2020-01-01

Weather Prediction: Improving Accuracy Using Data Mining And Forecasting Techniques

Pedro Marquez
University of Texas at El Paso

WEATHER PREDICTION: IMPROVING ACCURACY USING DATA MINING AND FORECASTING TECHNIQUES

PEDRO ALEJANDRO MARQUEZ

Master's Program in Industrial Engineering

APPROVED:

Jose F. Espiritu, Ph.D., Chair

Heidi A. Taboada, Ph.D.

Virgilio Gonzalez, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

Dedication

I dedicate every effort I made to complete this thesis project to my family. A special feeling of gratitude to my parents, Pedro and Blanca whose affection, love, encouragement and prayers day and night made me able to achieve success and honor them. Thank you to my brothers, Carlos and Andres, whose words of encouragement made it possible. Thank you to the staff and faculty that have guided me along my graduate journey.

WEATHER PREDICTION: IMPROVING ACCURACY USING DATA MINING AND FORECASTING TECHNIQUES

by

PEDRO ALEJANDRO MARQUEZ, B.S.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Industrial, Manufacturing and Systems Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

August 2020

Acknowledgements

I want to thank first and foremost my advisor Dr. Jose Espiritu for all of his support, patience, guidance, and understanding. Thank you Dr. Taboada for guiding me along the way and helping me achieve my goals. Also, I would also thank Pablo Bustamante for his guidance and sharing of his knowledge on data mining during the project.

Special thanks to Dr. Irma Lawrence, who is responsible for the USDA's NIFA HSI program, for giving me the opportunity to carry out this higher education program. Finally, I thank the University of Texas at El Paso for providing me with the tools I needed to achieve academic success.

Abstract

This study is focused to provide the insights of weather to understand the significance of weather changes in any parameter. Weather forecasting contributes to the social and economic welfare in many sections of the society. Weather is extremely difficult to predict because it is a complex and chaotic system. This means that small errors in the initial conditions of a forecast grow rapidly and affect predictability. Nowadays, massive real-time data is being generated by IoT devices, radars, weather stations, and satellites. The need to adopt big data analytics in IoT applications is compelling. These two technologies have already been recognized in the fields of IT and business. Data mining techniques and machine learning algorithms need to be considered and trained with big data to improve the accuracy of weather forecasts. The contribution to this problem is to analyze the accuracy and correlation between weather conditions with the use of different data mining and forecasting techniques to predict precipitation for the next year.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Chapter 1 will investigate the integration of two emerging concepts in the fields of information technology and business: 'Internet of Things' and 'Big Data'. Each concept was examined by their impact and their development of smart connected communities. This chapter analyses how the world has been revolutionized by these technologies by identifying and analyzing several opportunities and challenges presented by the capabilities to ingest and utilize huge amounts of 'Internet of Things' data, which includes applications in smart cities, smart agriculture, smart transportation, etc. Finally, this chapter approaches how the generation of real-time data from the 'Internet of Things' devices created an opportunity to apply data mining techniques and develop smart weather forecasts.

1.1 INTRODUCTION

In future years, it is expected that cities will face several challenges such as safety, sustainability, energy use, effective service delivery, effective transportation systems, etc. Advances in the network integration of information systems, sensing and communication devices, data sources, and artificial intelligence, are generating opportunities to develop smart and connected communities. Nowadays, people expect the power of sensors, cloud computing, high speed networks and data analytics with the use of a myriad of things like smart phones, cars, social networks, and transportation apps like Uber.

According to Cheng et al. (2016), statistics show that 500 billion devices are expected to be connected to the Internet by 2030. Each device includes sensors that collect data, interact with the environment, and communicate over the network without any manual intervention with the help of embedded technology. The Internet of Things (IoT) is the network of these connected devices.

On September 14, 2015, the United States government announced a new Smart Cities Initiative to support local communities tackle key challenges such as reducing traffic congestion, fighting crime, fostering economic growth, managing the effects of a changing climate, and improving the delivery of city services. On November 25, 2015, the Networking and Information Technology Research and Development (NITRD) Program announced the release of version 4 of a Smart and Connected Communities Framework (NITRD, 2015). The framework outline is that communities in all settings and at all scales have access to Internet of Things technologies and services to improve the health, safety and economy of their residents.

IoT is the primary enabler of a larger industry transformation called digital business. Digital business is one that uses technology as an advantage in its internal and external operations. Sensors can detect location, environment, presence, and more. Big Data Analytics (BDA) is emerging as a key to analyzing IoT generated data from "connected devices" which helps to take the initiative to improve decision making and provide competitive advantage. IoT securely connects devices and fuels applications that can be delivered as services. Organizations are increasingly looking to digital technologies to create or enhance their business models, processes and services; to empower workforce efficiency and innovation; and personalize the citizen, customer, or employee experience.

1.2 INTERNET OF THINGS

Today, people are almost completely dependent on computers and the internet for information. The problem arises when human beings have limited time, attention, and accuracy at capturing data about our surroundings. As a result, Internet of Things is providing a linked set of computer programs and sensors that do not experience the same limitations than people do. Internet

of Things (IoT) is an interconnection of several “smart devices” that are equipped with microchips, sensors and wireless communication capabilities to achieve a common goal.

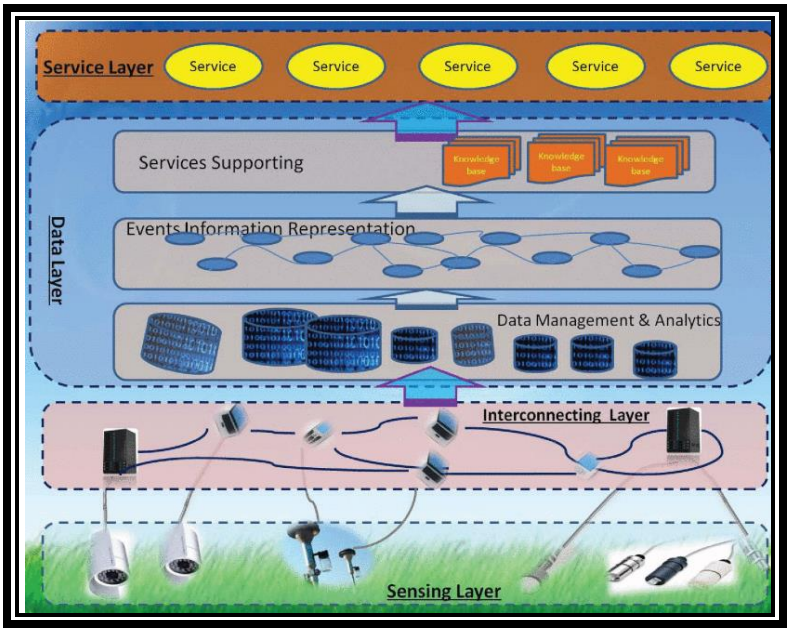


Figure 1.1: Function Architecture of Internet of Things for Smart (Sun et al. 2016).

The architecture of IoT consists of four different layers: sensing layer, interconnecting layer, data layer, and service layer (Sun et al. 2015). The sensing layer’s purpose is to perform ubiquitous sensing. Some examples of the most utilized sensors are smart phones, cameras, audios, accelerometers, GPS, gyroscope, compass, proximity, and ambient light. The interconnecting layer has the purpose of transmitting data and exchanging information among devices. Any IoT device is assigned an IP address, which identifies the network that the device is in. The Data layer is responsible for structuring and analyzing massive, trivial, heterogeneous data generated from different kinds of monitoring devices in the sensing layer. Meaningful and useful information is extracted from the data and represented in an efficient way that provides service support and actionable intelligence for the user. Finally, the service layer, or application layer, provides various services based on the data analyses.

As the number of connected IoT devices continues to grow, the amount of data generated by these devices will also grow. Some devices might just produce a minimum amount of data indicating only one metric, while others like video surveillance cameras generate huge amounts of data to examine, such as crowds of people surveillance. Immense data sets (petabytes or gigabytes) can demand advanced forms of information processing for better insights and decision making.

1.3 BIG DATA

Big Data has become important for any organization that generates a huge amount of heterogeneous data, which if captured, processed, and analyzed will reveal patterns and provide insights. Big Data concept emerged when large volumes of structured, semi-structured, and unstructured data posed a difficult task for processing using traditional methods and databases.

The size, speed, and format in which data is generated affect the quality of the information. Data can come from different sources such as business transaction systems, customer databases, mobile applications, websites, machines and real-time data sensors produced in IoT systems. This comes with challenges defined as 5Vs i.e. Volume, Variety, Velocity, Veracity, Value (Juneja & Das, 2019).

Gantz & Reinsel (2012) characterizes big data into three aspects: data sources, data analytics, and the presentation of the results of the analytics. Volume indicates the size of the data sets created at high frequency rates. Variety is present when there are different types of data types, such as structured, semi-structured or unstructured data. Velocity refers to the speed and frequency at which the data is created. Veracity deals with the accuracy, truthfulness and authenticity of the data. Value denotes the worthiness of data extracted from various raw data available, in other words, not all data extracted is useful.

According to Juneja & Das (2019), data flows through four phases in the Big Data System Lifecycle. These phases are Data Origin Identification, Data Acquisition, and Cleansing, Data Aggregation and Storage and Data Analysis as presented in Figure 1.2.

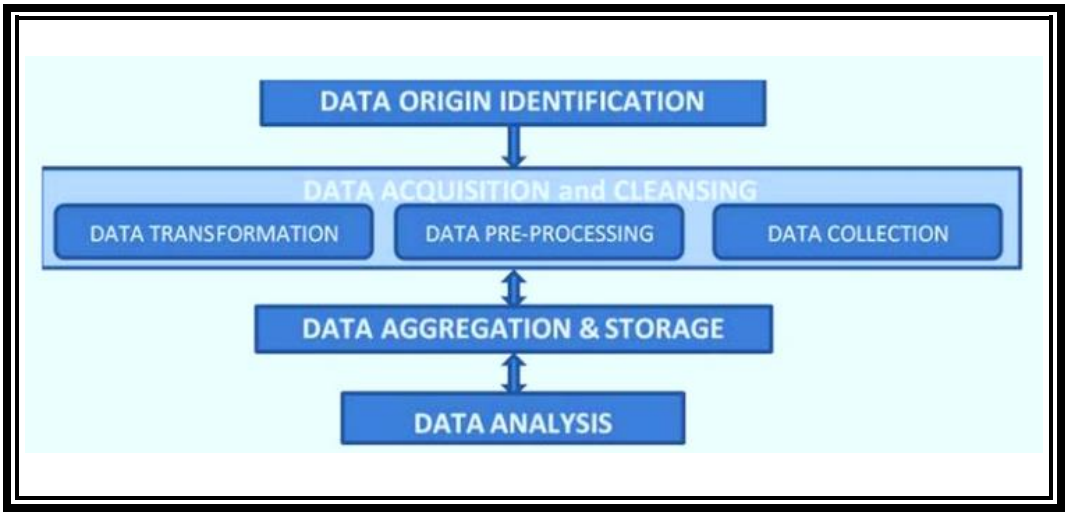


Figure 1.2: Big Data Lifecycle (Juneja & Das, 2019).

The first phase is ***Data Origin Identification***. This phase reviews the raw data being produced from multiple sources. The sources may include social sites, financial applications, customer relation applications, media web sites, images, etc. It is crucial to understand the veracity and reliability of data sources.

Secondly, ***Data Acquisition and Cleansing*** phase comprehends the data from many sources. This phase deals with the variety and value complexities. The raw data can be collected with anomalies, badly formatted, or a combination of structured, semi-structured and unstructured data. Such data requires to be cleaned and filtered, reformatted and structured, remove illegal values and compressed. These pre-processing steps are crucial to transform the data and eliminate the 5Vs variety and value complexities for analysis.

The third phase is ***Data Aggregation and Storage*** which confirms that all heterogeneous sources are stored and joined across databases.

Finally, the ***Data Analysis*** phase compares data characteristics to identify patterns usually by using high-level programming skills and methodologies. The results of this phase should inform the user about the final state, make forecasts, or provide an approach for decision making.

1.4 BIG DATA ANALYTICS IN IOT

The exponential growth of data produced by IoT devices has played a major role in the Big Data field. Big data analytics is quickly emerging as a significant IoT initiative to improve decision making. Big data analytics refers to the process of collecting, storing and analyzing large volumes of data sets to reveal trends, unseen patterns, hidden correlations, and new information (Golchha 2015). Big data analytics in IoT demands storing the data in several storage technologies. Big data implementations will require performing lightning-fast analytics with queries to allow organizations to gain rapid insights and make quick decisions. Depending on the requirements of the developed IoT applications, different analytic types have been discussed in this subsection under real-time, offline, business intelligence level, and massive level analytics (Chen & Zhang 2014). Furthermore, a comparison based on analytics types and their level is shown in Table 1.1.

Table 1.1: Comparison of Different Analytics Types and Their Levels (Marjani et al., 2017).

Analytic Type	Specified Use	Advantages
Real Time	To analyze the large amounts of data generated by the sensors	Parallel processing clusters using traditional databases memory-based computing platforms
Offline	To use for the Applications where there is no high requirements on response time.	Efficient data acquisition Reduce the cost of data format conversion
Memory Level	To use where the total data volume is smaller than the maximum. Memory of the cluster	-Real Time
Business Intelligence Data	To use when data scale surpasses the memory level	Both offline and online
Massive level	To use when data scale totally surpasses the capacity of business intelligence products and traditional databases.	Most are Offline

Historical data analysis uses a set of historical data for batch analysis. Real-time analytics instead visualizes and analyzes the data as it appears in the computer system. The role of big data

analytics in IoT is to process a massive volume of data on a real-time basis and take out the value by enabling high-velocity capture, discovery, and analysis. IoT big data processing follows four sequential steps: collecting, storing, mining and visualization. These four steps are described below.

1. Data generated by IoT devices is collected in the big data system. Big data is a large quantity of data characterized by the 3V's (volume, variety, velocity) definition proposed by Bayer (2011). Volume is the quantity of data produced. Variety refers to various forms of information that are retrieved like voice, texts, images, videos, document, sensor data, tweets, etc. Lastly, velocity approaches the high-speed at which these data is generated.
2. In the big data system, which is mainly a shared distributed database, a tremendous amount of data is stored in big data files. Big Data requires a parallel and distributed system architecture to store this data. Since data is provided from different sites and in different places, it needs to be stored in uniform manner.
3. Internet of Things data stored is used for analytics. The data from various sources is collected; they are refined and stored under uniform schemas. IoT applications involve data sets that may have a varied structure as unstructured, semi-structured and structured data sets. There may also be a significant difference in the data formats and types. Analyzing these large data allows people to discover the correlation, facts and other important information that lies in this large data set, which is impossible to be determined by human.
4. Finally, big data analytics generates reports, tables, graphs, diagrams or applications that uncover hidden patterns, unknown correlations, market trends, customer preferences or other useful business information. Usually, it allows the business executive to analyze all these varying sets of data using automated tools and software.

1.5 Applications, Features, and Products of IoT Data Analytics in Agriculture

A variety of IoT-based applications are being used in different sectors such as smart cities, smart home, smart business, agriculture, transportation, healthcare, logistics, etc. These applications have succeeded in providing enormous benefits to the users. These days, these applications have steadily increased and some of them are already deployed and being used at different levels (Want et al, 2015). IoT's applications require hardware, middleware and presentation. These applications have features such as interaction with the environment, interaction between people and devices, automatic routine tasks with less supervision, self-organized infrastructure and communication security (Javed et al 2018). Smart agriculture has been one of the most developed fields from IoT.

According to Elijah (2018), the world population is estimated to be about 9.7 billion in 2050 as such there will be great demand for food. Data analytics and IoT devices enables better and smarter agriculture that will allow people to overcome this demand. Some of the opportunities for implementing IoT devices in the agricultural sector are crop and livestock, machinery, irrigation and water quality monitoring, weather monitoring, soil monitoring, disease and pest control, automation and precision. Developing IoT and data analytics applications in the agricultural field enhance farmers' productivity, quality and profit.

The agriculture supply chain is composed of six different activities as shown in Figure 1.3. These six activities are input suppliers, farms, traders, processors, retailers, and consumers.

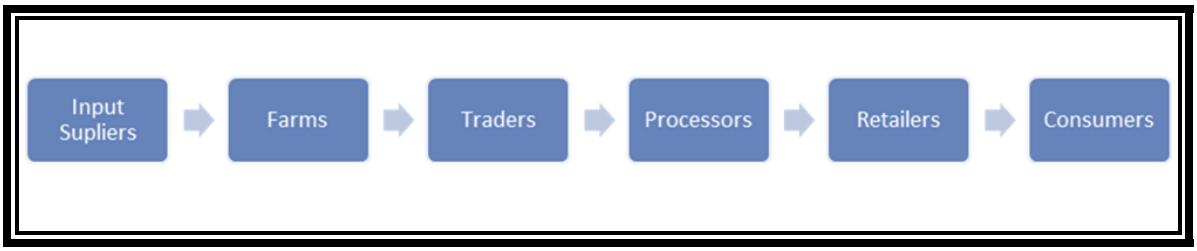


Figure 1.3: Supply Chain in Agriculture

IoT provides opportunities throughout the different activities in the supply chain in order to gain competitive advantage. For example, smart farms can offer insight on resource management. With the implementation of IoT and data analytics, farmers can make more effective purchases by optimizing and forecasting the use of machinery, fertilizers, seeds and chemicals from suppliers.

Sensors applied in smart agriculture generate data that assist farmers to monitor and optimize crops by adapting to changes in the environmental conditions. These sensors are placed on weather stations, drones, and machinery in the agriculture industry. Sensors provide valuable information on moisture soil, humidity levels, trunk diameter of plants, microclimate condition, as well as to forecast weather. Automatic climate control according to harvesting requirements, timely and controlled irrigation, and humidity control for fungus prevention are examples of actions performed based on big data analytics recommendations.

Agricultural gathering vehicles have been equipped with sensors, wireless communication, and dynamic programming has allowed the creation of autonomous vehicles. Furthermore, IoT sensors has been installed in the vehicles generate information that can be used to track current location and analyze efficiency in delivery times, fuel consumption, or delivery routes. IoT devices installed within the agricultural equipment will provide constant, accurate measurements of output to isolate sources of waste, gain process control, maximize productivity and ensure quality.

Connected devices and products gives retail companies the opportunity to optimize operations in their supply chain and improve the customer experience. One example of IoT technologies is RFIDs, which can precisely track inventory. Data visualization technologies allow employees to track products across the supply chain. Retail companies also count on the Internet of Things application development to improve self-checkout. Identifying person monitoring traffic

patterns in stores and trying to find a connection with trends can provide accurate data about how customers behave.

Like companies, government agencies are trying to deliver quality services in multifaceted environments. Smart infrastructure technologies can allow government agencies planners to measure and monitor traffic management, security, waste, energy, and water supplies to lower costs and improve services for the citizens (Meyers et al., 2015). At a federal level, agencies are more focused on scaling measurement capabilities: The Department of Defense uses RFID chips to monitor its supply chain (Defense Industry Daily, 2010), the US Geological Survey uses IoT devices to monitor the bacterial levels of rivers and lakes (Meyers et al., 2015), and the General Services Administration has initiated employing sensors to measure the energy efficiency of "green" buildings (Fowler et al., 2015). Regarding changes in the weather, IoT devices will provide information about temperatures, noise and air pollution to several agencies such as the National Weather Service and the Department of Agriculture.

Weather warning and forecasting applications seems to be one of the most useful applications for the government, industry, and the general public. The United States National Weather Service (NWS) supports all aspects of keeping the public safe from weather, water, and climate hazards by providing weather warnings and forecasting programs (weather.gov). In the agricultural field, sensors are retrieving useful information to forecast weather. Data analytics helps farmers to reduce the probability of having production risks and to create mitigation tactics for unexpected events. In addition, data analytics and IoT devices can significantly influence decisions and policymaking for food security.

1.6 Significance of Weather Forecasting

Weather forecasting is one of the most prominent topics that has been challenging scientists and engineers due to an expected increase of weather events and climate changes around the world. Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. Human beings have attempted to predict the weather since ancient times and started developing scientific forecast models since the nineteenth century. Nowadays, forecasters use more advanced methods and technologies to gather weather data, along with the world's most powerful computers. With the ability to launch satellites and supercomputers and harvest data from IoT devices, the new arrivals are advancing in information-gathering capabilities. Also, the use of data analytics techniques, as well as using machine learning, artificial intelligence and cloud-based warning systems; for example a system that indicates an airline company when to reschedule flights to avoid thunderstorms or a farmer when to irrigate a particular row of crops. These models can be programmed to predict how the atmosphere and the weather will change. Despite these advances, weather forecasts are still often incorrect. Weather is extremely difficult to predict because it is a complex and chaotic system.

Weather forecasting contributes to the social and economic welfare in many sections of the society. Weather forecast provides vital information to a wide range of fields: agriculture, aviation, energy, commerce, marine, advisories, insurance companies, etc. It can also significantly influence decision and policymaking, global food security, construction planning, productivity and environmental risk management (Wiston, 2018).

Weather forecasting is often used to predict and warn about natural disasters that are caused by abrupt change in climatic conditions. Catalyzed by climate change, extreme weather is an increasing liability to the economy, with approximately 10 weather and climate disasters costing

more than $1 billion each on previous years, according to NOAA. During the past two years alone, Western wildfires have cost more than $40 billion. Hurricanes are dumping more rainfall than they used to, and heat waves are more intense and frequent (Freedman, 2019). Forecasters are also responsible for effectively communicating the forecast and its anticipated impacts upon life and property to various stakeholders, including the public. The impacts of meteorological phenomena upon both life and property, whether it is a short or long effect in nature, are substantial. According to National Oceanic and Atmospheric Administration (NOAA, 2019) and the American Metrological Society (AMS, 2015), some of these impacts are:

- One-third of the US economy is sensitive to weather and climate.
- Approximately 90% of the emergencies declared by the Federal Emergency Management Agency are weather-related.
- Each year, the United States averages some 10,000 thunderstorms, 5,000 floods, 1,300 tornadoes and 2 hurricanes.
- Agriculture in the United States produces approximately $300 billion a year in commodities that are vulnerable to climate change.
- According to the U.S drought monitor in September 2019, moderate to extreme drought covers 14.3% of the United States
- More than 7000 road fatalities per year can be directly or indirectly attributed to weather. Approximately 70% of air traffic delays are caused by weather, at a cost of about $6 billion per year.
- Heat waves kill an average of about 175 people each year in the U.S.
- U.S. utilities save more than $150 million per year using 24-hour temperature forecasts to meet electricity demands most efficiently.

- Reducing the length of coastline under hurricane warnings involving the need for evacuations saves up to $1 million per coastal mile in evacuation and other preparedness costs.

The nation's Weather and Climate Enterprise is typically grouped into three sectors: government agencies, academic institutions, and the private sector. These three sectors play vital roles in providing products and services to the user community, which includes the general public, as well as weather-sensitive government agencies (e.g., the military, Homeland Security, Department of Commerce, Department of Agriculture, Environmental Protection Agency, numerous state and local offices, etc.) and private industries (e.g., energy, agriculture, transportation, etc.) (AMS, 2015). For example, the US Department of Agriculture develops and delivers science-based knowledge that empowers farmers, foresters, ranchers, landowners, resource managers, policymakers, and Federal agencies. The Department manages the risks, challenges, and opportunities of climate variability, which in turn informs decision-making and improves practices in environmental conservation.

The United States National Weather Service (NWS) supports all aspects of keeping the public safe from weather, water, and climate hazards and meeting the NWS mission to protect lives and property; and enhance the national economy. It provides weather warnings and forecasting programs to the government, industry, and the general public. Also, it monitors weather observations, provides public service, and monitors broadcasts over eleven NOAA Weather Radio-All Hazards Stations (NWS, 2019).

Nowadays, the National Weather Service is in an uncomfortable position as it tries to fulfill its mission of protecting lives and property. The agency is being challenged when it comes to weather and climate observations, since some private companies are obtaining better data. Private

weather forecasting is a $7 billion industry (and growing), according to a 2017 National Weather Service study. It's also increasingly testing the federal government's hold on weather data and warnings (Freedman, 2019).

