MACHINE LEARNING MODEL FOR PRECIPITATION FORECASTING IN KENYA

DAMARIS MUTHOKI MULINGE

A RESEARCH PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY IN THE SCHOOL OF COMPUTING AND MATHEMATICS IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE OF THE COOPERATIVE UNIVERSITY OF KENYA.

2025

DECLARATION

Declaration by the Candidate

This research project is my original work and has not been presented for the award of a degree in any university or for any other award.

Damaris Mulinge

Damaris Muthoki Mulinge
MDATC01/6065/2022

Date

Declaration the by Supervisors

We confirm that the work reported in this project was carried out by the candidate under our supervision and has been submitted with our approval as university supervisors.

Signed: _____ Date: 22/11/25

Prof. John M. Kihoro
Department of Mathematical Sciences
The Cooperative University of Kenya

Digitally Signed By Dr. Shadrack Stephen Madila
Sat Nov 22 12:16:03 EAT 2025

Signed: _____ Date: _____

Dr. Shadrack Madila
Department of Information & Communication Technology
Moshi Cooperative University.

DEDICATION

I dedicate this work to the Almighty God for His Grace, and Wisdom and to my mother Mrs. Agnes Henry for her encouragement and inspiration.

ACKNOWLEDGMENT

My deepest gratitude goes to God who provided all that was needed to complete this study. My sincere appreciation and honor to my supervisors Prof. John M Kihoro and Dr. Shadrack Stephen Madila for their remarkable contributions and constructive guidance, support, and invaluable feedback throughout this study that has been crucial to the success of this work, May the Almighty God reward them all. My thanks i also extend to Cooperative University of Kenya for provision of all necessary resources and a supportive environment for my study. To my family members especially my mother Mrs. Agnes Henry who has been prayerful continually throughout my study. My special gratitude to my priest and spiritual head. Pastor Benjamin Mwirigi for prayers, encouragement.

Table of Contents

List of Tables

List of Figures

Abbreviations

ARIMA Autoregressive Integrated Moving Average

ASALs Arid and Semi-Arid Lands

GCMs General Circulation Models

IPCC Intergovernmental Panel on Climate Change

KMD Kenya Meteorological Department

ML Machine Learning

RF Random Forest

SVM Support Vector Machine

XGB Extreme Gradient Boosting

Abstract

Accurate precipitation forecasting is important for mitigating the impacts of climate variability in Kenya, where erratic rainfall events considerably affect agriculture, water control, and disaster preparedness. Traditional methods such as ARIMA (Autoregressive Integrated Moving Average) and NWP (Numerical Weather Prediction) have been shown to struggle with complex weather patterns due to linearity assumptions, high computational demands and limited spatial resolution. This paper developed and evaluated an XGBoost-based machine learning model to enhance precipitation predictions both long-term and short-term. Utilizing a 20-year weather dataset (2004 - 2024) with 7300 daily data records sourced from online Visual Crossing Weather Data, key features include temperature, humidity, wind speed, lagged precipitation (1-7), rolling means and seasonal encoding to capture bimodal rainfall patterns of the months of march-May, and October-December. Data processing involved min-max normalization of 0-1 range, feature selection, sin/cosine transformations for seasonal patterns, and temperature- humidity interactions for connective modeling processes. The dataset used was split with 80% for training and 20% for testing and a temporal split ≤ 2020 for training and > 2020 for testing maintaining the chronological data order. The initial attempts exhibited poor performance with low R2 = 0.066 and a high RMSE=1.06. The model again was re-evaluated using XGBoost binary classification shift to predict the likelihood of rain/no-rain tomorrow. Bayesian optimization and GridSearchCV hyperparameter tuning was applied with default 0.5 threshold adjustment for improved rain class sensitivity using classification metrics and resulted 76.76% accuracy, 70.14% precision, 33.36% recall, 45.12% F1- Score and ROC-AUC 0.75. Post-tuning accuracy by reducing the threshold to 0.3 to capture missed rainfall events: 73% accuracy, no-rain precision and recall 81%, 53% rain precision, 54% recall, F1 Score 54%. Temperature-humidity interaction as the top predictor in feature importance. The results indicated that the XGBoost-based model with 73% accuracy and 54% recall in forecasting rain/no-rain occurrences forecasting to support agricultural planning, water resource management and early warning for disaster preparedness in Kenya's climate vulnerable regions.

CHAPTER ONE:INTRODUCTION

1.1 Introduction

Accurate precipitation forecasting is critical for mitigating the impacts of climate change, especially in Kenya, which is vulnerable to extreme weather events. Many areas in Kenya , face challenges such as food insecurity and water scarcity caused by unpredictable rainfall patterns. Due to ever-increasing uncertainty about global climate changes, precipitation variability poses a significant effect on the local communities dependent on natural resources, agricultural practices, and the region's socio-economic stability (IPCC, 2022) (KMD, 2023).

Kenya's Climatic and weather conditions extremely contribute to food insecurity and water scarcity and about 75% of Kenya's country's marks . Due to erratic and poor rainfall patterns, agricultural practices in these regions tend to be unreliable. Scarcity of rainfall contributes to worsening the situation by increasing frequent and severe droughts influenced by change of climate hence affecting water availability and agricultural productivity. Surprisingly, during the rainy seasons, various regions experience floods that disrupt livelihoods, and damage crops polluting water sources, hence leading to food challenges and water insecurity (Affoh et al., 2022). Traditional approaches to weather forecasting though valuable, have shown to fail due to historical reliance on statistical methods to establish correlations between rainfall and various meteorological factors, such as temperature, wind speed, and humidity, based on geographic coordinates. Weather prediction has seen a variety of approaches in recent years based on, traditional approaches Genetic Algorithms and Neural networks but these fail to capture the complex relationships between various factors that affect weather. However, rainfall dynamics' complex and non-linear nature presents inherent difficulties for accurate weather forecasting due to the complexity of capturing complications of the growing climate changes. Hence, the need to deploy the use of advanced machine learning models to develop an accurate weather forecasting model is essential to enhance precipitation forecast abilities in the regions. The task of unpredictability of climate events poses threats to

local communities and interferes with the communities' sustainable development. Machine learning models particularly XGBoost, offer a data-driven approach to enhance forecasting accuracy due to their ability to handle complex relationships and their ability to handle large datasets, which has shown promising outcomes in precipitation prediction across various areas. Despite these strides, there remains potential for refining the accuracy of this machine-learning algorithm. This proposal outlines the study to develop and evaluate an XGBOOST model to enhance precipitation forecasting in Kenya focusing on key climatic variables such as temperature, humidity, wind speed, lagged precipitations (1-7), rolling means and seasonal encoding.

