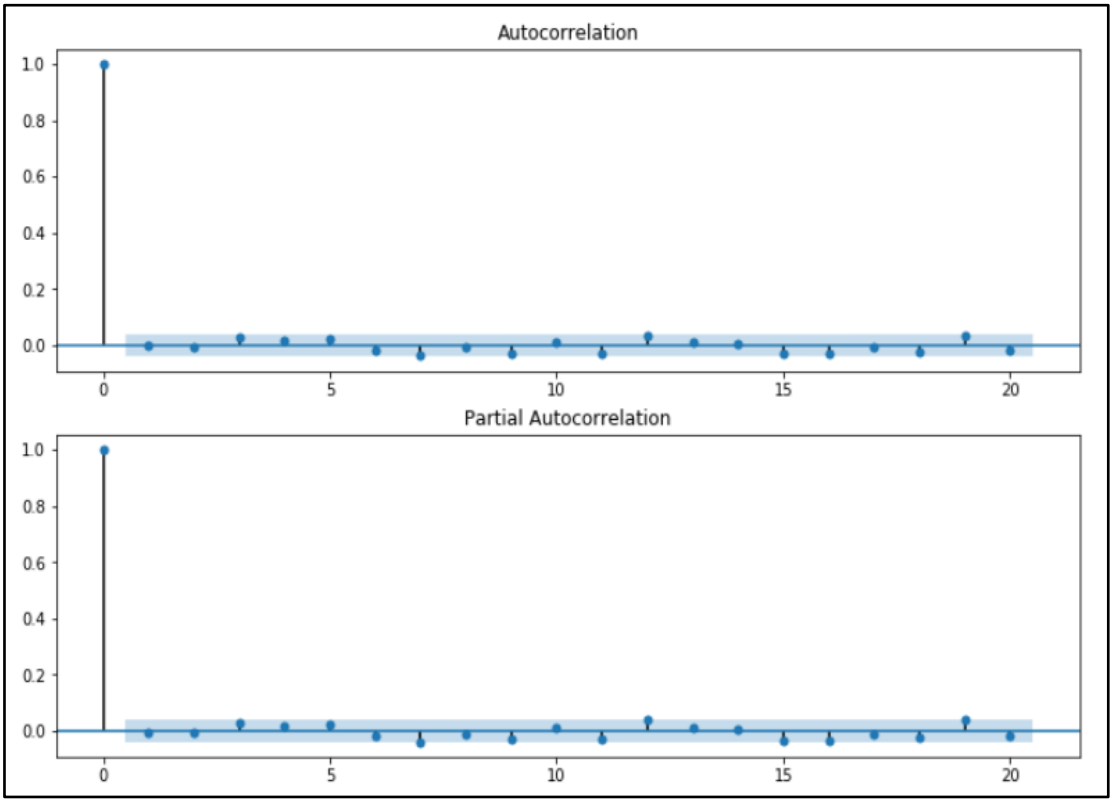


Figure 22: Output of the ACF and PACF plots for residuals for Average Temperature

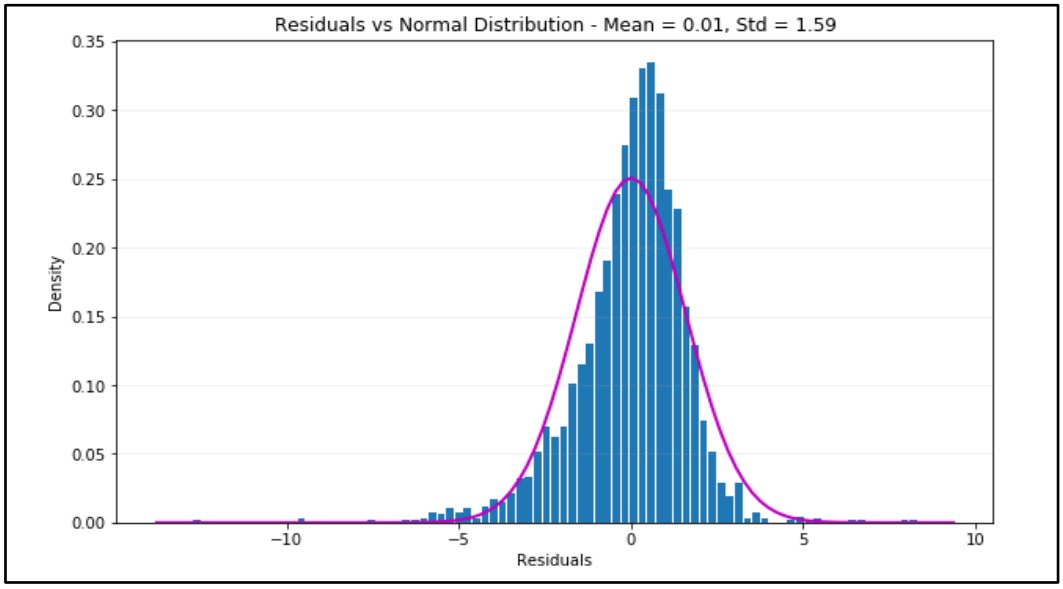


Figure 23: Histogram for residuals

According to the Figure. 4.8 residuals are close to the zero mean.

ARMA Model Results

Dep. Variable:	TEMP_AVG	**No. Observations:**	2599
Model:	ARMA(2, 3)	**Log Likelihood**	-4894.004
Method:	css-mle	**S.D. of innovations**	1.590
Date:	Sun, 21 Jun 2020	**AIC**	9802.007
Time:	11:01:46	**BIC**	9843.047
Sample:	0	**HQIC**	9816.877

	coef	std err	z	P>|z|	[0.025	0.975]
const	82.7646	0.359	230.664	0.000	82.061	83.468
ar.L1.TEMP_AVG	1.6072	0.083	19.322	0.000	1.444	1.770
ar.L2.TEMP_AVG	-0.6120	0.082	-7.471	0.000	-0.773	-0.451
ma.L1.TEMP_AVG	-1.1203	0.086	-13.020	0.000	-1.289	-0.952
ma.L2.TEMP_AVG	0.1027	0.050	2.063	0.039	0.005	0.200
ma.L3.TEMP_AVG	0.0742	0.035	2.092	0.037	0.005	0.144

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0128	+0.0000j	1.0128	0.0000
AR.2	1.6133	+0.0000j	1.6133	0.0000
MA.1	1.0852	+0.0000j	1.0852	0.0000
MA.2	2.4996	+0.0000j	2.4996	0.0000
MA.3	-4.9692	+0.0000j	4.9692	0.5000

Figure 24: ARIMA Model Results for Average Temperature

The following plot visualizes the actual values and the predicted values using the ARIMA ARIMA (2,0,3) model. The blue line denotes the actual average temperature values and the red line denotes the predicted values.

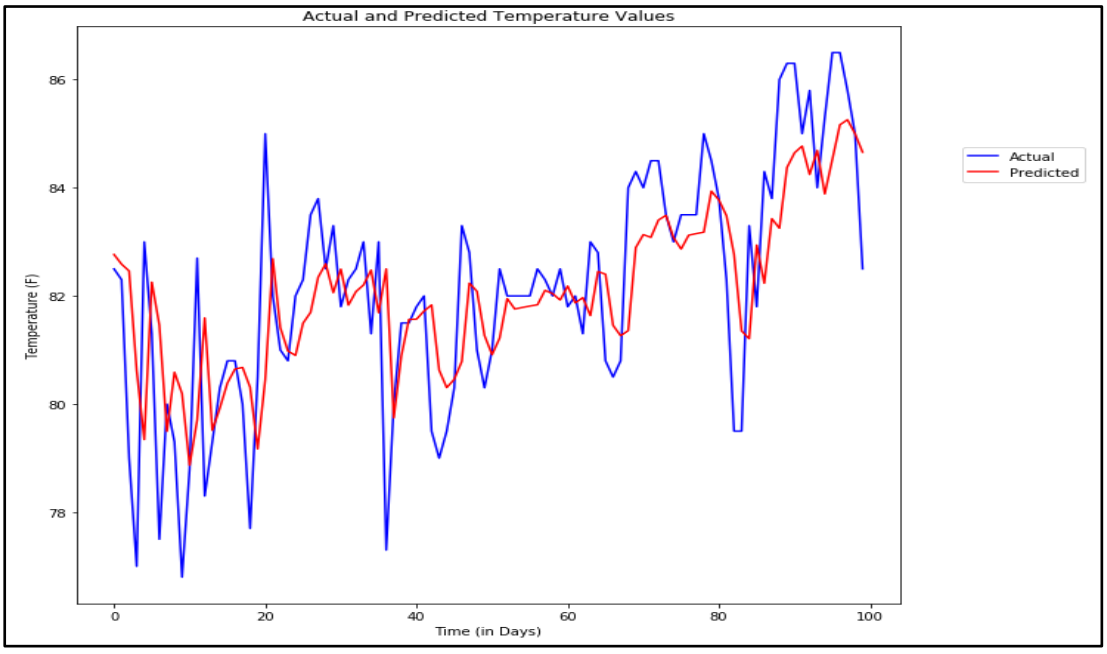


Figure 25: Actual and Predicted Values for Average Temperature

The accuracy of the model needs to be tested after finalizing the model. In this study Mean Squared Error and the Mean Absolute Error are used to test the accuracy of the model.

```
Mean Squared Error: 2.529886173875319
Mean Absolute Error: 1.1682059124537734
```

Figure 26: Test Results of Accuracy

Since the above results comparably small values we can proceed with the proposed model and predict the Average Temperature values for the future.

```
Last date of data set = 2020-09-30
2020-10-01 :  82.45571040726038 F
2020-10-02 :  82.21253364365289 F
2020-10-03 :  82.18687131353262 F
2020-10-04 :  82.17296411232391 F
2020-10-05 :  82.16672112957531 F
2020-10-06 :  82.16546326294925 F
2020-10-07 :  82.16743613459734 F
2020-10-08 :  82.17149105302924 F
2020-10-09 :  82.17687604423432 F
2020-10-10 :  82.18309897780621 F
2020-10-11 :  82.18983791505981 F
```

Figure 27: Predicted Values for Average Temperature

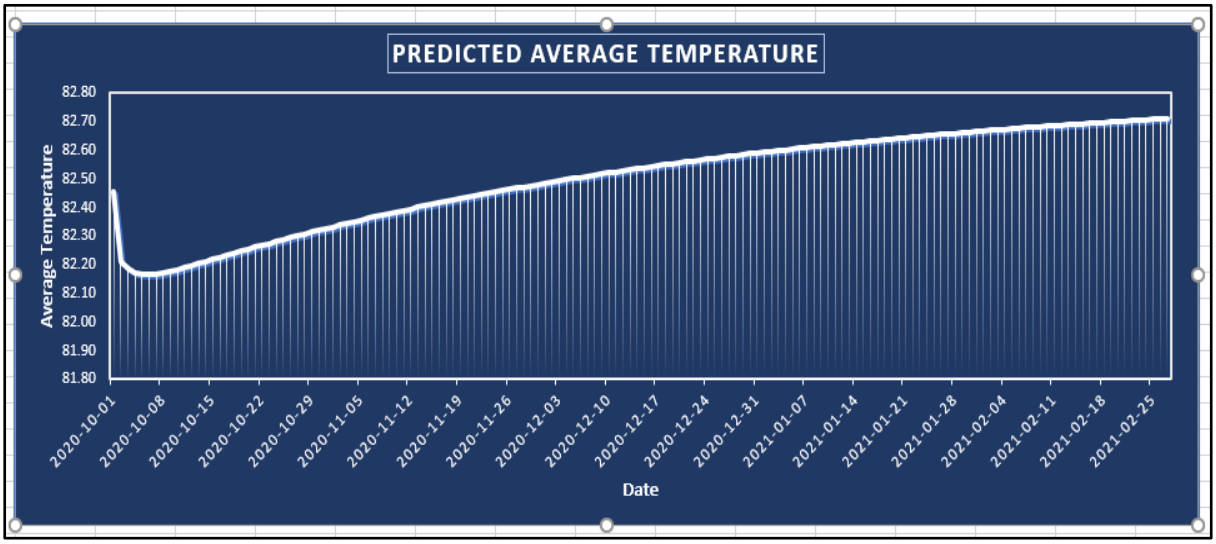


Figure 28: Graph of Average Temperature Prediction for Future

4.2.2.2 Test results for Average Wind Speed

Following visual plot indicate the average wind speed for the Colombo weather data set. By examining the plot, we can notice that there haven't any prominent trend associated with it. Therefore, there is no need for transforming the data to achieve the variance stability. Further by calculating the mean and the variance we can confirm the above conclusion. Mean and the variance is calculated in the below Figure 31.