1.8 CHAPTER CONCLUSION

In this chapter, the attractive fields of Internet of Things and Big Data were explored to see what a fully connected world can offer. Several IoT applications were reviewed and an opportunity for real-time analytics has been spotted in weather forecasting. It is becoming increasingly vital for scientists, agriculturists, farmers, disaster management, and related organizations to predict the atmospheric conditions and be prepared for any catastrophes. Massive weather real-time data from IoT devices are available with the use of new applications such as API's. By precisely analyzing the effects of observed weather conditions in real time, it is possible to anticipate the effects and trends similar such weather events will have in the future when their occurrence is expected to increase.

In this study, the objective is not only to correctly classify rainfall but also correctly predict weather conditions using various forecasted parameters. The focus of this research is to understand the effects of different meteorological conditions, along with an exploration of data mining techniques used for approaching weather forecasts. The main contribution is to find the best data mining and forecasting technique for several weather conditions that have a correlation with temperature and rainfall in order to predict precipitation for the next year. The following chapter will present a literature review of several weather predictions applying data mining techniques for weather prediction.

Chapter 2: Literature Review

This chapter is dedicated to reviewing available literature by different researchers that have developed and used algorithms considering data mining techniques for weather prediction. Also, various prediction analysis methods are studied and analyzed in terms of different parameters and trends. There are various algorithms of classification and prediction. Some of them are K-means clustering, Decision Trees, Artificial Neural Networks, Support Vector Machines (SVM), Bayesian Classification and Regression. There are several criteria for evaluating the prediction performance of algorithm. The work of these researchers has been reviewed and compared in a tabular form.

2.1 Weather Data Identification

Forecasting the weather begins by continuously observing the state of the atmosphere for a given location and its surrounding area. Temperature, air pressure, moisture, precipitation, evaporation, humidity, wind speed, wind direction, and other characteristics of the atmosphere must be measured, and the data collected. The World Meteorological Organization provides the framework of observing platforms, such as radars, satellites, and surface weather observations that support monitoring different weather variables. NOAA is the only authorized issuer of severe weather watches and warnings in the United States, and it still is widely viewed as the leader in accurate weather forecasts and lifesaving warnings.

Weather stations contain some type of thermometer and barometer. Other instruments measure different characteristics of the atmosphere such as wind speed, wind direction, humidity, and amount of precipitation. These instruments are placed in various locations so that they can check the atmospheric characteristics of that location. According to the WMO, weather information is collected from 15 satellites, 100 stationary buoys, 600 drifting buoys, 3,000 aircraft,

7,300 ships, and some 10,000 land-based stations. The official weather stations used by the National Weather Service is called the Automated Surface Observing System (ASOS).

Radiosondes is a balloon that measures atmospheric characteristics, such as temperature, pressure, and humidity as they move through the air. Radiosondes in flight can be tracked to obtain wind speed and direction. Radiosondes use a radio to communicate the data they collect to a computer. Radiosondes are launched from about 800 sites around the globe twice daily to provide a profile of the atmosphere.

Radar stands for Radio Detection and Ranging. A transmitter sends out radio waves that bounce off the nearest object and then return to a receiver. Weather radar can sense many characteristics of precipitation: its location, motion, intensity, and the likelihood of future precipitation. Doppler radar can also track how fast the precipitation falls. Radar can outline the structure of a storm and can be used to estimate its possible effects.

Weather satellites have been increasingly important sources of weather data since the first one was launched in 1952 and are the best way to monitor large scale systems, such as storms. At macro level, weather forecasts are commonly performed by gathering data from remote sensing satellites (Saxena et al., 2013). Satellites have the ability to take images that can project weather conditions such as maximum temperature, minimum temperature, extent of rainfall, cloud conditions, wind streams and their directions. The satellite-based systems are inherently costlier and require complete support system. Moreover, such systems are capable of providing only such information, which is usually generalized over a larger geographical area.

Moreover, satellites are able to record long-term changes, such as the amount of ice cover over the Arctic Ocean in September each year. They also observe all energy from all wavelengths in the electromagnetic spectrum. The National Weather Service relies in the Geostationary

Operational Environmental Satellites (GOES) information. There are three different types of GOES: visible, infrared, and water vapor. Visible produces light images that record storms, clouds, fires, and smog. Infrared images record clouds, water and land temperatures, and features of the ocean, such as ocean currents. The final type of GOES imagery is water vapor. This type of imagery looks at the moisture content in the upper half of the atmosphere for determining if clouds can grow to great heights. The other type of satellite commonly used in weather forecasting is called a Polar Orbiting Environmental Satellites (POES). These types of satellites fly much lower to the earth, only about 530 miles, and orbit the planet pole-to-pole.

Observations from private citizens, particularly of precipitation type and severe weather, are increasingly available through social media, platforms such as the mobile Precipitation Identification Near the Ground (mPING) initiative, and through organized efforts such as the National Weather Service's Cooperative Observer program and the Community Collaborative Rain, Hail and Snow (CoCoRaHS) network.

Data generated by radiosondes, weather stations, radars, satellites and even humans with the use of an IoT device is collected and stored in several servers. These servers contain billions of minute-to-minute observations that can be accessed today by anyone with the use of internet and other technologies. Several websites offer real-time weather data that can be utilized to keep training a machine learning model forecast.

2.2 WEATHER FORECAST METHODS

With the new advances in technology and data analytics techniques, people can be aware of extreme weather or any change in the atmosphere conditions, for example temperature, wind speed and direction, humidity, sunshine, cloudiness and precipitation. Deviations in these

conditions describe weather in terms of the state of the atmosphere at a specific location and time. According to Winston et al. (2018), weather forecasts generally follow three major steps:

1) Observation and data collection of the atmosphere, the ocean, and land surface
2) Assimilation, processing and analysis and extrapolation to predict the future state of the atmosphere.
3) The totality of observations, analysis, model and computer system constitute a forecast system.

The different methods used in weather forecasts are: Synoptic weather forecasting, Numerical methods, and Statistical methods.

i. *Synoptic weather prediction* is the traditional approach in weather prediction. Synoptic charts form the very basis of weather forecasts. Synoptic charts contain the analysis upon huge amounts of observational data of different weather elements collected from weather stations at a specific time. Meteorological centers prepare a series of synoptic charts every day to track weather changes, which forms the very basic of weather forecasts. From the careful study of weather charts over many years, certain empirical rules were formulated. These established rules have improved the forecasts estimating rate and direction of the movement of weather systems.

ii. *Numerical weather prediction (NWP)* is one of the most imperative operational tasks carried out by meteorological services around the world (Goger et al., 2016). NWP applies computer algorithms to forecast weather. Supercomputers run complex computer programs to provide predictions on many atmospheric conditions. This method involves a set of partial differential equations and other formulations describing the dynamic and thermodynamic processes of the earth's atmosphere. It is the comprising of equations, numerical approximations, parametrizations, domain settings as well as initial and boundary conditions.

iii. *Statistical weather prediction* is used along with the numerical methods. This method is based on past records of weather data and looks for factors that are good indicators for future events. The main purpose is to find out those aspects of weather that are good indicators of the future events. After establishing these relationships, correct data can be safely used to predict the future conditions. Only the overall weather can be predicted in this way. The variables defining weather conditions like temperature (maximum or minimum), relative humidity, rainfall etc., vary continuously with time, forming time series of each parameter and can be used to develop a forecasting model either statistically or using some other means like Artificial Neural Network (ANN) and Decision Tree algorithms.

2.3 LITERATURE REVIEW

Literature is replete with different data analytics techniques and statistical models for weather forecasting (Saba et al. 2014). In an attempt to find the best real-time data analytics technique to apply on weather forecasting, many researchers have applied and developed different prediction models either statistically or by using some other means like regressions, neural networks, decision trees, clustering, and other data mining techniques (Sheikh, 2016).

Shah et al. (2018) developed a research paper focused to comprehend the significance of changes in climate and atmosphere conditions. The main purpose of the paper is to predict precipitation with the use of estimated weather parameters introduced as input into machine learning techniques. Minimum temperature, maximum temperature, relative humidity and wind speed are used as input parameters in the study to predict rainfall. In order to prevent overfitting, data was split into two parts, 70% for training and 30% for testing. The machine learning techniques applied to the data were decision trees, random forest, K-nearest neighbor, neural

networks and Support Vector Regression (SVR). The prediction techniques were compared and selected based on the Root Mean Square Error (RMSE), while the machine learning technique is compared through Area Under Curve (AUC). These analyses conclude that ARIMA is the best method to predict temperatures; SVR is the best method to predict humidity and wind speed; and random forest method gives the best accuracy for rainfall prediction.

Chakrabotry et.al, (2014) suggested incremental K-mean clustering generic methodology for weather forecast. The authors demonstrated that tool clustering is used as different forecasting tools. K-means is used to group the data into clusters whose weather category has already been defined. The forecast analyzes air pollution of west Bengal. This purpose of this paper is to develop a weather forecast methodology to mitigate the impacts of air pollutions and launch focused modeling computations for prediction and forecasts of weather events. The authors have performed different experiments to evaluate the proposed approach. M. A. Kalyankar and S. J. Alaspurkar (2013) applied data mining techniques to acquire weather data and find the hidden patterns inside the large dataset. Time series, K-mean clustering and Naïve Forecast were coupled to classify and predict a weather condition. This data mining process is applied to extract knowledge from Gaza city weather dataset. This knowledge can be used to acquire valuable predictions for decision making, but dynamic data mining methods are required to build in order to learn dynamically. By creating a dynamic model, rapid changes in weather and sudden events can be forecasted.

Mahmood et al. (2019) analyzed and studied two cases of climate change in Chattisgarh, India region since climate changes happen frequently. The research proposes Cumulative Distribution Function (CDF) and analysis of complex data for weather prediction. The method proposed offers the better accuracy to climate changes and is useful for up to three-year predictions. Three cases were presented, predicted and evaluated.

These cases were analyzing abnormal high temperatures, low temperatures, and high precipitation. MATLAB was used to evaluate the metrics.

Abhishek Saxena et al. (2013) provided a review of weather forecast methodologies employed by different researchers using Artificial Neural Networks (ANN). Artificial neural network is a computer-based technique inspired in how human brains learn. This technique changes its structure based on the data that flows in the network. The research determined that an ANN is capable of predicting weather conditions like temperature, thunderstorms, rainfall and wind speed. The author concluded that Back Propagation (BP) and Multiple Linear Regression (MLR) models are suitable to forecast weather. Another researcher (Subhajini & Raj, 2010) introduced a different neural network architecture. Back propagation Algorithm, Radial Basis Function, Regression Neural Network, Optical Neural Network, and Fuzzy ARTMAP Neural Network were used to approach a weather forecasting rainfall problem. ARTMAP refers to a family of neural network architectures based on Adaptive Resonance Theory (ART) that is capable of fast, stable, on-line, unsupervised or supervised, incremental learning, classification, and prediction. The experiment was conducted with 10 years of historical data over 24 parameters (temperature, humidity, air pressure, wind speed, wind direction, cloudiness, precipitation, etc.). The study concluded that the performance of the Fuzzy ARTMAP network can give the best overall results in terms of accuracy and training. The Fuzzy ARTMAP network presented 97.45% accuracy.

Sawale and Gupta (2013) presented a neural network-based algorithm for the prediction of the atmosphere at a given location. The dataset was retrieved from the previous three years including more than 15000 instances. The weather attributes included in the model were temperature, humidity and wind speed. Back Propagation Neural Network (BPN) is used for initial

modeling. Followed by Hopfield Networks, which are fed with the result outputted by BPN model. Both models were combined effectively, the prediction error is very small and the learning process quick. This study was able to determine that the non-linear relationship that exists between weather attributes and predict what the weather will be in the future. Hemalata (2013) implemented data mining methods using a Global Positioning System (GPS) for guiding the path of the ships during sailing. The dataset is analyzed by ID3 and C4.5 classification methods to build a prediction model called Decision Tree. The attributes considered in the data were climate, temperature, humidity and stormy. Continuous attributes need to be transformed since they cannot directly fit the ID3 method. The weather report of the area traced is compared with the existing database and provides a decision notifying the ship which is the safest path based on the existing weather conditions.

A research by Kumar (2013) proposed the use of decision trees for weather forecasting. The author considered only three factors in his dataset (temperature, pressure and humidity). The study data was taken for one year. The data was divided into a training and a testing set. The training set consisted in 64 instances, while the testing set was prepared by randomly selecting 72 instances. Result shows that out of 72 test instances, 46 tests were classified properly which gave an accuracy of 63.9 percent. Results can be further improved by taking more attributes in the model and increasing the training set data.

Ali et al. (2019) implemented experiments and compared data mining techniques which include decision trees, Naive Bayes and KNN to predict different types of atmospheric dust. Data was collected from Cairo Airport reports, the variables analyzed in the data were pressure, temperature, humidity, dew point, and wind speed and direction. Data was processed using an open data science software platform that offers data preparation, machine learning, deep learning, text mining, and predictive analysis. The confusion matrix, correlation and root mean square error were used to

evaluate different models. The outcomes portrayed that the decision tree more successful in classifying and modelling data, followed by the Bayes theorem which is a classifier

Another study performed by Mekanik and Imteaz (2012), found that Australian rainfall is also affected by weather variables using an Artificial Neural Network model. Previous studies have demonstrated that rainfall is a complicated atmospheric phenomenon to predict using linear techniques. This research intends to find a nonlinear relationship between the rainfall in Victoria, Australia and the other indices affecting the region. Monthly rainfall data was obtained from January 1900 to December 2009. In this study, monthly values of NINO3, NINO4, and NINO3.4 were used as representation of ENSO. In addition to this Sea Surface Temperature and Surface Air Temperature related indices were also considered in this study. Implementing these indices in an ANN improved the model correlation up to 99%, 98% and 43% for the three case study stations of Horsham, Melbourne and Orbost, respectively.

Table 2.1: Literature review summary.

Authors	Applications	Techniques - Algorithms	Attributes	Dataset Time Period &Size	Accuracy	Advantages	Disadvantages
Shah et al. (2018)	Temperature, Rainfall, humidity and wind speed	Holt winter method, ARIMA model, Simple Moving Average model, Neural Network method, Seasonal Naive method	Max. and min. temperature, humidity and wind speed.	Daily data from 1/1/1979 to 7/31/2014	70.5	Good prediction and classification accuracy using limited parameters.	Depends on weather condition forecasts.
Chakrabotry , (2014)	Weather events Prediction (Hot/Cold, Smogy, Rainy)	K-mean Clustering	Pollution data (CO2, RPM, SO2, NOX)	10 Months	~83.3%	Good prediction accuracy and can be adapted to other prediction models.	Dynamic data mining methods required. No detailed information predicted.

Kalyankar, & Alaspurkar (2013)	Rainfall Prediction	K-mean Clustering	Temperature, humidity, rain, wind speed	8660 instances - 4 years		Good prediction accuracy and can be adapted to other prediction models.	Dynamic data mining methods required. No detailed information predicted.
Mahmood et al. (2019)	Abnormal Temperature Events	Empirical Cumulative Distribution Function (ECDF)	Temperature and other meteorological factors.	More than 3 years		Predicts weather events.	It can only predict the recurrence of event and percentage of area affected by this climate change.
Subhajini & Raj (2010)	Rainfall Prediction	Back Propagation Algorithm, Radial Basis Function, Regression Neural Network, Optical Neural Network, and Fuzzy ARTMAP Neural Network	Over 24 attributes (temperature, humidity, air pressure, wind speed, wind direction, cloudiness, precipitation, etc.)	10 years	- 97.45% on ARTMAP - 94.45% on GRNN	Overcomes the limitations of highly non-linear weight updates. ARTMAP requires fewer training epochs to converge and leads to more compact networks.	ARTMAP error tends to grow as the block size decreases.
Sawale and Gupta (2013)	Weather conditions	Back Propagation Neural Network	Temperature, wind speed, wind direction, humidity, precipitation	3 years, 15000 instances		Quick learning process and non-linear relationship.	Cannot incorporate to reflect global changes.
Hemalata (2013)	Weather prediction for ship navigation	Decision Tree	Climate, humidity, stormy, temperature	20-30 instances from 5 locations		Verifiable performance	Do not handle continuous range data directly.
Kumar (2013)	Sun, fog, rain and thunder events.	Decision Tree	Temperature, humidity and pressure.	136 instances	63.8%	Decision tree can be used in predicting dependent variables like fog and rain.	Requires several attributes and a big training data set.
Ali et al. (2019)	Atmospheric dust	Decision Tree, Naïve Bayes, KNN	Pressure, temperature, humidity, dew point, and wind speed and direction		97.45% on Decision Trees	Classifiers are effective for atmospheric dust prediction.	Assumes the attribute value on a given class is independent of the value of the other attributes.
Mekanik and Imteaz (2012)	Rainfall Forecasting	Neural Network	Sea surface temperature, Surface air temperature, Nino 3 and 4	19 years	~ 99%,	Able to provide higher non-linear correlations.	Not accurate for all locations tested

2.4 CHAPTER CONCLUSION

This chapter provided an overview of previous research in weather forecasting using data mining techniques. Together, the development of machine learning, data mining techniques and computers produce complex models that more accurately represent the conditions of the atmosphere. Also, the literature table reflects the factors, weather conditions and trends that influence each of the previous weather forecast models developed. The most common weather conditions or parameters analyzed on previous studies are temperature, humidity, cloudiness, rainfall, wind speed, wind direction, pressure and dew point. The confusion matrix, correlation and root mean square error (RMSE) are used to evaluate different models. It is evident that forecasting is often burdened with numerous problems such as inadequate data, environmental degradation and/or limited computer knowledge. The weather dataset is highly non-linear so different data mining techniques such as Naïve Bayes, Artificial Neural Network (ANN) and Decision Tree algorithms proves to have high weather forecasting accuracy.

Chapter 3: Technical Approach

This chapter is dedicated to introducing the methodology that will be used to collect, pre-process, and mining of data for weather condition predictions. Despite many weather conditions follow a linear trend, rainfall data is non-linear in nature (Agilan & Nanduri, 2016). Amount, frequency, and intensity are three main characteristics of rainfall time series. Weather conditions vary from place to place, day to day, month to month and year to year (Darji et al., 2015). On this study, we are focusing on predicting weather conditions for the next 365 days based on El Reno, Oklahoma. Each concept in the proposed architecture is explained in this chapter. Formulas and procedures used by JMP for each predictive and classification data mining model are explained. Finally, our proposed approach introduces the predictive and classification data mining models using JMP. The chapter describes the different evaluation methodologies that are going to be used to compare the forecasts and data mining techniques.

3.1 Data Collection

Today there are several websites that allow people to access weather observations or data from millions of global locations (weather stations, satellites, radars, etc.). Weather APIs are Application Programming Interfaces that allow anyone to connect to large databases of weather to retrieve real-time and historical information. Multiple mobile applications have been developed with the use of APIs and smartphones to provide hour-by-hour forecasts, severe weather alerts, and relevant weather information for any place in the world. Several APIs were examined to ensure they satisfy all the requirements.

The National Weather Service API (https://api.weather.gov) allows developers access to critical forecasts, alerts, and observations, along with other weather data. The API was designed

with a cache-friendly approach that expires content based upon the information life cycle. The API is based upon of JSON-LD to promote machine data discovery.

The OpenWeatherMap API (https://openweathermap.org) currently provides a wide variety of weather data including (but not limited to) current weather, forecasts, historical, weather stations, and weather alerts. This API contains information on different weather conditions such as temperature, pressure, wind, clouds, rain, snow, and air pollution. This API contains current and historical air pollution data such as carbon monoxide (CO), ozone (O3), sulfur dioxide (SO2), and nitrogen dioxide (NO2). The service allows you to regularly download current weather and forecast data in JSON format.

The Weatherbit API (https://www.weatherbit.io) delivers basic access to the Weatherbit.io Weather API. With just latitude and longitude coordinates, you can get weather forecast data returned in JSON format. The Weatherbit weather API, while not completely free, does offer a free basic plan that allows developers 150 requests/day.

The AccuWeather API (https://developer.accuweather.com) provides in depth current, historical, and forecasted weather data for any place in the world. AccuWeather is a good option for developers who are looking to build a wide range of innovative and engaging weather data applications. This API returns even information about flight delays, mosquito activity, stargazing, and dozens of other daily index values for a specific location. The API documentation is nicely designed, comprehensive, and includes interactive documentation to try out API endpoints and see the responses.

The Dark Sky API (https://darksky.net) lets developers to access Dark Sky's weather data through the API. This provides current weather conditions, forecasts by minute or hours, and severe weather alerts in the United States, Canada, and European Union nations. This API offers

developers two types of requests to select, Forecast and Time Machine. The Forecast Request contains the current weather forecast for the next week, and the Time Machine Request contains weather conditions (observed or forecast) for a given date. The Dark Sky API offers a full collection of meteorological conditions in 39 different languages, including: temperature, atmospheric pressure, cloud cover, dew point, humidity, liquid precipitation rate, moon phase, nearest storm distance, nearest storm direction, ozone, precipitation type, snowfall, sun rise/set, temperature, text summaries, UV index, wind gust, speed and direction.

While many of the weather observations delivered by these APIs are similar, there are differences in the days and time formats for weather forecasts, the number of years back for historical data, and the types of weather information provided. According to Fahey (2016), Dark Sky and AccuWeather seem to be the best weather apps at forecasting exact probabilities for rainfall. Dark Sky contains all the weather conditions that have been analyzed in previous research. Also, it is one of the most accurate sources of hyperlocal weather information.

Dark Sky API

Data is collected using an API named Dark Sky, which returns data in a JSON format. Each data observation contains various attributes, each representing the average (unless otherwise specified) of a specific weather phenomenon occurring during a given period. Data was retrieved into two datasets, one pulling observations by hour and the other by day. The attributes collected are shown in Table 3.1.

Table 3.1: Dark Sky API collected attribute description.

Attribute	Dataset	Unit	Data Type	Description
Apparent Temperature	Hourly	Celsius	Continuous	The apparent temperature in degrees Celsius. Temperature is a measure of how hot or cold the air is.
Apparent Temperature High	Daily	Celsius	Continuous	The daytime high apparent temperature.
Apparent Temperature High Time	Daily	UNIX time	Continuous	The UNIX time representing when the daytime high apparent temperature occurs.
Apparent Temperature Low	Daily	Celsius	Continuous	The overnight low apparent temperature.
Apparent Temperature Low Time	Daily	UNIX time	Continuous	The UNIX time representing when the overnight low apparent temperature occurs.
Apparent Temperature Max	Daily	Celsius	Continuous	The maximum apparent temperature during a given date.
Apparent Temperature Max Time	Daily	UNIX time	Continuous	The UNIX time representing when the maximum apparent temperature during a given date occurs.
Apparent Temperature Min	Daily	Celsius	Continuous	The minimum apparent temperature during a given date.
Apparent Temperature Min Time	Daily	UNIX time	Continuous	The UNIX time representing when the minimum apparent temperature during a given date occurs.
Cloud Cover	Hourly, Daily	Percentage	Continuous	The percentage of sky occluded by clouds, between 0 and 1, inclusive.
Dew Point	Hourly, Daily	Celsius	Continuous	The dew point in degrees Celsius.
Humidity	Hourly, Daily	Percentage	Continuous	The relative humidity, between 0 and 1, inclusive.
Icon	Daily	Text summary	Nominal	A machine-readable text summary of this data point, suitable for selecting an icon for display. If defined, this property will have one of the following values: clear-day, clear-night, rain, snow, sleet, wind, fog, cloudy, partly-cloudy-day, or partly-cloudy-night. (Developers should ensure that a sensible default is defined, as additional values, such as hail, thunderstorm, or tornado, may be defined in the future.)

