Rainfall Forecasting by Using Machine Learning Models: A Case Study of TRNC

Saeid Mahmoudi

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Civil Engineering

Eastern Mediterranean University
August 2023
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science in Civil Engineering.

Assoc. Prof. Dr. Dr. Eriş Uygar
Chair, Department of Civil Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Civil Engineering.

Prof. Dr. Mustafa Ergil
Supervisor

Examining Committee

1. Prof. Dr. Mustafa Ergil

2. Prof. Dr. Umut Türker

3. Asst. Prof. Dr. Bertuğ Akıntuğ

ABSTRACT

Rainfall forecasting is crucial to making decisions, managing irrigation resources, and agriculture, and even predicting floods. Mediterranean rain regime is effective on Turkish Republic of Northern Cyprus (TRNC). The use of machine learning methods is widespread in many fields, including engineering, agriculture, transportation, and for the prediction. Several machine learning procedures were used in this study to build daily rainfall prediction models, including Decision Trees, Random Forests, Bagging Regressions, and Stacking Regressions. Five climatic parameters, average temperature, specific humidity, relative humidity, wind speed, and wind direction datasets were compiled on daily bases from 1995 to 2022 and used as input parameters after training and test phases. A comparison between the actual rainfall data gathered from NASA and the predicted outcome rainfall data from the machine learning models were used to determine the appropriate model which was having maximum accuracy and minimum error. In order to evaluate them, the statistical measures, R^2, MSE, and MAE were used. It is determined that, the two most accurate models for predicting daily rainfall as a whole of TRNC, were Stacking Regression, and Random Forest with R^2, 95.66, and 95.43, MSE 0.0428 and 0.045, and MAE 0.0821 and 0.0891, respectively. By applying the similar approach, based on the selected meteorological station as a representative for each region of TRNC, two appropriate machine learning models were found to be the best two fitted models that are Stacking Regression and Bagging Regression.

Keywords: rainfall forecast, machine learning, decision tree, stacking regression, random forest, bagging regression.

ÖZ

Yağış tahmini, kararlar almak, sulama kaynaklarını ve tarımı yönetmek ve hatta selleri tahmin etmek için çok önemlidir. Kuzey Kıbrıs Türk Cumhuriyeti (KKTC) Akdeniz yağış rejimi etkisindedir. Makine öğrenimi yöntemlerinin kullanımı, mühendislik, tarım, ulaşım ve tahmin için olmak üzere birçok alanda yaygındır. Rastgele Ormanlar, Torbalama Regresyonları ve İstifleme Regresyonları dahil olmak üzere günlük yağış tahmin modelleri oluşturmak için literatürde çeşitli makine öğrenimi prosedürleri kullanılır. Beş iklim parametresi, ortalama sıcaklık, özgül nem, bağıl nem, rüzgar hızı ve rüzgar yönü veri setleri, 1995'ten 2022'ye kadar günlük olarak derlendi ve eğitim ve test aşamalarından sonra girdi parametreleri olarak kullanıldı. NASA'dan toplanan gerçek yağış verileri ile makine öğrenimi modellerinden tahmin edilen sonuç yağış verileri arasındaki karşılaştırma, maksimum doğruluk ve minimum hataya sahip uygun modeli belirlemek için kullanıldı. Bunları değerlendirmek için istatistiksel ölçümler, R^2, MSE ve MAE kullanıldı. KKTC genelinde günlük yağış tahmininde en doğru iki modelin sırasıyla R^2 95.66 ve 95.43, MSE 0.0428 ve 0.045 ve MAE 0.0821 ve 0.0891 ile İstifleme Regresyonu ve Rastgele Orman olduğu belirlenmiştir. Benzer yaklaşım uygulanarak, temsili olarak seçilen meteoroloji istasyonuna dayalı olarak KKTC'nin her bir bölgesini temsilen seçilen meteoroloji istasyonuna dayalı olarak iki uygun makine öğrenme modelinin İstifleme Regresyonu ve Torbalama Regresyonu olduğu bulundu.

Anahtar Kelimeler: yağış tahmini, makine öğrenimi, karar ağacı, istifleme regresyonu, rastgele orman, torbalama regresyonu.

ACKNOWLEDGMENTS

I could not have undertaken this journey without the support of Prof. Dr. Mustafa Ergil for his invaluable supervision, tolerance, assistance, and persistence throughout my thesis. I would like to extend my sincere thanks to him.

I am grateful to my family, who generously supported me at every step.

Additionally, this endeavor would not have been possible without the facilities that Civil Engineering Department supplied. I would like also to thank my classmates, and my friends for their patience and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Ω	Weights of the Base Models
ω_m	Weight of the Models
SS_{res}	Sum of Squared Residuals
SS_{tot}	Difference Between Actual Values and Mean of Predicted Values
$\bar{y}$	Average of the Predicted Values
y_i	Predicted Value for the i^{th} Observation
B_0	Linear Intercept
B_1	Linear Regression Coefficient
E	Error
$f(x_i)$	Actual Value for the i^{th} Observation
n	Number of Observations in the Dataset
R^2	Coefficient of the Determination of the Regression
X	Independent Variable
$X_{normalized}$	Normalized Value
X_{min}	Minimum Value of each Parameter
X_{max}	Maximum Value of each Parameter
$y_{p,i}$	Stacking Regression
AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
CV	Coefficient of Variation
IQR	Interquartile Range
MAE	Mean Absolute Error

ML	Machine Learning
MSE	Mean Squared Error
NASA	National Aeronautics and Space Administration
NWP	Numerical Weather Prediction
r	Correlation Coefficient
R^2	R-Squared
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
TRNC	Turkish Republic of Northern Cyprus

Chapter 1

INTRODUCTION

1.1 Background

All living things on the Earth, including flora, fauna, and people, require water, which is limited in quantity. As the population is increasing whereas the total amount of water is not increasing (based on the water cycle), it becomes crucial to use it beneficially. Rainfall is one of the precipitation types that happens more than others. It is a critical meteorological parameter that influences various aspects of human life, such as agriculture, water resource management, flood prediction, and urban planning. There are four types of rainfall based on intensity: extremely heavy, heavy, moderate, and light, where extremely heavy and heavy rainfall may cause flooding.

Accuracy of rainfall prediction helps to provide optimum decisions, funding strategies, and mitigated potential risks associated with extreme weather events. These methods include decision-making on saving water in heavy rainfalls and using some strategies in drought years. Despite the traditional methods for rainfall prediction, such as numerical weather prediction, with the availability of large-scale meteorological data and progress in machine learning methods, the use of these recent methods as forecasting tools are unavoidable.

Analyses of hydro-climatological variations, trend patterns, and intensity curves performed on TRNC climatic variables (Akıntug and Baykan 2000; Sharifi and Ergil 2006; Seyhun and Akıntug 2013; Payab and Türker 2018; Abdulkadir et al., 2020).

1.2 Problem Statement

Forecasting the amount of rainfall is a crucial task in meteorology and has significant implications for many sectors. In many regions, traditional weather prediction often faces limitations in accuracy and reliability, particularly at local or regional scales. Rainfall amounts are complex, non-linear patterns in meteorological data, leading to suboptimal prediction performance. The methods used in meteorology offices for predicting weather are not highly accurate. They use some traditional methods. As a first problem, they use limited data for forecasting, such as satellite and radar observations, which leads to uncertainty in prediction. Second, weather can change rapidly, particularly during severe weather events such as thunderstorms, tropical cyclones, or frontal passages. These rapid changes can be challenging to predict accurately and slight errors in initial conditions require 24/7 experts to forecast again. Third, weather forecasting comprises human interpretation of data, such as output data, which needs meteorologists' experience and judgment. Human error in interpreting data leads to inaccuracy in forecasting. Nowadays, because of progress in Artificial Intelligence (AI), it is becoming increasingly more popular for governments and researchers have focused on benefiting from technologies and increasing their knowledge. There are many types of machine learning, a subset of AI, which have shown success in various applications including climate forecasting.

1.3 Significance of this Study

In general, weather prediction has a significant impact on human life, animals, and plants.

1. Safety and Preparedness: Accurate weather forecasts enable people to manage their plans and prepare for severe weather events such as hurricanes, tornadoes, thunderstorms,

floods, and snowstorms. As a result, people take action and make decisions to protect lives, properties, and infrastructure, including stocking up on essential supplies and activating emergency response plans.

2. **Agriculture and Food Production:** Weather prediction plays a vital role in agriculture and food production, as it helps farmers make informed decisions about crop management, planting, and harvesting. They provide information about precipitation, temperature, humidity, and other meteorological factors that influence crop growth and yield based on a knowledge of the weather forecast. As a result, farmers are able to optimize their irrigation, fertilization, and pest control strategies, which leads to higher crop yields and greater food production. Drought primarily affects agriculture, since it can give rise to agricultural yield losses and water shortages (Anjum et al., 2010).

3. **Transportation and Travel:** Weather prediction is crucial for the transportation industry, including aviation, maritime, and ground transportation. Accurate weather forecasts help in planning routes, scheduling flights, and managing operations efficiently. Transportation and travel can experience delays, cancellations, and safety risks due to severe weather events such as storms, fog, and snowfall. Weather prediction enables better decision-making and mitigates risks.

4. **Energy and Utilities:** The energy and utilities sector rely heavily on weather prediction to regulate energy generation, transmission, and consumption. Climate forecasts provide insight into the temperature, wind speed, solar radiation, and other parameters affecting energy demand, renewable energy, and grid stability. As a result, energy resources can be planned and managed more efficiently and cost-effectively.

5. Economic Planning and Risk Management: The accurate prediction of weather is essential to economic planning and risk management for a range of industries, including construction, tourism, insurance, and retail. Accurate weather forecasts assist businesses in planning operations, managing risks, and optimizing resources according to expected weather conditions. For instance, construction companies can plan their schedules, tourism operators can adjust their offerings, insurance companies can assess risks, and retailers can manage inventory and pricing strategies based on weather forecasts.

6. Daily Life and Personal Planning: The weather has a significant impact on daily life and personal planning. A detailed weather forecast contributes to a person's ability to plan outdoor activities, sports, and gardening as well as the choice of clothing. Weather forecasts can affect personal safety by guiding decisions on clothing, hydration, and exposure to extreme weather.

1.4 Research Objectives

The aim of this study is to forecast daily rainfall by utilizing historical daily data through AI techniques, such as Decision Tree, Random Forest, Bagging Regression, and Stacking Regression. The algorithms must be compatible with the input data, so as to group and label them correctly. There are several features in the collected data where the target is to construct a model that has a high R-Squared (R^2), low Mean Squared Error (MSE) and Mean Absolute Error (MAE). This study has the following objectives:

1. Creating dependable and precise rainfall prediction models,
2. Investigating predictors and selecting important features,
3. Evaluating model performance,
4. Exploring the concept of model, and
5. The ability to generalize and transfer models.

1. Creating dependable and precise rainfall prediction models: The primary research goal in rainfall prediction with machine learning is to create models that can accurately and reliably forecast rainfall patterns on various temporal and spatial scales. The improvement of rainfall prediction models will be achieved by exploring different machine learning algorithms, feature engineering, and model optimization techniques. This is done by comparing R^2.

2. Investigating predictors and selecting important features: Another objective of this study is to discover and analyze the most dominant predictive variables or features that have a significant impact on rainfall patterns. This may involve analyzing various meteorological and environmental variables, such as average temperature, relative and specific humidity, wind speed, wind direction, and topography, and determining their relative importance in predicting rainfall. Feature selection techniques, such as statistical methods, dimensionality reduction, and machine learning algorithms, were used to identify the most relevant features for accurate rainfall prediction.

3. Evaluating model performance: Evaluation of rainfall prediction models' performance and uncertainty is a crucial research objective. A variety of performance metrics are used to evaluate the accuracy, reliability, and robustness of the developed models. In this study, R^2, MSE, and MAE are being used for comparison.

4. Exploring the concept of the model: It is important to investigate how to interpret machine learning models when predicting rainfall. In order to accomplish this, it is essential to understand how the developed models work. During this phase, it is necessary to interpret the relationships between input features and predicted rainfall, as well as to explain how the model reaches its conclusions. Machine learning models that

are easy to understand and interpret are critical to gaining trust and acceptance from users, decision-makers, and stakeholders.

5. The ability to generalize and transfer models: It is also important to assess how rainfall prediction models transfer and generalize across different spatial and temporal scales. An important part of this is evaluating how well the developed models can generalize their knowledge from one region or time period to another. Furthermore, it involves adapting to different climatic and environmental conditions. It is important to ensure that, rainfall prediction models can be applied in different regions with different data availability and characteristics, as well as ensure their practical applicability and usability.

1.5 The Study Region

Cyprus is an island located in the northeastern part of the Mediterranean Sea. It has a surface area of 9251 km^2 (Akıntug and Baykan, 2000).

There are about 1.258 million people living in Cyprus (until May 2023), of whom 313,600 live in TRNC. This number has probably risen to 500,000 since the population is expected to grow at 0.67% by 2023 (World Population Website).

Chapter 2

THEORY

2.1 Hydrological Cycle

The total global water resources constitute approximately 1.385 billion km^3, with 96.5% of their volume (1.338 billion km^3) contained in oceans, being the largest water storing volume (Shiklomanov and Rodda 2003).

The hydrological cycle is a continuous process where water circulates between the Earth's surface, atmosphere, and underground reservoirs. It is a vital natural process that regulates the distribution and the availability of water on the Earth. The hydrological cycle components are evaporation, condensation, precipitation, infiltration, runoff, and transpiration. When water droplets in clouds become heavy enough, they fall back to the Earth's surface as precipitation, which can be rain, snow, sleet, or hail as shown in Figure 2.1.

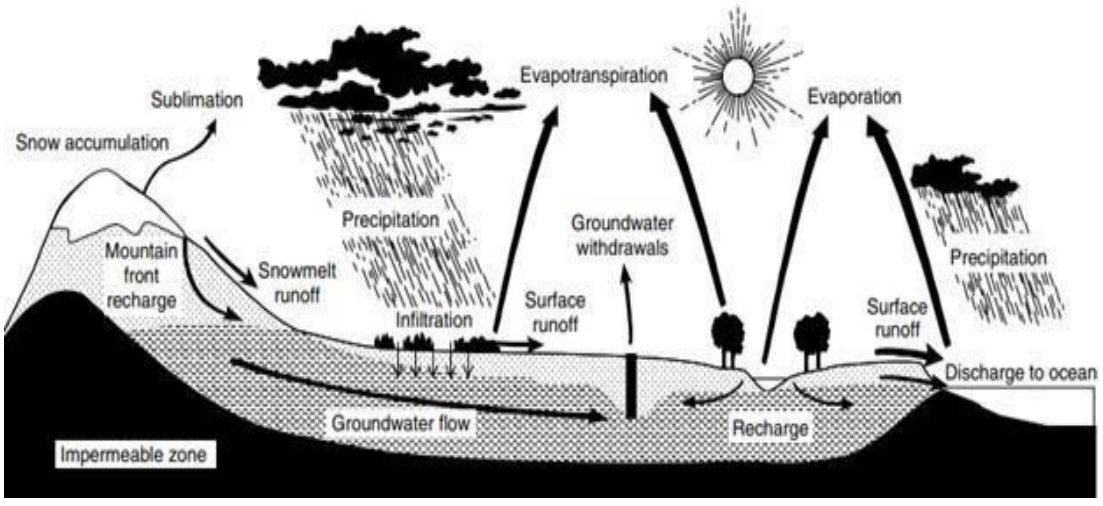

Figure 2.1: Hydrological (water) cycle (Pagano and Sorooshian, 2002)

2.2 Rainfall

Rainfall is the amount of precipitation in the form of water droplets that fall from the atmosphere to the Earth's surface. It is a vital component of the water cycle and plays a crucial role in the Earth's climate system. Rainfall can occur in various forms, such as drizzle, rain, showers, and thunderstorms. It can be influenced by many factors, including temperature, humidity, air pressure, and wind patterns.

Rainfall is essential for the survival of plants, animals, and humans, as it replenishes freshwater resources, supports agriculture, and sustains ecosystems. It also has significant impacts on human activities, such as transportation, construction, and recreation.

Rainfall is typically measured in length units, such as millimeters (mm) or inches (in) since is often expressed as the depth of water that would accumulate over 1 m^2 or 1 ft^2 on a flat surface during a specific time period, usually one day. Rainfall can vary greatly depending on location, climate, and weather patterns, with some areas receiving heavy rainfall and others experiencing droughts. The study of rainfall patterns, managing water resources, and predicting and mitigating extreme weather events such as floods and droughts.

2.2.1 Types of Rainfall

Rainfall can be categorized into different types based on various factors, such as its origin, its characteristics, and its formation process.

2.2.1.1 Convectional Rainfall

Convective rainfall, also known as convectional rainfall, is a type of precipitation that forms when warm air rises and cools. This leads to water vapor condensation into liquid water droplets or ice crystals. This process is typically associated with localized and short-lived weather events, such as thunderstorms or showers, and is driven by the vertical movement of air known as convection. As warm air rises, it cools and reaches its dew point, at this

point the moisture in the air condenses and falls to the ground as precipitation.

Convective rainfall is common in many parts of the world, particularly in tropical and subtropical regions where there is a sufficient amount of moisture and heat to fuel convection processes. Intensity can result in heavy rainfall over a short period of time, which occasionally leads to flash flooding. Convective rainfall plays a significant role in the Earth's water cycle, as it replenishes freshwater sources on the surface and helps regulate precipitation distribution around the globe. Figure 2.2 illustrates the instability of convectional rainfall clouds.