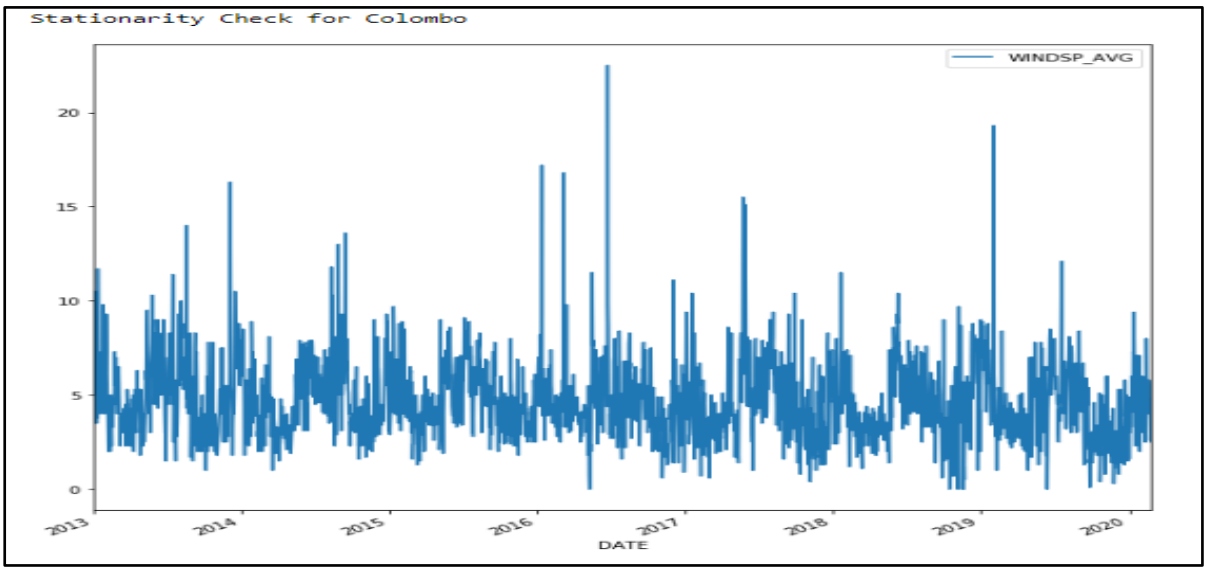


Figure 29: Variations of Average Wind Speed Data

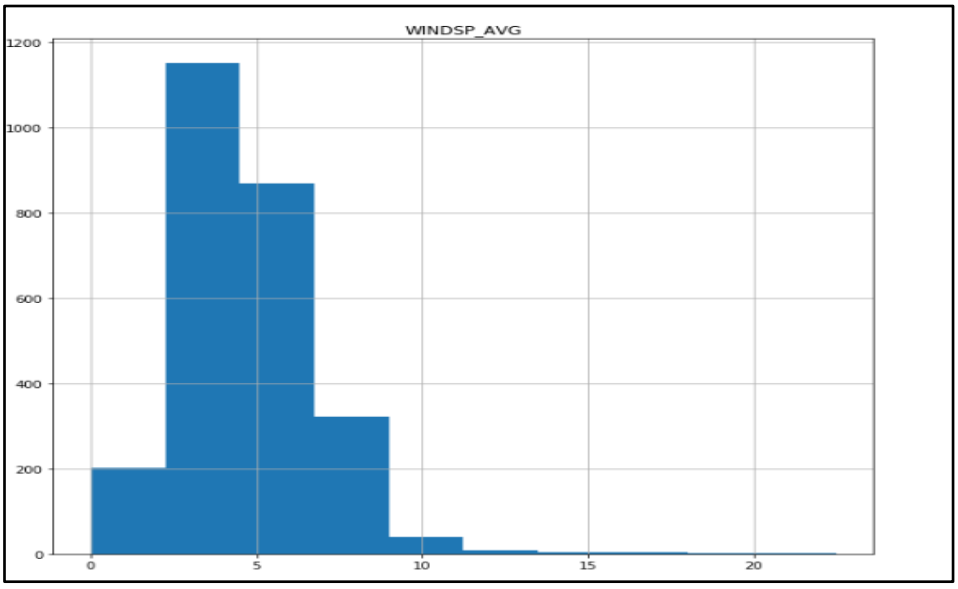


Figure 30: Histogram for Variations of Average Wind Speed

Since Augmented Dickey-Fuller test is performed to check the stationarity of the time series data, by analyzing the output results we can determine the stationarity of the Average Wind Speed weather variable.

```
mean1=4.892533, mean2=4.416462
variance1=4.061984, variance2=3.975252
Time Series may be Stationary, since means and variances vary only slightly.

Performing Augmented Dickey-Fuller Test to confirm stationarity...
ADF Statistic: -9.268841
p-value: 0.000000
Time Series is Stationary, since p-value <= 0.05
```

Figure 31: Output of the Augmented Dickey Fuller Test for Average Wind Speed

According to the output results of the Augmented Dickey Fuller (ADF) Test, we can notice that the ADF test statistic is -9.268841 and the calculated p-value is 0.000. Since p-value is <= 0.05, it indicates that the null hypothesis which means H_0: Time series is non-stationary, may be rejected.[33] Hence as a result we can conclude that data for the Average Wind Speed weather variable is stationary and there is no need for differencing or transformation to achieve the stationarity of the data. Therefore, the differencing parameter (d) in the ARIMA model is zero (d=0).

After identifying that the dataset is stationary, then the ACF, PACF autocorrelation plots need to draw for further analysis. Then based on the output results of the ACF and PACF autocorrelation plots, we can determine the most appropriate orders of the ARIMA models by evaluating the AIC criterion.

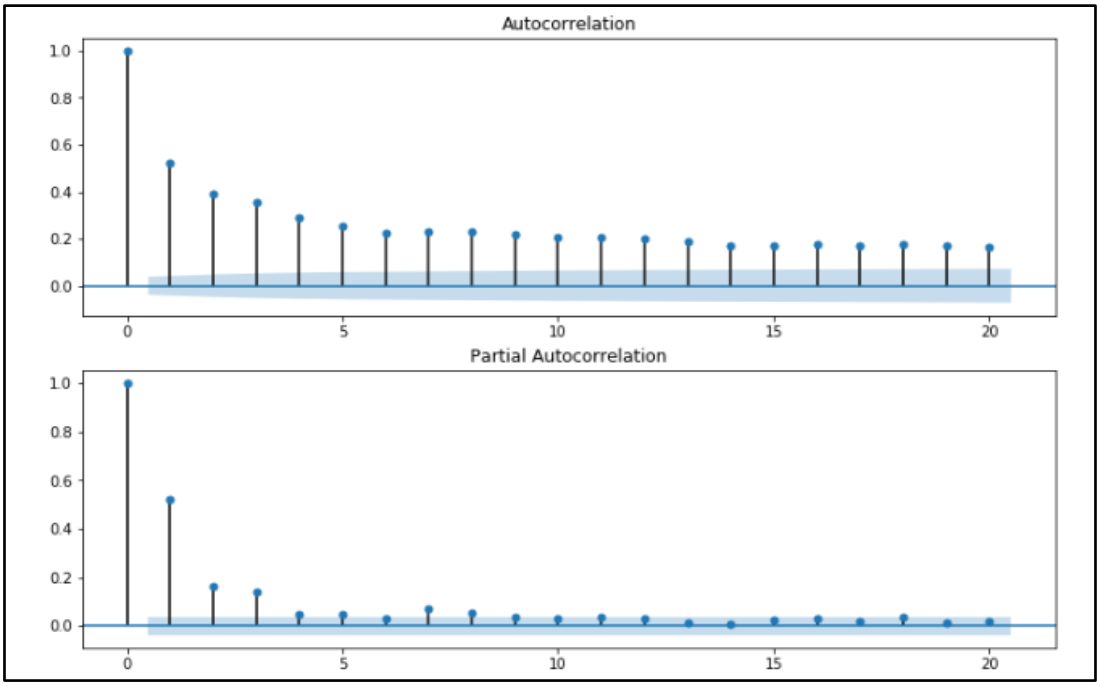


Figure 32: Output of the ACF, PACF Autocorrelation Plots for Average Wind Speed

Then based on the outputs of the ACF and PACF autocorrelation plots, the ARIMA model parameters need to be obtained.

```
Lowest AIC value: 10048.570569530142
(p,q) corresponding to lowest AIC score:  (3, 3)
```

Figure 33: ARIMA Model Parameters for Wind Speed

Based on the performance, the best fitted values of the ARIMA model p and q value are p = 3 and q=3. Since the data set is stationary, d = 0.

After estimating the model parameters, it needs to perform the diagnostic checking to measure appropriateness of the identified ARIMA model. This can be obtained through the model residuals of the ACF and PACF plots. According to the Figure.32 the spikes at the different lags are within the statistical confidence area of the ACF and PACF plots of residuals. Hence, we can determine that the ARIMA (3,0,3) model is appropriate for forecasting average wind speed data.

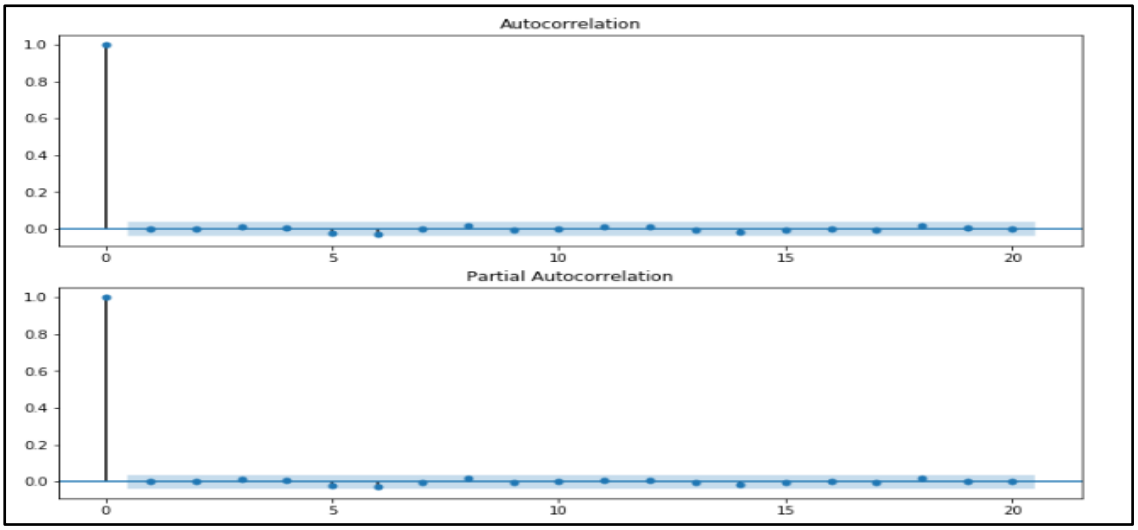


Figure 34: Output of the ACF and PACF plots for residuals

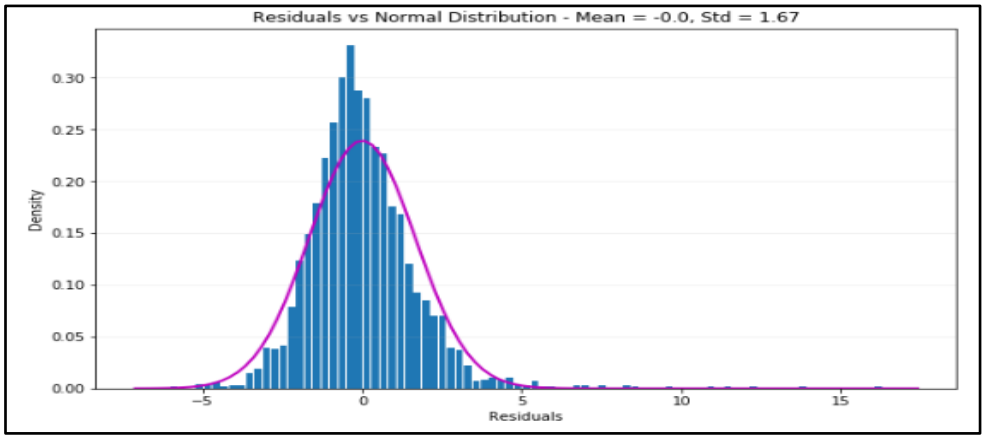


Figure 35: Histogram for residuals

According to the Figure. 35 residuals are having a zero mean.

ARMA Model Results

Dep. Variable:	WINDSP_AVG	No. Observations:	2599
Model:	ARMA(3, 3)	Log Likelihood	-5016.285
Method:	css-mle	S.D. of innovations	1.667
Date:	Sun, 21 Jun 2020	AIC	10048.571
Time:	12:29:11	BIC	10095.474
Sample:	0	HQIC	10065.565

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.6630	0.176	26.537	0.000	4.319	5.007
ar.L1.WINDSP_AVG	0.8940	0.217	4.116	0.000	0.468	1.320
ar.L2.WINDSP_AVG	0.4802	0.297	1.616	0.106	-0.102	1.062
ar.L3.WINDSP_AVG	-0.3945	0.103	-3.829	0.000	-0.596	-0.193
ma.L1.WINDSP_AVG	-0.5017	0.216	-2.318	0.021	-0.926	-0.078
ma.L2.WINDSP_AVG	-0.5980	0.208	-2.873	0.004	-1.006	-0.190
ma.L3.WINDSP_AVG	0.2103	0.054	3.903	0.000	0.105	0.316

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.4775	+0.0000j	1.4775	0.5000
AR.2	1.0315	+0.0000j	1.0315	0.0000
AR.3	1.6631	+0.0000j	1.6631	0.0000
MA.1	-1.3804	+0.0000j	1.3804	0.5000
MA.2	1.1042	+0.0000j	1.1042	0.0000
MA.3	3.1197	+0.0000j	3.1197	0.0000

Figure 36: ARIMA Model Results

The following plot visualizes the actual values and the predicted values using the ARIMA ARIMA (3,0,3) model. The blue line denotes the actual average temperature values and the red line denotes the predicted values.