Moon Phase	Daily	Fraction	Continuous	The fractional part of the lunation number during the given day: a value of 0 corresponds to a new moon, 0.25 to a first quarter moon, 0.5 to a full moon, and 0.75 to a last quarter moon. (The ranges in between these represent waxing crescent, waxing gibbous, waning gibbous, and waning crescent moons, respectively.)
Ozone	Hourly, Daily	Dobson	Continuous	The columnar density of total atmospheric ozone at the given time in Dobson units.
Precipitation Accumulation	Hourly, Daily	Inches (in)	Continuous	The amount of snowfall accumulation expected to occur (over the hour or day, respectively), in inches. (If no snowfall is expected, this property will not be defined.)
Precipitation Intensity	Hourly, Daily	Inches (in)	Continuous	The intensity (in inches of liquid water per hour) of precipitation occurring at the given time. This value is conditional on probability (that is, assuming any precipitation occurs at all).
Precipitation Intensity Max	Daily	Inches (in)	Continuous	The maximum value of precipitation intensity during a given day.
Precipitation Intensity Max Time	Daily	UNIX time	Continuous	The UNIX time of when the maximum precipitation intensity occurs during a given day.
Precipitation Probability	Hourly, Daily	Probability (0,1)	Continuous	The probability of precipitation occurring, between 0 and 1, inclusive.
Precipitation Type	Hourly, Daily	Type (clear, rain, snow, sleet)	Continuous	The type of precipitation occurring at the given time. If defined, this property will have one of the following values: "rain", "snow", or "sleet" (which refers to each of freezing rain, ice pellets, and “wintery mix”). (If precipitation intensity is zero, then this property will not be defined.
Pressure	Hourly, Daily	Millibars (hPa)	Continuous	The sea-level air pressure in milibars
Sunrise Time	Daily	UNIX time	Continuous	The UNIX time of when the sun will rise during a given day.
Sunset Time	Daily	UNIX time	Continuous	The UNIX time of when the sun will set during a given day.
Temperature	Hourly	Celsius	Continuous	The air temperature in degrees Celsius.
Temperature High	Daily	Celsius	Continuous	The daytime high temperature.

Temperature High Time	Daily	UNIX time	Continuous	The UNIX time representing when the daytime high temperature occurs.
Temperature Low	Daily	Celsius	Continuous	The overnight low temperature.
Temperature Low Time	Daily	UNIX time	Continuous	The UNIX time representing when the overnight low temperature occurs.
Temperature Max	Daily	Celsius	Continuous	The maximum temperature during a given date.
Temperature Max Time	Daily	UNIX time	Continuous	The UNIX time representing when the maximum temperature during a given date occurs.
Temperature Min	Daily	Celsius	Continuous	The minimum temperature during a given date.
Temperature Min Time	Daily	UNIX time	Continuous	The UNIX time representing when the minimum temperature during a given date occurs.
Time	Hourly, Daily	UNIX time	ID	The UNIX time at which this data point begins. minutely data point are always aligned to the top of the minute, hourly data point objects to the top of the hour, daily data point objects to midnight of the day, and currently data point object to the point of time provided all according to the local time zone.
UV Index	Hourly, Daily	UV index	Continuous	The UV Index is a linear scale, with higher values representing a greater risk of sunburn (which is correlated with other health risks) due to UV exposure.
UV Index Time	Daily	UNIX time	Continuous	The UNIX time of when the maximum uvIndex occurs during a given day.
Visibility	Hourly, Daily	Kilometers (km)	Continuous	The average visibility in kilometers (capped at 16 km).
Wind Bearing	Hourly, Daily	Degrees	Continuous	The direction that the wind is coming from in degrees, with true north at 0° and progressing clockwise. (If wind Speed is zero, then this value will not be defined.)
Wind Gust	Hourly, Daily	Meters/second	Continuous	The wind gust speed in meters per sec.
Wind Gust Time	Daily	UNIX time	Nominal	The time at which the maximum wind gust speed occurs during the day.
Wind Speed	Hourly, Daily	Meters/second	Continuous	The wind speed in meters per second.

Dark Sky API requests were placed through the python script shown in Figure 3.1. The inputs required to run the script are shown at the user settings line 14 of the script. GPS coordinates

refers to the latitude and longitude of the desired location. Latitude is used to express how far north or south a location is relative to the equator. The latitude in the equator is zero. Longitude represents the location in an east-west direction, relative to the Greenwich meridian. The latitude and longitude for El Reno is 35.53227 and -97.95505, respectively. Dark Sky API key is unique access key provided to each user by Dark Sky API. The API time defines a specific date in the past in UNIX timestamp when we want to finish collecting observations (i.e. 1/1/2020 0:00:00 GMT is 1577836800 in UNIX timestamp). Finally, the number of days you want to collect is required.

```python
# -*- coding: utf-8 -*-
"""
Created on Tue Feb 25 23:30:33 2020
You can make an API call to Dark Sky by typing in a URL into your browser in the following format:

https://api.darksky.net/forecast/[key]/[latitude],[longitude],[time]
@author: Pedro Marquez
"""

from urllib.request import urlopen
import json
import pandas as pd
import os
# --------- User Settings ---------
# Enter to https://darksky.net/dev/account and enter account
CITY = "El Reno"
GPS_COORDS = "35.53227,-97.95505"
DARKSKY_API_KEY = "96069efd2618f8281XXXXXXXXXXXX"
API_TIME = "1586667600" #UNIX TIMESTAMP - https://www.unixtimestamp.com/index.php
n = 250 # number of iterations, days?

# ---------------------------------


# Pull data function
def get_current_conditions(xtime):
    api_conditions_url = "https://api.darksky.net/forecast/" + DARKSKY_API_KEY + "/" + GPS_COORDS +','+ xtime +"?units=si"
    try:
        f = urlopen(api_conditions_url)
    except:
        return []
    json_currently = f.read()
    f.close()
    return json.loads(json_currently)


hourlyData = pd.DataFrame()
dailyData = pd.DataFrame()

while n > 0:
    cond = get_current_conditions(API_TIME)
    for i in range(len(cond['hourly']['data'])):
        hourlyData = hourlyData.append(cond['hourly']['data'][i], ignore_index = True)
    for j in range(len(cond['daily']['data'])):
        dailyData = dailyData.append(cond['daily']['data'][j], ignore_index = True)
    #Loop
    API_TIME = str(int(API_TIME) - 86400) #3600 means take 1 hour off
    n -= 1
```

Figure 3.1: Python script used to retrieve data from Dark Sky API.

Dark Sky API provides a wide and reliable data input that allows us to forecast weather. All weather conditions that were used in forecasting methods shown in the literature review can be retrieved from this data set. Daily data was collected for nine years (2011 – 2019) using Dark Sky API. The datasets retrieved from this API contains all the required data and is already in a structured form. Despite overcoming the challenge of structuring data, more pre-processing phases are required to improve the quality of our predictions.

3.2 PROPOSED ARCHITECTURE

In the first part of the proposed model, the use of Application Programming Interfaces (API) is considered to retrieve the data. Several weather API are available to access weather observations. Several APIs were analyzed and evaluated. Data is retrieved from global weather site provided by Dark Sky API which provides reliable and significant data. Daily data from 1/1/2011 to 12/31/2019 is collected from El Reno, Oklahoma.

After collecting and structuring the data, it is important to follow some data pre-processing procedures to ensure quality. These procedures are listed below:

- Data Consolidation and Integration - Consolidate homogeneously the data collected from all sources in order to form a single structured dataset.
- Data transformation –Reformat, normalize, aggregate, and even update data using regulatory standards.
- Data reduction - Minimize the amount of data so that it becomes non-redundant.
- Data discretization - Extract and segregate data into intervals so that it can be efficiently utilized within available mining algorithm and techniques.
- Data Cleansing - Identify and remove (or correct) inaccurate observations from a dataset.

Moreover, data is partitioned into two parts to prevent over-fitting: 70% for training and 30% for testing. This method is called cross-validation. Cross validation allows us to estimate model's performance on unseen data.

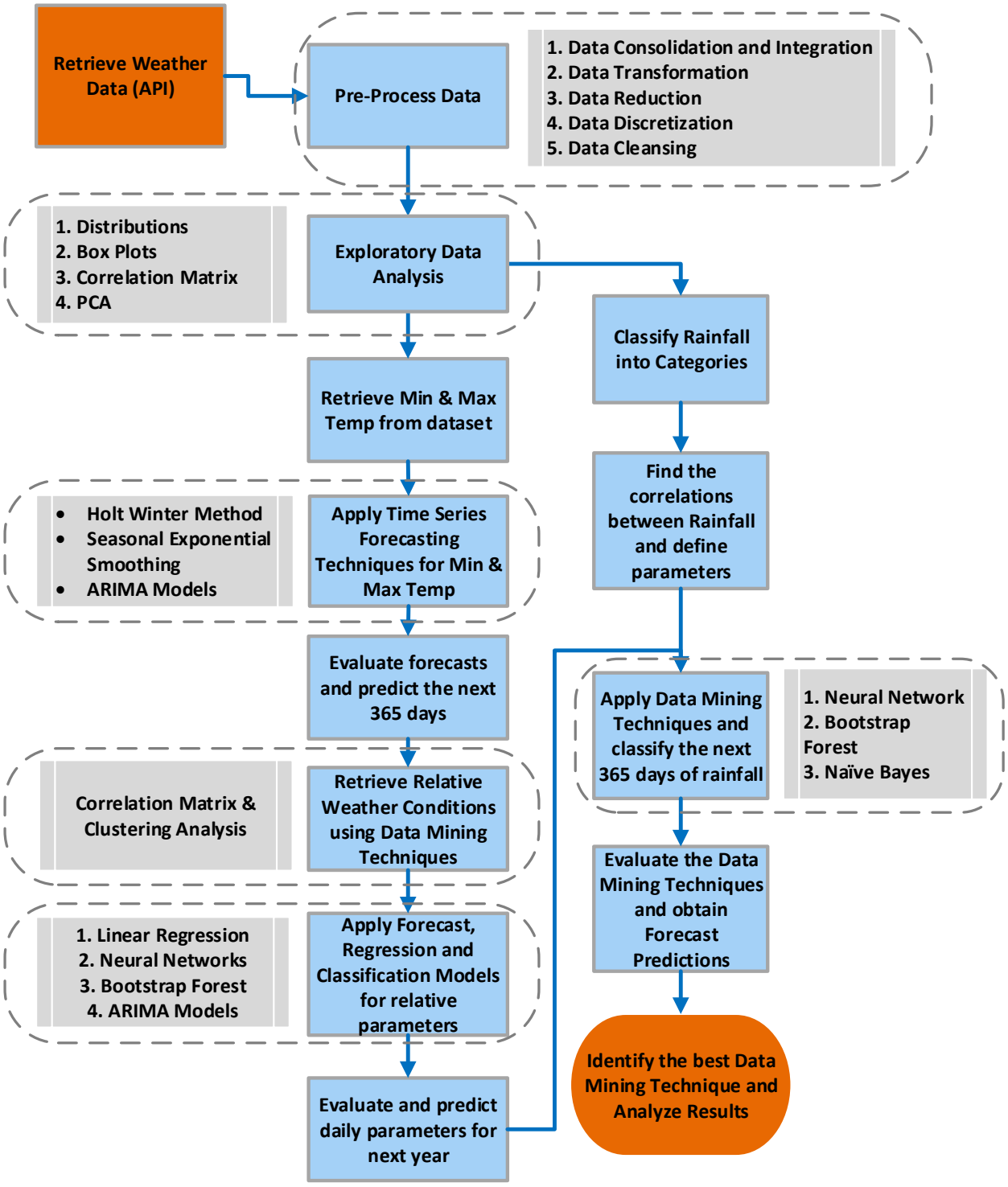


Figure 3.2: Technical proposed approach process diagram.

The second part of the model consists in data exploration. Data exploratory analysis is an approach that analysts use to visualize data and make assumptions. Several Exploratory Data Analysis (EDA) techniques are applied to the collected data in order to find the different weather

condition distributions, variances and correlations. These techniques are box plots, distribution graphs, correlation matrices and a Principal Component Analysis (PCA). PCA and correlations would allow us to identify which weather conditions affect more precipitation. Also, we look at time series graphs for each parameter to determine if they follow a linear trend and consequently, utilize time series forecasting methods to see the behavior of the data. On the other hand, non-linear trend weather conditions must use classification models such as Neural Networks, Bootstrap Forest or Naïve Bayes.

The next step will be to fit, evaluate and compare different time series forecasting methods for minimum and maximum temperature. The methods selected for predicting linear-trend conditions are Holts Winter Method, Seasonal Exponential Smoothing, ARIMA and Seasonal ARIMA. A detailed analysis will be performed in order to compare these methods and find the best-fitted model based on statistic performance measurements. The model performance measurement selected for temperature forecasts are Akaike Information Criterion (AIC). Subsequently, the rest of the weather conditions are retrieved from the temperature forecasted conditions using data mining techniques. Some weather conditions such as UV Index and dew point follow a linear trend, therefore time series forecasts are also being evaluated and compared to data mining techniques. The best model is chosen by using R Square as the performance measurement.

Finally, three different data mining techniques like Neural Networks, Bootstrap Forest, and Naïve Bayes were applied on the partitioned data (training and validation) in order to predict rainfall, the individual results were also analyzed and tuned. The detailed analysis of the best-fitted model and comparison of all methods based on performance is done for rainfall, taking

misclassification rate as the performance measurement. The forecast of the whole year is built introducing the forecasted weather conditions into the selected model.

3.3 DATA PRE-PROCESSING

Data Pre-Processing at the early stage of any big data system's lifecycle improves and refines the quality of the data (Juneja & Das, 2019). Data Pre-Processing enhances the accuracy, efficiency, and scalability of the classification and prediction models by applying the following sub-processes: data consolidation and integration, data transformation, data reduction, data discretization, and data cleansing.

Data Consolidation and Integration

Data may be retrieved from several locations that be may be in different forms: structured, semi-structured or unstructured. Data from all these sources requires to be consolidated homogeneously to form a single structured dataset to be used in the weather prediction. Data was retrieved from the same API, but different tables. These tables contain data collection by different times, one by hour and the other by day. The significant information from the hourly table was compiled into the daily dataset.

Data Transformation

Data transformation refers to the conversion of data from one format to another. Data transformation is an important part of most data integration and data management tasks. This process is used when data needs to be reformatted, normalized, aggregated, even updated using regulatory standards. The data may be transformed by normalization, particularly when neural networks or methods involving distance measurements are used in the learning step. Normalization is the process of scaling data attributes to fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0 (Sawale & Gupta, 2013). The retrieved data units for temperature were transformed

from Celsius to Fahrenheit to remove negative values, moreover, the time column was changed from UNIX timestamps into dates.

Data Reduction

Data reduction is a process that intends to minimize the amount of data so that it becomes non-redundant. It allows a more efficient data storage and a cost reduction by removing data that is not relevant while preserving only the significant parts for the prediction models. Duplicated values and irrelevant observations appeared during data collection and were discarded. Relevance analysis can be applied to detect attributes that do not contribute to the classification or prediction models and possibly mislead its learning step (Sawale & Gupta, 2013). These analyses were applied after obtaining the results of each prediction model.

Data Discretization

Data discretization is a process that extracts and segregates data into intervals so that it can be efficiently utilized within available mining algorithm and techniques. This process segregates values into intervals so that there are a limited number of possible circumstances (Thomas, 2018). The intervals themselves are treated as ordered and discrete values. Any attribute can be discretized regardless of whether its data is numeric and string. Precipitation intensity column was classified into two categories: rain and no rain. Rain is represented with a 1 and no rain with a 0. The intervals selected for each category are shown in Table 3.2.

Table 3.2: Rainfall data discretization into occurrence.

Grade	Rainfall Amount (mm)
No rain (0)	**0**
Rain (1)	**>0**

Data Cleansing

Data cleansing is the process of recognizing unfinished, unreliable, inaccurate or non-relevant parts of the data and restores, remodels or removes the dirty or crude data. There are four

critical steps for data cleaning: removing of unwanted observations, fixing structural errors, managing outliers, and handling missing data. Data cleansing is an essential part in data mining techniques that involve machine learning (Thomas, 2018). Machine learning (ML) is all about training and feeding data to algorithms to perform various compute intensive tasks.

1) The removal of unwanted observations consists of scrubbing duplicate, redundant or irrelevant information from the dataset.
2) Structural errors are errors that appear during measurement, transfer of data or other similar steps. This step requires to fix typos in the name of features, same attribute with different name, mislabeled classes, etc.
3) Detecting and managing outliers can be very informative about the subject and data collection process. Outliers are unusual values in your dataset, and they can distort statistical analyses and their assumptions. Deciding how to handle outliers depends on investigating their underlying cause. There are three causes for outliers: data entry or measurement errors, sampling problems and unusual conditions, and natural variation. Outliers increase the variability in your data, which sometimes decreases statistical power. Consequently, excluding outliers can cause your results to become statistically significant with certain types of models. Perhaps, linear regression models are less robust to outliers than classification models. For example, outliers can be a strong indicator in where to classify that observation in decision trees. Outliers were explored through quantile ranges. No values were found farther than 90% quantile range from the tail quantile at a 10% of significance level.

Table 3.3: Weather conditions outliers at a 10% significance level.

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
cloudCover	0.01	0.79	-2.33	3.13	0	
dewPoint	-4.549	20.629	-80.083	96.163	0	
humidity	0.47	0.85	-0.67	1.99	0	
ozone	263.24	324.52	79.4	508.36	0	
pressure	1007.8	1025.6	954.4	1079	0	
temperatureMax	45.5216	95.2862	-103.77	244.58	0	
temperatureMin	25.9016	72.428	-113.68	212.007	0	
uvIndex	3	11	-21	35	0	
visibility	12.5217	16.093	1.8078	26.8069	0	
windBearing	22	331	-905	1258	0	
windGust	5.221	14.44	-22.436	42.097	0	
windSpeed	2.01	6.479	-11.397	19.886	0	

4) Missing data requires to be handled carefully. Observations missing data are commonly handled in two different ways. One way is to delete observations with missing information. The other option is to predict the missing values based on previous observations. 'Precipitation Intensity' had some missing values where no precipitation occurred. These values where set to zero. Similarly, precipitation accumulation had missing values, but no data appeared for some heavily rainy days. The decision was to discard all the information of this parameter since it appears to be unreliable and there is not too much data to predict missing observations.

The purpose of this pre-processing framework is to overcome the challenge of how to tap and mine efficiently big data in a qualitative manner by following a standardized quality process that also cleanses and improves the data quality as part of the Big Data life cycle for weather prediction. To ensure quality of data, data may be assessed and transformed through numerous iterations to create a more reliable dataset.

3.4 EXPLORATORY DATA ANALYSIS (EDA)

In data mining, Exploratory Data Analysis (EDA) is an approach to analyzing datasets, typically by visual techniques with the purpose of maximizing insight. There are multiple

applications for EDA such as extracting important variables, testing assumptions, detecting outliers, determine optimal factor settings, etc. Most EDA techniques are visually representative in nature with a few quantitative techniques (Camacho et al, 2017). EDA techniques consist of various techniques plotting raw data, simple statistics and positioning of plots to maximize pattern recognitions. The EDA techniques used in this approach are histograms, simple statistics, boxplots, time series graphs and a PCA.

Explore Distributions and Outliers

Exploring and understanding outliers in the raw data is an important part of analysis. Outliers in data can be attributed to mistakes in data collection or reporting, measurement systems failure, or the inclusion of error or missing value codes in the data set. The presence of outliers can distort estimates. Therefore, any analyses that are conducted are biased toward those outliers. Outliers also inflate the sample variance. Sometimes retaining outliers in data is necessary, however, and removing them could underestimate the sample variance and bias the data in the opposite direction. Whether outliers are removed or retained, they must be located. There are many ways to visually inspect for outliers. Quantile ranges in our box plots are used to identify, explore, and manage outliers in our weather data. Box plots, histograms, and scatter plots can sometimes easily display these extreme values. Histograms and boxplots for each weather condition is shown in Figure 3.3.

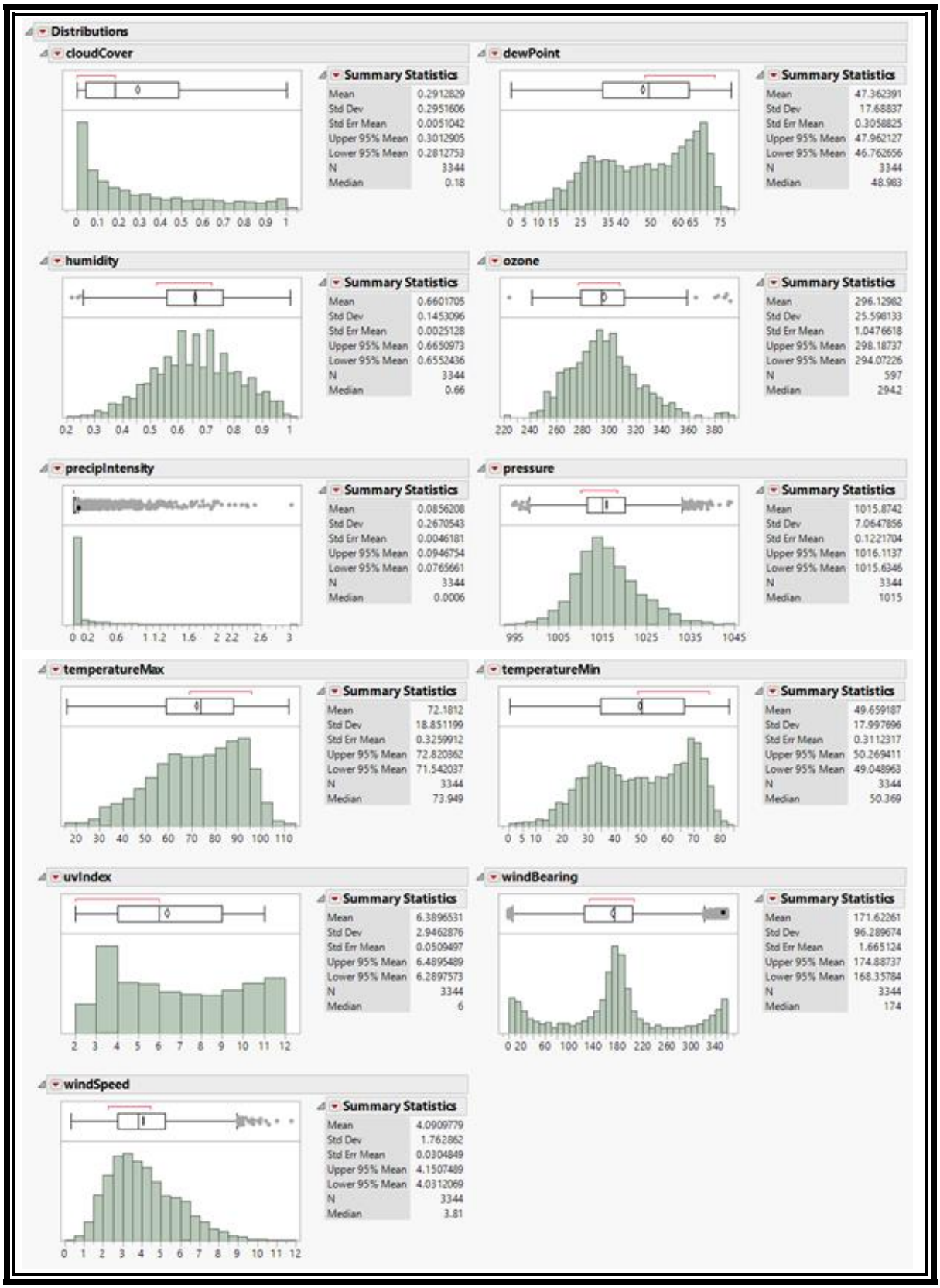


Figure 3.3: Box plots, distribution charts and summary statistics for all weather conditions.

All these weather conditions follow a normal distribution, except for cloud cover, precipitation intensity and wind bearing. Cloud cover describes the percentage of clouds in the sky, so data is already normalized to values from 0 to 1. Cloud cover data shows a right-skewed distribution, so we can infer that there is a higher probability that the sky at El Reno is clear. There is no reason to remove outliers in precipitation intensity since it follows an exponential distribution and these high values are significantly representative of extreme storms. Wind bearing distribution seems to follow a bimodal trend since we need to assume that 360° is the same as 0°. From this bimodal distribution we can say that wind direction tends to go towards east (0°) and west (180°).