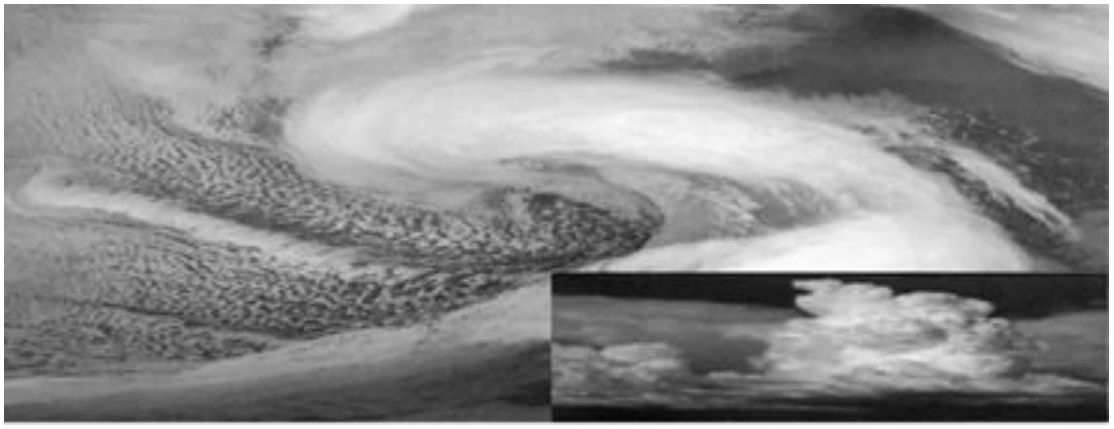

Figure 2.2: Formation of convective rainfall (Collier, 2003)

2.2.1.2 Orographic Rainfall

This type of rainfall occurs when moist air rises over elevated terrains, such as mountains or hills (Gray and Seed, 2006). As the air is lifted, it cools and condenses, leading to the formation of clouds and precipitation on the windward side of the mountains, also known as the rain shadow, often experiencing reduced rainfall as the air descends and warms, hence, resulting in drier conditions.

2.2.1.3 Cyclonic or Frontal Rainfall

This rainfall type is associated with air circulation around low-pressure systems, such as tropical cyclones or mid-latitude depressions. As the air circulates counterclockwise (in the Northern Hemisphere), it converges and rises, leading to the formation of clouds and precipitation in the vicinity of the low-pressure center. Cyclonic rainfall can be widespread and intense, with heavy precipitation and the potential for flooding. It is a log-normal distribution type that depends on the area of precipitation (Cheng and Qi, 2002). The formation of cyclonic rainfall is shown in Figure 2.3.

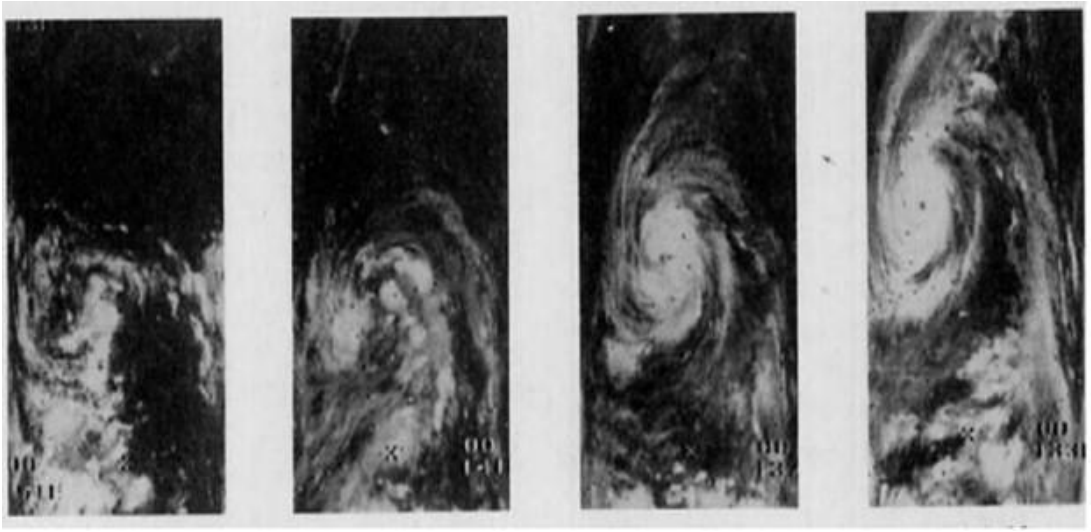

Figure 2.3: Cyclonic rainfall cloud formation (Rodgers and Adler, 1981)

2.2.2 Types of Rainfall Based on Amount

1. Light Rain

Light rain refers to gentle precipitation that falls in small droplets and does not result in heavy water accumulation. It is usually characterized by a low rainfall rate and is often described as drizzle or mist.

Light rain may not have significant impacts on the environment or human activities and may not require extensive protective measures. Typically, light rain is considered to be rainfall with an accumulation of up to 2.5 millimeters (0.1 inches) in 24 hours.

2. Moderate Rain

Moderate rain is characterized by a moderate intensity of precipitation, with larger droplets falling faster than light rain. It can cause a moderate accumulation of water on the ground and may have noticeable impacts on the environment and human activities. Moderate rain can cause wet roads, reduced visibility, and minor flooding in low-lying areas. Moderate rain is defined as rainfall with an accumulation between 2.5 millimeters (0.1 inches) and 7.6 millimeters (0.3 inches) in 24 hours.

3. Heavy Rain

Heavy rain is characterized by high intensity of precipitation, with large droplets falling quickly. It can cause a significant accumulation of water on the ground, causing flooding, erosion, and other impacts on the environment, infrastructure, and human activities. Heavy rain can lead to flash floods, landslides, and disruption of transportation, agriculture, and other sectors. Heavy rain is often defined as rainfall with an accumulation between 7.6 millimeters (0.3 inches) and 50 millimeters (2 inches) in 24 hours.

4. Extremely Heavy Rain

Extremely heavy rain is an extreme type of rainfall, characterized by an exceptionally high intensity of precipitation, with huge droplets falling at a quick rate. It can result in excessive water accumulation, causing severe flooding, widespread damage, and disruption of normal life.

Extremely heavy rain can be associated with extreme weather events, such as tropical cyclones, monsoons, or severe thunderstorms. Extremely heavy rain is typically defined as rainfall, with an accumulation exceeding 50 millimeters (2 inches) in 24 hours.

2.2.3 Rainfall Measurement

Rainfall measurement is a process of quantifying the amount of precipitation, typically in the form of rain, that falls over a specific area during a certain period of time. Rainfall measurements are important for a variety of applications, such as weather forecasting, hydrological modeling, agriculture, water resource management, and flood prediction. Several methods commonly used for rainfall measurement, are including.

2.2.3.1 Simple Rain Gauges

Simple rain gauges are simple devices that collect rainfall directly and measure the amount of water that accumulates it in its container. The measuring instrument has a diameter of 203 mm. Rain gauge measurement is less effective and less accurate than other methods. They typically consist of a cylindrical or funnel-shaped container with markings or a measuring scale to indicate the amount of rainfall collected. Simple rain gauges are typically placed in open areas away from buildings or trees to minimize interference from surrounding structures.

2.2.3.2 Tipping Bucket Rain Gauges

This type of rain gauge uses a tipping mechanism to measure rainfall. There is no doubt that, this method is more effective and more accurate than the simple method. Rainwater is collected in a funnel, and as it fills up, it tips a lever, which then empties the funnel and records the tip as a unit of rainfall. Details of tipping bucket rain gauges demonstrate in Figure 2.4.

Furthermore, tipping bucket rain gauges are widely used due to their accuracy and ability to provide real-time data.

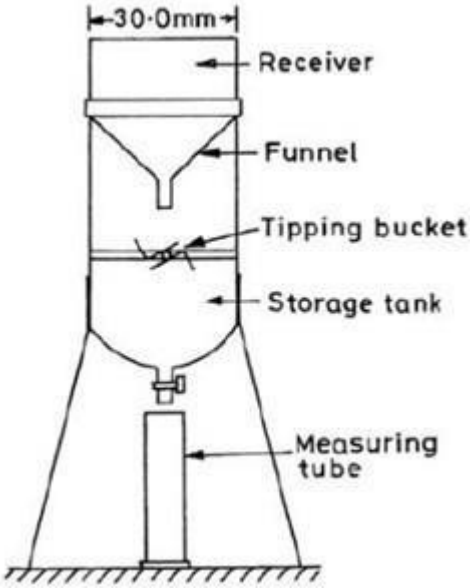

Figure 2.4: Schematic diagram of a typical tipping bucket rain gauge (Froehlich and Dhawan, 2017)

2.2.3.3 Radar and Satellite-Based Measurements

Radar and satellite-based measurements use remote sensing techniques to estimate rainfall from space. Radar measures the amount of precipitation by emitting radio waves and measuring their reflection of raindrops or snowflakes. Satellite-based measurements use sensors on satellites to estimate rainfall by analyzing the properties of clouds, such as their brightness and temperature. The process of radar and satellite measurements of rainfall is shown in Figure 2.5. These methods can provide rainfall data over large areas but may have limitations in accuracy and resolution.

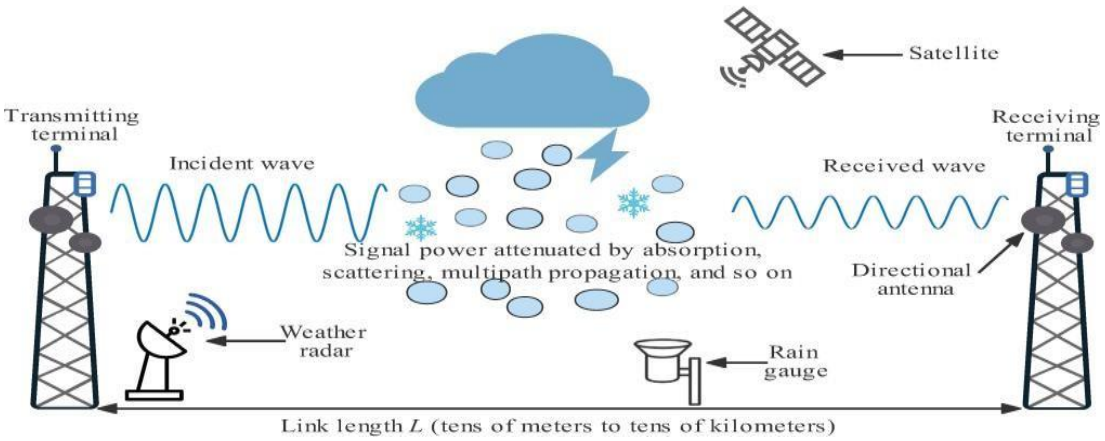

Figure 2.5: Basic operating principle of (Commercial Microwave Links) CML-based rainfall measurement (Lian et al., 2022)

2.2.4 Regime of Rainfall

The regime of rainfall can vary greatly depending on factors such as geography, topography, latitudes, and prevailing weather patterns, and it is an important consideration in understanding local climate and its impacts on various aspects of the environment and human activities.

The regime of rainfall refers to the pattern or characteristics of the precipitation. This includes the frequency, duration, intensity, and distribution of rainfall over a specific area or region. It is a significant aspect of weather and climate and can have significant impacts on ecosystems, agriculture, water resources, and human activities. There are six main rainfall regimes suggested by Haurwitz and Austin in 1944. These are:

i. Monsoon Rainfall Regime

In this regime type, a seasonal reversal of wind patterns leads to heavy rainfall during a particular season of the year. Monsoonal rainfall regimes are typically found in regions near large landmasses, such as South Asia, Southeast Asia, and Northern Australia.

ii. Tropical Rainfall Regime

This regime occurs in tropical regions near the Equator, where rainfall is generally abundant and evenly distributed throughout the year. These regions often have high temperatures and high humidity, leading to frequent showers or thunderstorms.

iii. Mediterranean Rainfall Regime

This type of regime is characterized by wet winters and dry summers. It is typically found in Mediterranean climate zones, such as California, central Chile, and the Mediterranean Basin. Most rainfall occurs during the cooler months, with little to no precipitation during the warm season. The Mediterranean rainfall regime is defined as zone 4 being extremely dry (Haurwitz and Austin, 1944). In TRNC, winters are cold and rainy, while summers are hot and humid, so TRNC falls under the Mediterranean climate regime.

iv. Continental Rainfall Regime

This regime occurs in large continental areas, where rainfall is usually moderate to low and distributed unevenly throughout the year. These regions often experienced distinct seasons, with higher rainfall during spring and summer, and drier conditions during fall and winter.

v. Maritime Rainfall Regime

This type of regime is characterized by very low rainfall, usually less than 250 mm (10 inches) per year and is typical of desert regions such as the Sahara, Arabian Peninsula, and Australian Outback. Rainfall in desert regimes is highly sporadic and unpredictable, with long periods of drought.

vi. Equatorial (Polar) Rainfall Regime

This regime occurs in polar regions, such as Arctic and Antarctic, where precipitation falls mostly as snow due to cold temperatures. Precipitation in the Equatorial regime is generally low and occurs throughout the year, with slightly higher amounts during the summer months.

Chapter 3

LITERATURE REVIEW

3.1 Artificial Intelligence in Civil Engineering Studies

Artificial Intelligence (AI) has the potential to greatly impact civil engineering studies by improving the efficiency, accuracy, and sustainability of various engineering processes. A wide range of applications of AI is used in civil engineering, including structural analysis and design, construction automation, geotechnical engineering, construction monitoring, traffic management, hydrological studies, sustainability, and green building.

AI algorithms can process large amounts of data quickly and provide insights into structural performance, leading to more efficient, cost-effective designs, by increasing safety on construction sites, improving project management, optimizing maintenance schedules, extending pavement life, identifying distress types, recommending treatment strategies, improving road safety, and reducing the carbon emissions by optimizing traffic signals and coordinating the traffic flows.

Recent studies show that, AI is increasingly used in civil engineering for sustainable development. More than 105 publications between 1995 and 2021 were utilized AI in civil engineering. An overview of the recent 26 years can be seen in Figure 3.1 (Manzoor et al., 2021).

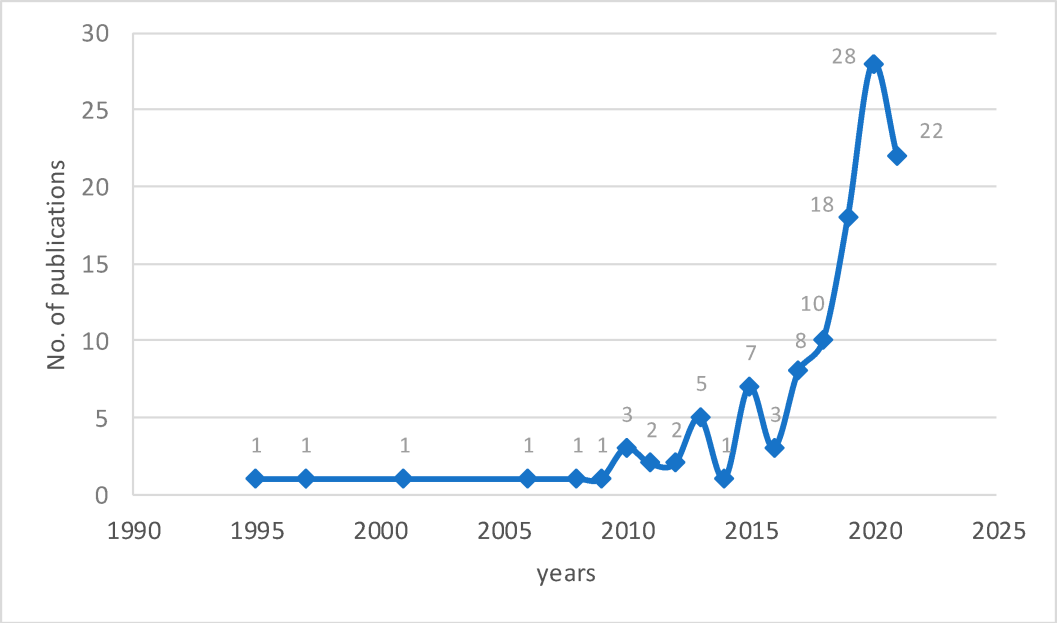

Figure 3.1: Annual publication trend of AI in civil engineering toward sustainable development (Manzoor et al., 2021)

3.1.1 Predictive Maintenance

AI can predict the maintenance needs of infrastructure assets, such as bridges, roads, and water systems. This is done by analyzing data on factors such as usage, weather, and wear and tear. This helps in planning and prioritizing maintenance activities, extending asset lifespan, and reducing downtime.

3.1.2 Hydrological Modeling

AI is capable of modeling and predicting hydrological processes such as rainfall, runoff, and river flow. Machine learning algorithms can analyze historical data on precipitation, evaporation, soil moisture, and other parameters to develop predictive models that can be used for flood forecasting, water resource management, and climate change impact assessment. In addition, AI can also help analyze and integrate data from remote sensing, weather stations, and other sources to improve hydrological modeling accuracy.

3.1.3 Flood Prediction and Management

Artificial intelligence can analyze rainfall, river flow, topography, and other factors to predict floods and assist in flood management. Machine learning algorithms can process real-time data from various sources and provide early warnings, flood forecasts, and flood risk assessments. AI can also help to optimize flood control strategies, such as reservoir operation, floodplain zoning, and emergency response planning by minimizing flood damage and protecting vulnerable areas.