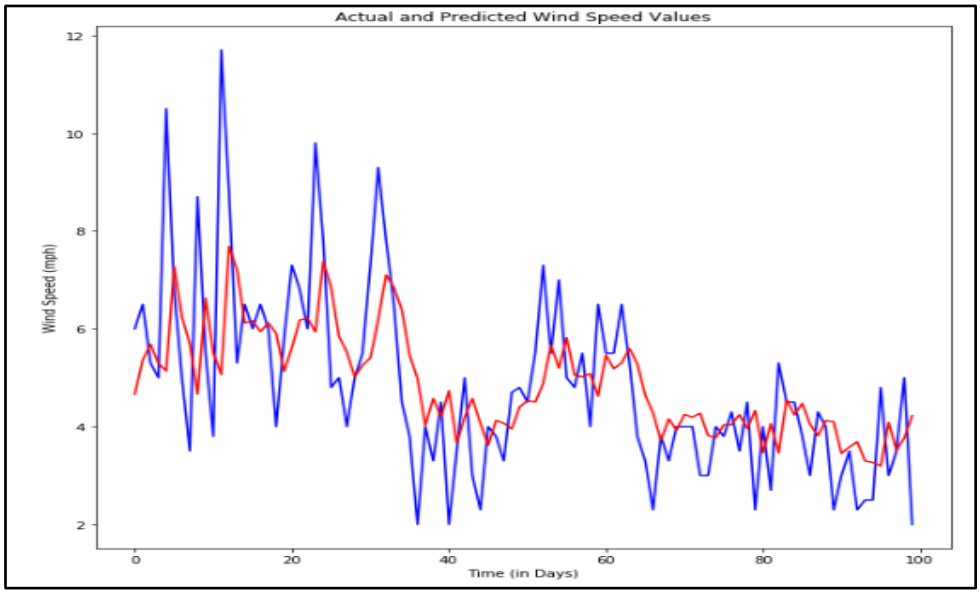


Figure 37: Actual and Predicted Values for Wind Speed

The accuracy of the model needs to be tested after finalizing the model. In this study Mean Squared Error and the Mean Absolute Error are used to test the accuracy of the model.

```
Mean Squared Error: 2.7795077863694746
Mean Absolute Error: 1.1977497080098443
```

Figure 38: Test Results of Accuracy

Since the above results comparably small values we can proceed with the proposed model and predict the Average Wind Speed values for the future.

```
Last date of data set = 2020-09-30
2020-10-01 :  4.290754950210557 mph
2020-10-02 :  4.122345232401296 mph
2020-10-03 :  4.232805374141508 mph
2020-10-04 :  4.173320363046834 mph
2020-10-05 :  4.224319996061039 mph
2020-10-06 :  4.207480848341287 mph
2020-10-07 :  4.234304952799692 mph
2020-10-08 :  4.233942425798055 mph
2020-10-09 :  4.250729938651179 mph
2020-10-10 :  4.256520565100192 mph
2020-10-11 :  4.268946604898753 mph
2020-10-12 :  4.276828696699855 mph
2020-10-13 :  4.287181770477202 mph
2020-10-14 :  4.295568316839434 mph
2020-10-15 :  4.304781835434371 mph
2020-10-16 :  4.313063499874572 mph
2020-10-17 :  4.32152825953653 mph
2020-10-18 :  4.32948145439591 mph
```

Figure 39: Predicted Values for Average Wind Speed

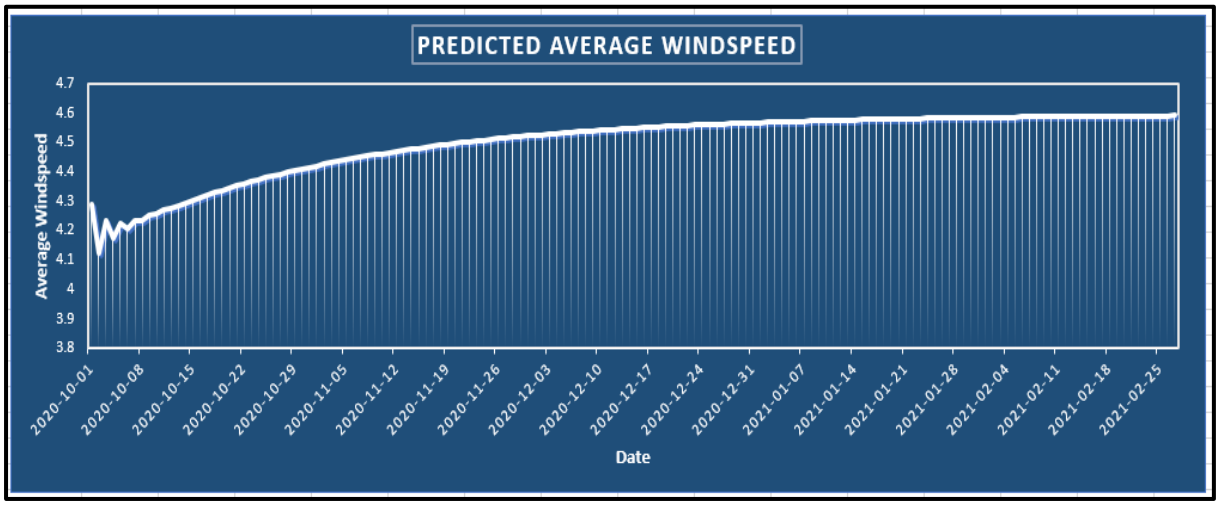


Figure 40: Graph of Average Wind Speed Prediction for Future

4.2.2.3 Test results for Relative Humidity

Following visual plot indicate the average humidity for the Colombo weather data set. By examining the plot, we can notice that there have not any prominent trend associated with it. Therefore, there is no need for transforming the data to achieve the variance stability. Further by calculating the mean and the variance we can confirm our assumption. Mean and the variance is calculated in the below Figure 43.

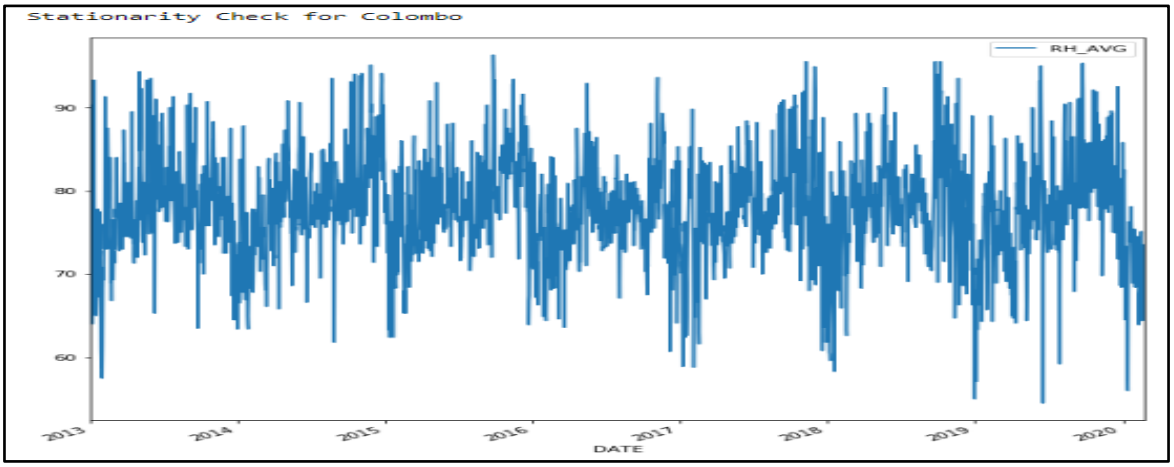


Figure 41: Variations of Relative Humidity

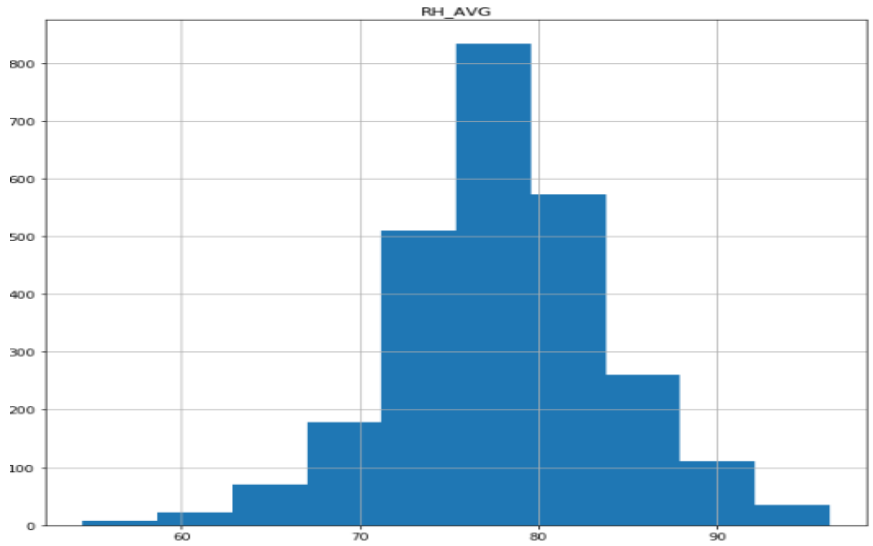


Figure 42: Histogram for Variations of Relative Humidity

Since Augmented Dickey-Fuller test is performed to check the stationarity of the time series data, by analyzing the output results we can determine the stationarity of the average humidity weather variable.

```
mean1=78.283988, mean2=77.820769
variance1=32.801822, variance2=39.262076
Time Series may be Stationary, since means and variances vary only slightly.

Performing Augmented Dickey-Fuller Test to confirm stationarity...
ADF Statistic: -5.449372
p-value: 0.000003
Time Series is Stationary, since p-value <= 0.05
```

Figure 43: Output of the Augmented Dickey Fuller Test for Relative Humidity

According to the output results of the Augmented Dickey Fuller (ADF) Test, we can notice that the ADF test statistic is -5.449372 and the calculated p-value is 0.000003. Since p-value is <= 0.05, it indicates that the null hypothesis which means H_0: Time series is non-stationary, may be rejected.[33] Hence as a result we can conclude that data for the average temperature weather variable is stationary and there is no need for differencing or transformation to achieve the stationarity of the data. Therefore, the differencing parameter (d) in the ARIMA model is zero (d=0).

After identifying that the dataset is stationary, then the ACF, PACF autocorrelation plots need to draw for further analysis. Then based on the output results of the ACF and PACF autocorrelation plots, we can determine the most appropriate orders of the ARIMA models by evaluating the AIC criterion.

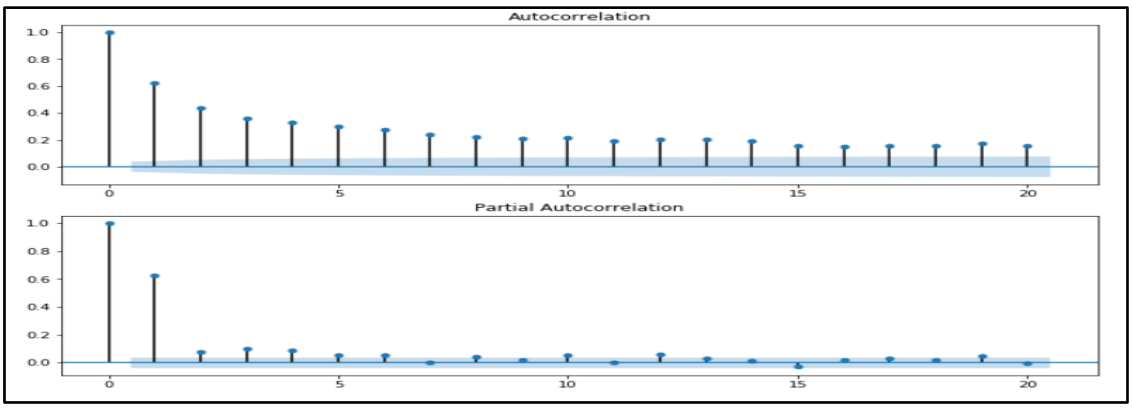


Figure 44:Output of the ACF, PACF Autocorrelation Plots for Relative Humidity

Then based on the outputs of the ACF and PACF autocorrelation plots, the ARIMA model parameters need to be obtained.

```
Lowest AIC value: 15326.197024936791
(p,q) corresponding to lowest AIC score:  (2, 1)
```

Figure 45:ARIMA Model Parameters

Based on the performance, the best fitted values of the ARIMA model p and q value are p = 2 and q=1. Since the data set is stationary, d = 0.

After estimating the model parameters, it needs to perform the diagnostic checking to measure appropriateness of the identified ARIMA model. This can be obtained through the model residuals of the ACF and PACF plots. According to the Figure.46 the spikes at the different lags are within the statistical confidence area of the ACF and PACF plots of residuals. Hence, we can determine that the ARIMA (2,0,1) model is appropriate for forecasting average humidity data.