Correlation Matrix

A correlation matrix is an EDA tool that shows correlation coefficients between variables. This tool is used as a diagnostic for advanced analysis. This matrix is symmetrical, which means that the same variables are shown in the rows and columns. The Pearson r correlation is used to measure how two variables are related to each other. For the Pearson r correlation, both variables should be normally distributed (normally distributed variables have a bell-shaped curve). The following formula is applied in JMP to calculate the Pearson r correlation:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}},$$

Where:

- $r_{xy} = Pearson\ r\ correlation\ coefficient\ between\ x\ and\ y$
- *n = number of observations*
- $x_i = value\ of\ x\ (for\ ith\ observation)$
- $y_i = value\ of\ y\ (for\ ith\ observation)$

Since precipitation intensity and cloud cover do not follow a normal distribution, another test called Spearman ρ (rho) is used to calculate the correlation coefficients. This formula does not

depend on any underlying assumption of normality. For a sample of size *n*, the *n* raw scores X_i, Y_i are converted to ranks rg (Xi), rg (Yi), and r_s is calculated as:

$$r_s = \rho_{rgX,rgY} \frac{cov(rg_X, rg_Y)}{\sigma_{rgX}, \sigma_{rgY}}$$

Where

- ρ denotes the usual Pearson correlation coefficient.
- $cov\ (rg_X, rg_Y)$ is the covariance of the rank variables.
- σ_{rgx} and σ_{rgy} are the standard deviations of the rank variables.

Table 3.4: Weather conditions correlation matrix.

	Cloud Cover	***Dew Point***	***Hu-midity***	***Ozone***	***Pres-sure***	***Max. Temp.***	***Min. Temp.***	***UV Index***	***Wind Bearing***	***Wind Speed***	***Precip. Inten-sity***
Cloud Cover	1	-0.016	0.686	-0.147	0.053	-0.394	-0.134	-0.390	-0.185	0.095	0.379
Dew Point	-0.016	1	0.254	-0.070	-0.589	0.859	0.947	0.674	-0.165	-0.035	0.165
Humidity	0.686	0.254	1	-0.218	-0.031	-0.238	-0.005	-0.256	-0.159	-0.085	0.377
Ozone	-0.147	-0.070	-0.218	1	-0.064	0.015	-0.044	0.094	0.204	-0.036	0.032
Pressure	0.053	-0.589	-0.031	-0.064	1	-0.583	-0.590	-0.400	-0.064	-0.303	-0.106
Max. Temp.	-0.394	0.859	-0.238	0.015	-0.583	1	0.918	0.792	-0.062	-0.034	-0.016
Min. Temp.	-0.134	0.947	-0.005	-0.044	-0.590	0.918	1	0.742	-0.133	0.012	0.095
UV Index	-0.390	0.674	-0.256	0.094	-0.400	0.792	0.742	1	-0.082	-0.060	-0.035
Wind Bearing	-0.185	-0.165	-0.159	0.204	-0.064	-0.062	-0.133	-0.082	1	0.144	-0.110
Wind Speed	0.095	-0.035	-0.085	-0.036	-0.303	-0.034	0.012	-0.060	0.144	1	0.032
Precip. Intensity	0.379	0.165	0.377	0.032	-0.106	-0.016	0.095	-0.035	-0.110	0.032	1

The integrated correlation matrix for all weather conditions is shown in Table 3.4. The numbers in the cells are in red or blue colors. Red represents a negative significant correlation and blue a positive significant correlation between variables. The greater positive correlation coefficients are among the temperatures, dew point and UV index. On the other side, pressure and cloud cover seems to be negatively correlated to maximum temperature, dew point and UV index.

Constructing a Principal Component Analysis and Cluster Model

Camacho et al. (2017) proposed a new framework for matrix factorization based on principal component analysis (PCA) where sparsity is imposed. The structure to execute sparsity is defined in terms of groups of correlated variables found in correlation matrices or maps. For this EDA we are adopting a similar framework in order to define which weather conditions are related and group them through clustering in order to find a sequential order to perform forecasts for each weather condition. The first weather parameter would be temperature, since it follows a linear trend. Variables with a linear trend can be predicted using time series forecasts. Time series forecasts do not require any related parameters to be considered in the models. Time series forecasting fits a model with the use of historical data to predict future observations.

Principal Component Analysis (PCA) is one of the EDA techniques commonly used to expose relationships among variables. It represents the variation in a set of variables in the matter of a minimum amount of independent linear combinations. PCA generates a visualization for multi-dimensional data that permits to see the arrangement of observations across many correlated variables. PCA can be based on either the covariance matrix or the correlation matrix (Jolliffe and Cadima, 2016).

The goal of PCA is to uncover the subspace of maximum variance in the dimensional variable space. In order to achieve this, principal components (PCs) need to be found. PCs are linear transformations of the original variables, which are orthogonal and explain reducing amounts of variance in the data. Using JMP, *p* principal components are created for *p* variables. A PCA was applied to two-way datasets, where 10 variables were measured for 9 years of daily observations.

Each principal component is calculated by taking a linear combination of an eigenvector of the correlation matrix with the variables. The eigenvalues represent the variance of each component. The created principal components are ranked from 1 to *p* based on the greatest possible variance across the linear combination of the standardized variables. The first principal component represents the linear combination of the variables that has the greatest possible variance and is uncorrelated with all previously defined components. Each subsequent principal component is the linear combination of the variables that has the greatest possible variance and is uncorrelated with all previously defined components.

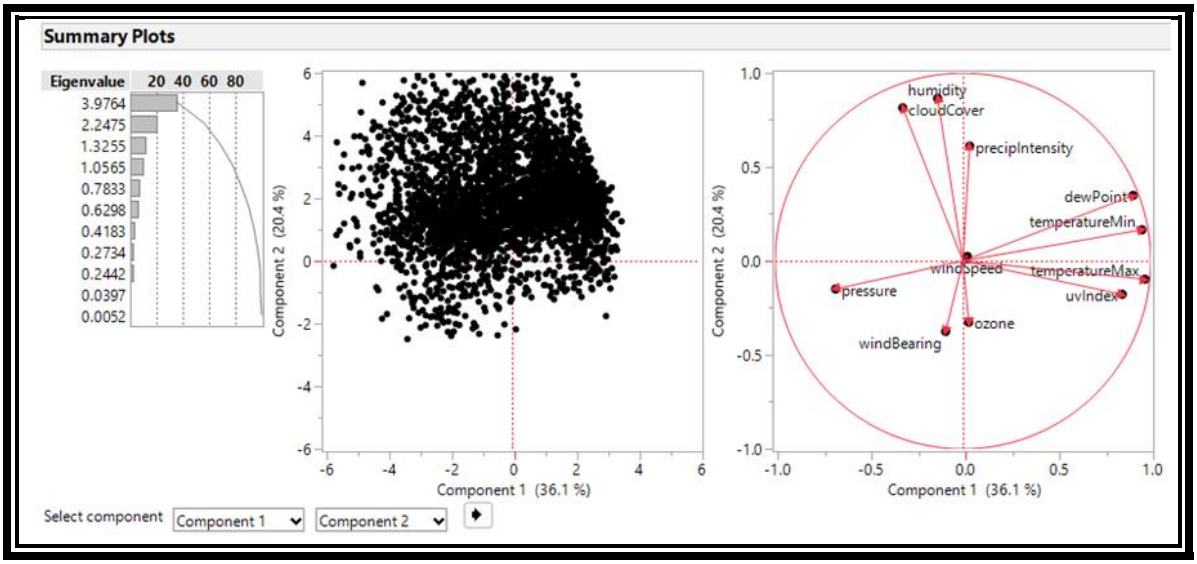


Figure 3.4: JMP Principal Component Analysis Report for all weather conditions.

Figure 3.4 shows the JMP output report after running the data on the PCA platform. This report provides the eigenvalues and a Pareto chart of the variation accounted by each principal component. In addition, a Score Plot and a Loadings Plot are given as well. The eigenvalues represent a partition of the total variation in the multivariate sample. They indicate the total number of components extracted based on the amount of variance contributed by each component. Eigenvalues are scaled to sum to the number of variables. The Pareto chart reflects that almost eighty percent of variance is reflected across the first four principal components. A PCA scatterplot

matrix containing these four PCs is shown in Figure 3.5. The PCA scatterplot matrix displays a matrix of score and loading plots in one frame for a defined number of principal components. The score plots are represented with a yellow shaded background and the loading plots are in a blue shaded background.

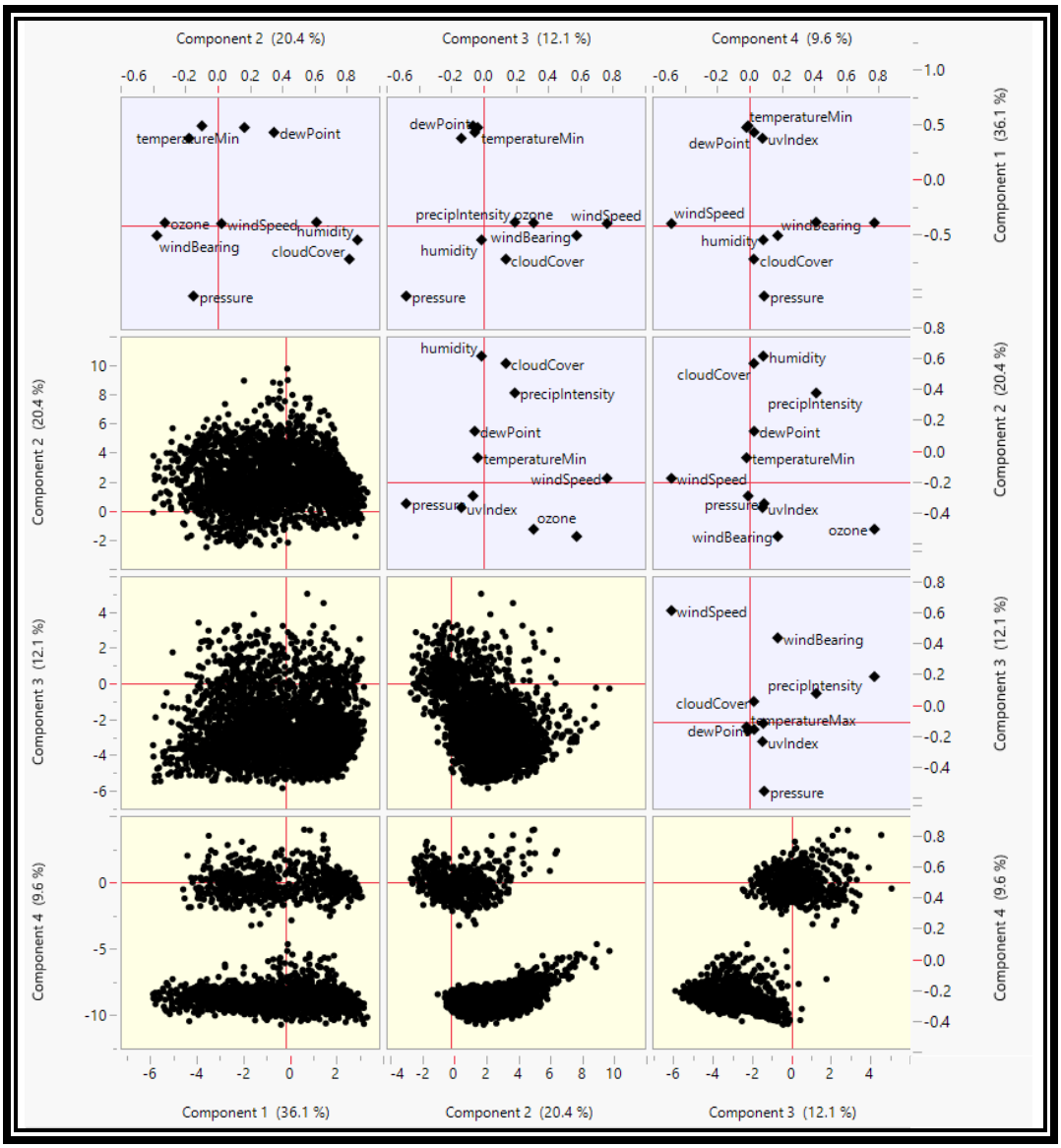


Figure 3.5: Principal Component Scatterplot Matrix.

After finishing the correlation matrix and PCA, a cluster analysis is completed on the variables by splitting the variables into non-overlapping clusters. Variable clustering offers a method for grouping related variables. Each cluster can then be represented by single or multiple

components. The component is a linear combination of all variables in the cluster. Figure 3.6 presents the Color Map on Correlations report for our data using JMP. The variables are organized in the same order as in the Cluster Members output report. This arrangement ensures that members of the same cluster are adjacent in the correlation plot. Variables in the same cluster incline to have a greater correlation than variables in different clusters. Cells are colored in red and blue colors to depict the correlation; the deeper colored cells represent a greater correlation. For that reason, a strong red color stands the diagonal in our colored map since it corresponds to a correlation within the same variable.

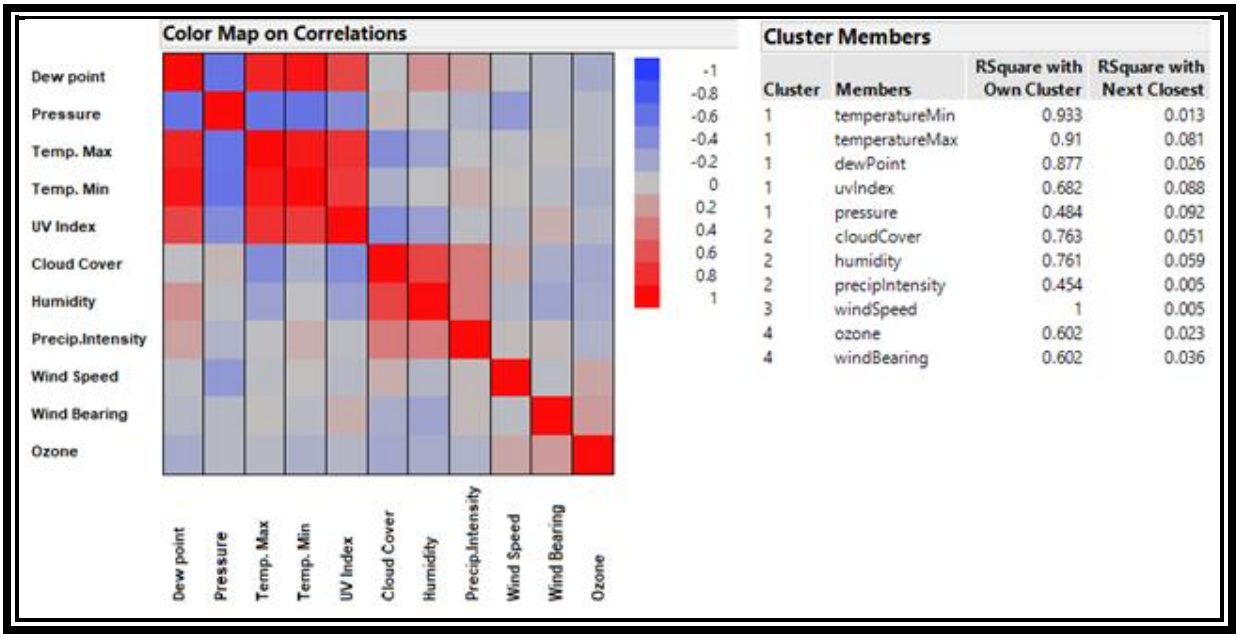


Figure 3.6: JMP Color Map on correlations report

The Color Map on correlations report shown in Figure 3.6 provides a more concise version of the information in the Correlation Matrix. The variables in top left corner of the plot has dark colors indicating that most of the variables are highly correlated. Cloud cover, Humidity and precipitation intensity seem to have a positive correlation among them. Wind bearing, wind speed and ozone have light colors across the matrix, indicating they are not very correlated with other variables. The table at the right of Figure 3.6 displays the cluster members and the R-Square

measurements with respect to its own cluster centroid. The first cluster is composed of dew point, UV index, pressure, maximum and minimum temperature. The second is represented by cloud cover, humidity and precipitation intensity. Based on this Color Map on correlations, we can conclude that the variables that impact precipitation intensity directly or indirectly are the variables in this two clusters. Also, the high correlations of temperatures with UV index, dew point and pressure demonstrate that regressions or data mining techniques can be used to retrieve predictions from temperature forecasts. Ozone, wind bearing, and wind speed variables lack of strong correlations with the rest of the variables, therefore, these three parameters will not be considered in this study.

3.5 TIME SERIES FORECASTS

A time-series is a set of sequential data observations over a certain successive time duration. Observations that are close together in time are typically correlated. This methodology predicts outcomes in the future by calculating the dependence between observations. The most common characteristics in time series are seasonality, trend and autocorrelation. Trend refers to long term movements of a series, such as gradual increases or decreases of values across time. Seasonality indicates the patterns that occur over a known period. Autocorrelation is the degree to which each point in a series is correlated with earlier values in the series (SAS Institute Inc., 2018).

There are many different models and forecasting methods available in the JMP Time Series platform. However, not all methods consider trend or seasonality. In order to choose an appropriate model, it is essential to define which characteristics exists in the series. Temperature prediction is a clear example in which all characteristics are applied to get the best fit. JMP time series platform

enables us to explore, analyze and forecast univariate time series. During this exploration phase, time series graphs were built for each weather condition in order to find a linear trend.

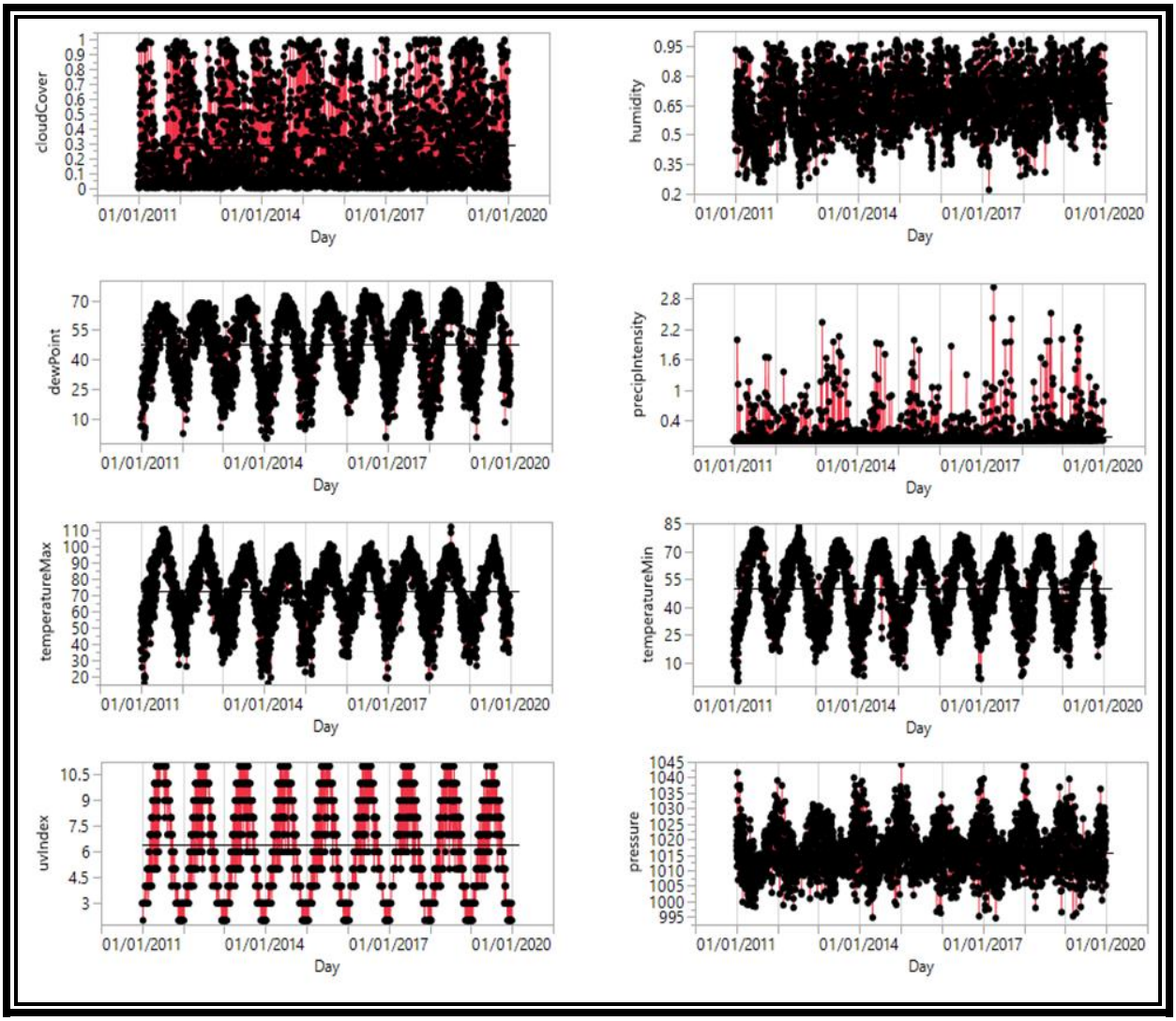


Figure 3.7: JMP Time Series Graphs for all weather conditions.

Figure 3.7 contains the time series graphs for all weather conditions. Maximum temperature, minimum temperature, UV index and dew point observations follow a linear and seasonal trend. This means that these variables can be analyzed with time series forecasts. The time series forecasts platform in JMP has the ability of enabling seasonality or general trends in data through several models such as Holt-Winter's Method, Seasonal Exponential Smoothing, and ARIMA models.

Smoothing Models

Smoothing models remove random variation and shows trends and cyclic components. These models analyze a level component, trend component and a seasonal component on a univariate dataset. They use weighted averages of past observations to forecast new values. Holt-Winters' Method and Seasonal exponential smoothing are two different forecasting techniques that consider seasonal trend and are based on this type of models. Smoothing models are defined by the following formula:

$$y_t = \mu_t + \beta_t t + s(t) + a_t$$

Where:

- μ_t is the time varying mean term
- β_t is the time varying slope
- s(t) is one of the time-varying seasonal terms
- a_t are the random shocks

Each smoothing model defines a set of recursive smoothing equations that describe the evolution of these estimators. The smoothing equations are written in terms of model parameters called smoothing weights:

- α is the level smoothing weight
- γ is the trend smoothing weight
- ϕ is the trend damping weight
- δ is the seasonal smoothing weight

While these parameters enter each model in a different way (or not at all), they have the common property that larger weights give more influence on recent data while smaller weights give less influence on recent data.

Seasonal Exponential Smoothing

Seasonal smoothing provides varying forecasts beyond the one-step-ahead available from simple smoothing. This model assumes horizontal, stationary series, so we assign any difference between the currently observed value y_t and the current level estimate, *Lt*, to the seasonal effect (SAS Institute Inc., 2018). This model considers a level and a seasonal component, as shown in the following formula.

$$y_t = \mu_t + s(t) + a_t$$

The smoothing equations in terms of smoothing weights α and δ are defined as follows"

$$L_t = \alpha(y_t - s_{t-s}) + (1-\alpha)L_{t-1}$$

$$S_t = \delta(y_t - L_{t-s}) + (1-\delta)S_{t-s}$$

Holt's Winter Method

Holt's winter method uses exponential smoothing to encode observations in the past. Holt-Winter's method is best for data with trend and seasonality that does not increase over time. This model is composed by a level component, a trend component, and a seasonal component. The number of periods per season needs to be specified. Establishing the right seasonality periods is crucial for the accuracy of our model. The model for the additive version of the Winters method is:

$$y_t = \mu_t + \beta_t t + s(t) + a_t$$

The smoothing equations for Holts Winter's Method in terms of weights α, γ, and δ are defined as follows:

$$L_t = \alpha(y_t - s_{t-s}) + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1}$$

$$S_t = \delta(y_t - L_t) + (1-\delta)S_{t-s}$$

The estimators for these time-varying terms are defined as follows:

- L_t is a smoothed level that estimates μt
- T_t is a smoothed trend that estimates βt
- S_{t-j} for j = 0, 1,..., s - 1 are the estimates of the s(t)

ARIMA Models

ARIMA, short for 'Auto Regressive Integrated Moving Average', is the most general class of forecasting models for stationary and non-stationary time series data. ARIMA models a non-stationary time series by applying differencing transformations in the data points. The use of lags of the dependent variables or lags of the forecast errors as regressors. The mathematical formulation of the ARIMA(p,d,q) model using lag polynomials is given below:

$$\phi(B)(\omega_t - \mu) = \theta(B)\alpha_t$$

Where:

- *t* is the time index
- B is the backshift operator defined as $By_t=y_{t-1}$
- $w_t = (1-B)^d y_t$ is the response series after differencing
- μ is the intercept or mean term
- ϕ(B) and θ(B) are the autoregressive operator and the moving average operator, respectively, and are written as follows:

$$\phi(B)=1- \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

$$\theta(B) = 1- \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$$

In the case of Seasonal ARIMA (SARIMA) modeling, the differencing, moving average operators are the product of seasonal and nonseasonal polynomials:

$$\omega_t = (1 - B)^d (1 - B^s)^D y_t$$

$$\phi(B) = (1 - \phi_{1,1}B - \phi_{1,2}B^2 - \ldots - \phi_{1,p}B^p)\ (1 - \phi_{2,s}B^s - \phi_{2,2s}B^{2s} - \ldots - \phi_{2,Ps}B^{Ps})$$

$$\theta\ (B) = (1 - \theta_{1,1}B - \theta_{1,2}B^2 - \ldots - \theta_{1,q}B^q)\ (1 - \theta_{2,s}B^s - \theta_{2,2s}B^{2s} - \ldots - \theta_{2,Qs}B^{Qs})$$

Where *s* is the number of observations per seasonal period. The first index on the coefficients is the factor number (1 indicates non-seasonal, 2 indicates seasonal) and the second is the lag of the term.