3.1.4 Water Quality Monitoring

AI can analyze data on water quality parameters such as pH, temperature, dissolved oxygen, and pollutant concentrations to monitor and predict water quality. Machine learning algorithms can process data from sensors, water quality databases, and other sources to detect water quality trends, identify pollution sources, and develop predictive models for water quality management. AI can also help in optimizing water treatment processes and identifying strategies for water pollution control.

3.1.5 Climate Change Impact Assessment

AI can analyze climate data, land use, and other factors to assess how climate change will affect hydrological processes and pavement performance. Machine learning algorithms can process data from climate models, remote sensing, and other sources to develop predictive models that can help in designing resilient pavements and water management strategies by adapting the changing climate condition.

Abdulkadir, et. al (2020) researched and compared the performance of some non- linear models in predicting daily rainfall at Ercan Airport in TRNC. The study used 10 years of meteorological data and applied three modeling techniques: Multi-Linear Regression (MLR), artificial neural network (ANN), and Adaptive Neuro-Fuzzy Inference System (ANFIS).

The results showed that, they could use all three models for rainfall prediction. However, the ANFIS model had the lowest mean square error and mean absolute percentage error, making it the most accurate. The study concluded that the developed ANFIS model could serve as are liable rainfall forecasting tool for Ercan Airport (Abdulkadir et al., 2020).

Tiwari, et. al (2020) discuss the importance of rainfall prediction in agriculture and analyzes the rainfall pattern by applying machine learning algorithms for Indian states. The authors tried various regression algorithms like XGBoost, linear regression, support vector regressor, gradient boosting regressor, ada boost, ridge regressor, lasso regressor, bagging regressor, and some ensemble methods like random forest regressor and elastic net to predict monthly and yearly rainfall. They found neural networks performed better than other machine learning algorithms, and trying different hyperparameters gave better results. The paper concludes with the analysis of the results obtained from the machine learning models (Tiwari and Singh, 2020).

Kumar, R. (2013) studied a decision tree as a method for predicting weather events like fog and rain. The author discussed different classification methods and their advantages and disadvantages. Kumar then used a decision tree to classify weather events used three parameters: the average temperature, the average humidity, and the sea level. The result shows that, out of 72 test instances, 46 tests were classified properly, which gave a kappa statistic of 0.0584. The author suggested that for further improvement of the results can be done by taking more attributes into the model and increasing the training of that dataset (Kumar, R. 2013).

Geetha, et. al (2014) studied using decision trees as a data mining technique for predicting weather phenomena like fog, rainfall, cyclones, and thunderstorms. The paper discusses the advantages of decision trees and compares them to other classification methods. The authors collected weather data from 2013 to 2014 and used RapidMiner, an open-source data mining tool, to implement their model. The paper concludes that, decision tree is an effective method for weather prediction and can be applied as a life-saving tool for informing the public about natural calamities. The accuracy rate for 2014 was 80.67%. The authors suggested that, further improvements can be made by incorporating other soft computing techniques, such as fuzzy, genetic algorithms, and artificial neural networks (Geetha and Nasira, 2014).

Chapter 4

MACHINE LEARNING

One of the subsets of Artificial Intelligence (AI) is Machine Learning (ML). It allows computers to learn and make predictions or decisions without being explicitly programmed. ML algorithms are designed to analyze large amounts of data, identify patterns, and learn from experience in order to improve their performance over time.

4.1 Use of Machine Learning

Machine learning has many applications across various domains, such as natural language processing, computer vision, speech recognition, recommendation systems, fraud detection, and autonomous vehicles. It has the potential to revolutionize many industries. It has been increasingly used in business, healthcare, finance, transportation, and other areas to improve decision-making, automate tasks, and enhance efficiency. However, it also raises ethical, privacy, and security concerns, as it relies on data and may have biases or unintended consequences if not carefully designed and monitored.

Artificial Neural Network (ANN) for rainfall prediction is one of the most suitable and reliable systems for the rainfall prediction that has already benefited the operators for rainfall prediction (Shaikh and Sawlani, 2017).

4.2 Types of Machine Learning

It includes: 1. supervised learning 2. unsupervised learning 3. semi-supervised learning

4. reinforcement learning.

4.2.1 Supervised Machine Learning

Supervised machine learning is a type of machine learning in which an algorithm learns to make predictions or decisions by analyzing labeled training data. In supervised learning, the algorithm is trained on a set of input/output pairs, where the inputs are the features or attributes of the data, and the outputs are the labels or target values that the algorithm is trying to predict.

The goal of supervised learning is to learn a model that can accurately predict the output of new, unseen input data. The model is trained by minimizing the difference between the predicted outputs and the actual outputs in the training data. This is typically done using a loss function and an optimization algorithm that adjusts the model's parameters to minimize loss (mean absolute error).

Examples of supervised learning algorithms include linear regression, logistic regression, decision trees, random forests, Support Vector Machines (SVMs), and neural networks.

4.2.2 Unsupervised Machine Learning

Unsupervised machine learning is a type of machine learning in which an algorithm learns to identify patterns or structures in unlabeled data. Unlike supervised learning, there are no target labels or outputs for the algorithm to predict. Instead, the algorithm must find meaningful patterns and relationships in the data on its own.

The goal of unsupervised learning is to identify hidden structures or groups in the data, such as clusters, associations, or patterns. This can be done through various techniques, such as clustering, dimensionality reduction, and anomaly detection.

Algorithms are used to group similar data points together into clusters based on their similarities in terms of their features. Dimensionality reduction algorithms are used to reduce the number of features in the data by preserving the variance.

Anomaly detection algorithms are used to identify outliers or anomalies in the data that do not fit normal patterns.

Examples of unsupervised learning algorithms include K-means clustering, hierarchical clustering, Principal Component Analysis (PCA), and auto-encoders.

4.2.3 Semi-Supervised Machine Learning

Semi-supervised machine learning is a type of machine learning that combines both supervised and unsupervised learning techniques to learn from both labeled and unlabeled data. In semi-supervised learning, the algorithm is trained on a small amount of labeled data and a large amount of unlabeled data.

The goal of semi-supervised learning is to use the labeled data to guide the learning process, while also leveraging the large amount of unlabeled data to discover patterns and structures in the data that may not be apparent from the labeled data alone. This can be particularly useful in situations where labeled data is scarce or expensive to obtain.

Semi-supervised learning algorithms typically work by first using the labeled data to train a supervised learning model to determined unlabeled data. The labeled and unlabeled data are then combined to train an adaptive model that incorporates both labeled and unlabeled data.

Examples of semi-supervised learning algorithms include self-training, co-training, and multi-view learning. Semi-supervised learning is commonly used in applications, such as natural language processing, image classification, and speech recognition, where labeled data is often limited and expensive to obtain.

4.2.4 Reinforcement Machine Learning

Reinforcement machine learning is a type of machine learning in which an algorithm learns to make decisions through trial and error by interacting with an environment. In reinforcement learning, the algorithm is not given labeled data or explicit instructions on what actions to take but instead learns by receiving feedback in the form of rewards or punishments for its actions.

The goal of reinforcement learning is to learn a policy, which is a mapping from states to actions, that maximizes the expected cumulative reward over time. Reinforcement learning algorithms are often modeled as Markov Decision Processes (MDPs) or Partially Observable Markov Decision Processes (POMDPs), which formalize the decision-making process as a sequence of states, actions, and rewards.

Examples are Q-learning, SARSA, and policy gradient methods. Reinforcement learning is commonly used in various applications, such as robotics, game playing, and autonomous driving.

4.3 Regression Models

Regression models are type of statistical analysis that is used in machine learning and statistical modeling so as to model the relationship between a dependent variable and one or more independent variables.

The goal of regression analysis is to estimate the parameters of the regression model in order to make predictions or infer relationships between variables.

In regression models, the dependent variable, also known as the response or target variable, is the variable being predicted or modeled. The independent variables, also called predictor variables or features, are the variables used to explain or predict the behavior of the dependent variable. The relationship between the dependent and independent variables is typically represented by an equation of a function, which is learned from the available data during the training phase.

4.3.1 Linear Regression

Linear regression assumes a linear relationship between the dependent and independent datasets by fitting a straight line that best represents these datasets.

Simple linear regression is formulated as:

$$y_i = B_0 + B_1 X + E \tag{4.1}$$

where:

- y_i is the dependent variable of the predicted value
- B_0 is the intercept when x is 0
- B_1 is the regression coefficient
- X is the independent variable
- E is the error of the estimate

4.3.2 Decision Trees Regression

A decision tree is a supervised machine learning algorithm that can be used for classification and regression tasks. It is a flowchart-like tree structure where an internal node represents a decision based on the value of a specific feature, and the branches illustrate the possible outcomes or decisions that can be made based on the values of the features. The tree leaves represent the final predicted outcome or value. Depending on the algorithm, each node may have two or more branches (Shaikh and Sawlani, 2017).

Typically, they are used for tasks involving decision-making or classification, where an input data point's features are used to predict its class or category.

Classification And Regression Trees (CART) is a data-mining algorithm suggested by Breiman et al., (1984). The decision tree has different algorithms which are: CART, C4.5, and ID3 (Singh et al., 2014). For regression tasks, where a continuous value is predicted for a given input, decision trees can also be used (Choubin et al. 2018).

In this study, some regression trees were used to forecast the amount of rainfall. Simple decision tree is less accurate (Breiman et al., 1984). The process of building a decision tree involves repeatedly splitting the data into subsets based on the values of the features. Then, it makes decisions at each node based on certain criteria, such as entropy of Gini impurity for classification tasks, and Mean Squared Error (MSE) or Mean Absolute Error (MAE) for regression tasks (Timofeev 2004).

After the decision tree is built, it is used for predicting new, unseen data points (Timofeev 2004). The predicted outcome or value at the leaf node is then used as the final prediction.

Decision trees can be further extended to other types of trees, such as random forests, which are ensembles of decision trees, and Gradient Boosting Machines (GBMs), which is an iterative model that combines multiple decision trees to improve prediction accuracy (Timofeev 2004). Decision trees and their variants are widely used in various applications, including classification, regression, anomaly detection, and recommendation systems (Seera et al., 2012).

Suppose that a training dataset: $P = \{(x_1, y_1), (x_2, y_2), \ldots . (x_i, y_i)\}$, $x_i \varepsilon R^n$.

$$Object = min\frac{1}{n} Loss(f(x_i) - y_i) \quad (4.2)$$

where:

- n is the number of observations in the dataset
- $f(x_i)$ is the actual value for the i^{th} observation
- y_i is the predicted value of the dependent variable for any given value of (x)

The goal of the regression is to minimize the loss function (error) (Dong et al., 2021).

4.3.3 Random Forest Regression

Random forest is a machine learning that uses multiple decision trees to make predictions. It combines the outputs of individual decision trees to improve prediction accuracy, robustness, and generalization performance compared to using a single decision tree.

The main idea behind random forest is to create a collection of decision trees, where each tree is trained on a random subset of the training data, and also a random subset of the features. This process is known as bootstrapping and feature bagging.

The building blocks of decision tree-based modelling approach, i.e., the Random Forest model is bootstrapping, and aggregation called bagging (Breiman, 1996; Schapire et al., 1998). The subsets of data and features are randomly sampled with replacement, meaning that, the same data point or feature can be included in multiple subsets. This introduces diversity and variability among the trees in the forest, making them less to overfitting and more robust to noise in the data.

The key steps in building a random forest are as follows:

1. Random sampling of training data: Randomly sampling a subset of the training data (with replacement) to train each decision tree in the forest. This is known as bootstrapping, and it creates multiple subsets of data for training different trees.
2. Random sampling of features: Randomly select a subset of features for each decision tree at each split during tree construction. This helps reduce the correlation between trees and encourages them to consider different feature combinations.
3. Decision tree construction: It is constructed by using bootstrapped data and randomly selected features. Decision trees are typically constructed using a process similar to regular decision tree algorithms, such as ID3, C4.5, or CART algorithms. However, they have the additional randomness introduced by bootstrapping and feature bagging.
4. Ensemble of decision trees: Once all decision trees are constructed, their outputs are combined to make a final prediction. For classification tasks, the outputs of the trees are combined using techniques such as majority voting, where the class with the most votes is chosen as the final predicted class. For regression tasks, the outputs of the trees are averaged to obtain the final predicted value (Verikas et al., 2011).

4.3.4 Bagging Regression

Bagging (Bootstrap Aggregating) is an ensemble machine learning technique which widely used for both classification problems and regression analysis (Breiman et al., 1984). It involves training multiple instances of the same regression algorithm on different subsets of the training data, obtained through bootstrapping (random sampling with replacement), and then aggregating their predictions to obtain a final prediction.

The training dataset is randomly sampled with replacement to create multiple bootstrap samples. This results in multiple base models, each trained on a different subset of the training data. For prediction, each base model makes predictions on the test dataset. These individual predictions are then aggregated, typically by taking the average of the predicted values, to obtain the final ensemble prediction. Optionally, different base models can be combined in different ways, such as using weighted averages, to further improve the prediction performance.

Variance of prediction can be reduced to $\frac{1}{N}$ (N is the number of learners) of the original variance (single learner) (Harrou et al., 2019). The idea behind bagging is that, training multiple instances of the same algorithm on different subsets of the training data so as to reduce overfitting and to increase the model's ability to generalize the unseen data.

Bagging can be used with various regression algorithms, such as decision trees, random forests, and gradient boosting, to improve their performance and robustness. The processes and structure of bagging regression represent in Figure 4.1.

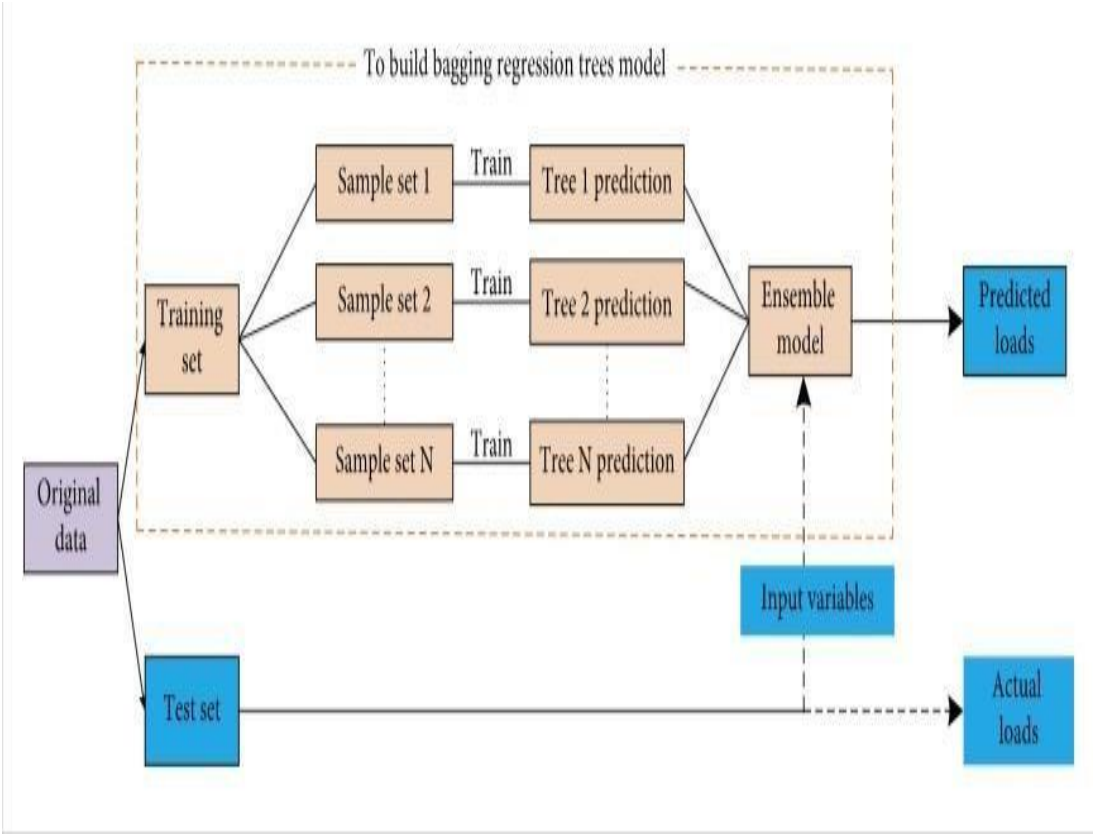

Figure 4.1: Flowchart of the basic idea of bagged regression trees prediction (Dong et al., 2021)

4.3.5 Stacking Regression

Stacking, also known as stacked generalization, was first used by Wolpert (1992) and is an ensemble machine learning technique used for regression tasks, similar to bagging. However, instead of using multiple instances of the same algorithm, stacking combines predictions from multiple different base regression models to obtain a final prediction.

The training dataset is split into two or more disjoint subsets. One subset is used for training the base regression models, and the other subset(s) is used for model validation.

Multiple base regression models, such as decision trees, Support Vector Machines (SVMs), or Neural Networks (NN), are trained independently on the first subset of the training data. Each base model produces predictions for validation data. In this study, linear regression, decision tree, random forest, and bagging regression were used (Figure 4.2).