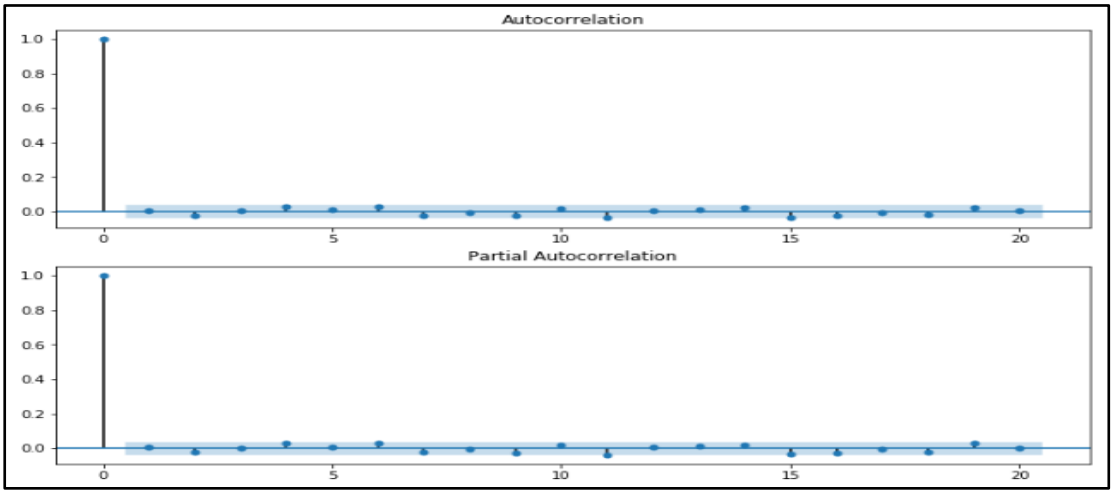


Figure 46: Output of the ACF and PACF plots for residuals for Relative Humidity

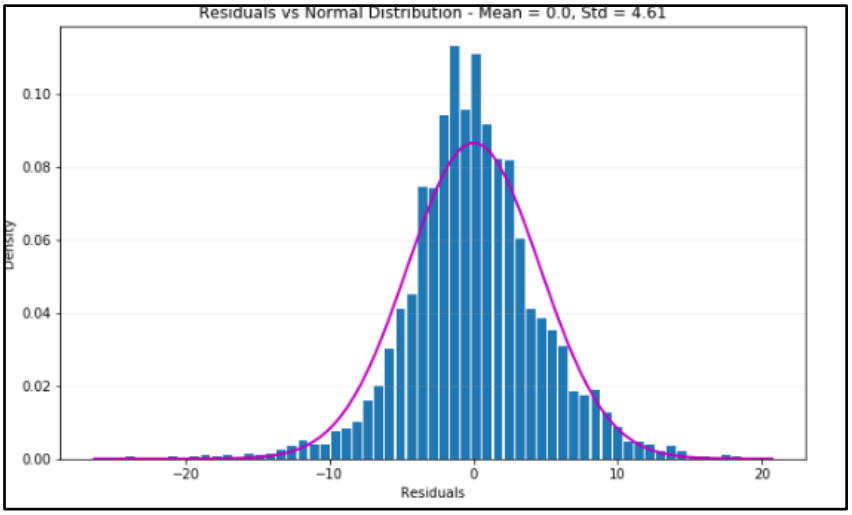


Figure 47: Histogram for residuals

According to the Figure. 47 residuals are having a zero mean.

ARMA Model Results

Dep. Variable:	RH_AVG	No. Observations:	2599
Model:	ARMA(2, 1)	Log Likelihood	-7658.099
Method:	css-mle	S.D. of innovations	4.607
Date:	Sun, 21 Jun 2020	AIC	15326.197
Time:	13:19:21	BIC	15355.511
Sample:	0	HQIC	15336.819

	coef	std err	z	P>|z|	[0.025	0.975]
const	77.9533	0.557	139.854	0.000	76.861	79.046
ar.L1.RH_AVG	1.4462	0.031	47.207	0.000	1.386	1.506
ar.L2.RH_AVG	-0.4615	0.027	-17.251	0.000	-0.514	-0.409
ma.L1.RH_AVG	-0.9046	0.021	-42.548	0.000	-0.946	-0.863

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0302	+0.0000j	1.0302	0.0000
AR.2	2.1032	+0.0000j	2.1032	0.0000
MA.1	1.1055	+0.0000j	1.1055	0.0000

Figure 48: ARIMA Model Results for Relative Humidity

The following plot visualizes the actual values and the predicted values using the ARIMA ARIMA (2,0,1) model. The blue line denotes the actual average humidity values and the red line denotes the predicted values.

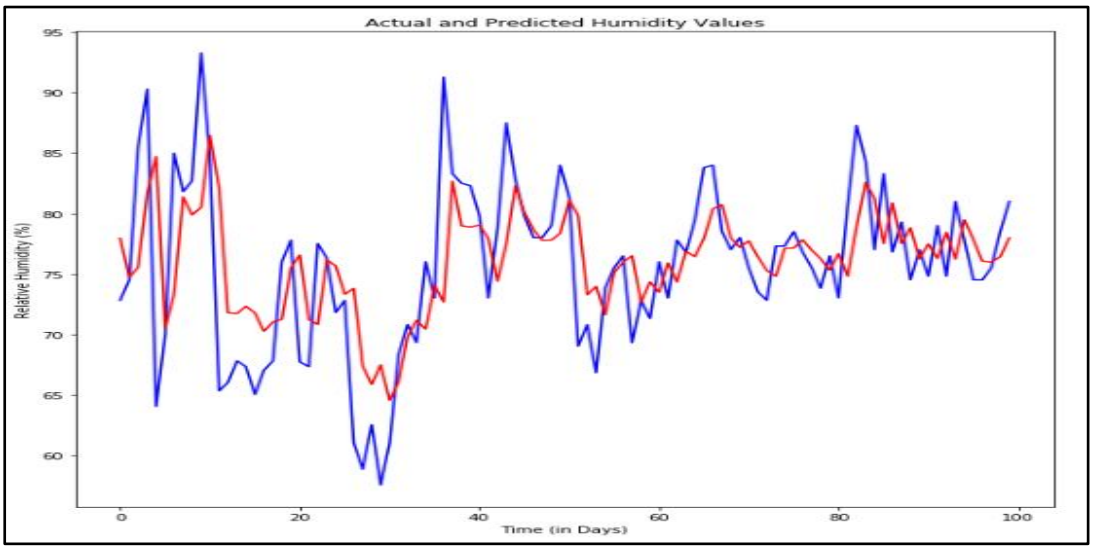


Figure 49: Actual and Predicted Values for Relative Humidity

The accuracy of the model needs to be tested after finalizing the model. In this study Mean Squared Error and the Mean Absolute Error are used to test the accuracy of the model.

```
Mean Squared Error: 21.23089498893645
Mean Absolute Error: 3.4523066545275602
```

Figure 50: Test Results of Accuracy

Since the above results comparably small values we can proceed with the proposed model and predict the average humidity values for the future.

```
Last date of data set = 2020-09-30
2020-10-01 :   80.64013800956192 %
2020-10-02 :   80.70632355980965 %
2020-10-03 :   80.69322294030637 %
2020-10-04 :   80.64468006456642 %
2020-10-05 :   80.580980322054 %
2020-10-06 :   80.51148925151303 %
2020-10-07 :   80.44051141814623 %
2020-10-08 :   80.37000638028405 %
2020-10-09 :   80.3008476973755 %
2020-10-10 :   80.23340653632638 %
2020-10-11 :   80.16782218271429 %
2020-10-12 :   80.10412741136798 %
2020-10-13 :   80.042306574981 %
2020-10-14 :   79.98232250369264 %
2020-10-15 :   79.92412894721897 %
2020-10-16 :   79.86767631577126 %
2020-10-17 :   79.81291431529763 %
2020-10-18 :   79.75979314380736 %
2020-10-19 :   79.7082640214141 %
```

Figure 51:Predicted Values for Relative Humidity

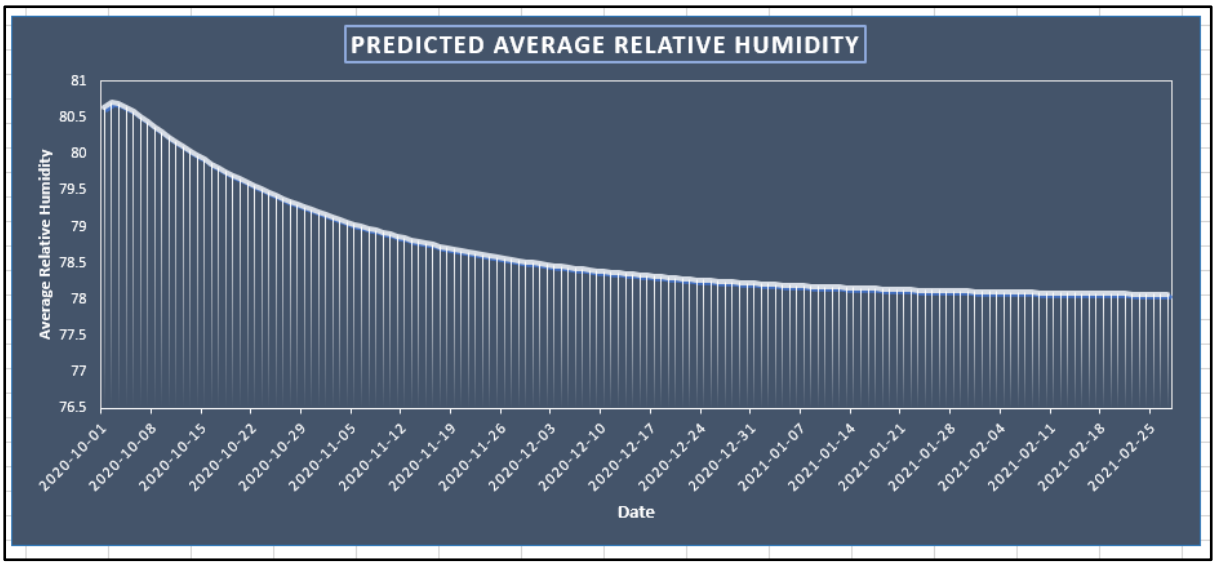


Figure 52: Graph of Relative Humidity Prediction for Future

4.2.3 Results related to the Time Series Analysis using Artificial Neural Networks

Neural networks also can be used to enhance the accuracy of the data obtained from the time series analysis. During this study we have use the time series weather data to train and fine tune several artificial neural networks. For this analyzing the neural networks we used python language with the TensorFlow library. Then based on the resulting neural network models we can predict the future values for the selected weather parameters.

In the following plot blue line represents the actual values and the orange line represents the corresponding predicted value using neural network.

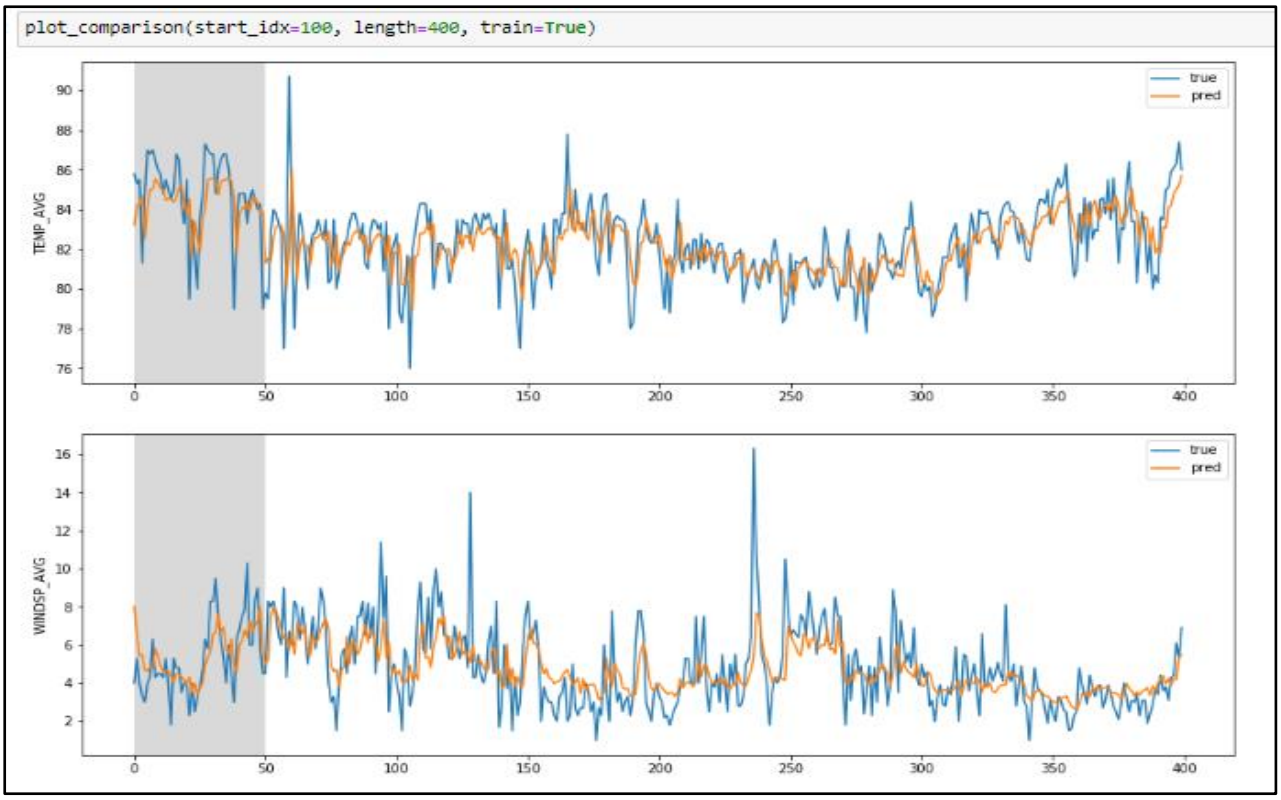


Figure 53: Analysis results for Average Temperature and Average Wind Speed with respect to the training data