1.2 Background of the Study

Forecasting precipitation plays a significant role in various sectors including agriculture, water resource management, and disaster preparedness. In Kenya, many regions frequently experience extreme weather events that lead to erratic rain patterns, including prolonged droughts and floods, which intensify food insecurity, hinder economic development, and disrupt livelihoods (SAMWEL, 2021). Climate change cannot be ignored because its impacts lead to changes in rainfall and weather patterns, water resources, agriculture, and ecosystems (Kogo et al., 2021). Traditional forecasting methods, such as (NWP) Numerical Weather Prediction and ARIMA models, often struggle with data limitations, high computational costs, and low accuracy in localized forecasts (Rojas-Campos et al., 2023). The increasing variability in rainfall patterns necessitates advanced data-driven approaches to improve precipitation prediction accuracy. Kenya is located in East Africa. It is part of the Eastern region of the African continent, located below the Sahara desert, defining the characteristics of sub-Saharan Africa. Most of its regions are susceptible to the effects of climate change which affect precipitation due to the mono-culture of rain-fed agriculture and minimal water resources (Pello et al., 2021) (Owino, 2022). For years, forecasts of this nature such as Autoregressive Integrated Moving Average (ARIMA) models and General Circulation

Models (GCMs) have been used severally for decades in weather patterns, particularly precipitation forecasting. ARIMA models are quite popular in the business domain mainly because they are easy to use and give better results in time series compared to other models however, they are very basic and limited for the non-linear climate dynamics and interactions (Kontopoulou et al., 2023). Since GCMs forecast the specifics of the global climate, they may provide complex simulations of climate change. However, because of their high computational requirements and lack of visibility for regional rainfall forecasting, they have several limitations. Utilizing historical weather datasets, the machine learning models seek to provide reliable climate forecasts to help in managing water resources, as well as agriculture planning for people living in these areas.

1.3 Statement of the Problem

Regardless of climate science advancement, many traditional forecasting models seem to struggle in incorporating various datasets and adjusting to the rapidly climate conditions changes. The application of the XGBoost modeling technique is crucial to improve climate predictions and facilitate effective preparedness. The ever-growing vulnerability to climate change and variability in Kenya poses threats to agricultural productivity and livelihoods due to water scarcity. In many cases, traditional water management and agriculture fail to adjust the rainfall pattern unpredictability, extreme droughts, and floods. The effects of weather variability and change in Kenya's regions lead to land degradation, insufficient weather data access, and poor infrastructure, which cause the communities to struggle to create sustainable strategies for water management, food security, and adaptation to weather change. As a result, communities in semi-arid regions struggle to develop sustainable strategies for water conservation, food security, and resilience to climate change. The competency by the currently employed traditional weather forecasting techniques like Numerical Weather Prediction (NWP), synoptic forecasting, and analog methods to sufficiently predict the short and medium weather patterns within the regions and understanding that the numerical

weather prediction is dependent on extensive observational data that may be lacking, its substantial computational demands making real-time prediction a problem, and its coarse spatial resolution, which usually fails to accurately capture localized weather phenomena. This problem presents a need to address this gap by applying the XGBoost Machine learning-based forecasting model to improve the reliability and accuracy of precipitation forecasts, hence aiding stakeholder decision-making and disaster preparedness in Kenya.

1.4 Project Objectives

1.4.1 General objective

To develop and evaluate XGBoost Machine Learning models to improve the accuracy of precipitation forecasting in Kenya.

1.4.2 Specific objectives

i. Evaluate the performance of the XGBoost model in precipitation prediction trends using temperature, humidity, wind speed, and lagged precipitation as key variables.

ii. Optimize the XGBoost model via hyperparameter tuning to enhance precipitation forecasting accuracy.

iii. Assess the implications of the enhanced forecasts in agriculture, water control management, and catastrophe preparedness in Kenya.

1.5 Research Questions

i. How can XGBoost be used to predict precipitation based on humidity, temperature, wind speed, and lagged precipitation as key variables?

ii. The performance of XGBoost in precipitation forecasting can be enhanced using which optimization method?

iii. Agricultural planning, water control management, and catastrophe preparedness can be informed how using the improved precipitation forecasts.

1.6 Significance of the Study

This research significantly contributes to the field of precipitation forecasting by addressing the traditional prediction techniques and promoting Machine Learning algorithms (ML). Enhanced forecasting capabilities can be can inform irrigation practices, schemes for crop rotation and measure water conservation hence improving food security. By identifying and forecasting the extreme weather occurrences such as drought and floods, the study also provides actionable insights for mitigating the effects of climate variability affecting vulnerable communities (Gleick and Cooley, 2021). The findings can aid policymakers in developing strong preparedness and strategies for allocating resources hence promoting resilience to climate change. Applying Machine Learning models offers a scalable solution that expands climate resilience to Kenya regions facing environmental challenges. The use of Machine Learning algorithms for precipitation forecasting in this study offers a better approach to overcoming the limitations of traditional techniques, setting a benchmark for the advanced machine learning models' performance in precipitation prediction, hence leading to insightful effectiveness for short-term, medium-term, and long-term predictions. The findings of this research will establish vital assumptions regarding agricultural planning and water control management in Kenya through precise precipitation forecasts from the XGBoost model which enhances agricultural planning by predicting weather factors that impact water resources and crop yields in targeted regions ((Sarma et al., 2024)). Mainly used in crop production and irrigation scheme decision-making processes Enhanced Prediction allows better decision support. This research establishes dependable precipitation forecasting to help regions in Kenya effectively manage their water resources since they practice rain-fed agricultural systems (Deo et al., 2022). The study results enable policymakers together with stakeholders to recognize why advanced forecasting methods are vital for building

climate resilience and disaster readiness (Abisha et al., 2022). The research evaluates how these models affect computational resources as well as resource constraints so scientists can create a practical implementation path for advanced prediction methods in areas of limited resources. The research study fills an existing gap in the literature review by demonstrating experimentally that Machine Learning algorithms deliver effective precipitation forecasting results in Kenya. The application of XGBoost Machine learning model in environmental science research becomes more comprehensive through this study. This fills the gap in the

literature review, providing empirical proof of the effectiveness of using Machine Learning algorithms for precipitation forecasting in Kenya. This provides a broader understanding of how the XGBoost Machine learning model can be applied in environmental science research.

1.7 Expected Outcomes of the Study

This study aims to develop an XGBoost-based model for precipitation forecasting to improve the precipitation accuracy of precipitation predictions in Kenya. The XGBoost model is expected to outperform the traditional and statistical approaches, hence providing dependable short-term and long-term precipitation forecasts.

i. Enhanced precipitation prediction accuracy.

ii. Optimize the model using Hyperparameter tuning.

iii. Improve agricultural and water resource planning.

iv. Scalability and applicability for real-world use.

v. Contribution of machine learning in precipitation prediction

1.8 Justification of the Study

There are gaps and challenges that exists in the study in precipitation prediction and strategies to adapt due to climate differences in Kenya regions. Traditional techniques fail to handle complex dynamic climate conditions, they rely on past weather events and are statistical based ARIMA, and struggle to predict short, and Medium-term precipitation patterns. Therefore XGBoost machine learning algorithm offers a data-driven solution to these challenges because of the models ability to handle non-linear interactions and enhance prediction accuracy.

1.9 Scope

The research utilizes a 20-year historical weather data with temperature, humidity, wind speed, lagged precipitations as the key variables. The study involves collecting and analysing historical weather data to develop and optimize the XGBoost Model. It will consider the possible limitations and challenges associated with Machine Learning (ML) models in the current weather prediction system. The study research imply the model development, training and evaluation employing historical weather data from Visual Crossing weather data web, to show hoe the XGBoost Model can enhance precipitation prediction successfully.