The validation data, along with the predictions from the base models, are used to train a meta-model that learns to combine the predictions from the base models to make the final prediction. The meta-model is typically trained using a different algorithm than the base model. Once the meta-model is trained, it can be used to make predictions on new, unseen data. The base models generate predictions on the test data, and these predictions are then combined with the meta-model to obtain the final ensemble prediction.

Mathematically it is given as:

$$y_{p,i} = \sum_{m=1}^{M} \omega_m . y_i \tag{4.3}$$

where $\omega_m (m = 1, 2, \ldots, M)$ represents the weight assigned,

y_i represents the prediction of the model m for the i^{th} observation. Also, $y_{p,i}$ is the Stacking Regression.

Optimal set of stacking weights was calculated by minimizing the mean square linear regression. Therefore, two limitations of the objective functions are:

$$\Omega = \text{argmin} \sum_{i=1}^{N} [f(x_i) - \sum_{m=1}^{M} \omega_m . y_i]^2 \tag{4.4}$$

$$\omega_m \geq 0 \ , \ m = 1, 2, \ldots, M$$

$$\sum_{m=1}^{M} \omega_m = 1\,, \quad m = 1,\ 2,\ \ldots\,,\ M$$

$\Omega = \{\omega_1, \omega_2, \ \ldots\,, \ \omega_m\}$ represents the assigned weights of the base models.

$f(x_i)$: is the actual value for the i^{th} observation.

Argmin: is the argument of the minimum where it implies the selection of the minimum among the local minima.

There are two limitations: First, the weights should be equal to or greater than zero. Second, the sum of the weights should equal one. This leads to a quadratic minimization problem (Frank & Wolfe, 1956).

Stacking allows for more complex interactions and combinations of models compare to bagging, as different algorithms can be used as base models, and the meta-model can learn to weigh their predictions based on their performance on the validation data.

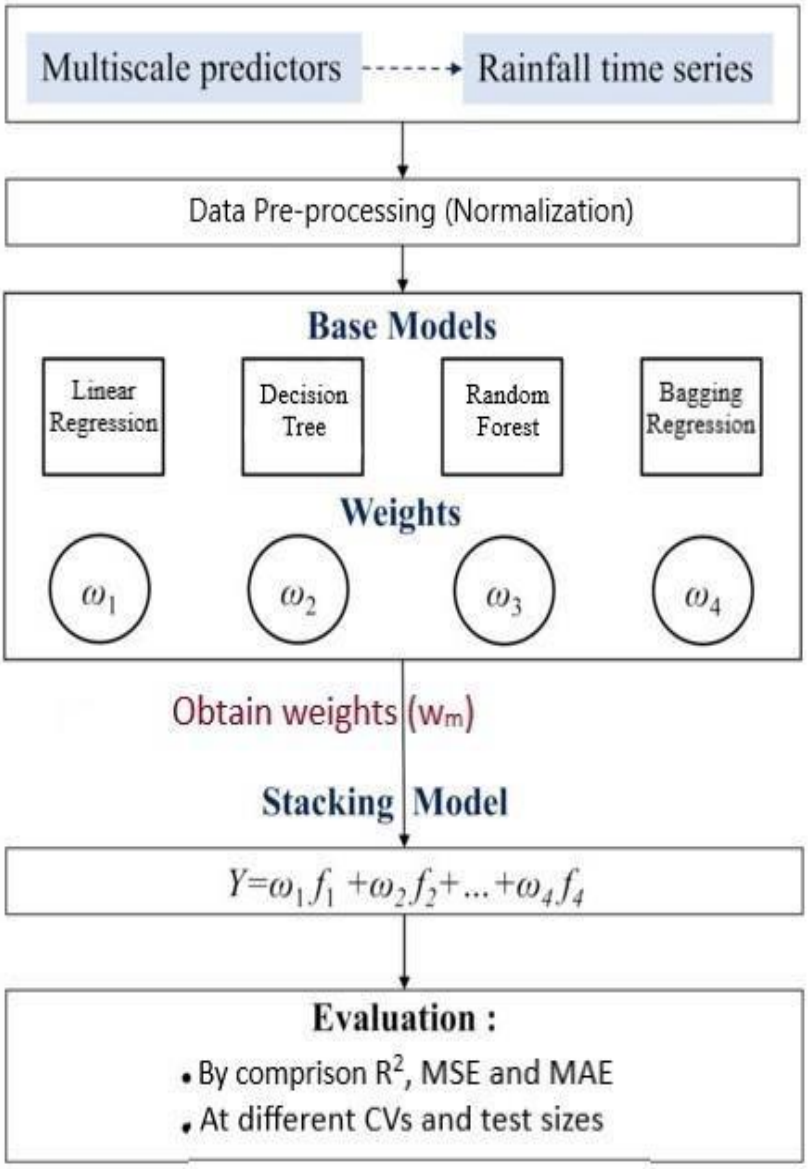

Figure 4.2: Stacking regression based on the methodology used in this study

Chapter 5

METHODOLOGY

5.1 Flowchart Rainfall Prediction

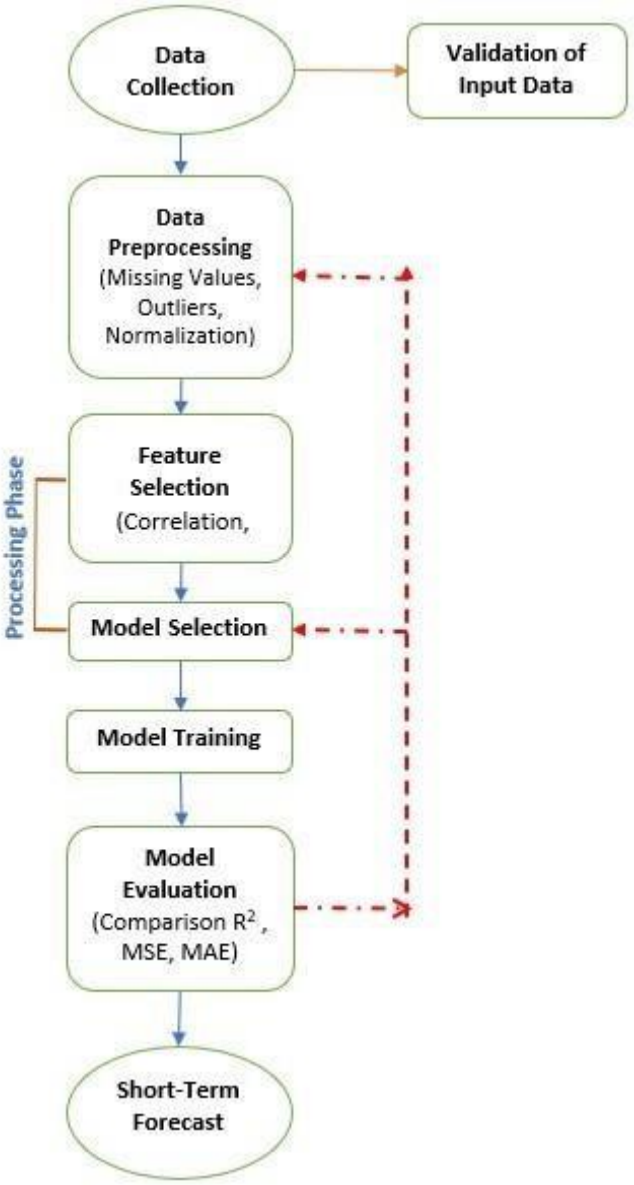

Figure 5.1: Flowchart of rainfall prediction

Rainfall forecasting using machine learning regression models typically involves some methods which exhibit in Figure 5.1 and the following steps.

5.1.1 Selection of Influencing Parameters

As the aim of this study is short term rainfall forecasting, other than the use of customary effective parameters like rainfall amount and its periodic variations, evaporation and temperature, some other meteorologic parameters like maximum and minimum temperatures, humidity, air pressure, wind speed and even the direction of the wind are proposed to achieve this scope.

5.1.2 Data Collection

In this study, datasets are gathered from two sources, the Meteorology Office of TRNC and the National Aeronautics and Space Administration (NASA) via the Internet. From the meteorology office, the data size is not sufficient. These are the monthly data sets from 1975 to 2021 for each region (6 regions) of TRNC that details the rainfall amount in millimeters (mm), the average air temperature in Centigrade (C), the average maximum temperature in Centigrade (C), the average minimum temperature in Centigrade (C), the average pressure in millibar (mbar) and the average wind speed in meters per second (m/s). They are for 46 years with a total of 552 datasets for each region. So, another data source NASA's Earth Observing System Data and Information System (EOSDIS) that gives a wealth of environmental data, including weather data collected from satellites and other instruments were gathered. The parameters that used for raw data at first are the amount of rainfall (mm), the average temperature (C), the minimum temperature (C), the maximum temperature (C), the wind direction (degrees), the wind speed (m/s), the relative humidity (%), the specific humidity (g/kg), the surface pressure (Pa) from 1995 to 2022 daily, for 27 years and have 10227 datasets in total for each region.

5.1.2.1 Validation of Input Data

Generally, there are six main geographical regions in TRNC according to meteorological classification shown in Figure 5.2:

1. North Coast and Beşparmaklar Mountains
2. West Mesaria
3. Central Mesaria
4. East Coast
5. East Mesaria
6. Karpaz.

Each of them is divided into sub-regions:

1. North Coast and Beşparmaklar Mountains:

   a. Girne, b. Lapta, c. Beylerbeyi, d. Esentepe, e. Tatlısu, f. Kantara, g. Alevkaya, h. Çamlıbel, i. Akdeniz, j. Kozanköy, k. Boğazköy, l. Taşkent, m. Değirmenlik.

2. West Mesaria:

   a. Yeşilırmak, b. Lefke, c. Yeşilyurt, d. Gaziveren, e. Güzelyurt, f. Yukarı Bostancı, g. Zümrütköy, h. Kalkanlı.

3. Central Mesaria:

   a. Alayköy, b. Lefkoşa, c. Ercan, d. Yakın Doğu Üni., e. Margo.

4. East Coast:

   a. Gazimağusa, b. Salamis, c. Iskele, d. Yeniboğaziçi.

5. East Mesaria:

   a. Serdarlı, b. Göndere, c. Geçitkale, d. Dörtyol, e. Beyarmudu, f. Çayönü.

6. Karpaz:

   a. Çayırova, b. Büyükkonuk, c. Ziyamet, d. Mehmetçik, e. Yenierenköy, f. Dipkarpaz, g. Zafer Burnu.

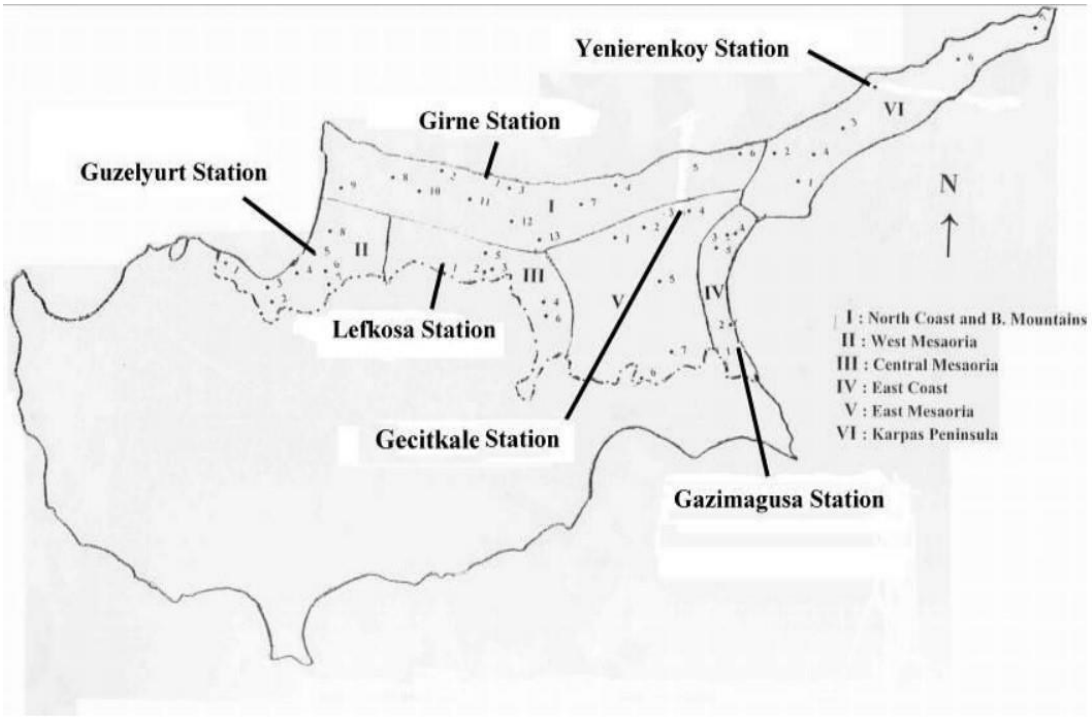

Figure 5.2: TRNC meteorological regions with stations

For verification of NASA's data and its trend, the annual amount of rainfall of each region from 1995 to 2021 were gathered from meteorological office and used for the correlation as detailed in Figures (5.3 to 5.8).

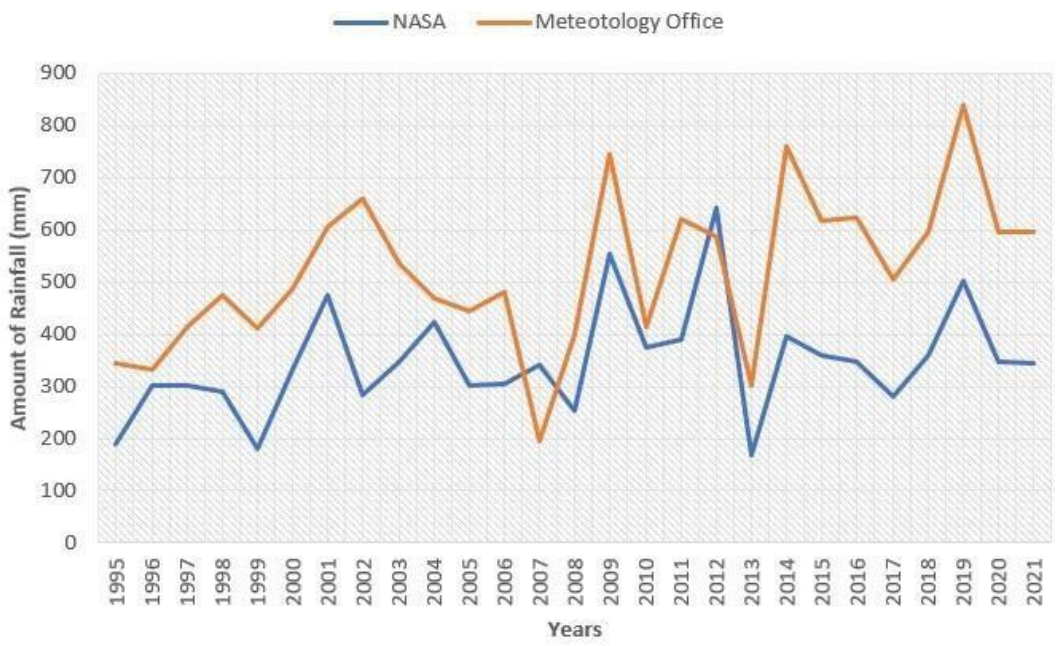

Figure 5.3: Amount of rainfall in North Coast and B. Mountains (Girne) region (yearly)

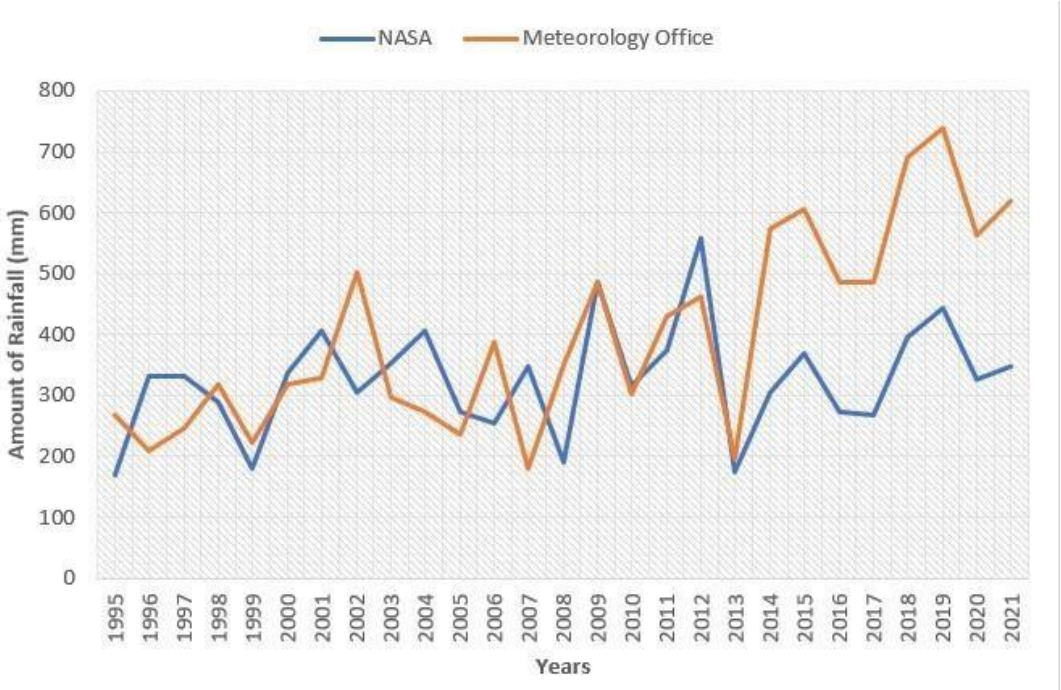

Figure 5.4: Amount of rainfall in West Mesaria (Güzelyurt) region (yearly)