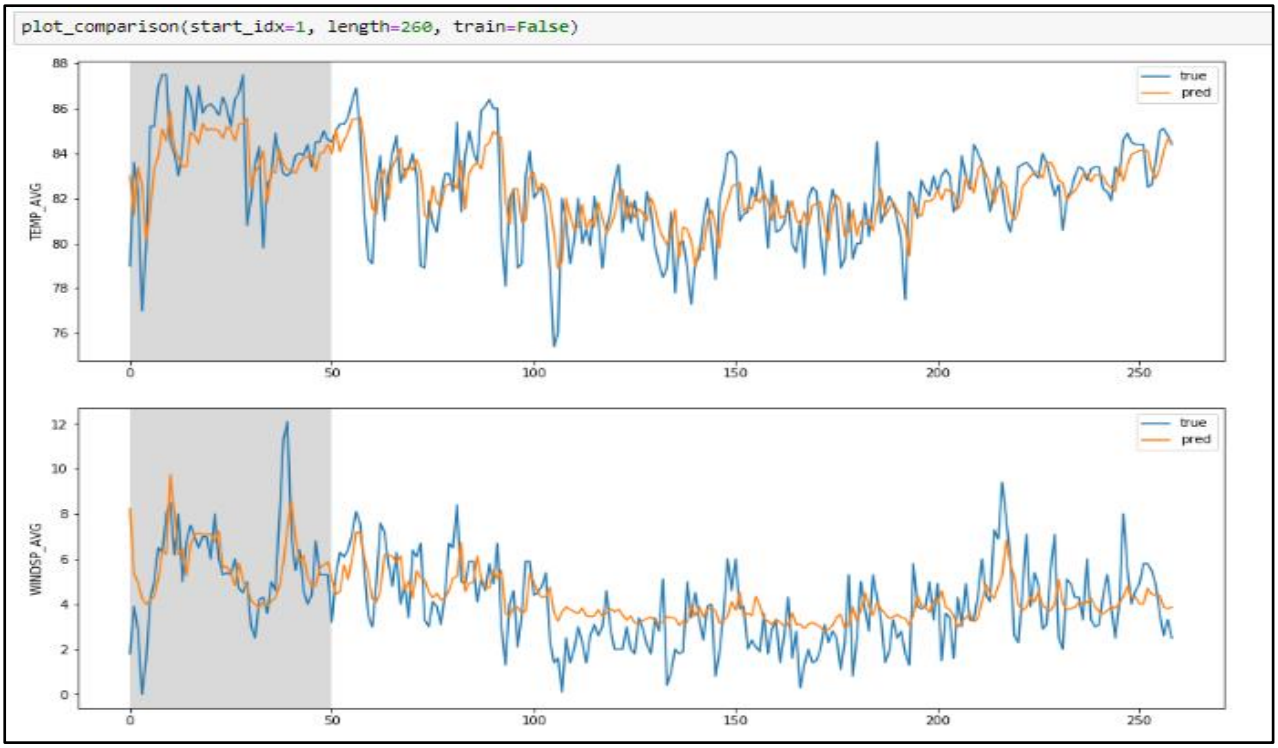


Figure 54: Analysis results for Average Temperature and Average Wind Speed with respect to the test data

4.2.4 Results related to the Regression Analysis

In this section we are going to check the efficiency and the accuracy level of the developed weather predictive model using the Regression Analysis. Following are the Test results achieved from regression analysis by considering average temperature, average wind Speed and relative humidity.

4.2.4.1 Test results for Average Temperature

Following plot indicates the visualization of relationships with dependent variable average temperature along with the Predictor variables using scatter plots.

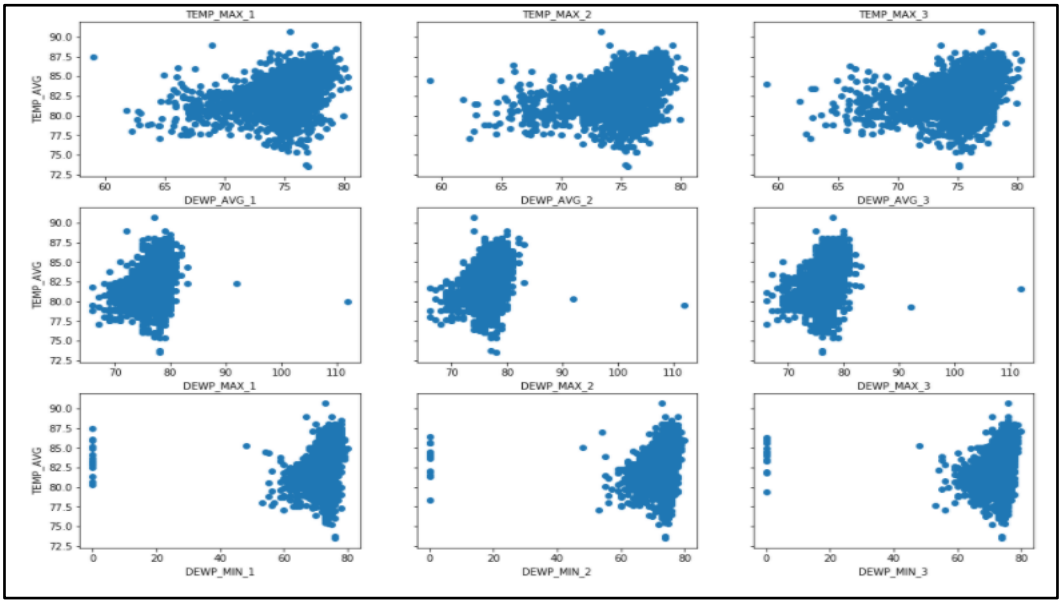


Figure 55: Scatter Plots for Average Temperature

	TEMP_AVG
RH_MAX_1	-0.251905
RH_AVG_1	-0.226970
RH_MAX_2	-0.155268
PRESSURE_MAX_1	-0.124677
RH_MAX_3	-0.123011
PRESSURE_MAX_3	-0.108080
PRESSURE_MAX_2	-0.106537
RH_AVG_2	-0.104817
PRECIPITATION_1	-0.091265
RH_AVG_3	-0.065995
PRESSURE_AVG_3	-0.043974
PRESSURE_AVG_2	-0.038972
PRECIPITATION_2	-0.035927
PRESSURE_MIN_3	-0.033211
PRESSURE_MIN_2	-0.027307
PRESSURE_AVG_1	-0.026383
PRECIPITATION_3	-0.018931
RH_MIN_1	-0.016291
PRESSURE_MIN_1	-0.014131
WINDSP_MAX_2	0.000224
WINDSP_MAX_3	0.004691
WINDSP_MAX_1	0.008667
RH_MIN_3	0.050500
RH_MIN_2	0.051508

Figure 56: Correlation Output for Average Temperature

OLS Regression Results

Dep. Variable:	TEMP_AVG	R-squared:	0.495
Model:	OLS	Adj. R-squared:	0.493
Method:	Least Squares	F-statistic:	281.6
Date:	Sun, 16 Aug 2020	Prob (F-statistic):	0.00
Time:	12:24:53	Log-Likelihood:	-4892.9
No. Observations:	2596	AIC:	9806.
Df Residuals:	2586	BIC:	9864.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	16.7296	1.498	11.167	0.000	13.792	19.667
TEMP_AVG_1	0.6600	0.033	20.250	0.000	0.596	0.724
TEMP_AVG_2	0.0483	0.022	2.214	0.027	0.006	0.091
TEMP_AVG_3	0.1187	0.020	6.029	0.000	0.080	0.157
TEMP_MIN_1	-0.0743	0.012	-6.182	0.000	-0.098	-0.051
TEMP_MAX_1	-0.0772	0.019	-4.003	0.000	-0.115	-0.039
DEWP_AVG_1	0.0720	0.026	2.763	0.006	0.021	0.123
DEWP_AVG_3	0.1065	0.026	4.155	0.000	0.056	0.157
DEWP_MIN_1	-0.0191	0.008	-2.512	0.012	-0.034	-0.004
DEWP_MIN_3	-0.0249	0.008	-3.307	0.001	-0.040	-0.010

Omnibus:	507.665	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2075.454
Skew:	-0.902	Prob(JB):	0.00
Kurtosis:	6.991	Cond. No.	1.13e+04

Figure 57: Final Result after applying the Backward Elimination for Average Temperature

The following conclusions can be made by observing the OLS regression results related to average temperature.

- All remaining predictors have p-values significantly below of 0.05.
- R-squared value which measure the overall variance of the model, 0.495 interpreted that final model explains about 50% of the observed variation in the outcome variable of average temperature.

```
[85.77186487, 83.21148333, 83.78286323, 81.15137308, 83.07406083,
 79.51420362, 84.82871578, 81.55112021, 81.47268418, 83.19859749,
 79.8186654 , 82.20908286, 84.27904355, 80.79766693, 83.57802488,
 82.05478074, 80.58830422, 85.80479131, 82.55576011, 84.02050707,
 81.16313592, 83.76550705, 81.58034407, 80.69784441, 83.62442789,
 83.40046243, 82.03826794, 81.38663373, 84.77041205, 81.66235831,
 81.96967933, 84.35868703, 83.58604952, 85.62308175, 82.14987847,
 81.65687818, 81.13751846, 82.61237396, 84.2352387 , 80.3878258 ,
 84.63039319, 80.2130127 , 82.32716384, 85.83261498, 83.19788044,
 84.89817632, 84.39853167, 80.9639096 , 81.29092618, 83.41714424,
 84.29460924, 82.55724048, 82.09093644, 85.94392864, 84.47110818,
```

Figure 58: Predicted Values for Average Temperature

4.2.4.2 Test results for Average Wind Speed

Following plot indicates the visualization of relationships with dependent variable wind speed along with the predictor variables using scatter plots.

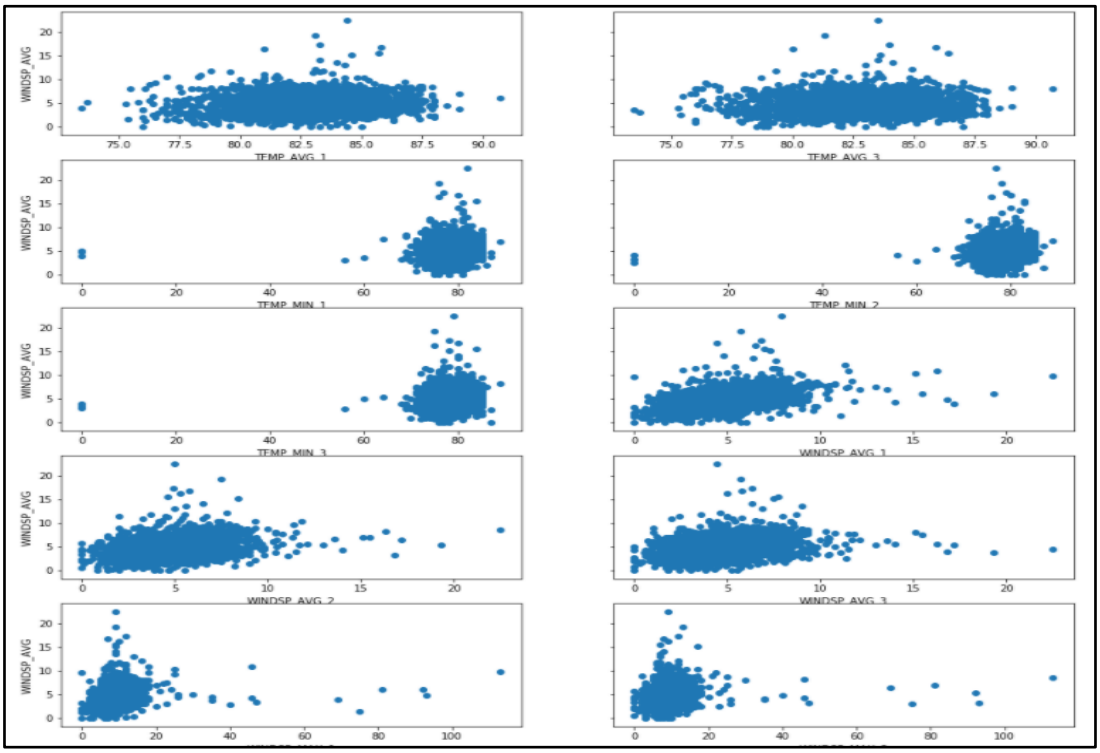
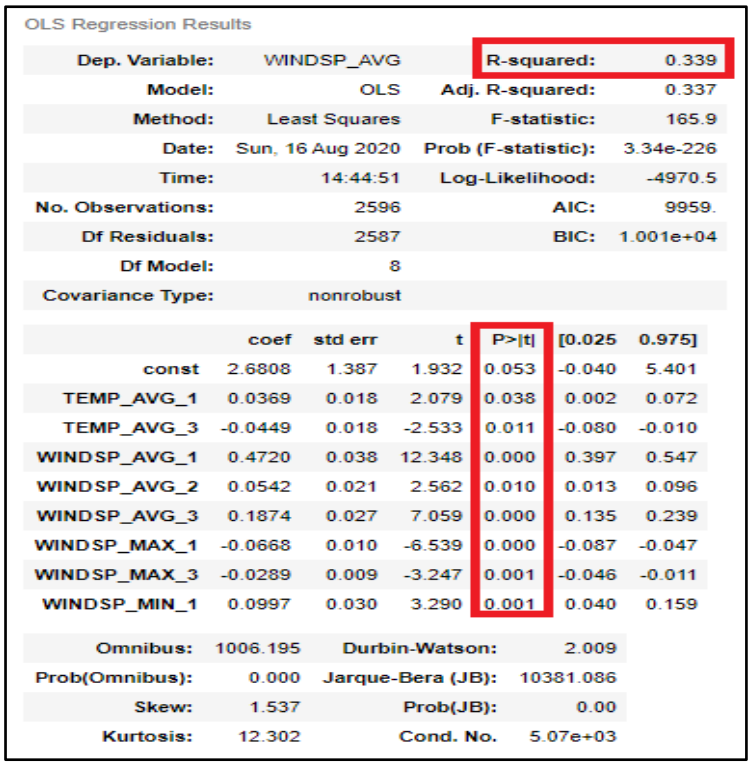