1.10 Limitations of the Study

While this study aims to improve precipitation forecasting, it is crucial to acknowledge the limitations that may affect the research outcomes. The accuracy of ML models depends on the quality and the totality of historical weather data, which may be incomplete or noisy. Machine Learning models require significant computational resources for training and evaluation. Limited access to such resources may restrict the complexity of the models that can be implemented and affect the study's ability to explore the full range of model configurations. The findings derived from the study may apply differently to different

regions due to climate disparities. Further research may be required to apply the results to different places with different precipitations patterns. The study is restrained by a specified period for collecting, model development, analysis. The research period may limit the scope of the study, affecting the interpretation of the analysis and the models that can be tested. The availability of time may not allow the thorough exploration of hyperparameters that are possible for the model. The interactions between the variables which include temperature, humidity, wind speed, and lagged precipitations can be tedious to model due to the complex interactions between these variables. Weather systems are inherently complex and influenced by many factors, making accurate predictions challenging. Capturing the effects of micro-climates or local weather variations may not be fully captured, as machine learning methods rely on historical data patterns, and may struggle with rare or high weather irregularities.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The main subject of this chapter involves the development process for machine learning models including the examination of the XGBoost Model Approach to boost precipitation forecast accuracy, and the XGBoost Model serves as the primary focus to develop precipitation prediction capabilities within Kenya. General Circulation Models and Autoregressive Integrated Moving Average (ARIMA) have been extensively used for over years but they have experienced challenges because they struggle when operating with complex datasets or complex climate patterns. Autoregressive models represent a common solution for time series forecasting because data linearity is one of their core assumptions as they make predictions by combining previous observations with error terms. The main limitations of Autoregressive models lead to substantial performance issues which occur during model predictions of non-linear climate changes together with complex meteorological system relationships (Khodakhah et al., 2022). Advanced modeling systems such as General Circulation Models (GCMs) serve as tools that stimulate climate processes in the Earth System, and their approach takes various physical sea and land phenomena into account. A detailed simulation of worldwide climate appears in GCMs yet their performance comes with large computational needs. Systems experience two primary limitations which include heavy computing needs and restricted ability to predict small geographic areas. The XGBoost machine learning model represents one of several data-driven tools that use XGBoost as their primary data-processing system. XGBoost enables processing of enormous data volumes and excels at detecting complex patterns so it produces better predictions. The machine learning models find applications as their domain expands through forecasting performance operations. This review investigates precipitation forecasts with XGBoost models on top of their ability to predict. It investigates modern machine models and their use in climate forecasting as well as their implementation in weather prediction systems of Kenya. These

algorithms handle specific challenges that occur throughout different regions of Kenya, which face unique difficulties caused by significant variations across this area.

2.2 Machine Learning Algorithms in Precipitation Forecasting

Machine Learning (ML) has transformed has transformed predictions by enabling the models to capture complicated relationships between precipitation variables, and making predictions more improved in terms of accuracy. Support Vector Machine (SVM) , Random Forest (RF) , and XGBoost machine learning models have critically expressed their potential in recording non-linear relationships and the improvement of prediction accuracy. But, the XGBoost model outperforms them as the leading model because it is more robust, more scalable, and excellent in handling tabular datasets (Mishra et al., 2024). On the other hand, combining multiple machine learning algorithms to make an ensemble model, the XGB model as a standalone which inherently incorporates Gradient Boosting to boost the model's performance, hence improving the prediction accuracy.

2.2.1 XGBoost in Machine Learning

XGB is an advanced implementation of gradient boosting that via system optimization and algorithm enhancement, increases performance and speed. (Sagi and Rokach, 2021). XGBoost machine learning algorithm has been defined by scholars as a high-performing machine learning algorithm that has been used in different field successfully including finance, healthcare and in metrology field, because of its ability to handle non-linear relationships, ability to handle missing data values efficiently, thus managing overfitting and providing accurate predictions, making it fit specifically for precipitation forecasting (Babu Nuthalapati et al., 2024). Studies show that XGBoost builds decision trees sequentially, with each tree correcting the errors of its predecessor and adjusting the prediction accordingly. This iterative process enables the gradient boosting to calculate the residual errors and fit new trees to minimize the errors while optimizing feature selection, hence making it a robust

choice for climate forecasting. Scientists have also pointed out the ability of the XGBoost to handle missing values effectively because it incorporates inbuilt imputation methods, which expand the degree of data integrity without requiring a specific imputation approach. L1 and L2 standardization techniques play an important role in controlling overfitting and hence securing the models' standardization to hidden weather data.

2.2.2 Identifying Climate Vulnerabilities and Risks

This sub-section identifies the particular vulnerabilities and risks associated with climate variability, using the forecasting results. The attention is on a focus on climate risks associated with droughts and floods, and other extreme weather events, and how the reliable forecasts can aid climate risk management strategies (Affoh et al., 2022).

2.2.3 Formulating Agricultural Adaptation Strategies

Forecasting data can also be applied to agricultural decision-making and, thus, formulating improvement strategies to maximize crop yield and food security is essential. Factors such as crop rotation, selecting drought resistant crops, and sustainable irrigation (based on climate forecast) are important for implementing viable strategies (Habib-ur Rahman et al., 2022).

2.2.4 Water Resource Management Strategies

Developing Water Resource Allocation Strategies Water resources required for rain-fed dependence are essential and require effective management strategies. This is a sub-section that provides viable and sustainable water use practices including developing resource allocation practices, irrigation practices and building water storage facilities based on precipitation forecasts (Gleick and Cooley, 2021).

2.2.5 Policy Development and Implementation

The information generated by using precipitation forecast data can provide increased resilience policies towards climate vulnerability. For example, useful frameworks that can be developed as part of policy frameworks would be disaster preparedness, resource efficiency and community (Parmesan et al., 2022). Reliable precipitation forecasting can provide increased resilience policies that can enhance decision making related to climate vulnerability. Reliable forecasting can stimulate and develop disaster preparedness, manage and ambiguous water resources and use crops and manage crop production according to the proper seasons and not in situations of urgency. Furthermore, it is widely accepted that reliable forecasting would provide enhanced early warning capability to communities and provide smart resource allocation as well as support the implementation of community engagement adaptation strategies. Studies have demonstrated that policy frameworks can focus on sustainable adaptation factors related to rainfall variability when further increasing the knowledge that has been generated from machine learning-based rainfall forecasting (Anwar et al., 2021) (Mishra et al., 2024). The various policy opportunity development can utilize water precipitation forecasts that includes innovative approaches using XGBoost as a single model to formalize policy tools using drought and flood-based climate predictions to ensure better data generation and evidence-based decisions.

2.2.6 Community-Based Initiatives

Local communities must actively participate in adaptation strategies for them to achieve success. The implementation of community-based programs features farmer training as well as joint local weather pattern assessment through Indigenous techniques and scientific measurements to strengthen resilience (Kilonzo, 2022) (Okedele et al., 2024). The local community has mapped regions that require precipitation forecasts to develop adaptation strategies (Oino and Musau, 2024). In addition, integrating Indigenous knowledge with scientific forecasts has been acknowledged as an important element in formulating effective

adaptations to weather emergencies in Africa, particularly in Kenya. This method improves the reliability of the forecast, ensuring that local communities can better anticipate and respond to irregular precipitation. Also, there has been community-led mapping that highlighted areas needing precipitation forecasts for the preparation of adaptation schemes.