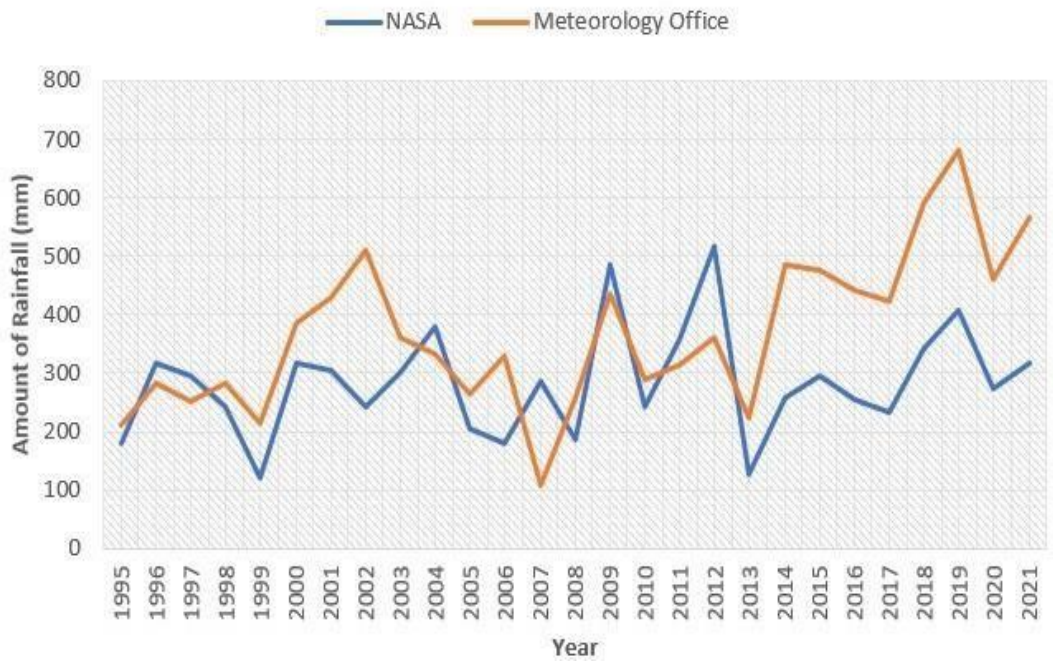

Figure 5.5: Amount of rainfall in Central Mesaria (Lefkoşa) region (yearly)

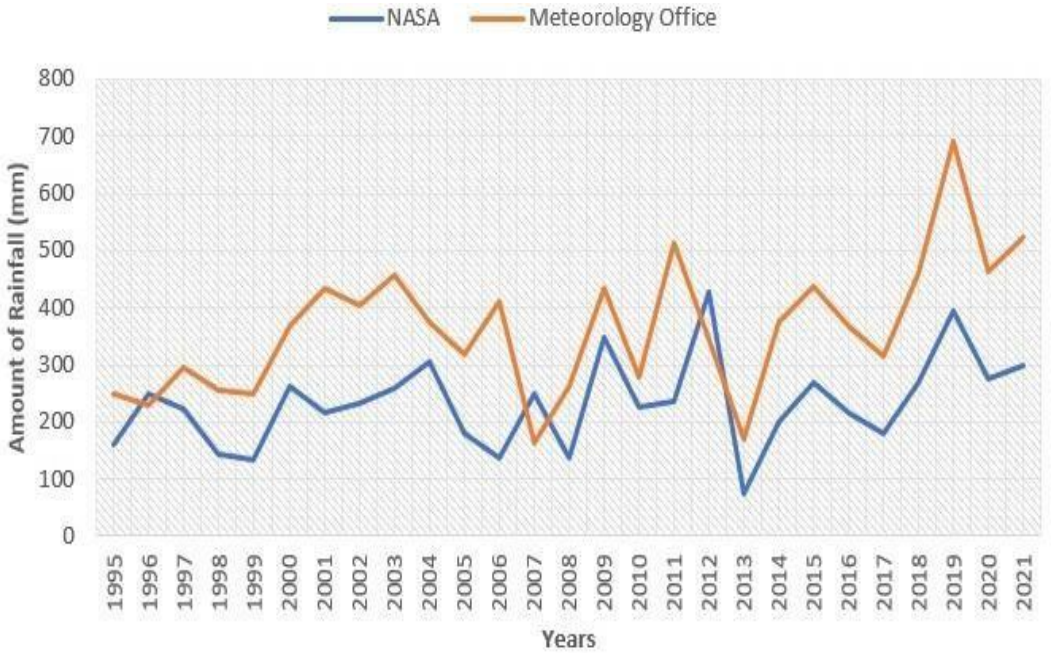

Figure 5.6: Amount of rainfall in East Coast (Gazimağusa) region (yearly)

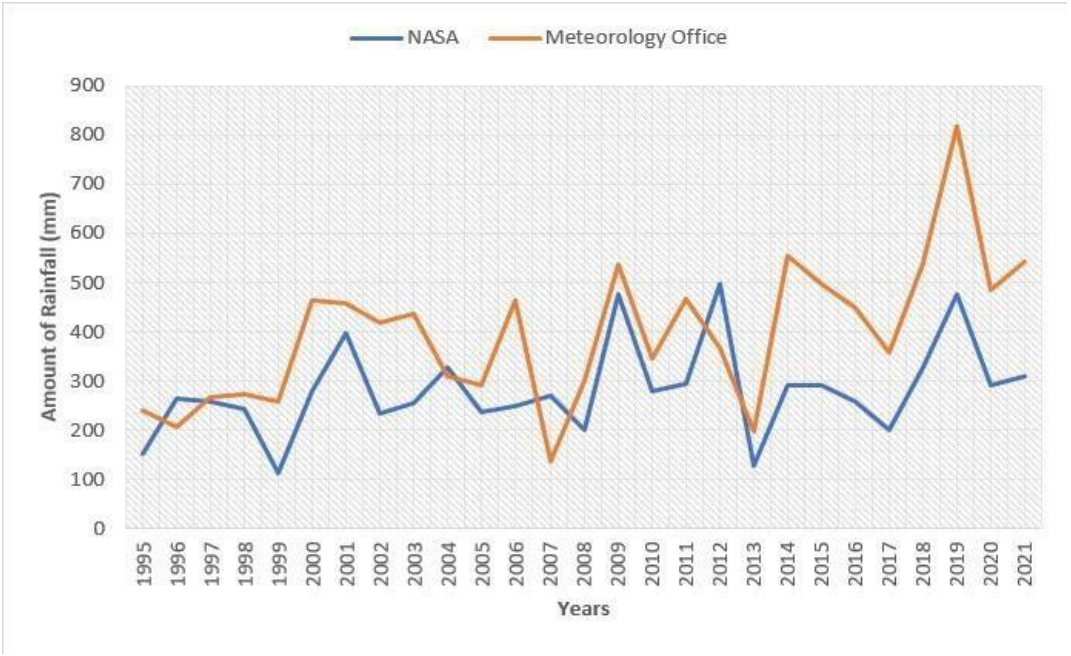

Figure 5.7: Amount of rainfall in East Mesaria (Geçitkale) region (yearly)

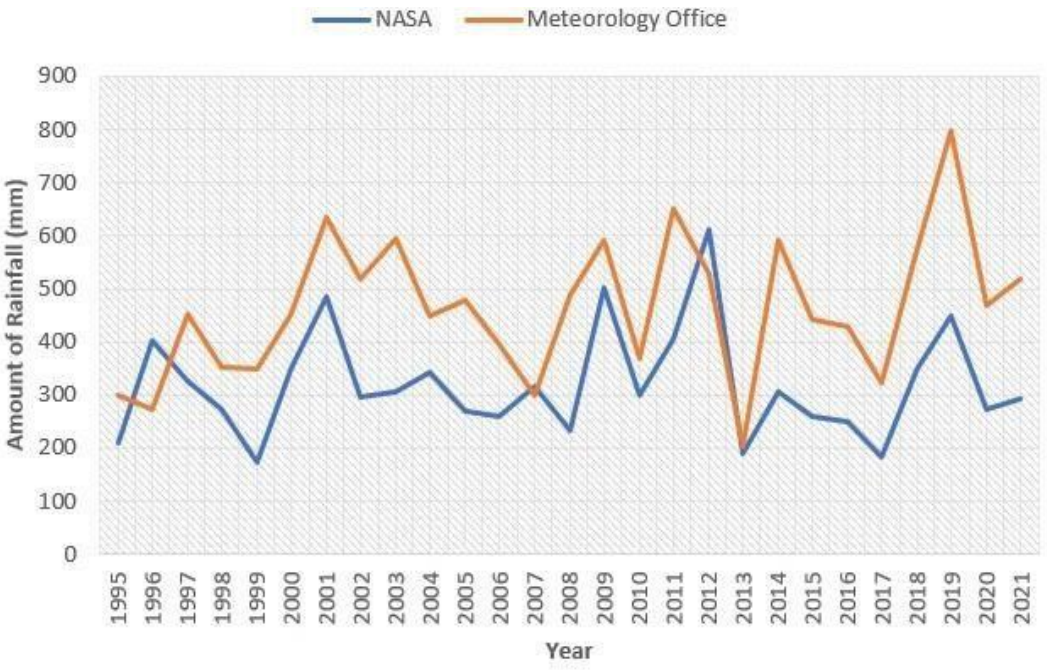

Figure 5.8: Amount of rainfall in Karpaz (Yenierenköy) region (yearly)

As it is obvious, the dataset of meteorology office has greater rainfall amounts than NASA in all TRNC regions, based on the representative station of each region. However, the trend is same for almost all years and regions hence implying the appropriateness of the NASA's dataset.

5.1.2.2 Data Used in this Study

In this study, for the daily rainfall prediction, six stations, (Girne, Güzelyurt, Lefkoşa, Gazimağusa, Geçitkale, and Yenierenköy), each representing different regions of TRNC, with 9 meteorological dataset that are supposed (believed) to be influencing, were gathered as a daily dataset from NASA's website between 1995 and 2022. As an initial attempt, the below detailed parameters were proposed:

- Rainfall Amount
- Minimum Temperature
- Maximum Temperature
- Average Temperature
- Specific Humidity
- Relative Humidity
- Wind Speed
- Wind Direction
- Surface Pressure

5.1.3 Data Preprocessing

Three distinct processes have to be done for this step.

a) During this step the defects of the datasets were investigated. Feature or variable creation may include filtering and scaling data to capture relevant patterns or relationships. The dataset containing missing values should be deleted if their existence

is less than 5 percent with the relevant dataset. In this study, if the dataset existence is more than 5 percent with respect to the relevant dataset filling the missing values by a variety of techniques have to be used, including the mean imputation, the median imputation, and interpolation. A trial-and-error procedure was applied until the above-mentioned criteria to be achieved for each parameter separately.

b) there are values in a dataset that are outliers, meaning that, they are significantly different from other values. Model performance can be affected by them, and they should be handled carefully. Detecting outliers is a critical step in data preprocessing. The splitting algorithm of CART (decision tree) easily isolates the outliers in a separate node (Loh 2011; Timofeev 2004). Some techniques for detecting outliers are:

    1) **Visual Inspection:** can be detected by creating a box plot, a histogram, or a scatter plot to visualize the data. Since the outliers are points that are far from the rest of the data, it is often quick and easy to identify these outliers with a visual inspection, especially when the dataset is small. To be able to achieve this, a trial-and-error procedure was applied until an acceptable accuracy compared with the relevant dataset is reached.

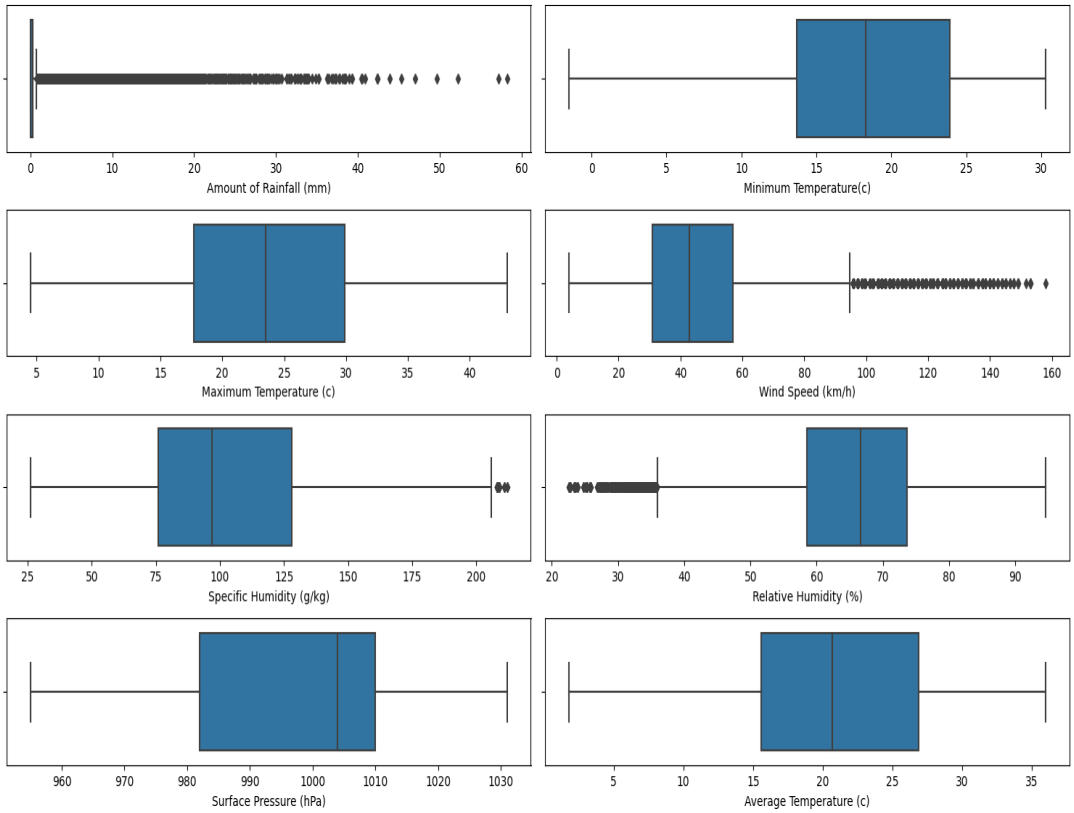

Figure 5.9: Box plot of continuous parameters of the raw dataset

2) **Statistical Methods:** Based on the statistical model that considers the distance of each data from the median or mean is used to identify the outliers. Outliers can be detected, for instance, using the z-score and the interquartile range (IQR). In general, an outlier is a value that is more than three standard deviations from the mean or outside the IQR. Applying this method, the outliers of the below given effective parameters were detected.

	Amount of Rainfall (mm)	Minimum Temperature (C)	Maximum Temperature (C)	Wind Speed (m/s)	Specific Humidity (g/kg)	Relative Humidity (%)	Surface Pressure (hPa)	Average Temperature (C)
Count	61362	61362	61362	61362	61362	61362	61362	61362
Mean	0.92	18.42	23.86	46.48	103.94	65.53	998.31	20.98
Standard Deviation	2.94	6.16	7.06	21.1	35.84	10.84	15.45	6.4
Minimum	0	-1.5	4.5	4	26	22.6	955	1.8
25%	0	13.7	17.7	31	76	58.5	982	15.6
50%	0	18.3	23.5	42.8	97	66.6	1004	20.7
75%	0.3	23.9	29.9	56.9	128	73.6	1010	26.9
Maximum	58.2	30.3	43	158	212	94.6	1031	36

Figure 5.10: Statistical analysis of the raw dataset

	Amount of Rainfall (mm)	Wind Speed (m/s)	Relative Humidity (%)	Surface Pressure (hPa)	Average Temperature (C)
Count	61296	61296	61296	61296	61296
Mean	0.9	46.44	65.52	998.3	20.99
Standard Deviation	2.84	21.08	10.84	15.45	6.39
Minimum	0	4	22.6	955	3.2
25%	0	31	58.5	982	15.7
50%	0	42.8	66.6	1004	20.8
75%	0.3	56.9	73.6	1010	26.9
Maximum	43.9	158	94.6	1031	35

Figure 5.11: Statistical analysis of the dataset after preprocessing

3) **Machine Learning Algorithms:** Outliers can be detected using machine learning algorithms such as isolation forest, Local Outlier Factor (LOF), and K-Nearest Neighbors (KNN). By learning the underlying patterns in the data, these algorithms can identify places that are significantly different from the rest. This method is not applied in this study.
4) **Domain Knowledge:** Identification of outliers can be based on domain knowledge or expert judgment. For example, a temperature less than 3 ˚C is identified as an outlier in this study during the summer period.

In the event that outliers are detected, one may either remove them from the dataset or transform them by replacing extreme values with non-extreme values. In any case, it is important to carefully consider the impact of removing or transforming outliers on the analysis and the results. Occasionally, outliers may contain valuable information, and removing them may affect the validity of the analysis. On the other hand, transformations play a central role in regression analysis (Cook and Weisberg, 1999).

c) data normalization is conducted to non-dimesionalizing the dataset groups so as to improve the performance accuracy of the model (Nourani et al., 2018).