Figure 59: Scatter Plots for Average Wind Speed

	WINDSP_AVG
RH_MAX_1	-0.380682
RH_MAX_2	-0.289320
RH_MAX_3	-0.236805
RH_AVG_1	-0.208334
RH_AVG_2	-0.132820
PRECIPITATION_1	-0.107202
PRESSURE_MAX_1	-0.101480
PRECIPITATION_2	-0.100429
RH_AVG_3	-0.086104
DEWP_MAX_1	-0.078285
TEMP_MAX_3	-0.077255
TEMP_MAX_2	-0.073280
PRECIPITATION_3	-0.069158
PRESSURE_MAX_2	-0.063083
PRESSURE_MAX_3	-0.057108
DEWP_AVG_1	-0.057020
DEWP_AVG_2	-0.043372
TEMP_MAX_1	-0.038154
DEWP_MAX_2	-0.036581
PRESSURE_AVG_3	-0.029703
DEWP_MAX_3	-0.024794

Figure 60:Correlation Output for Average Wind Speed



OLS Regression Results

Dep. Variable:	WINDSP_AVG	R-squared:	0.339
Model:	OLS	Adj. R-squared:	0.337
Method:	Least Squares	F-statistic:	165.9
Date:	Sun, 16 Aug 2020	Prob (F-statistic):	3.34e-226
Time:	14:44:51	Log-Likelihood:	-4970.5
No. Observations:	2596	AIC:	9959.
Df Residuals:	2587	BIC:	1.001e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.6808	1.387	1.932	0.053	-0.040	5.401
TEMP_AVG_1	0.0369	0.018	2.079	0.038	0.002	0.072
TEMP_AVG_3	-0.0449	0.018	-2.533	0.011	-0.080	-0.010
WINDSP_AVG_1	0.4720	0.038	12.348	0.000	0.397	0.547
WINDSP_AVG_2	0.0542	0.021	2.562	0.010	0.013	0.096
WINDSP_AVG_3	0.1874	0.027	7.059	0.000	0.135	0.239
WINDSP_MAX_1	-0.0668	0.010	-6.539	0.000	-0.087	-0.047
WINDSP_MAX_3	-0.0289	0.009	-3.247	0.001	-0.046	-0.011
WINDSP_MIN_1	0.0997	0.030	3.290	0.001	0.040	0.159

Omnibus:	1006.195	Durbin-Watson:	2.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10381.086
Skew:	1.537	Prob(JB):	0.00
Kurtosis:	12.302	Cond. No.	5.07e+03

Figure 61: Final Result after applying the Backward Elimination for Average Wind Speed

The following conclusions can be made by observing the OLS regression results related to average windspeed.

- All remaining predictors have p-values significantly below of 0.05.
- R-squared value which measure the overall variance of the model, 0.339 interpreted that final model explains about 33% of the observed variation in the outcome variable of average windspeed.

```
array([4.35975195, 4.76360773, 8.25176927, 3.70633873, 6.57307147,
       5.93305164, 5.38174876, 7.19995566, 3.1816921 , 4.77240089,
       2.68116687, 5.12112478, 6.57302091, 3.40081674, 5.47094795,
       2.90436849, 2.99449149, 6.55575696, 4.99187896, 5.23206116,
       5.00915724, 3.97676481, 6.80388001, 3.38959549, 3.7259713 ,
       3.55134529, 4.11034176, 4.54049898, 7.31683057, 5.90328971,
       3.42220776, 5.82799808, 3.86904057, 6.27857812, 3.76406319,
       3.91869932, 3.69501152, 3.31353485, 4.33453046, 2.75131689,
       4.57373424, 3.75822412, 4.04835119, 4.13020209, 3.29003597,
       7.83989657, 3.98833411, 5.31005739, 6.43144595, 3.92412366,
       4.15502374, 3.77153149, 4.00845778, 5.53576139, 9.54662213,
       3.63525002, 2.8527049 , 5.82911229, 2.39167484, 3.68883244,
       4.61422115, 5.19884783, 3.70714364, 3.68871782, 4.92619292,
       2.6147596 , 6.33285265, 3.4652311 , 2.58185543, 5.60298205,
       4.09650309, 4.74598682, 5.95528912, 3.71849116, 5.99290298,
       3.49990936, 5.78215598, 6.14802194, 5.6201054 , 4.32902397,
       6.71503166, 4.03298963, 4.38252918, 3.64986022, 6.15568133,
       6.19932352, 4.88922964, 5.32235203, 5.16981128, 4.12564117,
       6.80211545, 6.75236724, 4.52239227, 4.46646611, 6.2848345 ,
       4.43472787, 3.56969981, 5.15359439, 2.43003726, 8.57589857,
       4.96985051, 3.79134547, 7.18499248, 5.24904577, 5.24241007,
       4.24192382, 5.34402293, 6.24034167, 2.84081296, 5.38131113,
```

Figure 62:Predicted Values for Average Wind Speed

4.2.4.3 Test results for Relative Humidity

Following plot indicates the visualization of relationships with dependent variable relative humidity along with the Predictor variables using scatter plots.

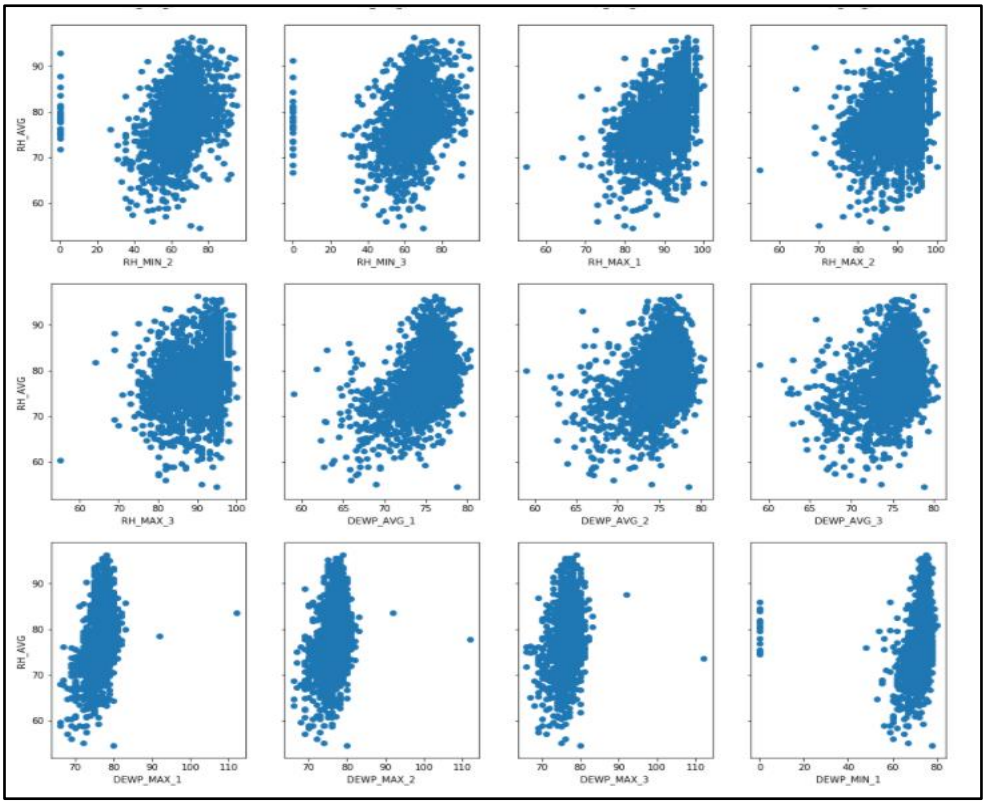
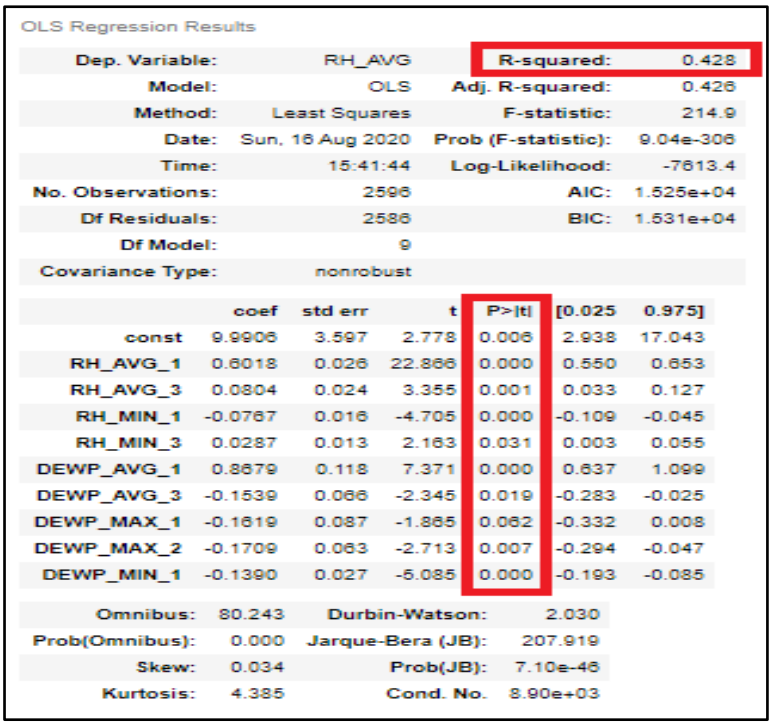


Figure 63: Scatter Plots for Relative Humidity

	RH_AVG
TEMP_MAX_1	-0.287776
PRESSURE_MAX_2	-0.218948
PRESSURE_MAX_1	-0.211055
TEMP_AVG_1	-0.203267
PRESSURE_MAX_3	-0.196126
WINDSP_AVG_1	-0.185821
TEMP_MAX_2	-0.184119
TEMP_MAX_3	-0.161527
WINDSP_MIN_1	-0.120798
WINDSP_AVG_2	-0.120695
WINDSP_AVG_3	-0.095895
TEMP_AVG_2	-0.093295
WINDSP_MAX_1	-0.089836
TEMP_AVG_3	-0.066463
WINDSP_MIN_2	-0.064978
WINDSP_MAX_2	-0.064969
WINDSP_MIN_3	-0.047383
WINDSP_MAX_3	-0.045897
PRESSURE_AVG_1	-0.024390

Figure 64: Correlation Output for Relative Humidity



OLS Regression Results

Dep. Variable:	RH_AVG	R-squared:	0.428
Model:	OLS	Adj. R-squared:	0.426
Method:	Least Squares	F-statistic:	214.9
Date:	Sun, 16 Aug 2020	Prob (F-statistic):	9.04e-306
Time:	15:41:44	Log-Likelihood:	-7613.4
No. Observations:	2596	AIC:	1.525e+04
Df Residuals:	2586	BIC:	1.531e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.9906	3.597	2.778	0.006	2.938	17.043
RH_AVG_1	0.6018	0.026	22.866	0.000	0.550	0.653
RH_AVG_3	0.0804	0.024	3.355	0.001	0.033	0.127
RH_MIN_1	-0.0767	0.016	-4.705	0.000	-0.109	-0.045
RH_MIN_3	0.0287	0.013	2.163	0.031	0.003	0.055
DEWP_AVG_1	0.8679	0.118	7.371	0.000	0.637	1.099
DEWP_AVG_3	-0.1539	0.066	-2.345	0.019	-0.283	-0.025
DEWP_MAX_1	-0.1619	0.087	-1.865	0.062	-0.332	0.008
DEWP_MAX_2	-0.1709	0.063	-2.713	0.007	-0.294	-0.047
DEWP_MIN_1	-0.1390	0.027	-5.085	0.000	-0.193	-0.085

Omnibus:	80.243	Durbin-Watson:	2.030
Prob(Omnibus):	0.000	Jarque-Bera (JB):	207.919
Skew:	0.034	Prob(JB):	7.10e-46
Kurtosis:	4.385	Cond. No.	8.90e+03

Figure 65: Final Result after applying the Backward Elimination for Relative Humidity

The following conclusions can be made by observing the OLS regression results related to relative humidity.