2.3 Theoretical Framework

This study uses the theoretical framework of supervised learning, which is trained with labeled historical weather data to make predictions of future precipitation. Gradient boosting theory underpins the work of XGBoost to show how sequential decision trees combined will perform better than a single decision tree, adding new sequential decision trees that correct each other's errors. Finally, feature selection theory informs the approach by emphasizing the importance of identifying the subset of important weather variables that can help optimize certain aspects of prediction accuracy (e.g., temperature, humidity, wind speed, and lagged precipitation).

2.4 XGBoost Theory

XGBoost is an enhanced ML algorithm that uses the framework of gradient boosting. XGBoost maximizes forecasting performance by building decision trees sequentially, where each successive tree corrects for errors made by the worse performing trees. At its core, XGBoost uses additional standardization methods such as L1 (lasso) and L2 (ridge) regularization techniques which help to avoid overfitting the model while also maximizing generalization during the model building process. The additional regularization used by XGBoost also allows for processing by handling missing values and can process more data efficiently than other methods, justifying the need to use it for precipitation prediction. XGBoost differs by reducing a loss function and uses gradient descent and adaptive learning to build the model to provide accurate and reliable precipitation forecasting based on historical weather data.

2.5 Conceptual framework

The conceptual framework determines the link between weather variables, ML methods, and predictions/outputs. Weather independent variables such as temperature, humidity, wind, and lagged precipitation, rolling means and seasonal encoding are input features to be generated into the model. The XGBoost feature is the predictions method for the processor of the inputs, which will be able to provide accurate precipitation forecasts. The dependent variable is the predicted precipitation parameter then can enable better decision making related to agricultural management practices, water resource management, and disaster management in the broader sense.

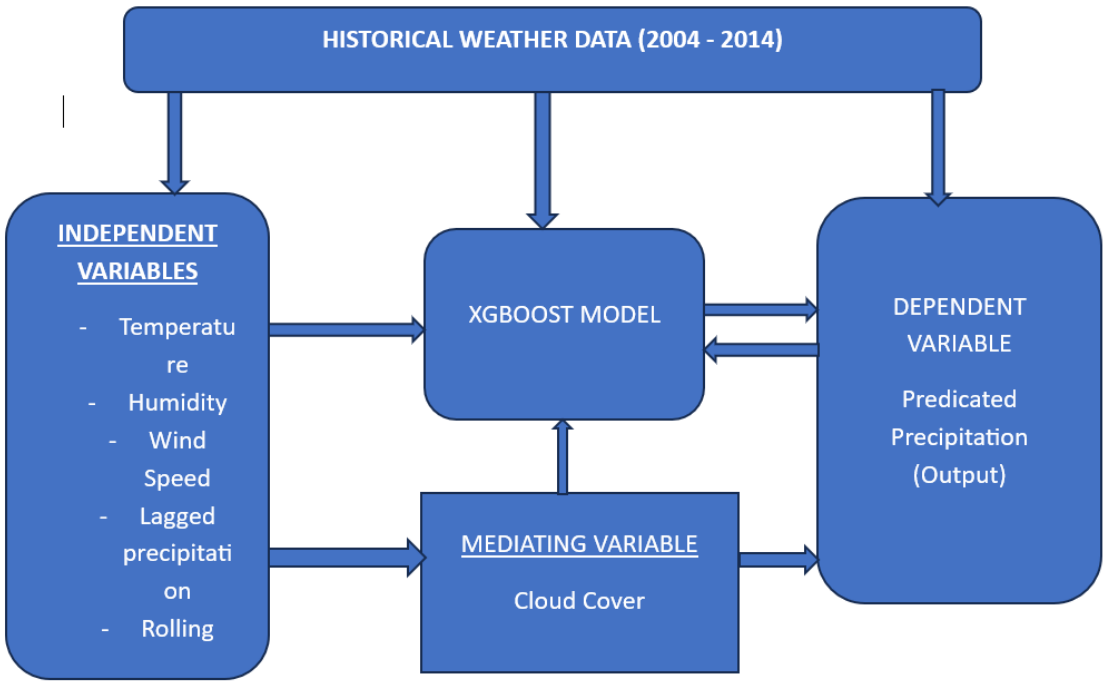

Figure 1: Conceptual Framework.

2.6 Research Gap

The analysis of existing precipitation forecasting systems shows considerable development and also significant gaps that this study aims to fill. Most of the current literature uses statistical approaches or machine learning models for the purpose of rainfall predictions. These systems have limitationsin their ability to manage the complexity and variability of weather data. For example, traditional time series approaches such as ARIMA or natural singular machine learning methods (the list ofwhich is long) may prove insufficient at the task of taking into account the various interactions between weather variables. Individual application of recent algorithms such as XGBoost, although very promising, also suffers from lack of accuracy and reliability. While powerful, it suffers from overfitting, missing data, and computational inefficiencies on ever-more-burgeoning big climate datasets. However, recent advances especially in machine learning

XGBoost offer a promising choice however have not been largely studied in the context of regional precipitation forecasting in Kenya. Therefore, this study seeks to bridge this gap by developing and optimizing an XGB model customized for Kenya's weather conditions while addressing forecasting accuracy and computational efficiency challenges.

2.7 Empirical Study

This section provides some relevant empirical studies regarding precipitation prediction in predicting the weather using advanced ML Models like RF, XGBoost, and SVM. The studies related to weather forecasting, discussing approaches, results, and the factors that played a role in model development.

Table 1: Summary of Empirical Study

Author(s) &Year	**Study Title**	**Methodology**	**Findings**	**Gap**
Aydin & Ozturk (2021) (2022	Performance analysis of XGBoost classifier with missing data	XGBoost-based imputation for handling missing values	Shown improved classification accuracy with XGBoost's native handling of missing data	Further validation is needed for large-scale meteorological weather datasets.
Anwar, M. &Winarno (2021) Rainfall prediction using extreme gradient	Rainfall prediction using extreme gradient boosting	Gradient boosting for weather forecasting	Discovered that XGBoost improves both short-term and long-term precipitation prediction accuracy	Need for additional comparison with deep learning models for long-term precipitation forecasting accuracy
Kontopoulou et al. (2023)	A Review of Traditional Models (ARIMA, GCMs) vs. Machine Learning Approaches for Time	Comparative analysis of Traditional Models and ML models	ML techniques, particularly XGBoost, generally outperform other ML model and traditional	Need for region specific model optimizations

	Serie		Models in precipitation prediction	
Mishra et al. (2024)	A Review of Traditional Models (ARIMA, GCMs) vs. Machine Learning Approaches for Time Series	Time series modeling using XGBoost	XGBoost provided high predictive accuracy for rainfall forecasting	Limited evaluation of feature selection impact on climate forecasting

3 CHAPTER THREE: METHODOLOGY

3.1 Introduction

This chapter details the approach of the methodology to achieve the study objectives, basically developing and implementing an XGBoost-based ML model for improving precipitation forecasting in Kenya. The study employs a quantitative, data-driven paradigm using historical weather data to predict daily precipitation amounts through regression-based forecasts and to classify whether events result in rain or no rain, utilizing a temporal split ($\leq 2020 > 2020$) for temporal realism. It details the research paradigm, emphasizing the data collection, preprocessing and analysis, population and sampling techniques, feature selection, model design, model development, evaluation criteria, and development, addressing the challenges faced in Kenya regions due to bimodal rainfall events of long rains during the months of may to march and October to December short rains . The chapter also outlines details on the basic learner used in XGBoost, the tools and software used in the implementation process utilized, and the Ethical considerations relevant to the research, ensuring a structured and reproducible workflow, compliant with the Kenya Data Protection Act (2019).