Seasonal ARIMA

In the seasonal ARIMA model, seasonal differencing of appropriate order is used to remove non-stationarity from the series. A first order seasonal difference is the difference between an observation and the corresponding observation from the previous year and is calculated as $z_t = y_t - y_{t-s}$. For monthly time series s = 12 and for daily time series s = 365. This model is typically termed as the SARIMA (p, d, q) × (P, D, Q) model. The orders that determine the model are:

- *p*, Autoregressive Order. The order p of the polynomial $\phi(B)$ operator.
- *d*, Differencing Order. The order d of the differencing operator.
- *q*, Moving Average Order. The order q of the differencing operator $\theta(B)$.
- *P*, Number of seasonal auto regressive terms
- *D*, Number of seasonal differences
- *Q*, Number of seasonal moving-average terms

The advantages of using ARIMA models are that it has a solid underlying theory, stable estimation of time-varying trends and seasonal patterns, relatively few parameters (SAS Institute Inc., 2018).

3.6 DATA MINING TECHNIQUES

Data mining is the process of exploring a large amount of data in order to find useful patterns for practical application. It is a powerful technology that uses machine learning, statistical and visualization techniques to discover and predict knowledge to the user. Prediction requires to develop a model able to discover relationships between dependent and independent variables. There are various types of data mining techniques such as: Association, Classification, Prediction, Text and Clustering. Data mining provides many tools by which big data can be analyzed automatically. Random forests, neural networks and Naïve Bayes are used in statistics, data analysis and machine learning.

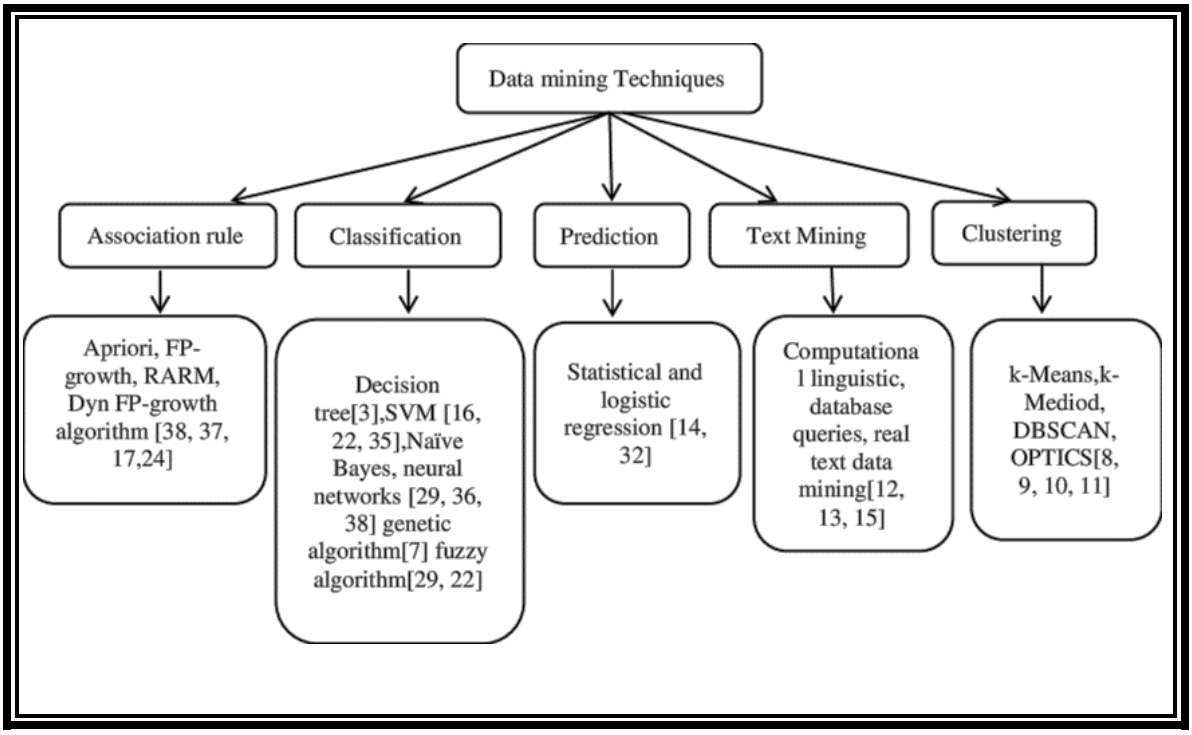


Figure 3.8: Classification of data mining techniques (Singh, 2015)

Data mining is of two types, Descriptive data mining, and Predictive data mining. Descriptive mining describes concepts or task-relevant data sets in concise, summarized, informative, discriminative forms, and Predictive mining based on data and analysis, constructs models for the database, and predicts the trend and properties of unknown data. This paper analyses

the effectiveness of predictive and classification algorithms for predicting rainfall and temperature. The data mining techniques are processed with the use of an advanced statistical analysis software called JMP.

JMP delivers different tools for visualizing, manipulating, interacting, comprehending, and analyzing the data found in JMP. JMP also provides predictive modeling with cross validation; advanced consumer research and reliability analysis; and modern statistical modeling and bootstrapping. Furthermore, JMP can be used for building predictive models and for data mining (SAS Institute Inc., 2018). These techniques are commonly used when the volume of data is large, or values are missing, or data includes outliers. Three classification techniques that can work accurately for predicting rainfall occurrence are Bootstrap Forest, Naïve Bayes, and Neural Networks.

Bootstrap Forest

The Bootstrap Forest model is a classification algorithm that is made up of a set of decision trees, that is fitted by averaging many decision trees each of which is fit to a random sunset of the training data. Each tree is grown on a bootstrap sample of the training data. A bootstrap sample is a random sample of observations, drawn with replacement. Decision trees is the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data to facilitate the machine learning (Kumar, 2013). Decision tree models' purpose is to predict the value of a targeted attributes based on several decision rules inferred from the input variables. A tree can be made to learn trough recursive partitioning. Recursive partitioning is the process that repeatedly divides the data samples into subsets depending on a parameter value test (Lin et al., 2006). The recursion of dividing subsets is completed when the last nodes have all the same value of the target variable or when they stop

adding value to the attributes. Cross-validation is a method that helps decision trees to prevent overfitting, especially in cases with many potential input variables.

The predicted probabilities for Bootstrap Forest are calculated as described below by the Probabilistic statistic formula. This formula represents the predicted probability for a given node in the tree. The method for calculating the probability for the i^{th} response level at a specific node is shown below:

$$Prob_i = \frac{n_i + n_{i-1}}{\sum(n_i + n_{i-1})}$$

Training the model is determined by testing conditions which are determined based on the data type of the variables. Data types can be either categorical or continuous. For continuous responses, Sum of Squares (SS) is the splitting criteria to fit means. SS is the change in the error sum-of-squares due to the split, which is denoted on the formula:

$$SS = \Sigma(xi + \bar{x})^2$$

$$SStest = SSparent - (SSright + SSleft) \text{ where } SS = s^2(n-1)$$

For categorical responses, the model is fitting the probabilities estimated for the response levels, minimizing the residual log-likelihood chi-square(G^2). G^2 is twice the change in the entropy. Entropy is Σ *-log(p)* for each observation, where *p* is the probability attributed to the response that occurred (SAS Institute Inc., 2018). The information gain split for categorical responses is calculated as follows:

$$G^2_{test} = G^2{}_{parent} - \left(G^2{}_{left} + G^2{}_{right}\right)$$

Bootstrap forest algorithm starts by drawing a bootstrap random sample for each tree. Then, it fits a large unpruned, classification and regression tree (CART) to this bootstrap sample using recursive partitioning. At each split in the tree only k variables are selected, instead of all. These steps are repeated for *m* times, until a stopping rule is met or until early stopping occurs. Finally,

the predictions for each created tree is averaged in order to predict observations of an output variable.

The number of observations to be selected for fitting the tree are called in-bag observations and they account for approximately 63% of the original sample. The rest of the observations (approximately 37%) are called out-of-bag and will never enter the learning sample. This stacked sample is used to teach the basic algorithms (in our case, decision trees). This is also done randomly: subsets (samples) of a certain length are taken and trained on a random subset of characteristics (attributes). The out-of-bag samples are used to measure the error rate of the model. Bootstrapping helps to remove some anomalies and ambiguity while making decision.

Artificial Neural Networks (ANN)

Artificial neural networks are based on how our biological neurons function. In the same way that our brain is made up of neurons interconnected with each other, an artificial neural network is made up of artificial neurons connected to each other and grouped together at different levels that we call layers. Layer is a general term that applies to a collection of 'nodes' operating together at a specific depth within a neural network. Figure 3.9 shows a neural network with three input variables and four layers.

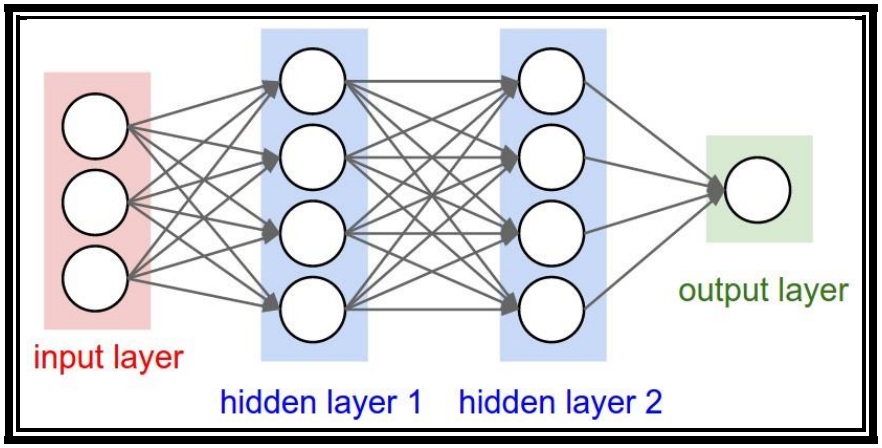


Figure 3.9: Neural Network Architecture Diagram

A Back-Propagation Network consists of at least three layers: an input layer, at least one intermediate hidden layer, and an output layer (Sawale & Gupta, 2013). The neurons in the first layer receive the actual data that feeds the neural network as input. Therefore, the first layer is known as the input layer. The output of the last layer is the visible result of the network, so the last layer is known as the output layer. The layers between the input layer and the output layer are known as hidden layers since we do not know both the input and output values. The hidden nodes are nonlinear functions of the original inputs. There is a weight associated for each node input. The functions applied at the nodes of the hidden layers are called activation functions. The activation function is a transformation of a linear combination of the input variables. They also have an output. The Deep Learning concept was born from using multiple hidden layers in networks.

Training a neural network consists in adjusting each of the input weights for all the neurons that are part of the neural network, so that the responses of the output layer fit as closely as possible to the original data. Learning in neural networks is done usually by backpropagation. This algorithm consists in calculating the gradient of the error function with respect to the neural network's weight. In other words, the errors propagate backwards from the output nodes to the inner nodes (Tiwari et al., 2013).

The functions applied at the nodes of the hidden layers are called activation functions. An activation function is a mathematical equation that determines the output of a neural network by creating a transformation of a linear combination of the input variables. TanH, Linear and Gaussian activation functions are available in JMP.

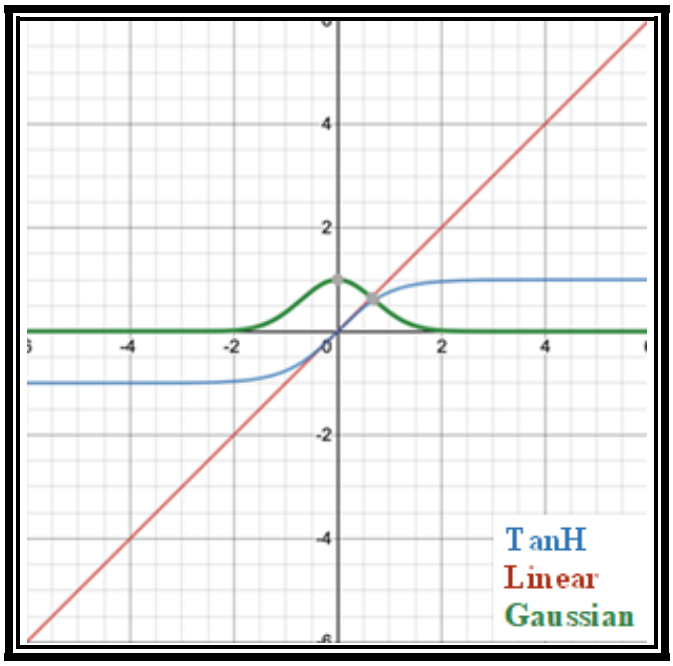


Figure 3.10: Neural Network activation functions plot (SAS Institute Inc., 2018).

The hyperbolic tangent function (TanH) is represented in blue in Figure 3.10. TanH function is a sigmoid function. This function scales input values to be between -1 and 1. The hyperbolic tangent function is:

$$\frac{e^{2x} - 1}{e^{2x} + 1}$$

where x is a linear combination of the X variables.

The linear activation function, represented as a red line in Figure 3.10, is also called identity function. The linear combination of the input variables is not transformed. The Linear activation function is most often used in conjunction with one of the non-linear activation functions. In this case, the linear activation function is placed in the second layer, and the non-linear activation functions are placed in the first layer. This is useful if you want to first reduce the dimensionality of the input variables, and then have a nonlinear model for the output variables.

For a continuous output variable, if only linear activation functions are used, the model for the Y variable reduces to a linear combination of the X variables. For a nominal or ordinal Y variable, the model reduces to a logistic regression.

$$A = cx$$

The Gaussian function is represented in green in Figure 3.10. Use this option for radial basis function behavior, or when the response surface is Gaussian (normal) in shape. The Gaussian function is where *x* is a linear combination of the X variables.

$$e^{-x^2}$$

The main advantage of a neural network model is that it can efficiently model different response surfaces. Given enough hidden nodes and layers, any surface can be approximated to any accuracy. The main disadvantage of a neural network model is that the results are not easily interpretable. This is because there are intermediate layers rather than a direct path from the X variables to the Y variables, as in the case of regular regression.

Naïve Bayes

Naive Bayes models are classification algorithms with machine learning that are based on Bayes theorem. This model classifies observations based on the levels of a categorical dependent variable. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). The algorithm calculates the conditional probability of each input variable value occurring. If the input variable is continuous, its marginal probability density is computed. Marginal probability refers to the probability of an event to happen without knowing the probability of other variables.

The algorithm is extremely fast since it estimates only one-dimensional density. The algorithm is usually employed in big data sets with many independent variables. The model requires to have non-missing observations to be able to calculate the conditional probabilities. The observation is assigned to a specific class depending on the high conditional probabilities taken from the input variables. Naïve scores are assigned to each observation for each class. An observation's naive score for a given class is the proportion of training observations that belong to that class multiplied by the product of the observation's conditional probabilities (SAS Institute Inc., 2018).

The naive probability that an observation belongs to a class is its naive score for that class divided by the sum of its naïve scores across all classes. The observation is assigned to the class for which it has the highest naive probability. Training this model requires a huge amount of observations. To prevent overfitting, data needs to be separated into two or more subsets. The larger subset is commonly used for training, and the others for validating the model.

The conditional probability that an observation belongs in the class C_k is calculated as follows:

$$P\left(C_k \middle| (x_1, \dots, x_p)\right) = \frac{(P(C_k)\prod_{i=1}^{p}[P(x_j|C_k)])/(R)}{\sum_{k=1}^{K}(P(C_k)(\prod_{i=1}^{p}[P(x_j|C_k)])/(R))}$$

Where:

- R is a regularization constant.
- C_K is class k
- X_p is the input variable p

In the previous equation, the conditional probability that an observation with $Xj = xj$ belongs to the class C_k, $P(x_j|C_k)$, is given as follows:

• If X_j is categorical:

$$P(x_j|C_k) = \frac{\#\ observations\ in\ x_j|C_k}{\#\ of\ observations\ in\ C_k}$$

• If x_j is continuous:

$$P(x_j|C_k) = \frac{1}{S_{jk}}\Phi(\frac{x_j - m_{jk}}{S_{jk}})$$

Where φ is the standard normal density function, m is the mean, and s is the standard deviation of the input variables values within the class C_k.

The unconditional probability that an observation belongs in class C_k, $P(C_k)$, is given as follows:

$$P(C_k) = \frac{\#\ observations\ in\ C_k + (\frac{0.5}{K})}{Total\ \#\ Observations + 0.5}$$

The Naive Predicted Formula classifies an observation into a class with the maximum conditional probability. In other words, where $P(C_k|(x_1,..., x_p))$ is the largest. $S(C_k)$ is the Naive Score formula for a specific class. Naive Score Sum formula, S, sums the Naive Score formulas over all classes. These formulas are shown below:

$$P\left(C_k\middle|(x_1,\ldots,x_p)\right) = \frac{S(C_k)}{S}$$

$$S = \sum_{k=1}^{K} S(C_k)$$

$$S(C_k) = \exp\left(\sum_{j=1}^{P} \ln\left(\frac{1}{s_{jk}}\right) + P\,(C_k) + \sum_{j=1}^{p} Normal\ Log\ Density(\frac{x_j - m_{jk}}{s_{jk}}) - \ln(R)\right)$$

3.7 Model Evaluation and Results

A model report is created for every time series forecast or data mining model that is analyzed. JMP provides multiple measures of fit to compare models such as R-Square, mean absolute deviation, misclassification rate, root mean square error and log likelihood. Measures of

fit appear for the training and validation sets. In addition, confusion statistics are shown for categorical responses like rainfall prediction.

For comparing time series forecasts, Akaike information criterion (AIC) is used as our model selection criteria. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model. AIC deals with both the risk of overfitting and the risk of under fitting. The preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The formula is shown below:

$$AIC = 2k - 2\ln(L)$$

where: k is the number of attributes and L is the maximum value of the Likelihood function of the model.

For predicting weather conditions with continuous variables like dew point or UV index, both time series, regressions and predictive data mining techniques are used. R-squared and RMSE are two measurements that can be used to compare these different types of models. R-squared is a statistical measure of how close the data are to the fitted regression line. R-squared measures the degree in which the independent variables explain the dependent variable. R-squared coefficients range from 0 to 1. A value of zero indicates that the model explains none of the variability of the response data around its mean. On the other side, a value of one indicates that all the variability of the response data is explained. In other words, the higher the R-squared, the better the model fits your data.

R-squared = Explained variation / Total variation

Rainfall requires another type of measurement for the comparison of the models since our response is categorical. The proposed models for rainfall prediction are Neural Network, Bootstrap Forest and Naïve Bayes. Misclassification rate and confusion matrices are two possible ways of comparing these models. Misclassification Rate is the rate for which the response category with the highest fitted probability is not the observed category. The model that best fit is the one with the lowest misclassification rate. Confusion matrix is a two-way classification of the actual response levels and the predicted response levels. The predicted level is the one with the highest predicted probability.

3.8 Chapter Conclusion

In this chapter, we defined the proposed architecture and its concepts. The goal of this planned framework is to find a way of accurately predict weather conditions. From the different weather APIs available, Dark Sky seems to be the best option since we can retrieve structured past data with a wide variety of weather conditions. There are iterative steps to ensure quality and accuracy in this proposed architecture of big data which are: exploring, pre-processing, and analyzing data.

After retrieving data, we propose a big data pre-processing framework to select only significant information and ensure the quality in our data. Before building a weather forecast model, we performed an Exploratory Data Analysis in order to determine what data processing techniques can help improve the accuracy, scalability and efficiency of the classification or prediction process. JMP allows us to create the different models explained in this chapter. The best model fit for each weather condition is defined by the response data type (categorical or continuous) and models compared. AIC, R-square and misclassification rate are the model

comparison measures applied to temperatures, weather conditions and rainfall occurrence, respectively.

Chapter 4: Development of Proposed Approach

This chapter explains and presents our findings during the development of the proposed approach. It contains the results of the forecasting and data mining techniques with the detailed explanation of their parameters. The detailed analysis of the best-fitted model and comparison of all methods based on performance measurements is completed using JMP. The forecasts provided daily predictions throughout 2020 for El Reno, Oklahoma.

4.1 Minimum and Maximum Temperature Forecasting Techniques

Temperature is comparatively easier to forecast compare to other meteorological condition since it follows a linear seasonal trend. JMP Time Series platform is designed to recognize a variety of time patterns. We use this platform to observe the series and predict the minimum and maximum daily temperature for the next 365 days. The Time Series graphs for minimum and maximum temperature show that the series is cyclic. Points that are 1 lag apart are positively correlated, with an auto-correlation value of 0.8923 for maximum temperature. As points become farther apart, they become negatively correlated, then positively correlated again, and then the pattern repeats. The Time Series graph and the autocorrelation chart provide evidence of seasonality.

Time series forecasts with a seasonal factor were used to predict minimum and maximum temperature. The forecasts applied were ARIMA, Holt's Winters Method and Seasonal Exponential Smoothing. AIC (Akaike Information Criterion) is the used as the performance measurement for these forecasts. Model Comparison tables summarize the fit statistics for each model and is used to compare several models fitted to the same time series. By default, the models are sorted in decreasing order by the AIC statistic, as shown in Table 4.1 and Table 4.2.

Table 4.1: Comparison of maximum temperature models.

Model	DF	Variance	AIC
Seasonal ARIMA (2, 0, 2)(1, 1, 1)-365	2908	60.9199	21075.755
Seasonal Exponential Smoothing (365)	2912	76.8393	21724.617
Winters Method (Additive)	2911	76.8657	21726.617
ARMA(2, 2)	3275	63.9800	22955.600

Table 4.2: Comparison of minimum temperature models.

Model	DF	Variance	AIC
Seasonal ARIMA (2, 1, 1)(1, 1, 1)-365	2908	43.3479	20083.490
Seasonal Exponential Smoothing (365)	2912	54.4059	20718.632
Winters Method (Additive)	2911	54.4245	20720.632
IMA(1, 1)	3277	53.5000	22356.825

The best model for both variables is the ARIMA model. ARIMA models are created for each possible combination of values among the different order elements. In this case, 216 ARIMA models were built to find the optimal solution. These order elements that constitute the model are p, d, q, P, D, and Q. JMP executes an automatic prediction for the elements of the ARIMA model. It iterates until selecting the appropriate values for the elements of model. The orders p, d, q, P, D, and Q are selected by means of a criterion of information such as the Akaike Information Criterion (AIC). The best ARIMA (p, d, q)(P, D, Q) model for high temperature is a seasonal ARIMA(2,0,2)(1,1,1) and for low temperature is ARIMA(2,1,1)(1,1,1).