In this study the below formulation is used for normalization of the datasets:

$$X_{normalized} = \frac{f(x_i) - X_{min}}{X_{max} - X_{min}} \quad (5.1)$$

where:

$X_{normalized}$: is the normalized value

f(x_i): is the actual value

X_{min}: is the minimum value of that parameter

X_{max}: is the maximum value of that parameter.

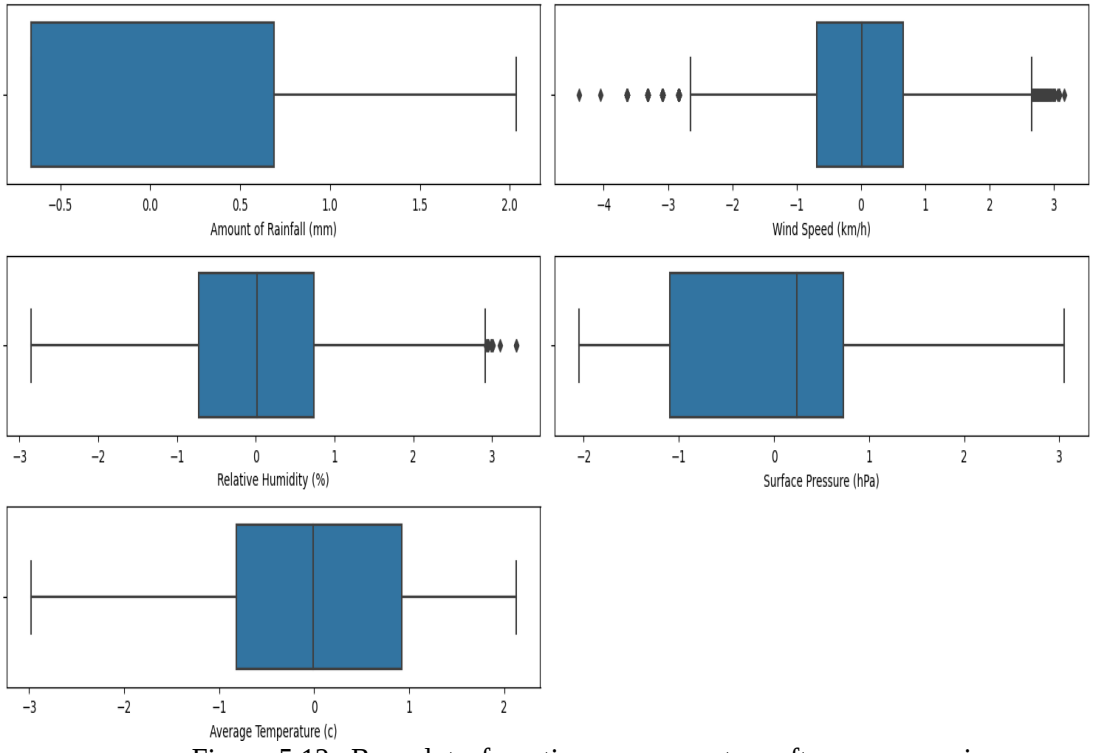

Figure 5.12: Box plot of continuous parameters after preprocessing

5.1.3.1 Wetness and Dryness

In this study, wetness and dryness have significant effects. It is calculated based on the annual average amount of rainfall. Whenever the daily rainfall a specific year exceeds the annual average of that year, that day is categorized as "wet", otherwise it is categorized as "dry". Figure 5.13 represents the percentage of wetness and dryness in this study.

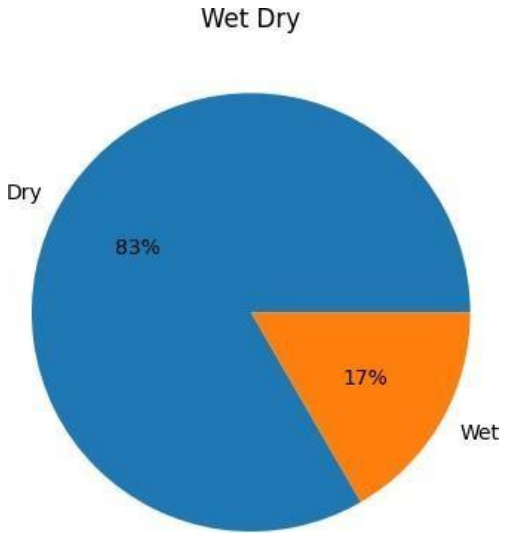

Figure 5.13: Wetness and dryness percentage in this study

5.1.3.2 Seasons

In general, rainfall rates vary by season, and in TRNC's most rainfall occurred in winter, hence, the selected data between 1995 and 2022 that are used in this study is shown in Figure 5.14.

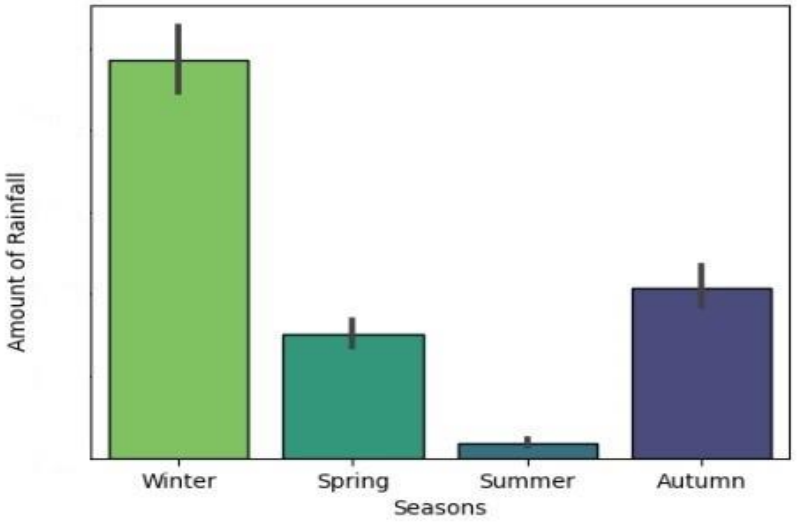

Figure 5.14: Amount of rainfall based on seasons in TRNC

5.1.4 Feature Selection

In this study as detailed earlier, to forecast the daily rainfall of each region, 9 effective meteorologic parameters were selected after normalization. For these suggested parameters to determine their interdependence among each other below given analyses were applied. Through ML models the most influential factors or variables likely to influence rainfall patterns were identified. By doing so, one will be able to detect dependent and independent parameters so as to improve the model's accuracy, reduce overfitting, and reduce the complexity of the computations. The following techniques can be used to detect the most significant features:

1) **Domain Knowledge:** Choosing the most relevant features requires domain knowledge or expert judgment. Based on meteorological knowledge, temperature, humidity, and wind speed may be considered important features in a climate dataset (Haurwitz and Austin,1944).
2) **Correlation Analysis:** The correlation between each effective parameter and the target variable (rainfall amount) were calculated as given in Fig 5.15. A model that includes characteristics by showing high correlation with the target variable is more likely to be relevant and be selected. On the other hand, if two or more features are highly correlated, choosing only one of it is applied as shown in Fig 5.16.

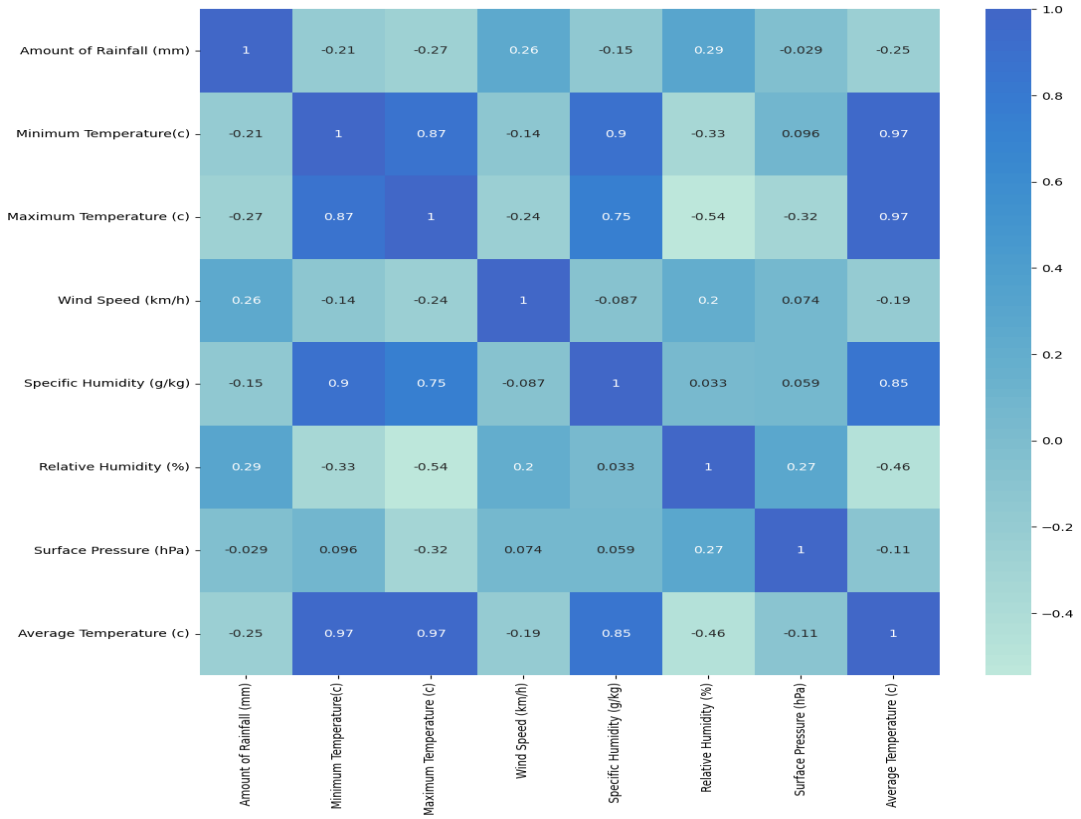

Figure 5.15: Correlation analysis of the raw dataset

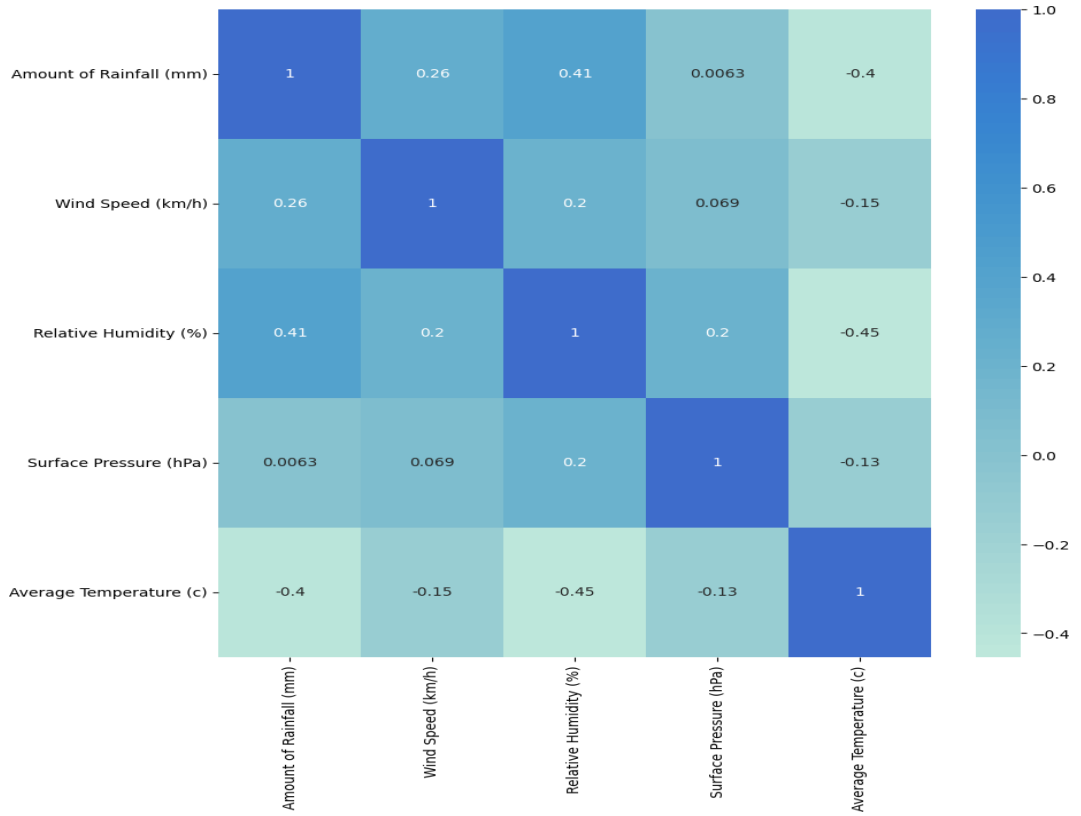

Figure 5.16: Correlation analysis of the dataset after preprocessing

3) **Variance Inflation Factor (VIF):** Regression analyses use the Variance Inflation Factor (VIF) to measure multicollinearity (correlation) between two or more predictor variables. The VIF calculates how much multicollinearity has inflated the regression coefficient variance. A regression model calculates the VIF by regressing each predictor variable against all the other predictor variables. Predictor variables are assigned a VIF by applying the following:

$$VIF = \frac{1}{1 - R^2} \tag{5.2}$$

where R^2 is the coefficient of determination of the regression model for that predictor variable, after regressing it against all the other predictor variables. A VIF value of 1 indicates no correlation between the predictor variable with the other predictor variables. In contrast, a VIF value greater than 1 shows some correlation. A VIF value of 5 or higher is often considered a sign of high multicollinearity. Depending on the severity, it may require further investigation or remedial action, such as removing one or more correlated predictor variables.

The value of VIF in regression analysis is its ability to diagnose problems caused by multicollinearity, such as unstable parameter estimates, inflated standard errors, and reduced statistical power.

In regression analysis, VIF can improve accuracy and reliability via the identification and removal of multi-collinearity. After preprocessing by eliminating the outliers, and some dependent effective variables through correlation analyses, the acceptable range of VIF values were obtained as tabulated below.

Table 5.1: VIF values between input parameters of the raw dataset

ID	Variables	VIF
1	Minimum Temperature (C)	149.01
2	Maximum Temperature (C)	127.65
3	Wind Speed (km/h)	1.12
4	Specific Humidity (g/kg)	58.03
5	Relative Humidity (%)	14.31
6	Surface Pressure (hPa)	3.01
7	Average Temperature (C)	417.7

Table 5.2: VIF values between input parameters of the dataset after preprocessing

ID	Variables	VIF
1	Wind Speed (km/h)	1.045468
2	Relative Humidity (%)	1.318045
3	Surface Pressure (hPa)	1.046080
4	Average Temperature (C)	1.266607

5.1.5 Model Selection

Model selection is the process of choosing the best-suited machine learning regression model for predicting the amount of rainfall based on the input parameters. Popular models for this task include linear regression, decision trees, random forests, and neural networks. Finding the best model that fits the input data is the hardest part. Selecting a model is based on many factors, such as the number of datasets, features, tasks, nature of the model, etc.

Two factors should be considered:

1. The rationale behind selecting a model is based on logical reasoning, and
2. The evaluation and comparison of the models' performance

Models can be selected as:

1. Type of data available:

(a) Images and Videos – CNN

(b) Text data or Speech data – RNN

(c) Numerical data – SVM, Regression Models, Decision trees, etc.

2. Based on the task, one has to carry out:

(a) Classification tasks – SVM, Logistic Regression, Decision trees, etc.

(b) Regression tasks – Linear Regression, Random Forest, Polynomial Regression, etc.

(c) Clustering tasks – K-means Clustering, Hierarchical Clustering

5.1.6 Model Training

The chosen model is then trained on the historical data, using a portion of the dataset for training and the remaining portion for validation and testing. This involves setting hyperparameters, such as the number of trees in a random forest and optimizing the model to minimize errors.

Training the selected model on the preprocessed data by applying a training algorithm. This involves finding the optimal values for the model's parameters to minimize the error between the predicted and the actual rainfall values. The largest subset of the original dataset used to train or fit the ML model is known as training data. Through ML algorithms this data is fed to teach them how to make predictions for the given task. The accuracy and prediction ability of the model is largely depending on the quality of the training data. Higher quality training data results in a better-performing model. Typically, training data constitute more than 70% of the total data for a ML project.

On the other hand, the test dataset is a separate subset of the original data and independent of the training dataset. Usually, the test dataset makes up approximately 20% of the total

original data for a ML project. Training and test (training: test) datasets are generally divided into 80:20, 70:30, or 90:10.

5.1.7 Model Evaluation

To assess rainfall forecasting accuracy, statistical methods were used to measure forecasting errors. These methods consist of the following evaluate the goodness-of-fit for the entire regression. The R^2 value falls within the range of 0-1, where a value closer to 1 indicates a higher level of goodness-of-fit of the regression line to the observed value, and conversely, a value closer to 0 indicates a poor fit.

Cross-validation: Cross-validation can be used to evaluate the performance of various regression models on the same dataset. Create two sets of data; a training set and a testing set, and train different models on the training set. Then, assess the performance of each model using statistical metrics such as Mean Square Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). It is determined that, the best model is the one with the lowest error and highest R^2. The MAE is less sensitive to extreme values than the RMSE (Fox, 1981).

5.1.7.1 Mean Squared Error (MSE)

In regression models, Mean Squared Error (MSE) is commonly used as a statistical metric to measure model performance. It is calculated as the average of the squared differences between the predicted values and the actual values of the dependent variable.