3.2 Research Design and Paradigm

The research project used a quantitative research design under the positivist paradigm, focusing on predictive modeling using the XGBoost algorithm. The approach involved training and testing the model on historical weather data to forecast rainfall occurrence (rain = 1, no rain = 0). This design enabled objective model evaluation based on performance metrics, including accuracy, precision, recall, and F1-score, and benchmarking against multiple linear (Ridge) regression and ARIMA to assess performance.

3.3 Population, Train-Test Split Technique

3.3.1 Population

In this time series dataset, population refers to all observations available for the research project, that include all historical records for temperature, wind speed, humidity, lagged precipitations, seasonal encoding, etc. with a given period of time. The population for this study comprised approximately 7,300 daily weather observations recorded between January 2004 and March 2025. To ensure temporal integrity, a chronological train–test split was applied. Approximately 80% of the data (2004–2020) was used for training, while the remaining 20% (2021–2025) served as the test set to evaluate model performance.

3.3.2 Train-Test Split Technique

The methodology employed a Chronological Split approach to maintain chronological order. The dataset was divided between train and test subsets as ¡=2020 train and ¿ 2020 test, preserving the temporal sequential weather observations.

3.4 Data Collection

Data collection involved sourcing a 20-year weather dataset from the historical Visual Crossing National Weather Database which provides both historical and real-time weather data (Visual Crossing, 2025). The data sourced includes daily temperature measurements, humidity, wind speed, and data points of previous precipitation that directly impact precipitation. The data was carried through legit meteorological report verification by confirming consistency and completeness.

3.5 Data Processing

3.5.1 Data Cleaning

Data Cleaning involved missing data handling using the intrinsic XGBoost imputation methods, outlier detection using box plots to maintain data quality, inconsistencies, and correction using the available patterns contained in the data to predict missing information (Aydin and Ozturk, 2021) This technique was used to maintain data integrity and improve the accuracy of predictions. Redundant records and sensor faults were verified and excluded for enhanced quality of data. It has an inbuilt process that can cope up with missing values during training as it does not involve explicit imputation, but it learns the optimal direction for missing trees when building decision trees automatically.This technique was used to maintain data integrity and improve the accuracy of predictions. Redundant records and sensor faults were verified and excluded for enhanced quality of data. It has an inbuilt process that can cope up with missing values during training as it does not involve explicit imputation, but it learns the optimal direction for missing trees when building decision trees automatically

3.5.2 Data Transformation

Data normalization process was applied to achieve data consistency for different climate variables. The normalization procedure reduces data distortions that stem from varying units and measurement scales between variables. Through min-max scaling techniques data values were transformed to a range from zero to one to prevent domain-biasing by maintaining equal contribution of all features during the learning process, and feature engineering using lag-1 to lag-7 to record persistence of rainfall, sin/cosine transformations of month and day of the year was applied to record the bimodal periodic variations of March to May and the October to December and the relationship such as Humidity-temperature to model the convective interactions.

3.5.3 Scaling and Normalization

Normalization via min-max scaling transforms values into a range from 0 to 1 applied to prevent high variables from controlling the model while maintaining feature equality for the learning process

3.6 Feature Selection:

The most crucial predictors for accurate climate predictions are selected through feature selection methods. Features will be engineered to record temporal, seasonal and meteorological patterns for predicting precipitation using 7-lagged features for bimodal model seasonality, the interaction feature will capture convection changes of rainfall, Regression handling precipitation and classification for binary whether it will rain tomorrow or not. The intrinsic feature importance ranking of XGBoost enables identification of critical variables which results in keeping only relevant features. The model becomes more efficient after successively removing irrelevant features (Awad and Fraihat, 2023) who describe this technique as a dimensionality reduction method. Seasonal The final set of parameters used in the model include temperature, humidity, wind speed, and precipitation, which are the most powerful variables for climate prediction.

3.7 Model Development

The precipitation (rainfall) variable was transformed to predict next day rain. The first step involved preprocessing through parsing the date column into a standard date-time format, setting it as the index, converting column names to lowercase, and removing the white-spaces trailing. All recorded zeros in the precipitation days were retained to maintain the balance between no-rain and rain categories. Thereafter, feature engineering was conducted to improve the predictive capability of the model. This involved creating lag variables to capture the persistence of rainfall over the previous week. Using the sine and cosine

transformation of the month and day-of-year, seasonal characteristics were represented to model the bimodal patterns of rainfall. A further interaction feature integrating temperature and humidity was added; this was key in reflecting their joint impact on convective rainfall. Through shifting the precipitation column backwards by one day and assigning 1 to values of precipitation above 0, the target variable was generated. Thereafter, rows containing missing values that resulted from lag creation and shifting were deleted. The numeric variables were then normalized to a range of 0 and 1 using MinMaxScaler. The set was then split into training 80% and testing 20% subsets without shuffling; this helped preserve temporal ordering.

3.7.1 XGBoost Model Performance Evaluation

The initial XGBoost model via regression metrics aimed to predict daily precipitation amounts, however it reflected low results thus low predictive capability using the same engineered features. The model was re-evaluated to XGBoost binary classification to forecast the occurrence of rain/no-rain and the XGBoost binary classification simplifies the forecasting while maintaining the practical applications such as agriculture and water management practices.

3.7.2 Classification Formulas Metrics

3.7.3 Classification Formulas and Metrics

The XGBoost via binary classification predicts rain probability using a logistic function based on the ensemble of decision trees. The probability $P(y = 1|X)$, calculated as:

$$P(y = 1|X) = \frac{1}{1 + e^{-f(X)}} \tag{1}$$

where $f(X)$ = boosted tree ensemble output; to classify rain 1 / no rain 0

$$\text{Predicted Class} = \begin{cases} 1 & \text{if } P > \theta \\ 0 & \text{if } P \leq \theta \end{cases} \tag{2}$$

The XGBoost model performance was evaluated using binary classification metrics, with F1-Score showing the balance between precision and recall for the imbalanced rain class.

- **Precision**: Measures the accuracy of positive predictions:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

- **Recall**: Measures the ability to detect all positive instances:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

- **F1 Score**: Balances Precision and Recall, critical for the imbalanced rain class:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

- ROC = Area Under the curve (True Positive Rate vs False Positive Rate)

3.8 Model Validation and Evaluation

The evaluation of machine learning models were made possible through the Python libraries Scikit-learn and also through Pandas and NumPy alongside Matplotlib and Seaborn. The essential package for this study is Scikit-learn because it provides tools for model evaluation and cross-validation and performance metrics functions. The data manipulation components use Pandas libraries while numerical operations run under NumPy libraries. To visualize results Matplotlib and Seaborn serve as the selected packages which help present work efficiently through Pandas data loading. The code implements a function for model testing

purposes. The XGBoost Binary Classifier model was validated using a chronological train-test split, with 80% of the 7,300 daily weather records (≤ 2020) allocated for training and 20% (> 2020) for testing. This method successfully preserved the temporal sequence of the data, ensuring the integrity of lagged precipitation features (Lag 1 - lag 7) and seasonal encoding, a critical achievement for accurate forecasting in Kenya's climate-vulnerable regions. Validation included threshold tuning to optimize the decision boundary for rain/no-rain classification, minimizing overfitting by maintaining temporal consistency and optimizing model complexity. Performance was evaluated using accuracy, precision, recall, F1 score, and ROC AUC to assess the model's ability to distinguish rain from no-rain events.