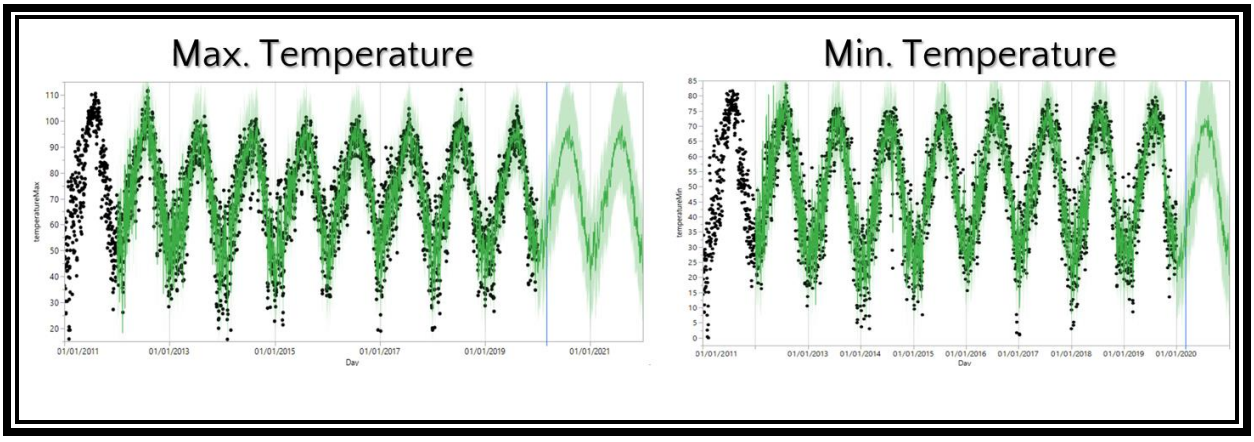


Figure 4.1: Maximum and Minimum Temperature ANOVA Time Series Forecast

The forecast plots in Figure 4.1 show the observed and predicted values by the Seasonal ARIMA model for the minimum and maximum time series. The plot is divided by a vertical line into two regions. To the left of the blue vertical line, the one-step-ahead forecasts are overlaid with the observed data points. To the right of the vertical line are the future predicted values by the model. The parameter estimates for maximum temperature is shown in Table 4.3, while minimum temperature is shown in Table 4.4.

Table 4.3. Parameter estimates for maximum temperature.

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	1.424021	0.0463524	30.72	<.0001*
AR1,2	1	2	-0.435656	0.0400366	-10.88	<.0001*
AR2,365	2	365	-0.023019	0.0214960	-1.07	0.2843
MA1,1	1	1	0.755447	0.0479174	15.77	<.0001*
MA1,2	1	2	0.191879	0.0312934	6.13	<.0001*
MA2,365	2	365	1.000000	0.0640861	15.60	<.0001*
Intercept	1	0	-0.144228	0.2372736	-0.61	0.5433

Table 4.4. Parameter estimates for minimum temperature.

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	0.671843	0.0189447	35.46	<.0001*
AR1,2	1	2	-0.148865	0.0188837	-7.88	<.0001*
AR2,365	2	365	-0.031317	0.0197719	-1.58	0.1133
MA1,1	1	1	0.986378	0.0057390	171.87	<.0001*
MA2,365	2	365	1.000000	0.1192333	8.39	<.0001*
Intercept	1	0	-0.000726	0.0012254	-0.59	0.5536

After analyzing different time series forecast models, ARIMA demonstrated to be the more accurate model to forecast temperature. Since most of the weather conditions analyzed have a correlation with minimum and maximum temperature, we can approach predictions for these weather conditions through regressions or classification data mining techniques. The next section

will evaluate different predictive models in order to find the next 365 days different weather conditions.

4.2 Related Weather Conditions Models

The framework for this section is build based on the correlation matrix. If there is a big positive or negative correlation between temperature and other weather condition, then we can develop data mining techniques to predict these weather conditions in the future with the use of minimum and maximum temperature as the input variables. The weather conditions analyzed on this section are dew point, humidity, UV index and cloud cover.

Dew Point

Dew point is the temperature to which air must be cooled to become saturated with water vapor. The correlation matrix supports this statement by demonstrating a very strong correlation with minimum and maximum temperatures. The correlation between dew point and both temperatures is above 0.850. Since Dew Point also follows a linear trend, time series forecasts are considered. In addition, Neural Networks, Bootstrap Forest and a regression are analyzed to find the best fit. We have used minimum temperature and maximum temperature as the input to these models and predicted relative dew point. Bootstrap Forest has the highest R-square value. Forty-eight decision trees are created in the forest.

Humidity

As the correlation between relative humidity and precipitation is significant (0.376). We have also forecasted relative humidity. We have used dew point, minimum temperature and maximum temperature as the input to the model and predicted relative humidity. Forecasted minimum and maximum temperature were given as the input instead of the measured temperature to get the final model accuracy and to predict the next 365 days. The result shows that neural

network works best by having the largest R-Square value (0.962). The proposed neural network contains 3 nodes in the hidden layer with TanH as the activation function.

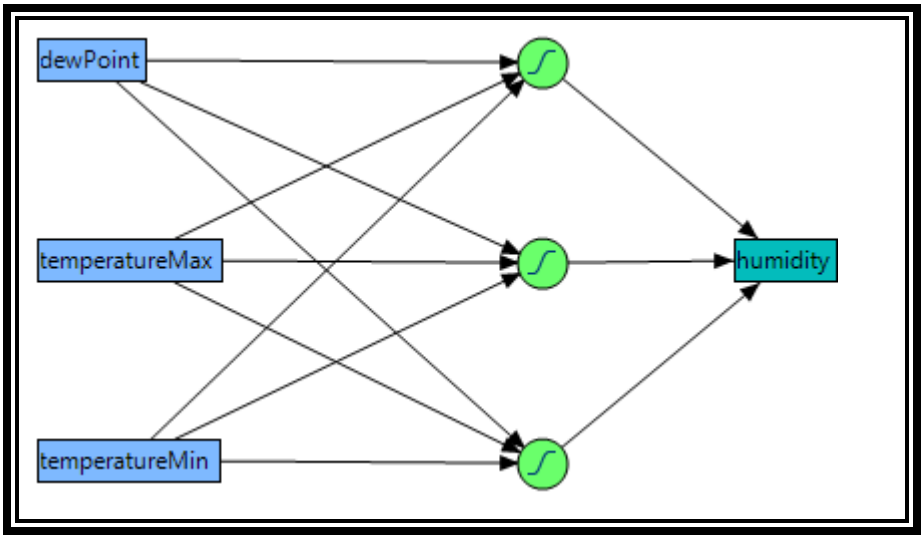


Figure 4.2: Developed Neural Network for Humidity.

UV Index

UV index is one of the important parameters to consider when predicting humidity and cloud cover as its correlation with them is above 0.25. UV index is correlated to almost every other weather condition or parameter considered in this study. UV index follows a seasonal linear trend; therefore, several time series forecasts will be analyzed. We have also applied the regression and data mining techniques for predicting UV index with five input parameters. The parameters that have been introduced to the models are dew point, humidity, minimum temperature and maximum temperature. As a result, neural network provides the highest R-Square compared to the other developed models. The neural network architecture consists in five inputs, five nods in the hidden layer with a TanH activation function and the response.

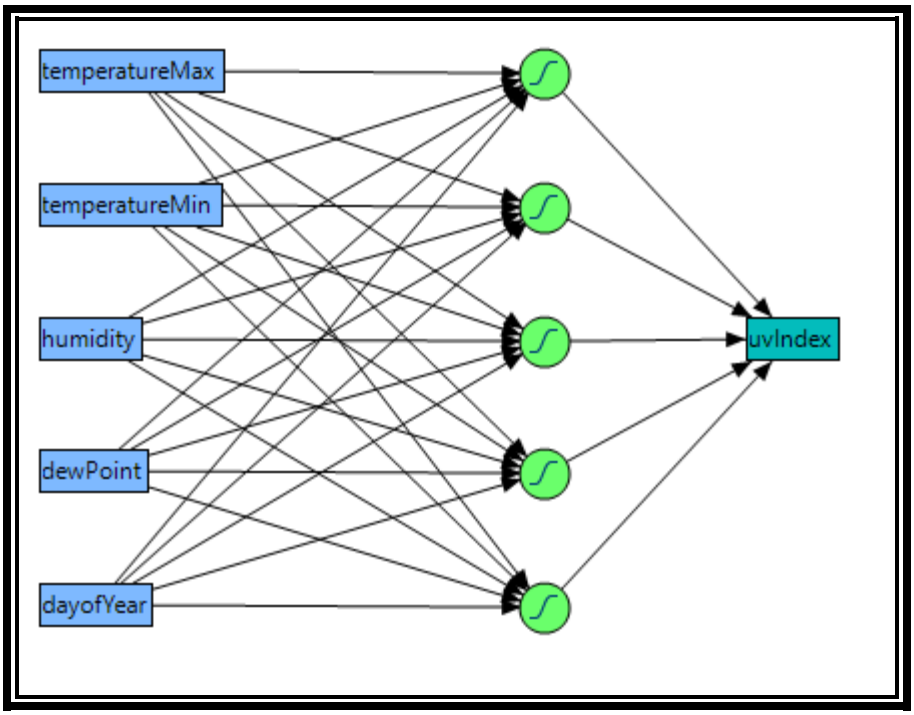


Figure 4.3: Developed Neural Network for UV Index.

Cloud Cover

Cloud cover presents a significant positive correlation with precipitation. The correlation coefficient between cloud cover and precipitation is 0.378. We have added several parameters as inputs. These parameters are dew point, humidity, UV index, minimum temperature and maximum temperature. The performance of neural networks, bootstrap forest and regressions has been analyzed and models have been compared. The result shows that neural network works best by having an R-Square value of 0.806. A neural network that with 6 nodes in a single hidden layer, and TanH activation function was proposed.

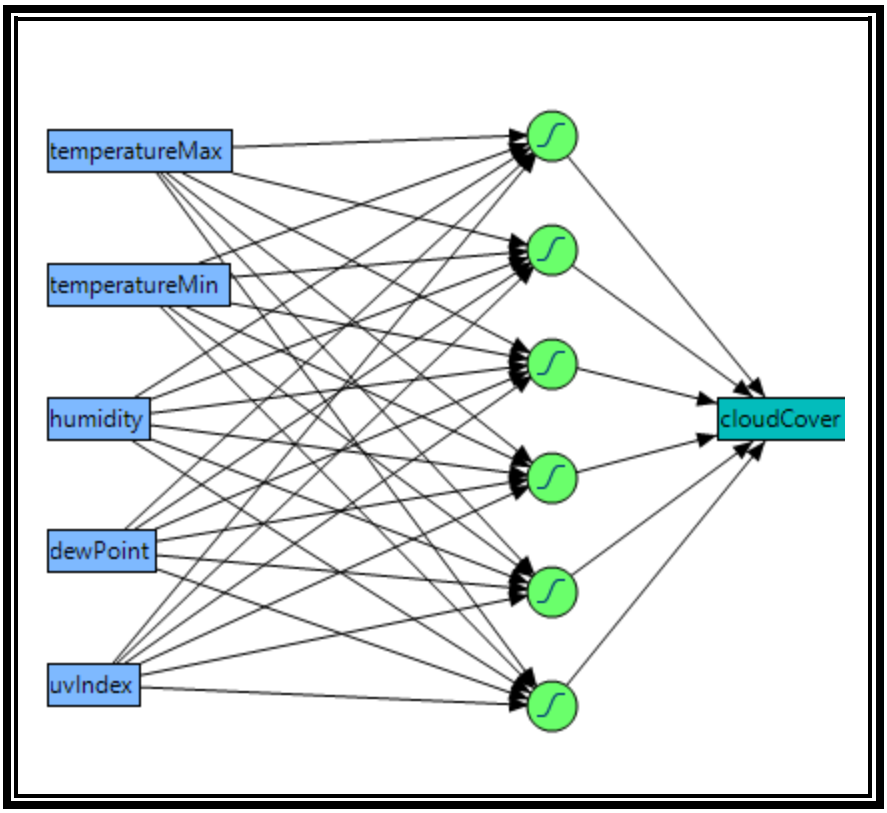


Figure 4.4: Developed Neural Network for Cloud Cover.

Atmospheric Pressure

Atmospheric pressure is the force exerted on a surface by the air above it as gravity pulls it to Earth. Atmospheric pressure drops as altitude increases. Atmospheric pressure in our data doesn't present too much variation since all observations were retrieved from one specific location; El Reno has an altitude of 1,358 ft. This weather indicator is negatively correlated to dew point, UV index, wind, precipitation, maximum and minimum temperature. Therefore, we can assume that when a low-pressure system moves into the area, it typically leads to higher temperatures, wind and precipitation. The fitted models for pressure have the lowest R-Square value. Neural network was the best fitted model with a R-Square value of 0.636. As per the fact that pressure prediction is not very accurate, rainfall prediction models will be created including and excluding pressure as an input variable. The results of such models will provide us a better understanding of how significant this parameter is for rainfall prediction.

Performance Measurements and Analysis

This section attempts to forecast precipitation related weather conditions in El Reno, Oklahoma using a fusion of forecasting and data mining techniques. The EDA provided us insight on the relationship among the different weather conditions. The PCA and cluster analysis indicated that the most significant variables to determine rainfall are maximum, minimum temperature, humidity, dew point, cloud cover, and UV index. Different data mining and forecasting models were compared using R-Square value as the performance measure criteria. The best fitted model for dew point is Bootstrap forest. While, neural networks proofed the best ability to learn and model non-linear and complex relationships in the rest of the weather conditions. Table 4.5 shows the created models for each weather condition and its R-Square value.

Table 4.5: Forecasted weather conditions R-Square.

Weather Condition	Model	R-Square
Dew Point	Bootstrap Forest	0.953
	Neural Network	0.925
	Generalized Regression	0.897
	Seasonal ARIMA (2, 0, 2)(0, 1, 1) -365	0.811
	Seasonal Exponential Smoothing (365)	0.753
	Winters Method (Additive)	0.753
	ARIMA(2, 1, 2)	0.85
Humidity	Neural Network	0.958
	Bootstrap Forest	0.94
	Generalized Regression	0.928
UV Index	Neural Network	0.916
	Bootstrap Forest	0.898
	Seasonal ARIMA (2, 0, 2)(0, 1, 1) -365	0.787
	Seasonal Exponential Smoothing (365)	0.769
	Winters Method (Additive)	0.769
	Generalized Regression	0.643
Cloud Cover	Neural Network	0.806
	Generalized Regression	0.634
	Bootstrap Forest	0.446

Pressure	Neural Network	0.636
	Bootstrap Forest	0.515
	Generalized Regression	0.388

4.3 RAINFALL PREDICTION MODELS

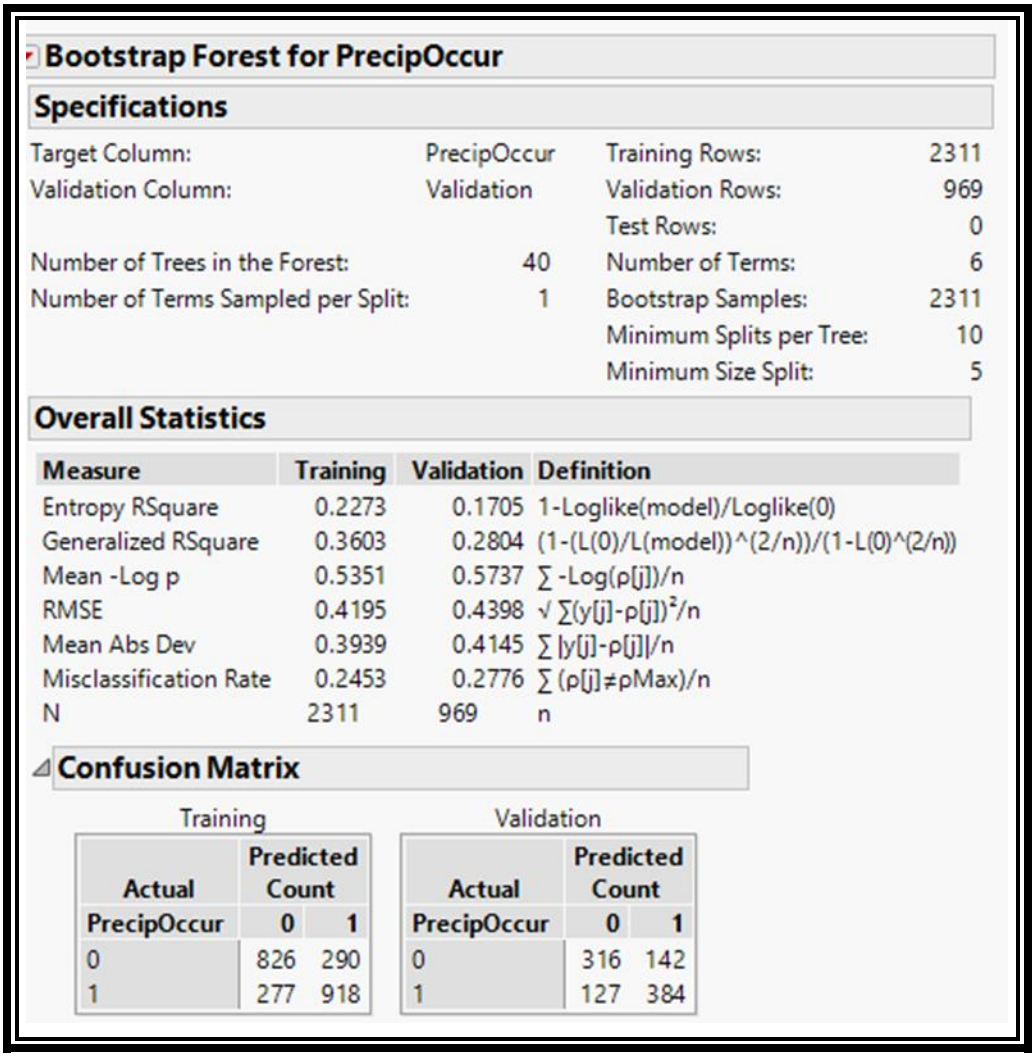
Models such as Neural Networks, Bootstrap Forest and Naïve Bayes allow us to predict the class or category for a given observation. These models are useful for predicting nominal responses. In order to predict rainfall, precipitation intensity values were classified into two categories: rain and no rain. Rain is represented with a value of 1 and no rain with 0. JMP is used to develop classification models that predict precipitation occurrence in the next 365 days based on weather conditions. The analysis uses the actual values of maximum temperature, minimum temperature, humidity, dew point, UV index and cloud cover as input parameters to the training and validation datasets.

In order to prevent overfitting, the original dataset was divided in 70% for training (2,411 observations) and 30% for validating the model (969 observations). The models are not trained with the forecasted values as input parameters since they contain some errors that will decrease the final accuracy of the model. But as we want to forecast the rainfall it is necessary to use the forecasted weather condition values for the next 365 days as input parameters to the trained model. Neural networks, bootstrap forest and Naïve Bayes models are developed to classify and predict rainfall prediction. The best fitted model is the one with the lowest misclassification rate. This is calculated as misclassifications divided by the total number of observations.

Bootstrap Forest

Bootstrap forest is a collection of several decision trees. According to Shah et al. (2018), bootstrap forest is characterized by their efficiency to deal with big data, relatively robustness for outliers and noise and ability to deal with highly correlated predictor variables. Different

specifications have been applied for tuning the model. The minimum split per tree is set to 10 and the minimum size split is 5. As in random forest case, one of these specifications is how many trees should be used to get the more accurate results. The model will continue to iterate until the error rate is constant. This model stopped after creating 40 decision trees. From the confusion matrix in Figure 4.5, it is found that 76.8% of rainy observations and 73.9% of no rain observations are accurately classified in the training set. The model is validated with a second set that accurately classified 72.2% of the observations.



Bootstrap Forest for PrecipOccur

Specifications

Target Column:	PrecipOccur	Training Rows:	2311
Validation Column:	Validation	Validation Rows:	969
		Test Rows:	0
Number of Trees in the Forest:	40	Number of Terms:	6
Number of Terms Sampled per Split:	1	Bootstrap Samples:	2311
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.2273	0.1705	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3603	0.2804	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5351	0.5737	∑ -Log(ρ[j])/n
RMSE	0.4195	0.4398	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3939	0.4145	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2453	0.2776	∑ (ρ[j]≠ρMax)/n
N	2311	969	n

Confusion Matrix

Training

Actual	Predicted Count	
PrecipOccur	0	1
0	826	290
1	277	918

Validation

Actual	Predicted Count	
PrecipOccur	0	1
0	316	142
1	127	384

Figure 4.5: Bootstrap forest report in JMP for rainfall prediction.

Neural Network

A model report is created for every neural network model. Four models were fitted and compared. Each created model contains a different number of nodes in the hidden layer, which varies from three to six nodes. The best fitted neural network showed in Figure 4.6 consists of 6 nodes in the hidden layer with TanH activation functions. TanH activation functions allow the nodes to predict the probability as an output. Measures of fit appear in Figure 4.7 for the training and validation sets. The misclassification rate statistic for the training set is 26.5% and the validation set is 25.8%, signifying that the model is predicting well on data not used to train the model.

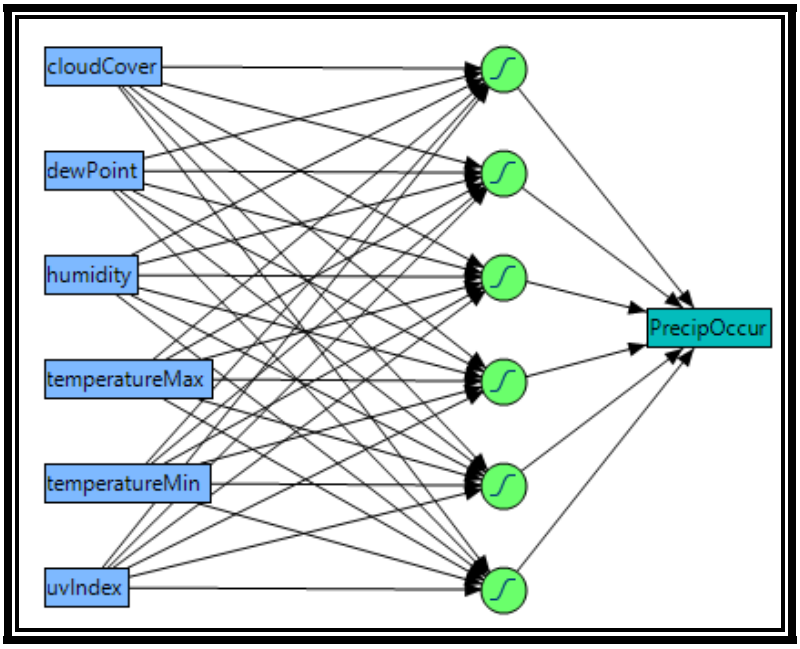
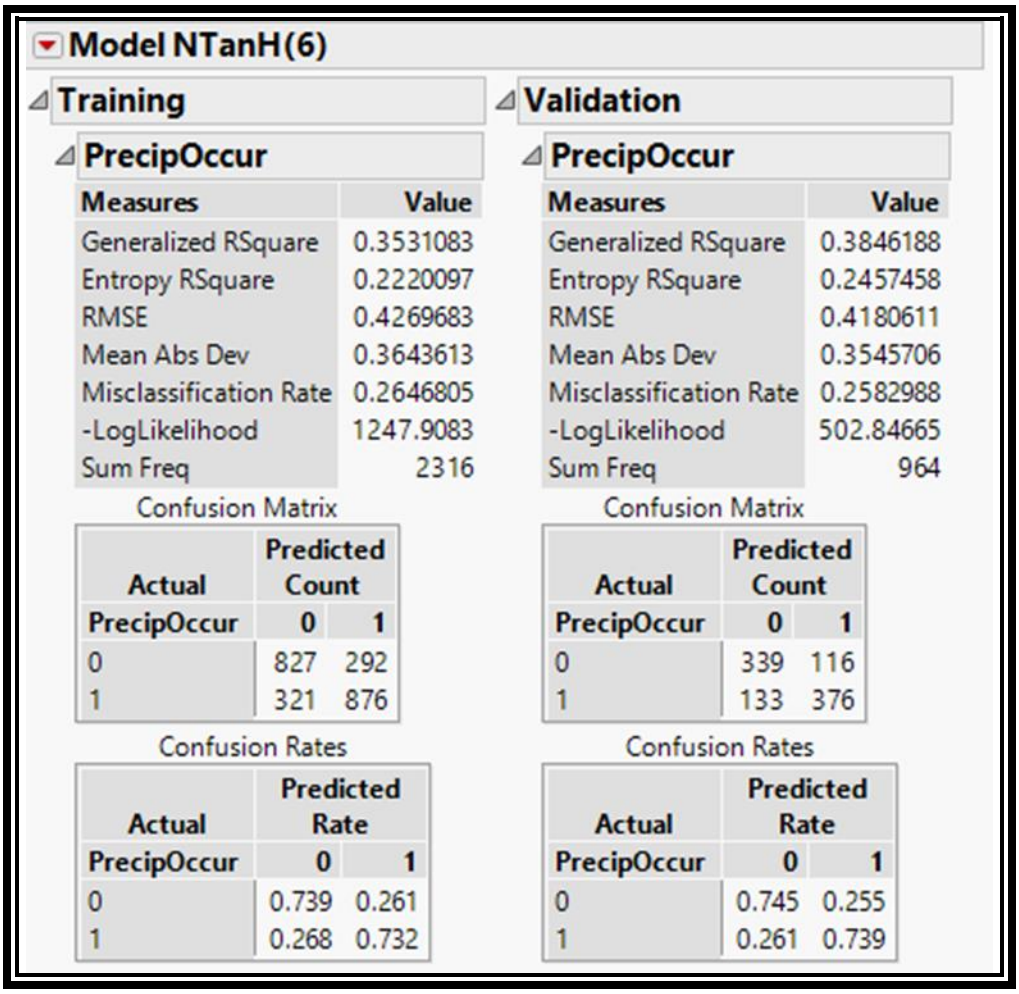


Figure 4.6: Neural network architecture for rainfall prediction.