MSE formula is:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}(\mathrm{y_i} - \mathrm{f}(x_i))^2 \tag{5.3}$$

where:

- *n* is the number of observations in the dataset.
- y_i is the predicted value of the dependent variable for the i^{th} observation

- $f(x_i)$ is the actual value of the dependent variable for the i^{th} observation.

In other words, the MSE is a statistical measure which uses predicted and actual values. The lower the MSE, the better the regression model's performance. It is pertinent to note that, the MSE is sensitive to the scale of the dependent variable can influence outliers.

5.1.7.2 R-Squared (R^2)

Based on the independent variables of the regression model, the coefficient of determination is a measure of how much variance can be predicted from the dependent variable. R^2 is often used as a measure of the goodness of fit of a regression model. The formula for R^2 is:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(f(x_i) - y_i)^2}{\sum(f(x_i) - \bar{y})^2} \quad (5.4)$$

where:

- SS_{res} is the difference between the predicted and actual values of the dependent variable measured by the sum of squared residuals (also known as the sum of squared errors)
- SS_{tot} can be thought of as the difference between the actual values and the mean of a dependent variable
- y_i is the predicted value of the dependent variable for the i^{th} observation
- $f(x_i)$ is the actual value for the i^{th} observation
- $\bar{y}$ is the average of the prediction values.

In other words, R^2 measures the proportion of the total variance in the dependent variable explained by the regression model. A value of R^2 close to 1 shows a great fit, meaning a

lot of the variation in the dependent variable is determined by the independent variables. A value of R^2 close to 0 means there is no fit. This means that the independent variables do not explain much of the variation in the dependent variable. However, it is critical to note that R^2 should not be used as the sole criterion for evaluating the quality of a regression model, as it can be misleading in certain situations, such as when the model is overfitting the data.

5.1.8 Forecast

Once the model is trained and validated, it can be used to make rainfall forecasts for future time periods, using new weather and climate data as input. Short range forecasting lasts for 1-2 days. Since weather greatly affects human activities, food production, and personal comfort, accurate forecasting plays a crucial role in planning current and future activities. Although there are various aspects that may impact the forecasting outcome, it is a valuable tool for different analyses.

Medium Range Forecasting lasts from 3-4 days to 2 weeks and is a necessity for small strategic decisions related to business nature. Medium-term forecasts play a vital role in business budgeting and development. Inaccurate forecasting can have serious impacts on the rest of the organization, leading to unsold stock and overspending on production. It is crucial to make accurate forecasts to prevent an organization from going bankrupt. Medium-term forecasts are usually for one year.

Long-range forecasts are for times longer than four weeks and are made for major upcoming strategic decisions. Organizations focus on general ongoing trends rather than specific items and try to predict revenue-generating sales over periods greater than two years.

Accurate predictions might be needed for a decade or more for huge industries to tackle the changes. However, the disadvantage of such forecasts is that they cannot be more than unclear. Prediction planners receive criticism when things go wrong, which is entirely opposite to what was predicted.

Chapter 6

ANALYSES AND RESULTS

6.1 Training Criteria

This section discusses the critical criteria for training the model. The provided dataset that has a total of 61,362 rows before preprocessing with 8 effective meteorological parameters, get reduced to 5 effective meteorological parameters (including, Average Temperature, Relative Humidity, Surface Pressure, Wind Speed, and Wind Direction) on average daily bases. After preprocessing (deleting outliers and organizing the data), the dataset was reduced to 61,296. To prepare the dataset for forecasting, the whole dataset was divided into two subsets:

- the training set
- the validation set.

For this study, to achieve the forecasting amount of rainfall on daily bases for each region, with the help of the above mentioned preprocessed meteorological parameters, first the training part that is composed of train: test ratio should be studied. To achieve that, 5 different ratios were selected arbitrarily and tested (90:10, 85:15, 80:20, 75:25, and 70:30) where R^2, MSE, and MAE for each model were calculated separately and given in Figs. 6.1 - 6.3. The results are interestingly common where for each model being highest R^2, lowest MSE and lowest MAE the train: test ratio is 90:10. Implying that, keeping nearly 90% of each station for fitting the models (training), and the remaining roughly 10% for evaluating their prediction skill (testing) (Hofmann et al., 2008; Marsland 2015).

While the proper train: test ratios were detected, the algorithm was scripted in the Python programming language (Python 3.10.11) and four widely used regression methods were given used in this study:

1. Decision Tree
2. Bagging Regression
3. Random Forest
4. Stacking Regression.

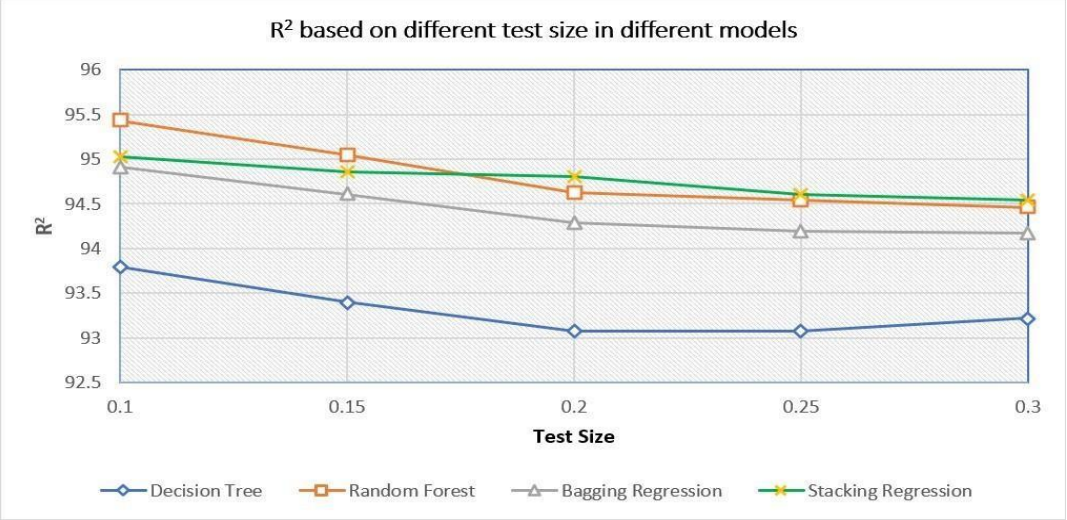

Figure 6.1: R^2 based on different test sizes in different models of TRNC

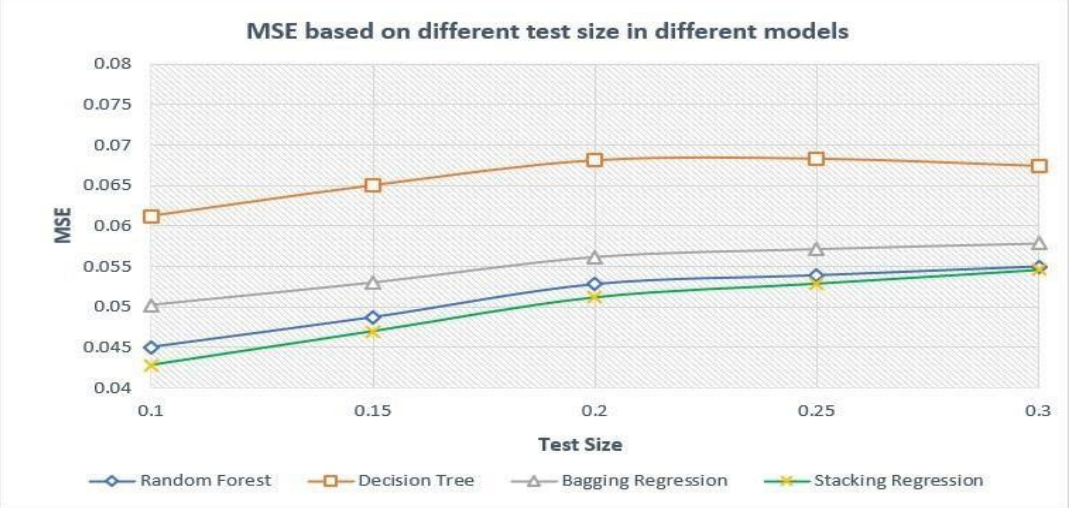

Figure 6.2: MSE based on different test sizes in different models of TRNC

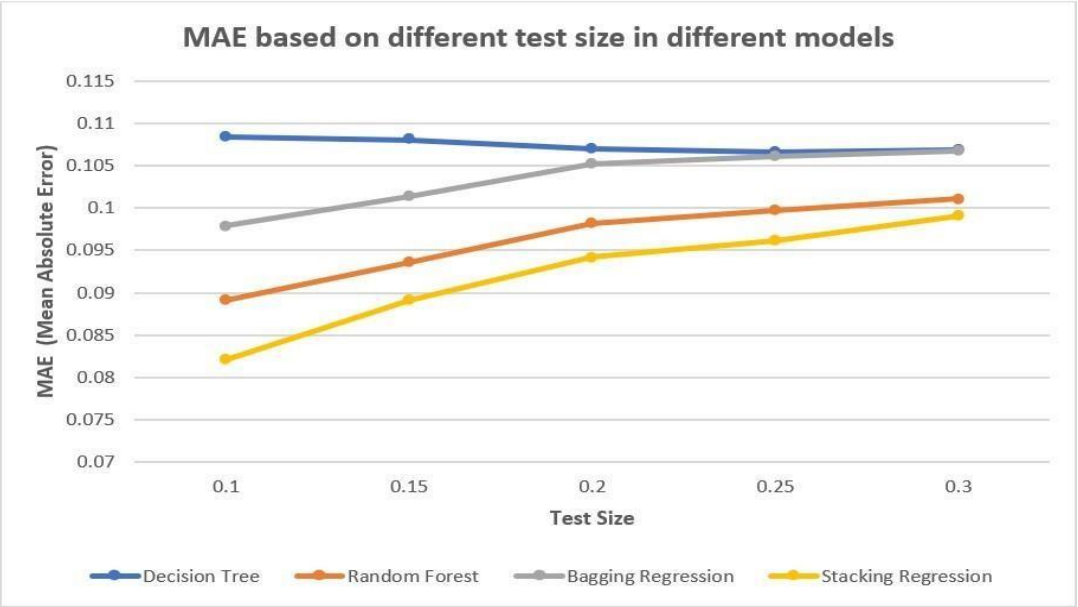

Figure 6.3: MAE based on different test sizes in different models of TRNC

To adjust the algorithm's hyperparameters that are the variable part of the decision tree model code where each regression type has different codes with specific hyperparameters as detailed below, whereby keeping the other parameters constant and changing the remaining one parameter arbitrarily and one at a time, the trials continue until a higher precision achieved based on R^2. Hence, among several trials, the appropriate value for each hyperparameter having the highest R^2 was determined for each regression type. The hyperparameters for different regression types used in this study are:

- Max Depth, Min Samples Split, and CV for Decision Tree Regression,
- Number of Estimators and CV for Random Forest Regression,
- Number of Estimators, Selected Final Estimator Type and CV for Bagging Regression,
- Number of Estimators, Combined Estimators (Linear Regression, Decision Tree, Random Forest, and Bagging Regression), Selected Final Estimators Type and CV for Stacking Regression.

6.1.1 Decision Tree Regression

After several trials, the appropriate hyperparameter values for this study's datasets were found to be:

- Max Depth = 10
- Min Samples Split = 10

And the best R^2 is 93.26 for CV = 10 and 15, (both giving the same value) so in this study, CV=10 was selected to reduce the running time and the cost.

6.1.2 Random Forest Regression

After several trials, the appropriate hyperparameter values for this study's datasets were found to be:

- number of estimators = 150

And the best R^2 is 94.8 for CV = 25, hence, CV = 25 was selected.

6.1.3 Bagging Regression

After several trials, the appropriate hyperparameter values for this study's datasets were found to be:

- number of estimators = 150
- selected final estimator type = Random Forest

And the best R^2 is 94.42 for CV = 25, so again CV =25 was selected.

6.1.4 Stacking Regression

- number of estimators = 150
- combined estimators = Linear Regression, Decision Tree, Random Forest, and Bagging Regression
- selected final estimators' type = Bagging Regression with Random Forest Regression

the best R^2 is 95.03 for CV = 25 and 30, (both giving the same value) so in this study, CV=25 was selected to reduce the running time and the cost.

Figure 6.1 illustrates how the R^2 results have changed by increasing CV during training stage for different models where and the detailed values are given in Table 6.1.

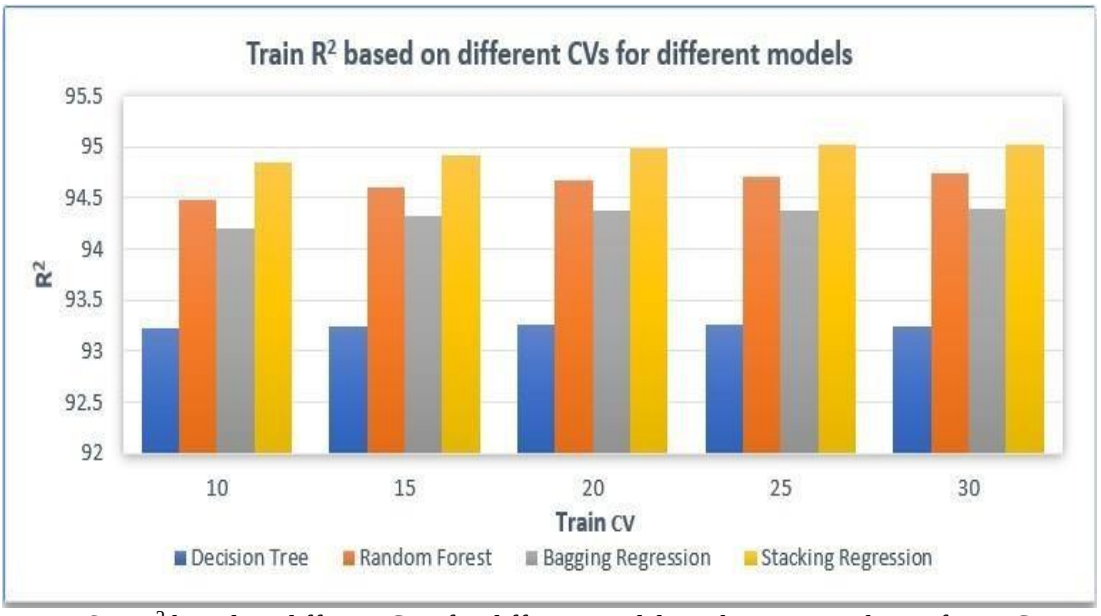

Figure 6.4: R^2 based on different CVs for different models in the training phase of TRNC

Table 6.1: Values of R^2 based on different CVs in different models in the training phase of TRNC

MODEL	R^2 during training						
	CV						
	5	8	10	15	20	25	30
Stacking Regression			94.85	94.92	94.99	95.03*	95.03
Random Forest	94.48	94.61	94.68	94.71	94.75	94.8*	94.78
Bagging Regression	94.2	94.32	94.37	94.38	94.4	94.42*	94.41
Decision Tree	93.23	93.25	93.26*	93.26	93.24		

* Suggested CV and R^2 values for that regression model

6.2 Validation of Predicted Values

After the training and test the datasets for the suggested regression models with their determined estimator values, each model performance based on its predicted rainfall value was compared with the measured rainfall value and assessed through three statistical measures (Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-Squared (R^2)).

6.2.1 Validation Results of TRNC

A detailed account of the validation results obtained after evaluating the above mentioned 3 statistical metrics for TRNC are:

- Stacking Regression is the best with R^2=95.66 and Random Forest Regression with R^2=95.43.
- The least MSE is 0.0428 for Stacking Regression whereas 0.045 for Random Forest.
- For MAE, the most accurate method is Stacking Regression with 0.0821 and the next was Random Forest Regression with 0.0891.

6.2.2 Validation Result of Meteorological Regions of TRNC

In this part, regions were separated, and trained separately (as detailed in the previous sections). The dataset was separated as training part (90 percent) for learning the models and as testing part (10 percent) for validating (checking) the models' performance.

During the training phase, more accurate hyperparameters were used and in the testing phase, it was checked by assessing some metrics such as R-Squared (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE). The results are shown in Table 6.2, from the most accurate to the least accurate.

For almost all meteorological regions, Stacking Regression is the most accurate model. Then, Bagging Regression gave the highest result. However, for TRNC, when it considered TRNC as a whole, the Random Forest provided higher accuracy. After the Bagging

Regression, the Random Forest model performed the best. Among all the models which tested, the Decision Tree had the lowest accuracy where Karpaz was an exception, where the Decision Tree had 92.25 for R^2 compared to the Random Forest's 91.96.