3.9 Optimizing the XGBoost Model via Hyperparameter Tuning

Hyperparameter tuning was performed via GridSearchCV and adjusting the parameters along with the number of estimators and learning rate, with maximum depth and a subsample of 0.8. The classification threshold was tuned to improve rain detection, balancing the precision and recall considering the dataset imbalance. The XGBoost binary classification model demonstrated good performance in distinguishing between rain and no-rain cases.

Practical Implications: The practical use of XGBoost machine learning technology for climate forecasting extends across multiple commercial industries. Individuals engaged in agriculture employ improved precipitation forecasting to set both planting dates and irrigation periods which minimize crop damage from weather uncertainties. Water management reservoirs benefit from exact predictions through their use to allocate water resources properly and develop protective water-saving strategies. Severe weather alerts in disaster preparedness allow authorities to take immediate action which minimizes the effects of floods droughts and storms. Various industry stakeholders can achieve sustainable and precipitation-resilient goals through using reliable prediction information.

3.10 Ethical Considerations

The study adhered to ethical guidelines and legal policies using secondary data from Visual Crossing weather data. Data privacy was ensured by complying with the terms and use and proper citing. Permission for conducting the research was obtained from the National Commission for Science, Technology and Innovation (NACOSTI). To minimize bias, the model (XGBoost Binary Classifier) was developed using min-max normalization for fair scaling, ensuring transparent and reliable precipitation predictions for Kenya's climate-vulnerable regions.

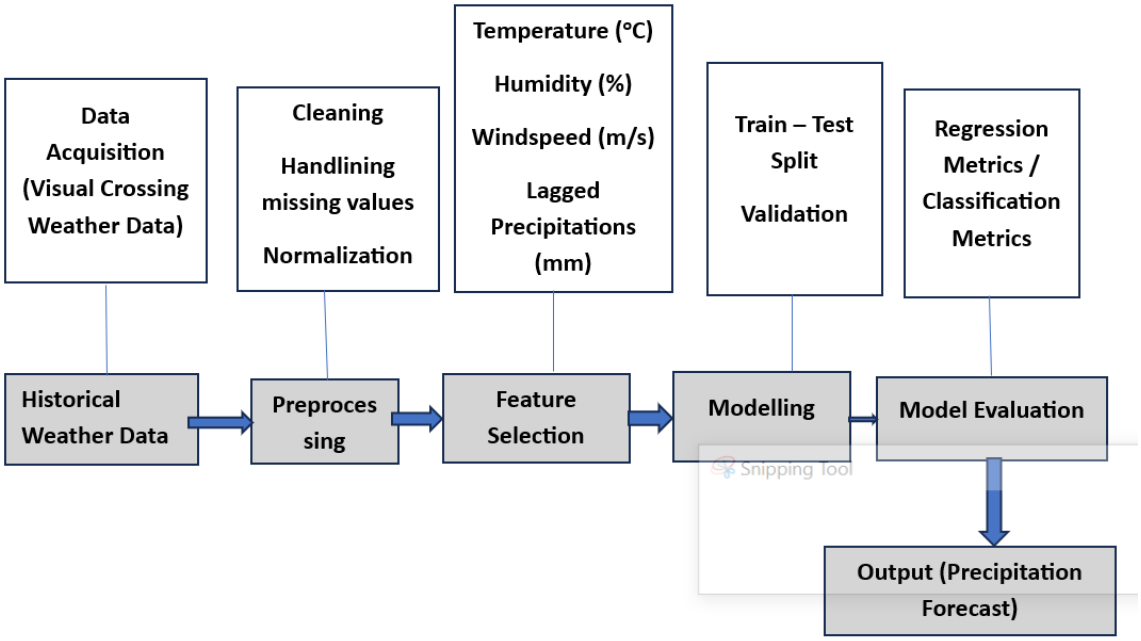

Figure 2: Model Development Process

4 CHAPTER FOUR: MODEL DEVELOPMENT, ANALYSIS AND RESULTS

4.1 Introduction

The chapter presents the findings of the machine learning model used to forecast precipitation patterns in Kenya, covering between 2004 and 2024, with 7300 daily data records sourced from online Visual Crossing Weather Data, and including the key meteorological features: Temperature, Humidity, Wind speed and Lagged Precipitation values, seasonal encoding and rolling means. This study applied a binary classification approach to predict rainfall occurrence on the next day using historical weather data. The original goal was to forecast daily precipitation amounts using XGBoost for time series regression. However, preliminary experiments revealed limitations in the model, particularly in capturing extreme rainfall events and seasonal dynamics within the dataset. Evaluation of the model performance was done using four primary metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R-squared (R^2). Regression-based forecasts exhibited low R-squared values and high root mean squared error (RMSE), making it less suitable for operational decisions in disaster preparedness and agricultural planning. Due to poor outcomes using XGB regression, the model shifted to XGBoost binary classification to predict the likelihood of rain, that is, whether it will rain tomorrow or not. This shift improved model interpretability, robustness, and relevance for real-world decision-making.

4.2 Data Analysis

4.3 Data Cleaning

Dates were parsed and converted into a proper date-time format to enable time-based operations, and the dataset was sorted chronologically. Only records with precipitation greater than zero were retained to focus on rainfall events. The date-time column was set as

the index to facilitate time series analysis, and column names were standardized by stripping spaces and converting them to lowercase for consistency. Missing values were checked, and any incomplete records were removed to ensure data quality before modeling.

4.4 Exploratory Data Analysis

4.4.1 Descriptive Statistics

The first step involved summary statistics for the four variables of interest. As expressed in table 1, the average temperature for the country was 66.94 (SD = 2.41) with the minimum and maximum records being 58.20 and 76.20 respectively. The small deviation, though showing steadiness in the value of temperature, is large enough to express worry in terms of climate change. The second variable, humidity, recorded 77.03 (SD = 7.56) on average with maximum and minimum values of 37.60 and 96.40 correspondingly. In terms of precipitation, the average was .29 (SD = .60) with the large standard deviation expressing irregularity in rainfall patterns, suggesting the need to establish accurately predicting models. The final wind speed, recorded an average of 15.68 (SD = 6.13) with a large variation recorded in minimum and maximum values of 2.90 and 117.40, a factor attributed to the different wind patterns within Kenya.

Table 2: Summary Statistics of Weather Dataset

Statistic	**Temp**	**Humidity**	**Precipitation**	**Windspeed**
Count	2246.000000	2246.000000	2246.000000	2246.000000
Mean	66.940338	77.030855	0.288359	15.675334
Std	2.409488	7.559530	0.603790	6.130008
Min	58.200000	37.600000	0.004000	2.900000
25%	65.500000	72.600000	0.020000	12.100000
50%	66.900000	77.900000	0.079000	15.000000
75%	68.500000	82.400000	0.311000	18.200000
Max	76.200000	96.400000	8.661000	117.400000

4.4.2 Time Series Analysis

The second analysis involved time series analysis for the four variables of interest. As expressed in figure 1, there were inconsistent patterns in the values of temperature, humidity, precipitation, and wind speed. The level of irregularity, specifically in terms of precipitation, suggest the need of models that capture complex and non-linear patterns, justifying the use of XGBoost ML modeling in prediction.