Model NTanH(6)

Training

PrecipOccur

Measures	Value
Generalized RSquare	0.3531083
Entropy RSquare	0.2220097
RMSE	0.4269683
Mean Abs Dev	0.3643613
Misclassification Rate	0.2646805
-LogLikelihood	1247.9083
Sum Freq	2316

Confusion Matrix

Actual	Predicted Count	
PrecipOccur	0	1
0	827	292
1	321	876

Confusion Rates

Actual	Predicted Rate	
PrecipOccur	0	1
0	0.739	0.261
1	0.268	0.732

Validation

PrecipOccur

Measures	Value
Generalized RSquare	0.3846188
Entropy RSquare	0.2457458
RMSE	0.4180611
Mean Abs Dev	0.3545706
Misclassification Rate	0.2582988
-LogLikelihood	502.84665
Sum Freq	964

Confusion Matrix

Actual	Predicted Count	
PrecipOccur	0	1
0	339	116
1	133	376

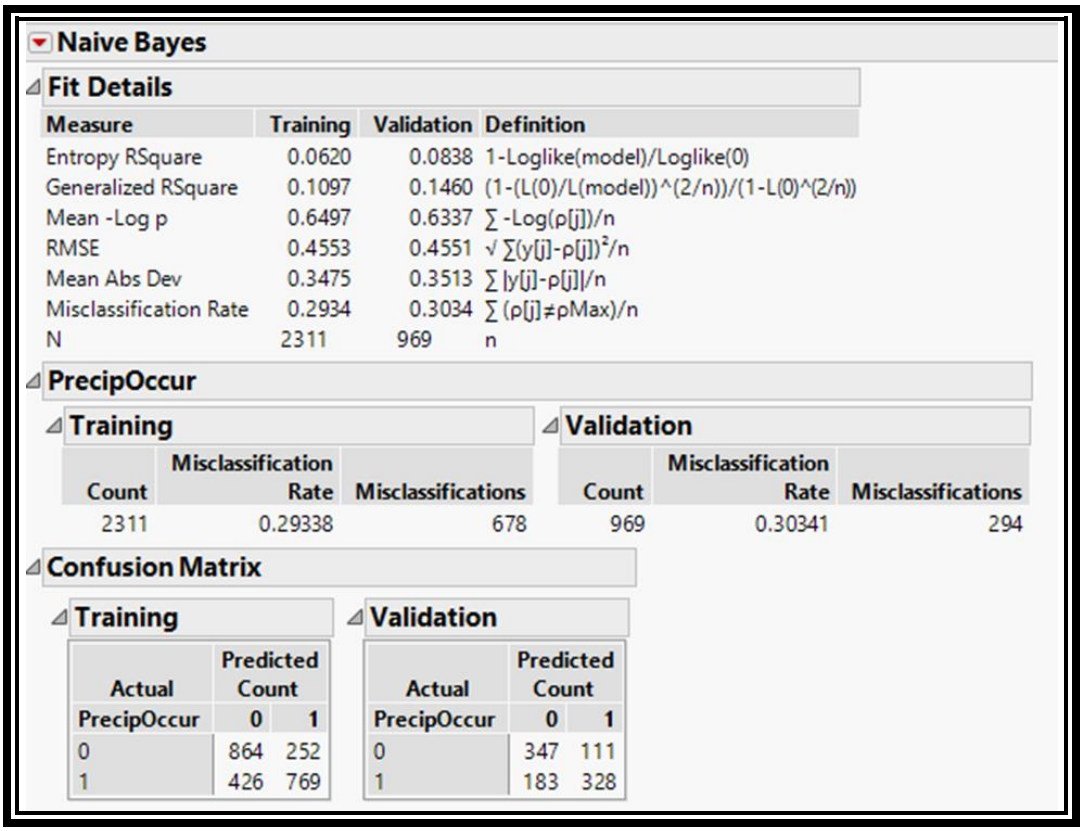
Confusion Rates

Actual	Predicted Rate	
PrecipOccur	0	1
0	0.745	0.255
1	0.261	0.739

Figure 4.7: Neural network report in JMP for rainfall prediction.

Naïve Bayes

A Naïve Bayes model is used on the training and validation sets. Figure 4.8 shows that misclassification rates are between 29.3% and 30.3%. The confusion matrices for all the sets suggest that the largest source of misclassification is the classification of precipitation as no precipitation. A confusion matrix is a two-way classification of actual and predicted responses.



Naive Bayes

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.0620	0.0838	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1097	0.1460	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6497	0.6337	∑ -Log(ρ[j])/n
RMSE	0.4553	0.4551	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3475	0.3513	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2934	0.3034	∑ (ρ[j]≠ρMax)/n
N	2311	969	n

PrecipOccur

Training

Count	Misclassification Rate	Misclassifications
2311	0.29338	678

Validation

Count	Misclassification Rate	Misclassifications
969	0.30341	294

Confusion Matrix

Training

Actual PrecipOccur	Predicted Count 0	1
0	864	252
1	426	769

Validation

Actual PrecipOccur	Predicted Count 0	1
0	347	111
1	183	328

Figure 4.8: Naïve Bayes report in JMP for rainfall prediction.

Performance Measurements and Analysis

Bootstrap forest outperforms compared to Neural Networks and Naïve Bayes. The data set used contains 9 years of daily weather data observations in El Reno, Oklahoma. The input parameters used to train the models were minimum temperature, maximum temperature, humidity, dew point, cloud cover, and UV index. It is important to consider that forecasted parameters also have their own error, so if we put the forecasted value as input parameter to this classification technique there are chances to decrease the final accuracy of the model. Table 4.6 is the resultant misclassification rates for the created random forest, neural network and Naïve Bayes training and validation data sets.

Table 4.6: Misclassification rates in rainfall prediction models.

Dataset	Model Selected	Misclassification Rate	N
Training	Bootstrap Forest	0.2453	2311
Validation	Bootstrap Forest	0.2776	969
Training	Neural Network	0.2647	2311
Validation	Neural Network	0.2583	969
Training	Naïve Bayes	0.2934	2311
Validation	Naïve Bayes	0.3034	969

As the bootstrap forest model gives the best accuracy, Table 4.7 shows the confusion matrices for the 70% training and 30% validation data sets. Diagonal shows the correctly classified category. It is found that bootstrap forest shows good accuracy for both precipitation and no precipitation.

Table 4.7: Bootstrap Forest confusion matrix for rainfall prediction.

Training

Actual	Predicted Count	
PrecipOccur	0	1
0	826	290
1	277	918

Validation

Actual	Predicted Count	
PrecipOccur	0	1
0	316	142
1	127	384

4.4 Chapter Conclusions

Table 4.8 shows the daily weather condition prediction values for 2020. The obtained predictions for minimum temperature, maximum temperature, humidity, dew point, cloud cover, and UV index are used as input parameters on the trained decision tree model in order to predict rainfall. Based on the 9-year data used to train the model, precipitation has occurred in 52% of the days in El Reno, Oklahoma. For 2020, rainfall is predicted for 30% of the days. The maximum forecasted temperature is expected to be in May 22nd and has a value of 97 degrees Fahrenheit. On the other side, the lowest forecasted temperature is 17 degrees Fahrenheit and it is anticipated in October 31st. Regardless of the technique used to predict the

values, all weather conditions show a seasonal trend. For example, humidity is predicted to be high in winter and low in summer. Rainfall is more likely to happen in the months of February, March, August and October.

Table 4.8: Weather conditions and rainfall prediction for the next 365 days.

Day	Max Temp.	Min Temp.	Dew Point	Humidity	UV index	Cloud Cover	Rainfall
1/1/2020	55.778	30.896	28.766	0.621	2.523	0.133	0
1/2/2020	55.760	31.725	29.192	0.623	2.587	0.142	0
1/3/2020	54.863	30.691	29.377	0.644	2.609	0.164	0
1/4/2020	58.938	30.053	28.313	0.587	2.826	0.028	0
1/5/2020	61.984	34.002	30.629	0.568	2.911	0.027	0
1/6/2020	63.517	37.124	32.458	0.562	2.768	0.079	0
1/7/2020	63.399	42.875	38.173	0.627	2.808	0.211	1
1/8/2020	60.546	40.503	36.976	0.653	2.873	0.237	1
1/9/2020	63.144	38.153	33.029	0.568	3.176	0.053	0
1/10/2020	64.159	37.066	32.121	0.549	3.345	0.005	0
1/11/2020	62.408	36.712	32.014	0.568	3.181	0.056	0
1/12/2020	66.839	39.350	36.124	0.581	3.255	0.048	0
1/13/2020	67.358	38.966	36.146	0.579	3.197	0.055	0
1/14/2020	69.405	42.763	40.170	0.605	3.374	0.078	0
1/15/2020	68.050	44.542	39.172	0.585	3.415	0.103	0
1/16/2020	67.502	42.755	38.733	0.597	3.444	0.094	0
1/17/2020	68.326	44.665	39.127	0.580	3.519	0.099	0
1/18/2020	65.810	42.990	37.990	0.598	3.472	0.127	0
1/19/2020	69.732	44.331	40.443	0.594	3.615	0.089	0
1/20/2020	70.656	48.871	44.322	0.623	3.550	0.173	0
1/21/2020	68.036	45.699	39.476	0.581	3.636	0.130	0
1/22/2020	64.697	44.302	37.531	0.589	3.567	0.172	0
1/23/2020	64.830	39.045	35.201	0.586	3.716	0.083	0
1/24/2020	62.162	38.670	33.499	0.583	3.684	0.117	0
1/25/2020	57.743	37.711	32.308	0.613	3.546	0.211	1
1/26/2020	65.009	37.227	31.993	0.536	3.866	0.049	0
1/27/2020	64.692	43.109	37.716	0.603	3.786	0.169	0
1/28/2020	67.781	41.920	38.508	0.596	3.898	0.111	0
1/29/2020	68.688	45.652	39.615	0.577	4.051	0.127	0
1/30/2020	68.615	46.873	41.844	0.612	3.917	0.187	1
1/31/2020	72.286	43.385	39.257	0.552	4.294	0.049	0
2/1/2020	73.187	45.927	38.987	0.517	4.368	0.055	0
2/2/2020	69.973	42.973	39.739	0.589	4.291	0.093	0

2/3/2020	67.004	44.664	38.791	0.587	4.212	0.157	0
2/4/2020	66.920	43.010	38.363	0.593	4.271	0.142	0
2/5/2020	70.983	45.453	39.918	0.561	4.467	0.102	0
2/6/2020	72.996	47.130	41.737	0.562	4.522	0.104	0
2/7/2020	72.956	46.640	40.044	0.534	4.668	0.082	0
2/8/2020	75.807	49.486	43.401	0.546	4.755	0.089	0
2/9/2020	74.763	49.157	43.163	0.555	4.725	0.108	0
2/10/2020	67.992	46.866	41.872	0.619	4.542	0.209	1
2/11/2020	69.321	47.032	42.747	0.621	4.689	0.189	1
2/12/2020	71.670	44.399	39.260	0.550	4.889	0.091	0
2/13/2020	72.273	49.966	44.422	0.599	4.949	0.167	0
2/14/2020	73.406	51.043	43.553	0.561	5.050	0.150	0
2/15/2020	70.143	49.062	44.579	0.631	4.878	0.221	1
2/16/2020	67.681	47.709	43.075	0.638	4.686	0.267	1
2/17/2020	70.129	47.217	42.922	0.614	4.987	0.194	1
2/18/2020	67.087	45.822	39.217	0.585	5.042	0.199	1
2/19/2020	70.231	46.732	40.889	0.578	5.028	0.179	0
2/20/2020	72.450	47.709	42.267	0.574	5.277	0.149	0
2/21/2020	69.837	46.333	40.684	0.581	5.355	0.165	0
2/22/2020	77.805	51.976	43.766	0.512	5.715	0.098	0
2/23/2020	71.675	48.863	43.624	0.598	5.445	0.187	1
2/24/2020	70.110	44.767	39.742	0.573	5.127	0.181	1
2/25/2020	72.507	49.986	44.422	0.596	5.083	0.236	1
2/26/2020	73.030	50.766	43.994	0.576	5.454	0.204	1
2/27/2020	67.304	49.657	45.381	0.671	5.142	0.357	1
2/28/2020	74.630	48.565	43.162	0.561	5.996	0.125	0
2/29/2020	75.177	51.586	43.490	0.537	6.091	0.144	0
3/1/2020	72.932	51.064	43.864	0.572	6.062	0.182	1
3/2/2020	74.534	47.018	41.667	0.546	6.240	0.105	0
3/3/2020	75.776	48.385	42.210	0.533	6.258	0.108	0
3/4/2020	78.459	50.875	43.934	0.518	6.497	0.091	0
3/5/2020	78.879	53.659	46.100	0.533	6.559	0.118	0
3/6/2020	78.340	55.629	46.878	0.537	6.542	0.151	1
3/7/2020	78.944	56.036	47.591	0.541	5.786	0.216	1
3/8/2020	80.328	60.607	50.406	0.544	6.600	0.181	1
3/9/2020	81.400	57.481	48.633	0.525	6.896	0.122	0
3/10/2020	79.344	57.608	48.514	0.542	6.909	0.154	1
3/11/2020	79.343	56.825	48.101	0.541	6.950	0.149	0
3/12/2020	77.304	54.590	46.719	0.553	6.782	0.174	1
3/13/2020	75.119	55.232	45.918	0.555	6.198	0.262	1
3/14/2020	74.775	54.557	45.823	0.562	6.617	0.232	1
3/15/2020	77.039	53.967	46.206	0.551	7.014	0.170	1
3/16/2020	76.775	54.268	45.933	0.546	7.009	0.182	1

3/17/2020	74.588	54.085	46.339	0.578	6.720	0.243	1
3/18/2020	74.690	52.984	44.286	0.546	7.072	0.201	1
3/19/2020	80.539	57.177	48.320	0.530	7.468	0.142	0
3/20/2020	81.049	60.952	50.894	0.543	7.108	0.205	1
3/21/2020	79.072	61.843	51.395	0.566	7.459	0.214	1
3/22/2020	78.530	58.186	49.180	0.559	7.669	0.178	1
3/23/2020	82.994	60.421	50.130	0.513	7.879	0.129	0
3/24/2020	84.003	63.995	51.427	0.498	7.894	0.146	0
3/25/2020	83.850	63.364	51.439	0.505	7.164	0.208	1
3/26/2020	82.078	61.262	50.808	0.529	6.601	0.267	1
3/27/2020	83.341	60.317	49.999	0.508	6.925	0.219	1
3/28/2020	86.099	63.619	51.255	0.476	7.219	0.189	1
3/29/2020	85.696	64.771	50.889	0.464	7.379	0.191	1
3/30/2020	84.659	64.265	51.476	0.490	8.196	0.146	0
3/31/2020	82.516	61.924	51.138	0.525	8.339	0.151	1
4/1/2020	85.008	63.498	50.829	0.480	8.481	0.124	0
4/2/2020	82.684	63.751	51.670	0.518	8.210	0.182	1
4/3/2020	84.924	62.332	50.393	0.482	8.576	0.120	0
4/4/2020	87.975	62.054	50.732	0.460	8.579	0.088	0
4/5/2020	88.312	63.905	51.100	0.448	8.627	0.096	0
4/6/2020	88.162	63.182	51.073	0.455	8.889	0.080	0
4/7/2020	88.007	63.847	51.150	0.453	8.672	0.105	0
4/8/2020	86.311	64.870	50.883	0.456	8.882	0.117	0
4/9/2020	86.407	65.067	50.807	0.452	8.980	0.113	0
4/10/2020	89.538	66.011	51.760	0.431	9.053	0.084	0
4/11/2020	90.192	66.550	51.633	0.417	9.157	0.075	0
4/12/2020	87.410	67.694	51.690	0.437	9.137	0.114	0
4/13/2020	89.466	67.362	51.765	0.420	9.265	0.085	0
4/14/2020	91.837	68.507	52.602	0.403	9.384	0.062	0
4/15/2020	89.867	70.318	53.262	0.421	9.227	0.104	0
4/16/2020	88.387	68.378	52.804	0.443	9.272	0.116	0
4/17/2020	90.725	67.623	52.056	0.411	9.535	0.070	0
4/18/2020	91.493	68.521	52.757	0.409	9.587	0.066	0
4/19/2020	90.930	68.088	52.677	0.417	9.553	0.077	0
4/20/2020	90.747	66.599	51.742	0.413	9.110	0.106	0
4/21/2020	89.784	69.963	53.262	0.425	9.573	0.101	0
4/22/2020	89.440	68.095	52.646	0.432	9.727	0.091	0
4/23/2020	89.211	67.788	52.183	0.428	8.271	0.203	1
4/24/2020	92.198	68.903	52.596	0.396	8.404	0.162	0
4/25/2020	93.274	69.266	52.748	0.384	9.578	0.073	0
4/26/2020	91.742	70.188	52.913	0.396	9.342	0.112	0
4/27/2020	93.631	69.710	52.771	0.377	10.057	0.044	0
4/28/2020	93.322	70.487	52.719	0.373	10.009	0.053	0

4/29/2020	92.993	69.690	52.660	0.382	10.042	0.057	0
4/30/2020	93.430	70.673	52.719	0.370	9.768	0.075	0
5/1/2020	92.044	68.828	52.596	0.398	10.100	0.068	0
5/2/2020	90.742	67.059	51.860	0.412	10.207	0.071	0
5/3/2020	91.205	66.779	51.751	0.407	10.232	0.066	0
5/4/2020	93.056	67.982	52.039	0.384	10.270	0.050	0
5/5/2020	93.585	70.032	52.771	0.375	10.347	0.049	0
5/6/2020	92.690	70.140	52.660	0.381	10.339	0.061	0
5/7/2020	92.765	69.708	52.660	0.384	10.321	0.064	0
5/8/2020	93.853	69.823	52.461	0.368	10.409	0.048	0
5/9/2020	93.461	69.546	52.780	0.381	10.312	0.063	0
5/10/2020	95.221	71.786	52.406	0.337	9.113	0.123	0
5/11/2020	94.871	72.320	52.447	0.337	10.198	0.058	0
5/12/2020	95.969	72.550	52.366	0.322	10.519	0.027	0
5/13/2020	95.066	71.123	52.406	0.344	10.574	0.037	0
5/14/2020	94.919	70.352	52.328	0.350	10.637	0.037	0
5/15/2020	92.616	69.503	52.681	0.388	10.537	0.072	0
5/16/2020	92.811	69.558	52.681	0.385	10.655	0.064	0
5/17/2020	93.764	70.291	52.771	0.371	10.515	0.067	0
5/18/2020	91.210	70.369	52.880	0.399	9.817	0.149	0
5/19/2020	94.502	69.805	52.316	0.359	9.329	0.138	0
5/20/2020	93.668	69.395	52.780	0.380	10.659	0.065	0
5/21/2020	95.917	70.851	52.398	0.337	10.768	0.035	0
5/22/2020	97.190	70.699	52.312	0.324	10.520	0.038	0
5/23/2020	96.476	71.952	52.391	0.322	10.201	0.068	0
5/24/2020	96.395	71.366	52.391	0.328	10.800	0.034	0
5/25/2020	96.029	71.354	52.398	0.332	10.727	0.044	0
5/26/2020	94.083	71.736	52.410	0.349	10.464	0.084	0
5/27/2020	93.999	71.082	52.410	0.355	10.550	0.081	0
5/28/2020	91.060	68.814	52.792	0.412	10.713	0.099	0
5/29/2020	91.423	69.406	52.902	0.405	10.671	0.102	0
5/30/2020	89.176	68.959	52.810	0.430	10.621	0.129	0
5/31/2020	92.028	66.681	51.448	0.393	10.717	0.086	0
6/1/2020	94.374	70.147	52.316	0.357	10.950	0.059	0
6/2/2020	94.888	70.562	52.328	0.349	10.959	0.055	0
6/3/2020	97.706	72.215	52.354	0.306	10.965	0.028	0
6/4/2020	96.118	73.610	52.396	0.312	10.943	0.044	0
6/5/2020	94.522	71.324	52.406	0.348	10.965	0.063	0
6/6/2020	98.422	72.001	52.301	0.300	11.011	0.023	0
6/7/2020	96.559	71.366	52.391	0.326	11.004	0.043	0
6/8/2020	94.973	66.505	51.144	0.359	11.055	0.045	0
6/9/2020	92.679	71.801	52.707	0.369	10.876	0.092	0
6/10/2020	89.770	66.394	51.870	0.427	10.543	0.133	0

6/11/2020	89.712	62.869	51.214	0.444	11.003	0.088	0
6/12/2020	89.439	65.556	50.927	0.419	10.892	0.109	0
6/13/2020	90.853	65.902	51.364	0.410	10.933	0.095	0
6/14/2020	91.893	68.586	52.658	0.403	10.928	0.098	0
6/15/2020	91.512	66.324	51.463	0.402	11.018	0.086	0
6/16/2020	90.859	68.378	52.792	0.418	10.930	0.110	0
6/17/2020	92.684	68.572	52.605	0.394	10.893	0.094	0
6/18/2020	88.635	66.635	51.704	0.434	10.852	0.132	0
6/19/2020	88.306	65.113	50.986	0.436	10.910	0.125	0
6/20/2020	89.823	65.903	51.830	0.430	10.876	0.116	0
6/21/2020	91.133	67.205	51.831	0.406	10.805	0.112	0
6/22/2020	90.711	67.932	52.352	0.414	10.814	0.121	0
6/23/2020	89.669	67.746	52.420	0.428	10.873	0.128	0
6/24/2020	92.367	67.733	51.923	0.390	10.980	0.091	0
6/25/2020	94.090	70.020	52.461	0.364	10.965	0.082	0
6/26/2020	90.207	68.359	52.902	0.427	10.865	0.128	0
6/27/2020	90.482	66.324	51.643	0.416	10.849	0.115	0
6/28/2020	91.360	66.589	51.702	0.406	10.944	0.101	0
6/29/2020	91.683	66.779	51.798	0.403	10.907	0.101	0
6/30/2020	91.434	66.628	51.702	0.405	10.894	0.103	0
7/1/2020	91.656	67.537	51.994	0.401	10.867	0.107	0
7/2/2020	92.237	68.634	52.596	0.397	10.810	0.110	0
7/3/2020	91.685	67.977	52.290	0.403	10.861	0.110	0
7/4/2020	90.826	66.651	51.665	0.410	10.849	0.113	0
7/5/2020	90.833	66.098	51.334	0.408	10.785	0.113	0
7/6/2020	89.337	64.986	50.881	0.424	10.696	0.128	0
7/7/2020	86.738	64.438	51.091	0.460	10.697	0.157	1
7/8/2020	88.678	63.026	51.062	0.451	10.695	0.125	0
7/9/2020	87.993	63.553	50.886	0.450	10.600	0.141	0
7/10/2020	87.769	62.064	50.655	0.460	10.692	0.127	0
7/11/2020	83.879	59.414	48.976	0.490	10.629	0.156	1
7/12/2020	85.945	59.008	48.456	0.461	10.689	0.122	0
7/13/2020	86.262	59.505	48.397	0.453	10.675	0.122	0
7/14/2020	86.506	62.110	50.541	0.471	10.561	0.147	0
7/15/2020	87.269	65.192	50.861	0.443	10.538	0.158	1
7/16/2020	83.605	63.992	51.551	0.505	10.365	0.210	1
7/17/2020	82.277	62.454	51.520	0.530	10.313	0.220	1
7/18/2020	82.893	63.538	51.593	0.516	10.286	0.220	1
7/19/2020	84.377	63.049	51.272	0.499	10.373	0.188	1
7/20/2020	83.799	63.463	51.376	0.504	10.301	0.202	1
7/21/2020	83.736	62.656	50.919	0.502	10.282	0.195	1
7/22/2020	85.181	61.855	50.039	0.477	10.358	0.162	1
7/23/2020	84.891	59.981	48.491	0.465	10.307	0.150	1