- All remaining predictors have p-values significantly below of 0.05.
- R-squared value which measure the overall variance of the model, 0.428 interpreted that final model explains about 43% of the observed variation in the outcome variable of average relative humidity.

```
array([78.37862692, 78.82689772, 73.05147631, 81.79321127, 77.2923559 ,
       67.88341045, 72.85903901, 78.92710193, 78.4093715 , 76.76016138,
       83.5993224 , 81.17954409, 76.28146518, 73.70412596, 78.29840596,
       80.76421504, 83.78505428, 76.42239438, 81.07389982, 76.53584116,
       70.28628556, 78.07084978, 77.02845453, 86.17676402, 77.77741127,
       80.39157259, 77.72451148, 75.17539279, 77.18969421, 64.29346768,
       80.86475665, 76.75973383, 80.04575497, 78.5833549 , 78.45337498,
       83.41996   , 80.61011863, 78.64251054, 80.42650677, 84.68256306,
       76.32141403, 82.50680537, 80.8953779 , 77.76794404, 79.63208458,
       79.84471283, 80.29545841, 71.52614923, 73.23036461, 75.93732509,
       76.82540457, 73.61547933, 83.18141721, 76.51345454, 78.54745027,
       79.91812189, 75.36740372, 81.34013476, 85.20026137, 73.12774284,
       74.56520118, 78.77525141, 72.54070402, 77.2459565 , 81.70170848,
       84.74727712, 69.87340699, 78.89844811, 78.20556629, 70.64595652,
       79.12635608, 73.42553701, 73.60770097, 76.27816602, 68.84161498,
       81.86827051, 78.50929026, 80.13481115, 76.40397414, 80.73886121,
       76.05127018, 77.87646335, 78.62443752, 78.82322848, 63.66555929,
       61.34713321, 80.61601156, 79.3508908 , 72.0754302 , 83.5740984 ,
       76.78964885, 75.24317095, 81.08582692, 76.85225546, 76.42249345,
       79.42768241, 81.52379569, 77.98835371, 83.5109861 , 77.0443108 ,
       80.99313423, 78.78926082, 80.62051355, 71.81752813, 75.72885483,
       79.52808841, 82.87835692, 71.2047275 , 83.49780313, 74.91797766,
       75.43892814, 79.33563957, 78.38323286, 76.74225548, 83.72618013,
       82.75901422, 86.92339558, 78.63122005, 79.11978581, 84.63750849,
```

Figure 66: Predicted Values for Relative Humidity

4.2.5 Results related to the Decision Tree Analysis

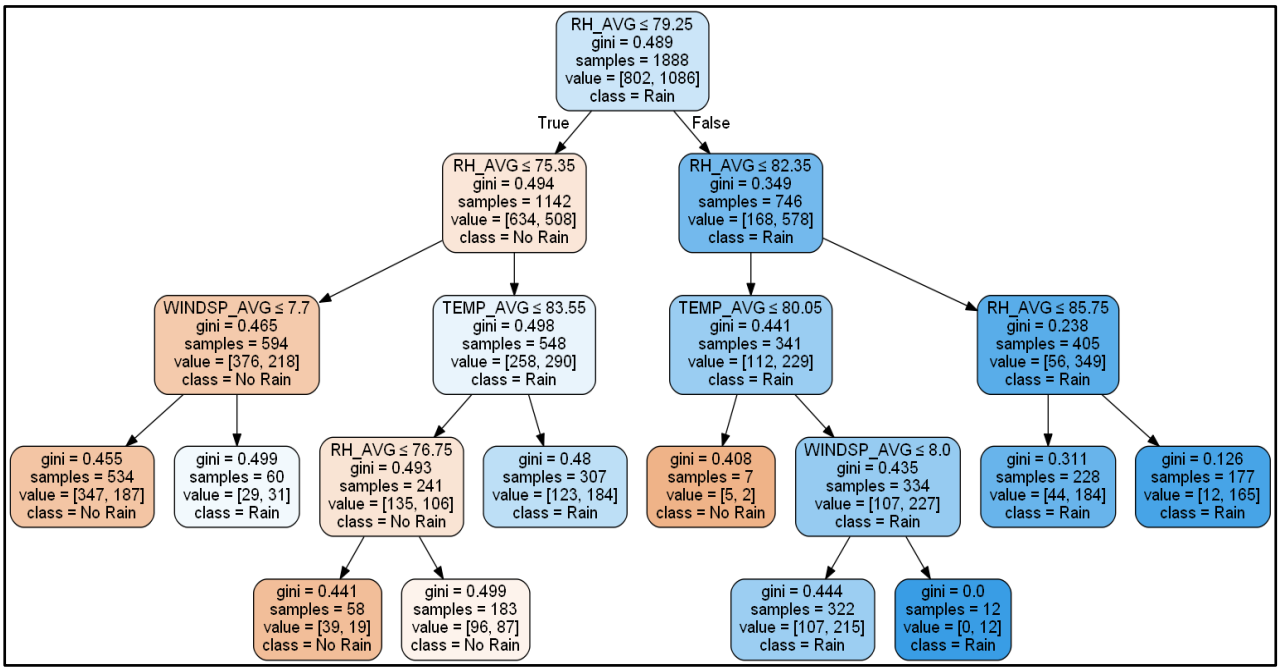


Figure 67:Decision Tree for determining the conditions for day being a Rainy day or a Not Rainy day.

```
|--- RH_AVG <= 79.25
|    |--- RH_AVG <= 75.35
|    |    |--- WINDSP_AVG <= 7.70
|    |    |    |--- class: 0
|    |    |--- WINDSP_AVG >  7.70
|    |    |    |--- class: 1
|    |--- RH_AVG >  75.35
|    |    |--- TEMP_AVG <= 83.55
|    |    |    |--- RH_AVG <= 76.75
|    |    |    |    |--- class: 0
|    |    |    |--- RH_AVG >  76.75
|    |    |    |    |--- class: 0
|    |    |--- TEMP_AVG >  83.55
|    |    |    |--- class: 1
|--- RH_AVG >  79.25
|    |--- RH_AVG <= 82.35
|    |    |--- TEMP_AVG <= 80.05
|    |    |    |--- class: 0
|    |    |--- TEMP_AVG >  80.05
|    |    |    |--- WINDSP_AVG <= 8.00
|    |    |    |    |--- class: 1
|    |    |    |--- WINDSP_AVG >  8.00
|    |    |    |    |--- class: 1
|    |--- RH_AVG >  82.35
|    |    |--- RH_AVG <= 85.75
|    |    |    |--- class: 1
|    |    |--- RH_AVG >  85.75
|    |    |    |--- class: 1
```

Figure 68: Tree Rules

4.3 Evaluation

In this section it describes how evaluation process is carried out when developing the optimal weather prediction model for ensure the efficiency and the accuracy.

4.3.1 Evaluation of Time Series Analysis

4.3.1.1 Average Temperature

When estimating the values for the model parameters in the ARIMA (p, d, q) model, we need to test the AIC and BIC values for different combinations of the p (the order of autoregression) and q (the order of the moving average) values. AIC and BIC values are mainly used to select between different models and the lowest score is the most preferable. In this study we are mainly concern with the AIC value for optimal model selection.

Following tables show the different combinations of the p and q values as well as the AIC and BIC values associate with ARIMA (p, 0, q) model.

Models	**AIC value**	**BIC value**
ARIMA (0,0,0)	11576.438	11576.438
ARIMA (0,0,1)	10597.237	10614.825
ARIMA (0,0,2)	10328.094	10351.546
ARIMA (0,0,3)	10197.294	10226.609
ARIMA (1,0,0)	10035.423	10053.012
ARIMA (1,0,1)	9919.949	9943.400
ARIMA (1,0,2)	9833.376	9862.691
ARIMA (1,0,3)	9816.410	9851.588
ARIMA (2,0,0)	9973.999	9997.450
ARIMA (2,0,1)	**9802.108**	9831.422
ARIMA (2,0,2)	9803.784	9838.961
ARIMA (2,0,3)	**9802.007**	9843.047
ARIMA (3,0,0)	9915.239	9944.553
ARIMA (3,0,1)	**9803.853**	9839.030
ARIMA (3,0,2)	9805.916	9846.956

ARIMA (3,0,3)	9803.878	9850.781

Table 3: AIC and BIC value representation of Average Temperature fitted models

According to the above table, the lowest AIC value can be obtained from ARIMA (2,0,3) model for the average temperature variable. Therefore, ARIMA (2,0,3) is selected for forecasting values.

After selecting the optimal ARIMA model for the average temperature, it needs to perform the model evaluation to ensure the accuracy of the model. Hence the model evaluation can be performed by calculating the Mean Square Error (MSE) and the Mean Absolute Error (MAE). Following are the values obtained from the ARIMA (2,0,3) model.

```
Mean Squared Error: 2.529886173875319
Mean Absolute Error: 1.1682059124537734
```

Figure 69: Error calculation of ARIMA (2,0,3) model for average temperature

According to the Table 2 the next suitable model is ARIMA (2,0,1) Since it has an AIC value of 9802.108. Therefore, we can consider the MSE and MAE values of the ARIMA (2,0,1) as well for determining the accuracy of the model. Following are the values obtained from the ARIMA (2,0,1) model.

```
Mean Squared Error: 2.5339666199925164
Mean Absolute Error: 1.1687106371909006
```

Figure 70: Error calculation of ARIMA (2,0,1) model for average temperature

Since lower the values for the Mean Square Error (MSE) and the Mean Absolute Error (MAE) will generate the better prediction model, according to the above outputs ARIMA (2,0,3) has a high accuracy comparing with the ARIMA (2,0,1) model related to the average temperature variable.

4.3.1.2 Average Wind Speed

Models	AIC value	BIC value
ARIMA (0,0,0)	11031.068	11042.794
ARIMA (0,0,1)	10458.336	10475.9249
ARIMA (0,0,2)	10326.778	10350.229
ARIMA (0,0,3)	10221.338	10250.652
ARIMA (1,0,0)	10207.330	10224.918
ARIMA (1,0,1)	10090.612	10114.063
ARIMA (1,0,2)	10065.104	10094.419
ARIMA (1,0,3)	10063.670	10098.847
ARIMA (2,0,0)	10139.304	10162.755
ARIMA (2,0,1)	10055.801	10085.116
ARIMA (2,0,2)	**10050.844**	10086.021
ARIMA (2,0,3)	**10051.031**	10092.071
ARIMA (3,0,0)	10091.876	10091.876
ARIMA (3,0,1)	10053.407	10088.584
ARIMA (3,0,2)	10051.769	10092.81
ARIMA (3,0,3)	**10048.570**	10095.473

Table 4: AIC and BIC value representation of Average wind speed fitted models

According to the above table, the lowest AIC value can be obtained from ARIMA (3,0,3) model for the average wind speed variable. Therefore, ARIMA (3,0,3) is selected for forecasting values.

After selecting the optimal ARIMA model for the average wind speed, it needs to perform the model evaluation to ensure the accuracy of the model. Hence the model evaluation can be performed by calculating the Mean Square Error (MSE) and the Mean Absolute Error (MAE). Following are the values obtained from the ARIMA (3,0,3) model.

```
Mean Squared Error: 2.7795077863694746
Mean Absolute Error: 1.1977497080098443
```

Figure 71: Error calculation of ARIMA (3,0,3) model for average wind speed

According to the Table 2 the next suitable model is ARIMA (2,0,2) Since it has an AIC value of 10050.844. Therefore, we can consider the MSE and MAE values of the ARIMA (2,0,2) as well for determining the accuracy of the model. Following are the values obtained from the ARIMA (2,0,2) model.

```
Mean Squared Error: 2.7862392692541587
Mean Absolute Error: 1.2010257274941332
```

Figure 72: Error calculation of ARIMA (2,0,2) model for average wind speed

Since lower the values for the Mean Square Error (MSE) and the Mean Absolute Error (MAE) will generate the better prediction model, according to the above outputs ARIMA (3,0,3) has a high accuracy comparing with the ARIMA (2,0,2) model related to the average wind speed variable.