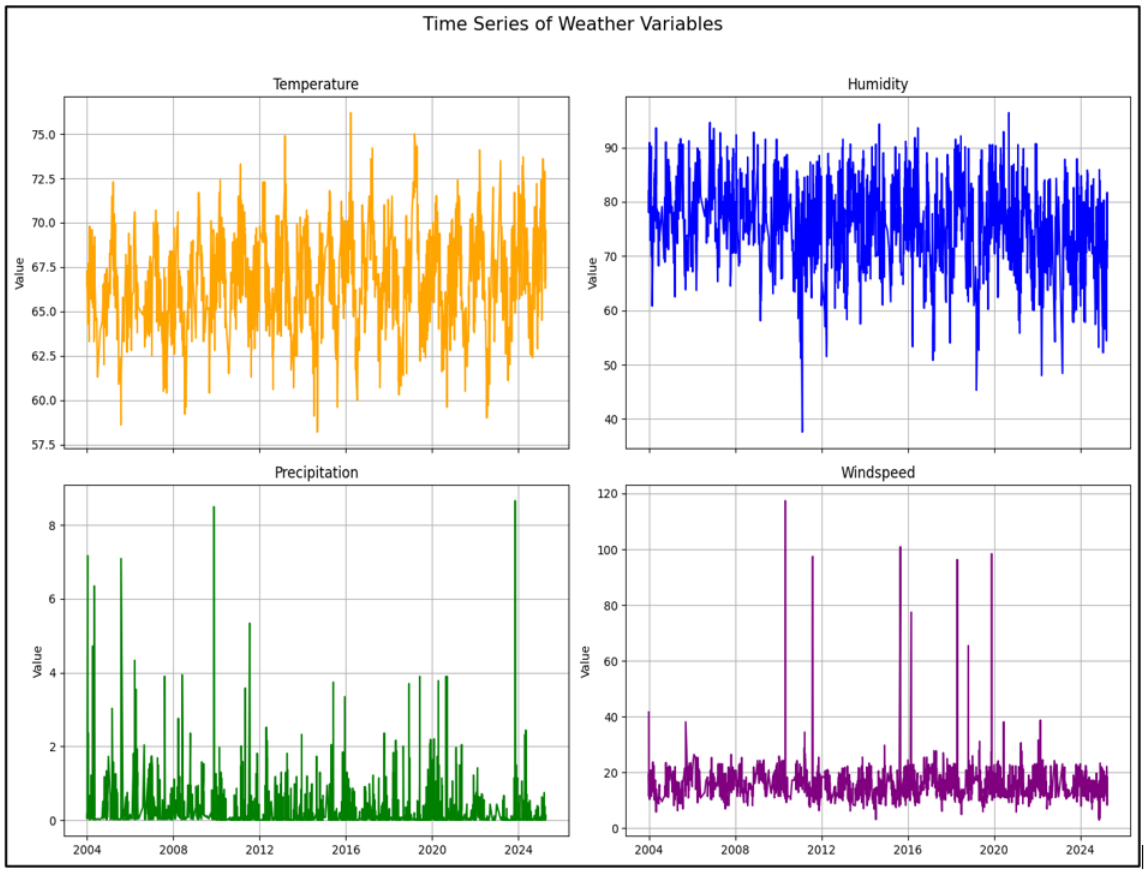

Figure 3: Time series analysis for Weather Variables

Precipitation was highly skewed to the right, showing that most values were low while temperature values were skewed on both tails. Humidity on the other hand, was skewed to the left showing that most values were high, while wind speed values were mainly low as observed from the upper tail skewness.

4.4.3 Data diagnostic and transformation

Having understood the data patterns, the next step involved checking for outliers, extreme values within the data. This was important in confirming whether standardization is needed to enhance precision. he existence of outliers called for Min-Max scaling which was conducted prior to model development. Other than outliers, the data was checked for missing values which were 0 for all variables.

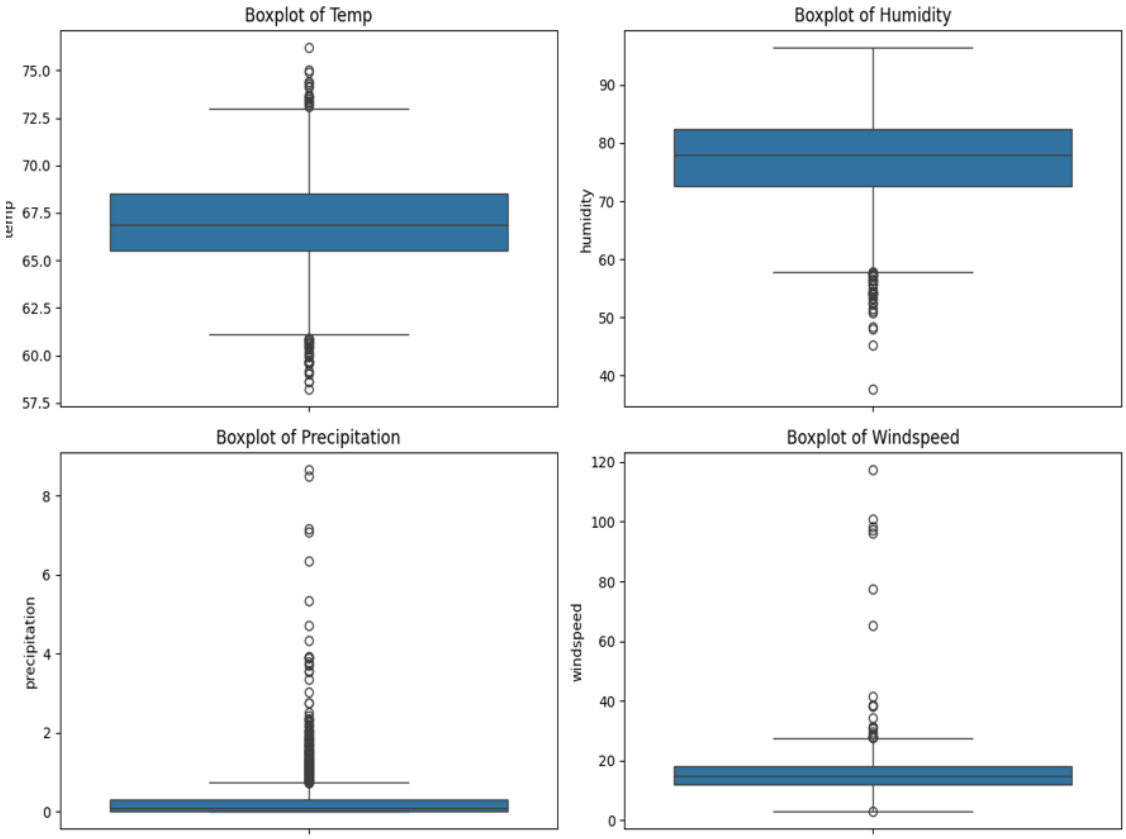

Figure 4: Outlier Detection

4.5 Model Development

The precipitation (rainfall) variable was transformed to predict next day rain. The first step involved preprocessing through parsing the date column into a standard date-time format, setting it as the index, converting column names to lowercase, and removing the white-spaces trailing. All recorded zeros in the precipitation days were retained to maintain the

balance between no-rain and rain categories. Thereafter, feature engineering was conducted to improve the predictive capability of the model. This involved creating lag variables to capture the persistence of rainfall over the previous week. Using the sine and cosine transformation of the month and day-of-year, seasonal characteristics were represented to model the bimodal patterns of rainfall. A further interaction feature integrating temperature and humidity was added; this was key in reflecting their joint impact on convective rainfall. Through shifting the precipitation column backwards by one day and assigning 1 to values of precipitation above 0, the target variable was generated. Thereafter, rows containing missing values that resulted from lag creation and shifting were deleted. The numeric variables were then normalized to a range of 0 and 1 using MinMaxScaler. The set was then split into training 80% and testing 20% subsets without shuffling; this helped preserve temporal ordering. XGBoost was then developed and configured with 300 estimators, a learning rate of 0.05, a maximum tree depth of 5. Both columns and rows sub-sampling rates were set at 0.8. Log loss and used as the assessment metric during training.