Table 6.2: Results of the meteorological regions of TRNC in training and testing phase

Location	MODEL	Train			Test			Dataset*
		R^2	MSE	MAE	R^2	MSE	MAE	
Girne	Stacking Regression	92.71	0.0748	0.126	93.17	0.0566	0.1064	before
	Bagging Regression	92.37	0.0782	0.1299	92.78	0.0265	0.0734	10227
	Random Forest	92.01	0.0819	0.1322	92.53	0.0103	0.0453	after
	Decision Tree	91.58	0.0863	0.1366	92.14	0.0581	0.1066	10223
Guzelyurt	Stacking Regression	94.15	0.0563	0.1006	93.62	0.0618	0.1072	before
	Bagging Regression	93.81	0.0596	0.1032	93.37	0.0247	0.0604	10227
	Random Forest	93.69	0.0608	0.1032	93.03	0.0095	0.0418	after
	Decision Tree	92.99	0.0675	0.1045	92.35	0.0503	0.0933	10221
Gecitkale	Stacking Regression	93.92	0.0612	0.1102	93.75	0.0497	0.0995	before
	Bagging Regression	93.46	0.0658	0.1122	93.62	0.0231	0.066	10227
	Random Forest	93	0.0706	0.1163	93.38	0.0089	0.0409	after
	Decision Tree	92.96	0.0709	0.1155	93.18	0.0521	0.096	10226
Karpaz	Stacking Regression	93.08	0.0696	0.126	92.63	0.0646	0.1221	before
	Bagging Regression	92.54	0.075	0.1305	92.22	0.0289	0.0811	10227
	Decision Tree	92.25	0.0779	0.1358	91.74	0.0647	0.1191	after
	Random Forest	91.96	0.0808	0.1335	91.97	0.011	0.0499	10226
Lefkosa	Stacking Regression	93.98	0.058	0.1014	93.62	0.0615	0.1063	before
	Bagging Regression	93.82	0.0595	0.1031	93.37	0.0247	0.0647	10227
	Random Forest	93.72	0.0605	0.1031	93	0.0096	0.0418	after
	Decision Tree	93.16	0.0659	0.1027	92.29	0.0498	0.0929	10221
Gazimagusa	Stacking Regression	94.06	0.0583	0.1111	93.67	0.0465	0.0966	before
	Bagging Regression	94.04	0.0585	0.1106	93.55	0.0566	0.0682	10227
	Random Forest	93.97	0.0592	0.1111	93.25	0.0092	0.0422	after
	Decision Tree	93.29	0.0659	0.1155	92.71	0.0519	0.0977	10227

* The number of datasets before and after deleting outliers

6.3 Running Time of Models

The prediction process is heavily dependent on time as a key parameter. The time taken by each model varies, with less time resulting in reduced costs. However, sometimes with lots of datasets, the time is rising and furthermore, accuracy is increasing. The result of this study for TRNC is in Figure 6.5, and separately region by region in Figure 6.6.

Moreover, Table 6.3, shows the meteorological stations of TRNC models, and the amount of time for each model.

Figure 6.5: Running time based on different models in minutes of TRNC

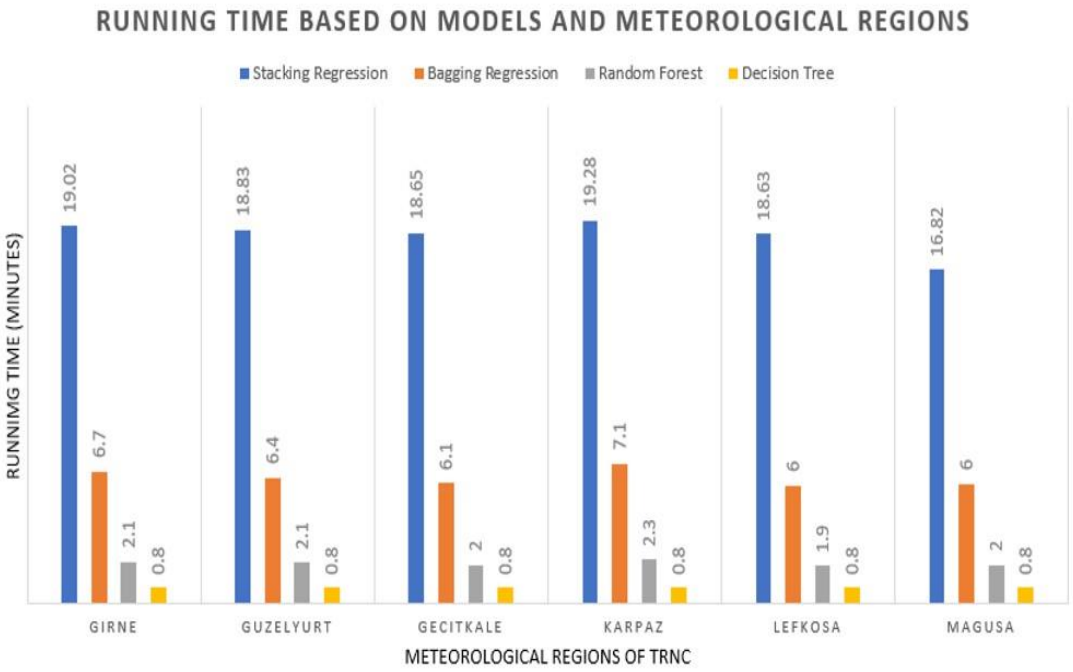

Figure 6.6: Running time based on different models for meteorological regions of TRNC

Table 6.3: Running time in minutes for meteorological stations of TRNC

Location	MODEL	Time (min)
Girne	Stacking Regression	19
	Bagging Regression	6.7
	Random Forest	2.1
	Decision Tree	0.8
Guzelyurt	Stacking Regression	18.8
	Bagging Regression	6.4
	Random Forest	2.1
	Decision Tree	0.8
Gecitkale	Stacking Regression	18.7
	Bagging Regression	6.1
	Random Forest	2
	Decision Tree	0.8
Karpaz	Stacking Regression	19.3
	Bagging Regression	7.1
	Decision Tree	0.8
	Random Forest	2.3
Lefkosa	Stacking Regression	18.6
	Bagging Regression	6
	Random Forest	1.9
	Decision Tree	0.8
Gazimagusa	Stacking Regression	16.8
	Bagging Regression	6
	Random Forest	2
	Decision Tree	0.8

6.4 Short-Term Forecasted Results of TRNC

The short-term (daily) forecasted (predicted) and the measured (actual) values of the available dataset (61296 data) of TRNC was compared and among the 10% (6129) forecasted value, 20 randomly selected values that were obtained through Stacking Regression is illustrated in Table 6.4, and similarly from that obtained set 100 randomly selected values were presented in Figure 6.7 so as to demonstrate the error between them.

ID: is the ID number of the prediction model

Actual Value: is the real (measured) amounts of rainfall in millimeters

Prediction Value: is the predicted amounts of rainfall in millimeters

Table 6.4: Some actual and prediction values through stacking regression of TRNC

ID	Actual Values	Prediction Values
53334	0.0	0.000
18644	1.4	1.346
50494	0.7	0.645
12153	0.0	0.000
43512	1.2	1.225
40278	0.0	0.000
1555	0.7	0.554
50261	0.0	0.000
44171	1.3	1.306
36712	0.0	0.000
19517	0.0	0.000
40987	0.0	0.000
4069	0.0	0.000
17394	0.3	0.322
36705	0.0	0.000
11420	0.0	0.000
38409	0.0	0.000
14463	0.0	0.000
31282	0.0	0.000
46018	0.0	0.000

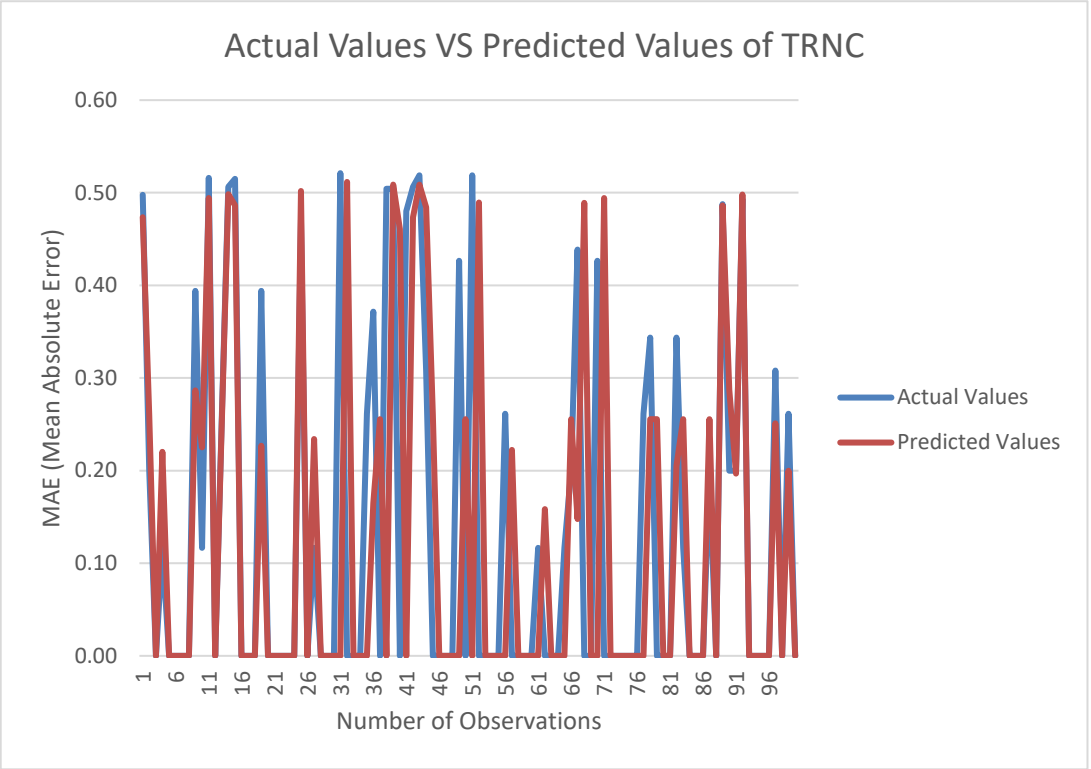

Figure 6.7: Some actual and predicted values through stacking regression of TRNC

Chapter 7

CONCLUSION AND RECOMMENDATIONS

This study discusses the rainfall dataset of TRNC and 6 different stations datasets that were used for short-term prediction by applying machine learning models. In this regard, an analysis was conducted with most prominent techniques, strategies, and algorithms in machine learning technology that was used to forecast rainfall amounts. This was accomplished in order to produce an accurate, robust, and reliable rainfall prediction model. Decision methods (regression models) such as Decision Tree, Random Forest, Bagging Regression, and Stacking Regression were used in this study to develop an efficient forecasting model for the entire TRNC and 6 representative datasets (Girne, Guzelyurt, Lefkosa, Gazimagusa, Gecitkale, and Karpaz), separately from 1995 to 2022 by using 5 different daily meteorological parameters (average temperature, average wind speed, average wind direction, average humidity, and average surface pressure), so as to generate short-term rainfall amount on daily basis as an output (prediction). The dataset was separated into two distinct components as the training (90 percent) and the testing (10 percent). In order to validate the trained data, the actual values of the rainfall gathered from NASA, and the predicted values obtained from the appropriate machine learning models that were selected in this study for each region separately, were compared. In addition, to create a model, the dataset was cleaned (not to have any missing value), classified (organized accordingly), and normalized (by applying power transformation technique) before it was used.

The proposed model is optimized using data normalization techniques, considering seasons, and calculating wetness and dryness through Python programming language. After evaluating the performance of each decision model mentioned above, based on the statistical measure values, for all regions separately and as a whole of TRNC, the Stacking Regression model, was found to be the most appropriate model, where R^2, MSE, and MAE reached 95.66, 0.0428, and 0.0821 for the whole TRNC.

From the experienced gained from this study, the below given recommendations are:

1. It will be a good attempt to group the rainfall amounts based on different sub-groups and used them as an output, so as to develop a model for short-term and long-term forecasts.
2. As an input, solar radiation data and evaporation values could be amended as an input of effective meteorological data so as to achieve higher accuracies.
3. This proposed short-term model forecasts data on daily bases due to the daily based dataset used as an input, if a longer period data prediction is required, the similar methodology could be applied however, it is necessary to organize the dataset (by averaging them) based on those time interval that we want to forecast.

REFERENCES

Abdulkadir, R. A., Ali, S. I. A., Abba, S., & Esmaili, P., (2020). Forecasting of daily rainfall at Ercan Airport Northern Cyprus: a comparison of linear and non-linear models. *Desalination Water Treat, 177*, 297–305.

Akıntug, B., & Baykan, O. (2000). Isohytes, iso-evapotranspiration, iso-drought, and iso-severity intensity curves of TRNC. Advances in Civil Engineering.

Anjum, S. A., Wang, L., Salhab, J., Khan, I., & Saleem, M., (2010). An assessment of drought extent and impacts in agriculture sector in Pakistan. *Journal of Food, Agriculture & Environment, 8*(3/4 part 2), 1359–1363.

Bartlett, P., Freund, Y., Lee, W. S., & Schapire, R. E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics, 26*(5), 1651–1686. Retrieved from https://www.cc.gatech.edu/~isbell/ tutorials/boostingmargins.pdf

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123–140.

Cheng, M., & Qi, Y. (2002). Frontal Rainfall-Rate Distribution and some conclusions on the threshold method. *Journal of applied meteorology, 41*(11), 1128–1139.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., & Kløve, B. (2018). River suspended sediment modelling using the CART model: A comparative study of machine learning techniques. *Science of the Total Environment, 615*, 272–281. doi: https://doi.org/10.1016/j.scitotenv.2017.09.293

Collier, C. (2003). On the formation of stratiform and convective cloud. *Weather*, *58*(2), 62–69.

Cook, R., & Weisberg, S. (1999). Applied regression including computing and graphics Wiley. *New York, NY*.

Dong, H., Gao, Y., Fang, Y., Liu, M., & Kong, Y. (2021). The short-term load forecasting for special days based on bagged regression trees in qingdao, China. *Computational Intelligence and Neuroscience*, *2021*, 1–16.

Fox, D. G. (1981). Judging air quality model performance: a summary of the AMS workshop on dispersion model performance, woods hole, Mass., 8–11 September 1980. *Bulletin of the American Meteorological Society*, *62*(5), 599– 609.

Frank, M., & Wolfe, P. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, *3*(1-2), 95–110.

Froehlich, D., & Dhawan, D. A. K. (2017). Guidelines for Instrumentation of Large Dams. Central Water Commission. Retrieved from https://www.researchgate.net/publication/321299610

Geetha, A., & Nasira, G. (2014). Data mining for meteorological applications: Decision trees for modeling rainfall prediction. In *2014 IEEE international conference on computational intelligence and computing research* (pp. 1–4).

Gray, W., & Seed, A. (2000). The characterisation of orographic rainfall. *Meteorological Applications: A journal of forecasting, practical applications,*

training techniques and modelling, *7*(2), 105–119.

Harrou, F., Saidi, A., & Sun, Y. (2019). Wind power prediction using bootstrap aggregating trees approach to enabling sustainable wind power integration in a smart grid. *Energy Conversion and Management*, *201*, 112077.

Haurwitz, B., & Austin, J. M. (1944). Climatology. New York; London: Mcgraw-Hill Book Company.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning.

Kumar, R. (2013). Decision tree for the weather forecasting. *International Journal of Computer Applications*, *76*(2), 31–34.

Lian, B., Wei, Z., Sun, X., Li, Z., & Zhao, J. (2022). A Review on Rainfall Measurement Based on Commercial Microwave Links in Wireless Cellular Networks. *Sensors*, *22*(12), 4395.

Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14–23.

Manzoor, B., Othman, I., Durdyev, S., Ismail, S., & Wahab, M. H. (2021). Influence of artificial intelligence in civil engineering toward sustainable development—A systematic literature review. *Applied System Innovation*, *4*(3), 52.

Marsland, S. (2015). *Machine learning: an algorithmic perspective*. CRC press.

Nourani, V., Elkiran, G., & Abba, S. (2018). Wastewater treatment plant performance analysis using artificial intelligence–an ensemble approach. *Water Science and Technology*, *78*(10), 2064–2076.

Pagano, T., & Sorooshian, S. (2002). Hydrologic cycle. *Encyclopedia of Global Environment Change*.

Payab, A. H., & Türker, U. (2018). Analyzing temporal-spatial characteristics of drought events in the northern part of Cyprus. *Environment, development and sustainability*, *20*, 1553–1574.

Rodgers, E., & Adler, R. (1981). Tropical cyclone rainfall characteristics as determined from a satellite passive microwave radiometer. *Monthly Weather Review*, *109*(3), 506–521.

Seera, M., Lim, C. P., Ishak, D., & Singh, H. (2011). Fault detection and diagnosis of induction motors using motor current signature analysis and a hybrid FMM– CART model. *IEEE transactions on neural networks and learning systems*, *23*(1), 97–108.

Seyhun, R., & Akıntuğ, B. (2013). Trend analysis of rainfall in north Cyprus. *Causes, impacts and solutions to global warming*, 169–181.

Shaikh, L., & Sawlani, K. (2017). A rainfall prediction model using artificial neural network. *International Journal of Technical Research and Applications*, *5*(2), 45–48.

Sharifi, Y., & Ergil, M. (2006). Hydro-climatological variations and trends in TRNC. M.Sc. Thesis.

Shiklomanov, I. A., & Rodda, J. C. (2003). *World water resources at the beginning of the twenty-first century*. Cambridge University Press.

Singh, J., Knapp, H. V., Arnold, J., & Demissie, M. (2005). Hydrological modeling of the Iroquois river watershed using HSPF and SWAT 1. *JAWRA Journal of the American Water Resources Association*, *41*(2), 343–360.

Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin*, *54*.

Tiwari, N., & Singh, A. (2020). A novel study of rainfall in the indian states and predictive analysis using machine learning algorithms. In *2020 International Conference on Computational Performance Evaluation (ComPE)* (pp. 199–204).

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern recognition*, *44*(2), 330–349.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259.