7/24/2020	81.105	59.202	50.105	0.542	10.135	0.205	1
7/25/2020	80.285	56.207	48.195	0.538	10.170	0.182	1
7/26/2020	81.151	56.848	48.567	0.531	10.132	0.176	1
7/27/2020	82.371	56.552	48.316	0.516	10.175	0.152	1
7/28/2020	85.277	59.417	48.387	0.464	10.242	0.135	0
7/29/2020	83.153	58.379	49.449	0.515	10.167	0.155	1
7/30/2020	79.731	56.662	48.202	0.540	9.985	0.193	1
7/31/2020	80.992	58.597	50.113	0.548	9.888	0.199	1
8/1/2020	80.414	57.650	48.713	0.535	9.814	0.198	1
8/2/2020	77.437	54.903	46.727	0.549	9.815	0.210	1
8/3/2020	77.023	53.794	46.206	0.553	9.788	0.202	1
8/4/2020	79.438	54.291	46.828	0.536	9.759	0.170	1
8/5/2020	81.759	54.984	47.237	0.514	9.928	0.134	0
8/6/2020	74.240	50.486	43.691	0.560	9.589	0.206	1
8/7/2020	73.674	48.372	42.105	0.552	9.618	0.181	1
8/8/2020	70.196	48.365	43.643	0.618	8.997	0.281	1
8/9/2020	73.130	48.319	42.692	0.570	9.014	0.217	1
8/10/2020	74.543	48.961	43.714	0.570	8.509	0.230	1
8/11/2020	72.709	50.095	44.415	0.593	8.810	0.257	1
8/12/2020	72.016	47.465	42.402	0.583	8.614	0.237	1
8/13/2020	73.324	48.960	43.553	0.579	8.923	0.215	1
8/14/2020	75.380	50.884	43.336	0.538	9.227	0.184	1
8/15/2020	74.197	52.717	44.461	0.557	9.052	0.236	1
8/16/2020	76.991	51.162	43.966	0.531	9.185	0.158	0
8/17/2020	74.591	49.991	43.963	0.566	9.119	0.181	1
8/18/2020	73.879	44.796	38.732	0.514	9.224	0.109	0
8/19/2020	70.751	47.538	42.939	0.606	8.915	0.220	1
8/20/2020	69.110	42.860	39.618	0.597	8.644	0.188	1
8/21/2020	68.898	46.031	39.979	0.579	8.623	0.231	1
8/22/2020	69.894	47.158	42.893	0.617	8.745	0.229	1
8/23/2020	75.212	47.585	42.025	0.541	9.000	0.119	0
8/24/2020	74.725	48.230	41.712	0.535	8.731	0.144	0
8/25/2020	69.287	46.354	40.840	0.589	8.545	0.218	1
8/26/2020	71.005	45.185	39.918	0.563	8.639	0.156	0
8/27/2020	70.304	42.980	39.705	0.585	8.699	0.133	0
8/28/2020	61.814	41.773	38.197	0.653	8.092	0.316	1
8/29/2020	59.742	39.016	34.505	0.625	8.105	0.290	1
8/30/2020	68.523	42.305	38.667	0.589	8.485	0.148	0
8/31/2020	69.377	45.061	40.132	0.585	8.389	0.175	0
9/1/2020	67.046	41.698	38.245	0.601	8.275	0.168	0
9/2/2020	71.133	46.564	40.344	0.559	8.259	0.160	0
9/3/2020	66.232	43.234	38.333	0.598	7.405	0.247	1
9/4/2020	62.462	42.837	38.208	0.638	7.596	0.310	1

9/5/2020	59.391	39.114	34.784	0.633	7.535	0.297	1
9/6/2020	61.368	38.927	35.788	0.634	7.648	0.245	1
9/7/2020	60.742	35.463	32.350	0.602	7.688	0.174	0
9/8/2020	61.320	35.445	32.411	0.597	7.603	0.162	0
9/9/2020	61.842	36.765	32.186	0.577	7.451	0.169	0
9/10/2020	61.407	37.214	32.464	0.583	7.316	0.188	0
9/11/2020	58.226	35.758	30.865	0.596	7.371	0.213	0
9/12/2020	62.811	37.186	32.411	0.568	7.656	0.127	0
9/13/2020	59.208	33.859	28.634	0.558	7.127	0.152	0
9/14/2020	60.527	34.047	30.812	0.586	6.553	0.175	0
9/15/2020	62.959	38.431	33.388	0.575	6.154	0.225	0
9/16/2020	62.424	37.255	32.095	0.565	6.484	0.182	0
9/17/2020	59.629	32.025	28.406	0.565	6.542	0.132	0
9/18/2020	56.399	30.446	28.459	0.612	6.653	0.161	0
9/19/2020	61.137	33.393	30.098	0.571	6.988	0.094	0
9/20/2020	61.703	34.699	30.814	0.568	6.859	0.106	0
9/21/2020	62.292	37.263	32.216	0.568	6.191	0.174	0
9/22/2020	62.981	35.522	32.320	0.578	6.481	0.117	0
9/23/2020	55.595	33.836	29.518	0.613	6.288	0.224	1
9/24/2020	58.401	35.596	30.660	0.592	6.225	0.192	1
9/25/2020	57.644	31.685	28.793	0.596	6.385	0.125	0
9/26/2020	54.359	35.245	32.449	0.672	5.978	0.322	1
9/27/2020	53.865	33.496	29.926	0.642	6.077	0.258	1
9/28/2020	51.916	31.473	27.829	0.638	6.134	0.242	1
9/29/2020	55.881	31.048	28.415	0.612	6.411	0.122	0
9/30/2020	55.957	28.464	25.834	0.582	6.244	0.073	0
10/1/2020	54.193	30.672	28.964	0.643	6.155	0.168	0
10/2/2020	56.510	31.860	28.594	0.602	6.233	0.111	0
10/3/2020	53.210	32.440	28.863	0.637	5.961	0.213	0
10/4/2020	55.213	33.377	29.968	0.630	5.848	0.190	1
10/5/2020	49.720	31.490	27.737	0.658	5.361	0.312	1
10/6/2020	45.798	25.144	23.267	0.664	5.107	0.279	1
10/7/2020	45.090	24.213	22.805	0.670	4.705	0.310	1
10/8/2020	53.207	27.510	24.837	0.599	4.778	0.167	0
10/9/2020	53.424	30.871	27.949	0.630	4.768	0.230	0
10/10/2020	49.054	30.001	27.278	0.669	4.541	0.343	1
10/11/2020	49.810	29.214	26.992	0.662	4.840	0.270	1
10/12/2020	45.770	26.891	25.292	0.689	5.032	0.305	1
10/13/2020	49.149	26.871	25.132	0.652	4.961	0.204	0
10/14/2020	53.949	31.220	28.973	0.641	5.016	0.180	0
10/15/2020	54.869	32.612	29.772	0.636	5.048	0.175	0
10/16/2020	50.628	28.988	26.891	0.653	4.633	0.228	0
10/17/2020	48.177	28.973	27.078	0.682	4.750	0.283	1

10/18/2020	52.849	29.369	27.472	0.639	4.860	0.151	0
10/19/2020	56.396	31.898	28.747	0.606	4.763	0.119	0
10/20/2020	53.773	34.711	30.723	0.648	4.572	0.253	1
10/21/2020	53.705	30.582	28.392	0.638	4.401	0.182	0
10/22/2020	49.103	30.070	27.278	0.668	4.099	0.308	1
10/23/2020	47.716	29.321	27.552	0.693	4.143	0.340	1
10/24/2020	48.405	29.010	27.037	0.679	4.162	0.292	1
10/25/2020	46.361	25.421	23.000	0.651	4.175	0.222	0
10/26/2020	43.368	25.107	22.703	0.678	3.735	0.354	0
10/27/2020	42.976	24.907	23.085	0.692	3.914	0.342	0
10/28/2020	43.243	20.539	16.761	0.601	4.095	0.146	0
10/29/2020	41.305	21.819	19.429	0.663	3.993	0.255	0
10/30/2020	38.570	19.268	15.220	0.630	3.848	0.243	0
10/31/2020	39.740	17.823	15.069	0.627	3.879	0.182	0
11/1/2020	48.074	24.356	22.778	0.638	3.900	0.155	0
11/2/2020	43.738	23.223	21.472	0.666	3.902	0.226	0
11/3/2020	42.842	22.279	20.555	0.665	3.843	0.225	0
11/4/2020	46.980	22.667	21.922	0.646	3.895	0.135	0
11/5/2020	45.003	23.353	22.195	0.666	3.755	0.208	1
11/6/2020	47.005	21.681	20.051	0.618	3.749	0.106	0
11/7/2020	47.015	23.622	22.103	0.642	3.603	0.167	0
11/8/2020	49.030	26.249	24.174	0.639	3.654	0.163	0
11/9/2020	45.188	25.123	23.170	0.669	3.543	0.246	0
11/10/2020	44.168	25.414	22.178	0.657	3.545	0.253	0
11/11/2020	46.049	25.754	23.873	0.668	3.646	0.215	0
11/12/2020	46.434	25.969	24.048	0.666	3.752	0.189	0
11/13/2020	45.988	24.615	23.012	0.662	3.500	0.200	1
11/14/2020	44.448	22.795	21.727	0.667	3.385	0.212	0
11/15/2020	47.738	23.629	21.989	0.632	3.451	0.127	0
11/16/2020	48.930	22.743	21.833	0.624	3.437	0.092	0
11/17/2020	51.183	27.888	25.101	0.622	3.392	0.133	0
11/18/2020	57.266	28.380	25.712	0.567	3.523	0.018	0
11/19/2020	56.124	28.494	25.834	0.580	3.577	0.023	0
11/20/2020	52.106	27.564	24.553	0.604	3.475	0.077	0
11/21/2020	47.603	26.301	24.408	0.658	3.186	0.208	1
11/22/2020	48.915	22.022	20.632	0.607	3.282	0.067	0
11/23/2020	57.157	26.609	25.083	0.571	3.423	0.002	0
11/24/2020	51.876	26.242	23.358	0.594	3.398	0.050	0
11/25/2020	51.844	23.506	21.147	0.574	3.277	0.024	0
11/26/2020	55.183	25.775	24.023	0.577	3.176	0.031	0
11/27/2020	55.153	25.974	24.045	0.576	3.130	0.035	0
11/28/2020	59.536	27.746	25.653	0.548	3.167	0.000	0
11/29/2020	60.896	31.998	29.686	0.577	3.073	0.050	0

11/30/2020	59.161	34.006	28.625	0.558	3.126	0.063	0
12/1/2020	56.319	33.885	29.054	0.596	3.211	0.099	0
12/2/2020	48.005	28.173	26.144	0.672	2.911	0.249	1
12/3/2020	42.124	20.354	16.968	0.619	2.976	0.146	0
12/4/2020	44.604	18.880	16.426	0.595	3.027	0.062	0
12/5/2020	42.848	22.259	20.555	0.665	2.830	0.226	1
12/6/2020	44.655	23.705	22.095	0.665	2.816	0.219	1
12/7/2020	52.926	24.934	22.509	0.578	2.941	0.039	0
12/8/2020	49.992	26.366	23.761	0.620	2.997	0.102	0
12/9/2020	48.838	26.724	24.719	0.648	2.884	0.163	0
12/10/2020	52.436	27.464	24.663	0.604	2.993	0.069	0
12/11/2020	51.645	26.974	24.771	0.618	2.818	0.110	0
12/12/2020	47.063	26.833	25.430	0.679	2.728	0.245	1
12/13/2020	48.498	24.749	22.986	0.634	2.855	0.118	0
12/14/2020	51.777	29.302	27.546	0.652	2.696	0.185	1
12/15/2020	55.543	29.064	27.903	0.622	2.702	0.109	0
12/16/2020	53.126	28.320	25.373	0.603	2.710	0.104	0
12/17/2020	57.803	30.546	28.519	0.598	2.643	0.091	0
12/18/2020	61.115	33.123	29.618	0.564	3.010	0.002	0
12/19/2020	60.352	35.389	32.196	0.604	2.990	0.056	0
12/20/2020	58.779	34.210	29.050	0.568	2.838	0.059	0
12/21/2020	57.713	35.795	31.333	0.611	2.878	0.108	0
12/22/2020	58.890	34.557	29.029	0.564	2.759	0.069	0
12/23/2020	57.814	30.651	28.562	0.598	2.489	0.107	0
12/24/2020	52.766	26.377	23.799	0.593	2.616	0.074	0
12/25/2020	50.513	27.347	24.931	0.630	2.684	0.122	0
12/26/2020	52.078	25.315	21.515	0.564	2.539	0.063	0
12/27/2020	54.416	28.866	27.592	0.629	2.455	0.140	0
12/28/2020	51.178	31.540	28.052	0.649	2.424	0.240	1
12/29/2020	54.148	28.582	26.140	0.606	2.335	0.141	0
12/30/2020	52.834	30.444	27.568	0.632	2.310	0.204	1

Chapter 5: Conclusions and Future Work

The proposed work is an attempt to forecast different weather conditions using a fusion of different forecasting and data mining techniques. Even though the rainfall is dependent on many parameters, the proposed model was able to get an impressive classification accuracy using limited parameters. Validations for time series forecasts are done using AIC, while the rest of the forecasted conditions are done using R-Square measure. Empirical results show that ARIMA provides the best fit for maximum and minimum temperatures; Neural Networks for humidity, UV index and cloud cover; and Bootstrap Forest for dew point prediction. Validation of classification is measured through accuracy, precision and recall. Miss-classification rates showed that random forest works best for rainfall classification. Table 5.1 provides a summary of the best fitted models for each weather condition.

Table 5.1: Summary of the best fitted models for each weather condition.

Weather Condition	Best Fitted Model	Evaluation Measurement	Input Variables
Maximum Temperature	ARIMA	AIC	N/A
Minimum Temperature	ARIMA	AIC	N/A
Dew Point	Bootstrap Forest	R-Square	Max Temp, Min Temp
Humidity	Neural Network	R-Square	Max Temp, Min Temp, Dew Point
UV index	Neural Network	R-Square	Max Temp, Min Temp, Dew Point, Humidity
Cloud Cover	Neural Network	R-Square	Max Temp, Min Temp, Dew Point, Humidity, UV Index
Pressure	Neural Network	R-Square	Max Temp, Min Temp, Dew Point, Humidity, UV Index
Rainfall	Bootstrap Forest	Misclassification Rate	Max Temp, Min Temp, Dew Point, Humidity, UV Index, Cloud Cover

The growth in IoT technology provides opportunities to retrieve real-time data from multiple smart devices. This data can be immediately processed and analyzed for predicting future weather variables with the implementation of data mining and machine learning techniques. In agriculture, weather prediction enables farmers to make informed decisions on planning and operation. The requirements for irrigation and crop growth are affected by different weather

conditions such as temperature, sunlight and rainfall are major effects on the crops. Further work is needed to develop science-based, region-specific information and technologies for agricultural and natural resource managers that enable climate-smart decision-making and transfer the information and technologies to users. The integration of IoT devices and the selected time series and data mining techniques can provide farmers with smart real-time monitoring and forecasting applications that will enable them to enhance crop growth, developments and yield.

References

Agilan, V & Nanduri, U. (2016). Modelling nonlinear trend for developing non-stationary rainfall intensity–duration–frequency curve. International Journal of Climatology. 37. 10.1002/joc.4774.

Ali M. F., Asklany S. A., El-wahab M. A., & Hassan M.A. (2019). Data Mining Algorithms for Weather Forecast Phenomena Comparative Study. *International Journal of Computer Science and Network Security*, *19(9)*, 76-81.

C. L. P. Chen & C.-Y. Zhang (2014). Data-intensive applications challenges techniques and technologies: A survey on big data. *Inf. Sci., vol. 275, pp. 314-347.*

Camacho, J., Rodriquez-Gomez, R. A., & Saccenti, E. (2017). Group-wise Principal Component Analysis for Exploratory Data Analysis. Journal of Computational and Graphical Statistics, 26(3), pp. 501-512, doi: 10.1080/10618600.2016.1265527

Cheng, X., Jing, W., Song, X., & Lu, Z. (2019). 5th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2019, Guilin, China, September 20–23, 2019, Proceedings, Part I. Pages 217-218. ISBN: 978-981-15-0117-3

Darji, M. P., Dabhi, V. K., and Prajapati, H. B. (2015). Rainfall forecasting using neural network: A survey, *2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad*, pp. 706-713, doi: 10.1109/ICACEA.2015.7164782.

Fahey, M. (2016, Aug 25). This weather app will give you the most accurate forecasts. *CNBC.* Retrieved from https://www.cnbc.com/2016/08/25/this-weather-app-will-give-you-the-most-accurate-forecasts.html

Fowler, K.M., Rauch, E.M., Henderson, J. & Kora, A.R. (2010). Re-Assessing Green Building Performance: A Post Occupancy Evaluation of 22 GSA Buildings.

Freedman, A. (2019). Weather is turning into big business. And that could be trouble for the public. *The Washington Post*. Retrieved from https://www.washingtonpost.com/business/2019/11/25/weather-is-big-business-its-veering-toward-collision-with-federal-government/

Gantz, J.& Reinsel, D. (2012). The digital universe in 2020: Big data bigger digital shadows and biggest growth in the far east. International Data Corporation.

Goger, B., Rotach, M., Gohm, A., Stiperski, I. & Fuhrer, O. (2016). Current Challenges for Numerical Weather Prediction in Complex Terrain: Topography Representation and Parameterizations. International Conference on High Performance Computing & Simulation (HPCS), Innsbruck, 2016, pp. 890-894, doi: 10.1109/HPCSim.2016.7568428.

Golchha N. (2015). Big data - The information revolution. *International Journal of Applied Research.* 1, 791-794.

Han, J., Kamber M. & Pei, J. (2000). Data Mining: Concepts and Techniques. Morgan & Kaufmann. ISBN: 9780123814807

Hemalata, P. (2013). Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System. International Journal of Computational Engineering Research 3(3).

International Data Corporation (2019). The Growth in Connected IoT Devices Is Expected to Generate 79.4ZB of Data in 2025, According to a New IDC Forecast. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS45213219

Javed, F., Afzal, M. K., Sharif, M., & Kim, B.(2018). Internet of Things (IoT) Operating Systems Support, Networking Technologies, Applications, and Challenges: A Comparative

Review. *IEEE Communications Surveys & Tutorials*, 20(3), pp. 2062-2100, doi: 10.1109/COMST.2018.2817685.

Jolliffe I.T. & Cadima J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374 (2065), doi:10.1098/rsta.2015.0202.

Juneja, A. & Das, N.N. (2019). Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, pp. 559-563. doi: 10.1109/COMITCon.2019.8862267

Kalyankar, M.A. & Alaspurkar, S.J. (2013). Data Mining Technique to Analyze the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering 3(2), 114-118.

Kumar R. (2013). Decision Tree for the Weather Forecasting. International Journal of Computer Applications 76(2):31-34.

Lin N., Noe D., & He X. (2006). Tree-Based Methods and Their Applications. Springer Handbook of Engineering Statistics. Springer Handbooks. Springer, London. doi:10.1007/978-1-84628-288-1_30

Mahmood M. R., Patra R. K., Raja R., and Sinha G.R. (2019). A Novel Approach for Weather Prediction Using Forecast Analysis and Data Mining Techniques. Innovations in Electronics and Communication Engineering, 479-489.

Marjani, M. et al. (2017) Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. IEEE Access, Vol. 5, pp. 5247-5261, doi: 10.1109/ACCESS.2017.2689040.

Mekanik, F. & Imteazmn, M.A. (2012). Artificial Neural Networks (ANN) modelling of spring rainfall using dual-climate indices for Victoria, Australia. Hydrology and Water Resources Symposium 2012, 557-561.

Meyers, M., Niech, C., & Eggers, W.D. (2015). Anticipate, sense, and respond: Connected government and the Internet of Things. Deloitte University Press, 2-7.

Networking and Information Technology Research and Development (NITRD) Program. (2015). Smart and Connected Communities Framework. Retrieved from https://www.nitrd.gov/sccc/materials/ scccframework.pdf.

Saba, T., Rehman, A. & AlGhamdi, J.S. (2017). Weather forecasting based on hybrid neural model. Applied Water Science 7, 3869–3874, doi:10.1007/s13201-017-0538-0.

Saba, T., Rehman, A., Altameem A., & Uddin M. (2014). Annotated comparisons of proposed preprocessing techniques for script recognition. Neural Comput Appl 25(6):1337–1347. doi:10.1007/s00521-014-1618-9

Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "Weather Forecasting using Incremental K-means Clustering", Vol. 8, 2014, pp. 142-147.

SAS Institute Inc. (2018). JMP® 14 Documentation Library. Cary, NC: SAS Institute Inc.

Sawale G. J. & Gupta S. R. (2013). Use of Artificial Neural Network in Data Mining For Weather Forecasting. *International Journal of Computer Science and Applications, 6(2).*

Saxena, A., Verma, N. & Tripathi K. C. (2013). A Review Study of Weather Forecasting Using Artificial Neural Network Approach. *International Journal of Engineering Research & Technology* (IJERT), 2(11).

Shah, U. & Garg, Sanjay & Sisodiya, Neha & Dube, Nitant & Sharma, Shashikant. (2018). Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques. doi: 776-782. 10.1109/PDGC.2018.8745763.

Sheikh, F., Karthick, S., Malathi, D., Sudarsan, J.S, & Chinnathambi, A. (2016). Analysis of Data Mining Techniques for Weather Prediction. Indian Journal of Science and Technology. 9(38), doi:10.17485/ijst/2016/v9i38/101962.

Singh, H. (2015). A State of the Art Survey of Data Mining Techniques for Software Engineering Data. International Journal of Applied Engineering Research. 10(55). 1512-1522. ISSN 0973-4562.

Sun Y., Song H., Jara A.J., and Bie R. (2016). "Internet of Things and Big Data Analytics for Smart and Connected Communities" in *IEEE Access, Vol. 4*, 766-773, doi: 10.1109/ACCESS.2016.2529723

Thomas, S. (2018). Data Cleaning in Machine Learning: Best Practices and Methods. *Einfochips. AI & Machine Learning*.

Tiw, M.A. (2013). Comparative Study of Backpropagation Algorithms in Neural Network Based Identification of Power System. International Journal of Computer Science & Information Technology (IJCSIT), 5(4).

Tsai, C.W.(2015) Big data analytics: A survey, Journal of Big Data, vol. 2(1), 1-32, doi: 10.1186/s40537-015-0030-3.

Want, R., Schilit, B.N., & Jenson, S. (2015). Enabling the Internet of Things. Computer, 48(1), 28-35, doi: 10.1109/MC.2015.12.

Wiston, M. & Mphale, K. (2018). Weather Forecasting: From the Early Weather Wizards to Modern-day Weather Predictions. Journal of Climatology & Weather Forecasting. Vol. 6, doi:10.4172/2332-2594.1000229.

Vita

Pedro A. Marquez was born in El Paso, Texas. The first children of Pedro Marquez and Blanca Ibarra. She studied her fundamental education and high school in Mexico to later pursue higher education in the United States. He graduated with a Bachelor of Science in Industrial Engineering in December 2016 from the University of Texas at El Paso. During his graduate school years, he was a research assistant in the Industrial Manufacturing and Systems Engineering department. His research was focused towards the Internet of Things and its applications.