4.3.1.3 Relative Humidity

Models	**AIC value**	**BIC value**
ARIMA (0,0,0)	16699.469	16711.195
ARIMA (0,0,1)	15806.281	15823.869
ARIMA (0,0,2)	15583.706	15607.158
ARIMA (0,0,3)	15512.290	15541.604
ARIMA (1,0,0)	15412.791	15430.379
ARIMA (1,0,1)	15393.184	15416.6364
ARIMA (1,0,2)	15345.329	15374.643
ARIMA (1,0,3)	15332.920	15368.097
ARIMA (2,0,0)	15400.681	15424.133
ARIMA (2,0,1)	**15326.197**	15355.511
ARIMA (2,0,2)	15327.571	15362.748
ARIMA (2,0,3)	**15327.471**	15376.370
ARIMA (3,0,0)	15376.370	15405.685
ARIMA (3,0,1)	15327.646	15362.823
ARIMA (3,0,2)	15329.325	15370.365

ARIMA (3,0,3)	**15326.401**	15373.30

Table 5: AIC and BIC value representation of Average humidity fitted models

According to the above table, the lowest AIC value can be obtained from ARIMA (2,0,1) model for the average humidity variable. Therefore, ARIMA (2,0,1) is selected for forecasting values.

After selecting the optimal ARIMA model for the average humidity, it needs to perform the model evaluation to ensure the accuracy of the model. Hence the model evaluation can be performed by calculating the Mean Square Error (MSE) and the Mean Absolute Error (MAE). Following are the values obtained from the ARIMA (2,0,1) model.

```
Mean Squared Error: 21.23089498893645
Mean Absolute Error: 3.4523066545275602
```

Figure 73: Error calculation of ARIMA (2,0,1) model for average humidity

Since there is a considerably high value for the Mean Squared Error with Mean Absolute error, we can further calculate the error term using other methods such as Relative Mean Square error (RMSE), Relative Mean Square error (RMSE) and Mean Absolute Percent error (MAPE).

4.3.2 Evaluation of Recurrent Neural Networks (RNN) for Rainfall Prediction

```
The Mean Squared Error: 18.908 mm
The Mean Absolute Error: 9.89  mm
The Median Absolute Error: 3.28 mm
```

Figure 74: Error calculation for Recurrent Neural Network

4.3.3 Evaluation of Decision Tree Classifier for Rainfall Prediction

```
Accuracy of the Decision Tree Classifier :  0.613978494623656
```

```
The Mean Squared Error: 0.386 mm
The Mean Absolute Error: 0.39  mm
The Median Absolute Error: 0.00 mm
```

Figure 75: Error calculation for Decision Tree Analysis

4.3.4 Evaluation on Linear Regression

4.3.4.1 Average Temperature

```
The Explained Variance: 0.51
The Mean Absolute Error: 1.16 degrees fahrenheit
The Mean Squared Error: 1.16 degrees fahrenheit
The Median Absolute Error: 0.98 degrees fahrenheit
```

Figure 76: Error calculation for Average Temperature using Linear Regression

According to the Figure 76, the model is able to **explain about 51% of the variance** observed in the Average temperature variable.

4.3.4.2 Average Wind Speed

```
The Explained Variance: 0.39
The Mean Absolute Error: 1.14 miles per hour
The Mean Squared Error: 2.33 miles per hour
The Median Absolute Error: 0.90 miles per hour
```

Figure 77: Error calculation for Average Wind Speed using Linear Regression

According to the Figure 77, the model is able to **explain about 39% variance** observed in the outcome variable which is Average Wind Speed variable

4.3.4.3 Average Relative Humidity

```
The Explained Variance: 0.44
The Mean Absolute Error: 3.32 percent
The Mean Squared Error: 19.73 percent
The Median Absolute Error: 2.67 percent
```

Figure 78: Error calculation for Relative Humidity using Linear Regression

According to the Figure 78, the model is able to explain about 44% variance observed in the outcome variable which is Average Relative Humidity variable.

Explained variance measures the discrepancy between a model and actual data. Generally, value for explained variance should not be less than 60% and not maximum of 100%. [4] According to the evaluation results of linear regression for average temperature, wind speed and relative humidity; the explained variance achieved is less than 60%. Therefore, it shows that there may be a need for a refinement in analysis process and linear regression model is not the optimal solution for the prediction of weather variations.

CHAPTER 5: DISCUSSION, CONCLUSION AND FUTURE WORK

5.1 Discussion and Conclusions

This section concludes this document by providing the final comments, thoughts related to the study and future works that can be carried out by extending this research. The main objective of this study is to develop a reliable weather prediction model using data mining and machine learning approaches. This study mainly carried out using Artificial Neural Network, Time Series Analysis, Regression Analysis and Decision Tree Analysis as the main data mining and machine learning approaches and predicted the future values for rainfall, average temperature, windspeed and relative humidity, dew point and atmospheric pressure weather variables depending on the approach.

During this study artificial neural network has been applied for predicting the rainfall as the output variable while taking average temperature, dew point, relative humidity, windspeed and atmospheric pressure as the input variables. Although this study was carried out only for prediction the rainfall, it can be further enhanced for prediction of other meteorological parameters as well.

Further time series analysis can be applied not only for the temperature, wind speed and relative humidity attributes but also for the other meteorological parameters such as atmospheric pressure, dew point and rainfall as well. Depend on the availability of the data related to the above weather parameters we can apply time series analysis.

During this study we have use the ARIMA (p, d, q) model for forecasting the values for temperature, wind speed and relative humidity weather variables. The data used in this work was collected from the 'Weather Underground'[6] website. Since the raw data is always affecting with the missing values, noise, and inconsistency, it is necessary to pre-process data before applying the machine learning techniques. After preparing the dataset, we have performed the time series analysis on the refined weather data set. As the initial step of the time series analysis we have checked the stationarity of the weather dataset by performing the Augmented Dickey-Fuller test and by using the inspection results of the ACF, PACF autocorrelation plots. Based on the output results, we have noticed that the data for average temperature, average windspeed and relative humidity are having seasonal stationarity but there is no trend associated with it. Since time series is stationary it is not necessary to perform

the differencing. Hence the non- seasonality parameter (d) in the ARIMA model becomes zero. (d = 0). Then the most appropriate order of the ARIMA model has determined using the AIC criterion. AIC is mainly used to select between different models where the lowest score is the most preferable. After estimating the values for the model parameters, we have developed the ARIMA model for predicting the future values for the selected weather parameters.

Finally, based on the evaluation results of this study, it can conclude that **Recurrent Neural Network** and **Time Series Analysis** with **ARIMA model** will generate an **optimal solution model** for predicting the weather variations in Sri Lankan context.

5.2 Limitations of the Study

Since this study is conducted for the Sri Lankan context, we use the meteorological data related to Sri Lanka. There are more than 20 meteorological stations located in Sri Lanka. But due to unavailability of the data, incompleteness of the data as well as the time and cost constraints we have only used the data collected from the Colombo meteorological station.

5.3 Future Work

As the future work we are planning to apply other machine learning algorithms such as Recurrent Neural Network, Regression Analysis and Deep Learning approaches for developing a model for weather predictions.

REFERENCES

[1] A. H. Nury, M. Koch, and M. J. B. Alam, “Time Series Analysis and Forecasting of Temperatures in the Sylhet Division of Bangladesh,” p. 4.

[2] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, “An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives,” *Expert Systems with Applications*, vol. 85, pp. 169–181, Nov. 2017, doi: 10.1016/j.eswa.2017.05.029.

[3] A. R. Ganguly and K. Steinhaeuser, *2008 IEEE International Conference on Data Mining Workshops Data Mining for Climate Change and Impacts*. .

[4] A. A. Shafin, “Machine Learning Approach to Forecast Average Weather Temperature of Bangladesh,” p. 11, 2019.

[5] A. Perera and U. Rathnayake, “Rainfall and Atmospheric Temperature against the Other Climatic Factors: A Case Study from Colombo, Sri Lanka,” *Mathematical Problems in Engineering*, vol. 2019, pp. 1–15, Dec. 2019, doi: 10.1155/2019/5692753.

[6] “Local Weather Forecast, News and Conditions | Weather Underground.” https://www.wunderground.com/ (accessed May 16, 2020).

[7] M. A. Nayak and S. Ghosh, “Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier,” *Theor Appl Climatol*, vol. 114, no. 3–4, pp. 583–603, Nov. 2013, doi: 10.1007/s00704-013-0867-3.

[8] “Distributed Data Mining Solution for Detecting Behavior of Climate Changes.doc.” .

[9] “NASA,” *NASA*. http://www.nasa.gov/index.html (accessed Oct. 24, 2019).

[10] M. Das and S. K. Ghosh, “Detection of climate zones using multifractal detrended cross-correlation analysis: A spatio-temporal data mining approach,” in *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, Kolkata, India, Jan. 2015, pp. 1–6, doi: 10.1109/ICAPR.2015.7050702.

[11] “The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.” .

[12] G. V. Krishna, “An Integrated Approach for Weather Forecasting based on Data Mining and Forecasting Analysis,” 2015.

[13] H. Dai, “Machine Learning of Weather Forecasting Rules from Large Meteorological Data Bases,” *Adv. Atmos. Sci.*, vol. 13, no. 4, pp. 471–488, Nov. 1996, doi: 10.1007/BF03342038.

[14] E. Abrahamsen, O. M. Brastein, and B. Lie, “Machine Learning in Python for Weather Forecast based on Freely Available Weather Data,” Nov. 2018, pp. 169–176, doi: 10.3384/ecp18153169.

[15] Zhaoxia Wang *et al.*, “Disclosing Climate Change Patterns Using an Adaptive Markov Chain Pattern Detection Method,” in *2013 International Conference on Social Intelligence and Technology*, State College, PA, May 2013, pp. 72–79, doi: 10.1109/SOCIETY.2013.15.

[16] Y. Jararweh, I. Alsmadi, M. Al-Ayyoub, and D. Jenerette, “The analysis of large-scale climate data: Jordan case study,” in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, Doha, Qatar, Nov. 2014, pp. 288–294, doi: 10.1109/AICCSA.2014.7073211.

[17] A. Culclasure, “Using Neural Networks to Provide Local Weather Forecasts,” p. 73.

[18] B. Wang *et al.*, “Deep Uncertainty Quantification: A Machine Learning Approach for Weather Forecasting,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA, Jul. 2019, pp. 2087–2095, doi: 10.1145/3292500.3330704.

[19] H. Kremer, S. Gunnemann, and T. Seidl, “Detecting Climate Change in Multivariate Time Series Data by Novel Clustering and Cluster Tracing Techniques,” in *2010 IEEE International Conference on Data Mining Workshops*, Sydney, TBD, Australia, Dec. 2010, pp. 96–97, doi: 10.1109/ICDMW.2010.39.

[20] Mashfiqul Huq Chowdhury 1, Somaresh Kumar Mondal 1, and Jobaidul Islam *2, “Modeling And Forecasting Humidity In Bangladesh: Boxjenkins Approach,” Apr. 2018, doi: 10.5281/ZENODO.1241452.

[21] “Box G E P & Jenkins G M. Time series analysis: forecasting and control.San Francisco, CA: Holden-Day. (1970) 1976. 575 p.,” p. 1.

[22] A. H. Nury, K. Hasan, and Md. J. B. Alam, “Comparative study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh,” *Journal of King Saud University - Science*, vol. 29, no. 1, pp. 47–61, Jan. 2017, doi: 10.1016/j.jksus.2015.12.002.

[23] R. S. Somasundaram and R. Nedunchezhian, *Missing Value Imputation using Refined Mean Substitution*. .

[24] X. Yi, Y. Zheng, J. Zhang, and T. Li, “ST-MVL: Filling Missing Values in Geo-Sensory Time Series Data,” p. 7.

[25] D. N. Fente and D. K. Singh, “Weather Forecasting Using Artificial Neural Network,” p. 5, 2018.

[26] “Time series,” *Wikipedia*. Oct. 10, 2020, Accessed: Nov. 18, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Time_series&oldid=982786780.

[27] “Augmented Dickey–Fuller test,” *Wikipedia*. Jul. 03, 2020, Accessed: Nov. 18, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Augmented_Dickey%E2%80%93Fuller_test&oldid=965832014.

[28] “Unit root test,” *Wikipedia*. Oct. 30, 2020, Accessed: Nov. 18, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Unit_root_test&oldid=986188646.

[29] “Introduction to ARIMA models.” https://people.duke.edu/~rnau/411arim.htm (accessed Jun. 20, 2020).

[30] “Using Machine Learning to Predict the Weather: Part 2,” *Stack Abuse*. https://stackabuse.com/using-machine-learning-to-predict-the-weather-part-2/ (accessed Nov. 18, 2020).

[31] “Decision tree learning,” *Wikipedia*. Sep. 22, 2020, Accessed: Oct. 09, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=979741036.

[32] K. Mittal, D. Khanduja, and P. C. Tewari, “An Insight into ‘Decision Tree Analysis,’” p. 6.

[33] Mashfiqul Huq Chowdhury 1, Somaresh Kumar Mondal 1, and Jobaidul Islam *2, “Modeling And Forecasting Humidity In Bangladesh: Boxjenkins Approach,” Apr. 2018, doi: 10.5281/ZENODO.1241452.