4.5.1 Data Preprocessing

This included data cleaning and data standardization to prepare it for analysis. This includes parsing the date column into a standard date-time format and set as the index to maintain the temporal sequence. Column names were converted to lowercase, and trailing white spaces were removed for consistency. All zero-precipitation records were retained to preserve the balance between rain and no-rain categories, preventing class imbalance that could bias the model.

4.5.2 Feature Engineering

Feature engineering was conducted to improve the predictive capability of the model. This involved creating lag variables to capture the persistence of rainfall over the previous week. Using the sine and cosine transformation of the month and day-of-year, seasonal

characteristics were represented to model the bimodal patterns of rainfall. A further interaction feature integrating temperature and humidity was added; this was key in reflecting their joint impact on convective rainfall. Through shifting the precipitation column backwards by one day and assigning 1 to values of precipitation above 0, the target variable was generated.

Feature Importance: The final model was further analyzed to display strong predictors of rainfall. As shown in figure 4, indicated that temperature-humidity interaction was the top precipitation predictor followed by humidity, and temperature, aligning with local weather cycles. The lag, cosine and sine variables created were poor predictors of precipitation, suggesting that previous rainfall patterns do not necessarily determine the future incidences.

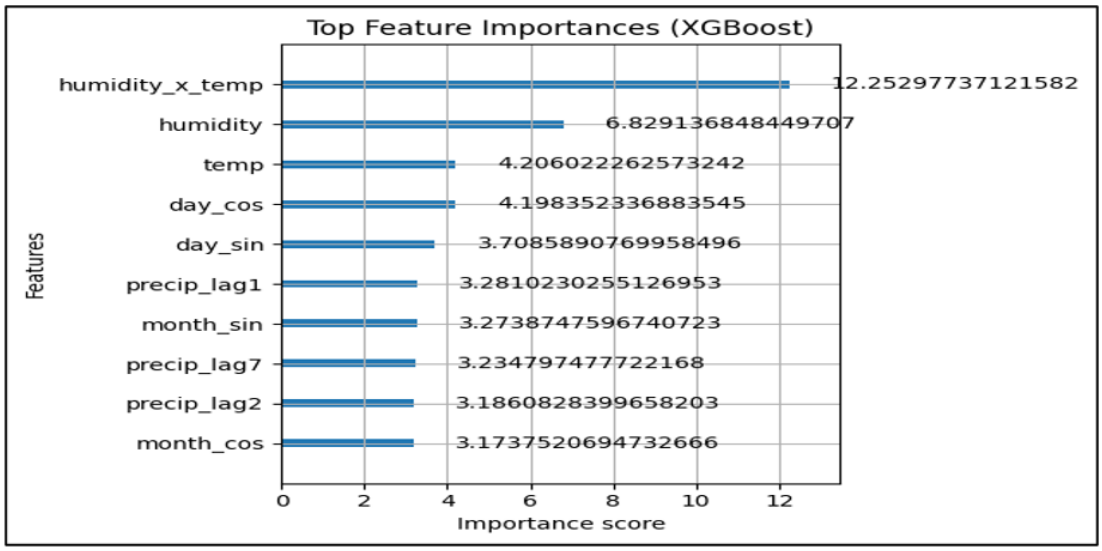

Figure 5: Feature Importance Results

4.5.3 Handling Missing Values and Scaling

Rows containing missing values, which resulted from lag creation and the one-day shift of the target variable, were removed to ensure data integrity. The numerical features were then normalized to a uniform range between 0 and 1 using the MinMaxScaler technique. This scaling step ensured that no variable dominated others due to differences in magnitude,

improving model convergence and stability during training.

4.5.4 Data Splitting

Data Splitting After cleaning dataset was divided into 80% train and 20% test without shuffling to preserve the chronological order of weather records, which is an important characteristic of time-series data. Train-test chronological split method enabled the model to learn from past weather patterns and be evaluated on future data, while ensuring real performance assessment.

4.5.5 Model Training

The model was configured with 300 estimators, a learning rate of 0.05, and a maximum tree depth of 5. Both column and row subsampling rates were set at 0.8 to reduce overfitting and improve generalization. The log loss function was employed as the evaluation metric during training since the task involved binary classification of rain and no-rain events

4.6 XGBoost Model Evaluation Performance

The initial regression approach using regression metrics aimed to predict daily precipitation amounts, however it reflected low results hence indicating low predictive capability using the same engineered features (Temperature, humidity, wind speed, lagged precipitation, temperature-humidity interaction, rolling means and seasonal encoding). The low-poor results reflected the regression approach irrelevant for the prediction.

```
Tuned XGBoost - R²: 0.06579746431625888
Tuned XGBoost - MSE: 1.124557246672206
Tuned XGBoost - RMSE: 1.06045143531998
```

Figure 6: XGB (Regression showing low prediction results)

Due to suboptimal results of the XGB via regression approach, the model was re-evaluated to XGBoost binary classification to forecast whether there will be occurrence of rain the next day or no-rain and the shift to XGBoost binary classification simplifies the forecasting while maintaining the practical applications such as agriculture and water management practices.

XGBoost Binary Classification Via binary classification approach, the XGBoost model was evaluated on the test dataset using accuracy, precision, recall, F-1 score, and ROC AUC as performance evaluators. The model before tuning attained an accuracy of 76.76%, correctly classifying no-rain and rain periods in over 75% of cases. The resulting precision was 70.14%, suggesting that model correctly predicted around seven out of 10 cases. The value of recall was 33.26%; this indicated that the model recognized only about one-third of actual rainfall events, meaning data imbalance was due to unequal or too many number of no-rain events in the dataset, which biases the model in predicting no-rain patterns frequently and thus expressing under-performance in detecting all positive cases. The high precision but low recall shows a conservative behavior fo the XGB model. The next metric, F-1 score, was 45.12%, showing a balance between precision and recall while the ROC AUC score of 0.75 suggesting a moderate capacity to differentiate between rain and no-rain classes. Whereas the model expressed strong precision and overall accuracy, the imperatively low recall shows the potential for missed rainfall forecasting. This constraint may be substantial in high-stakes usage such as flood early warning systems, where the cost of false negatives is high. As such, there was need for further hyperparameter tuning to enhance the model performance.

```python
from xgboost import XGBClassifier

clf = XGBClassifier(
    n_estimators=300,
    learning_rate=0.05,
    max_depth=5,
    subsample=0.8,
    colsample_bytree=0.8,
    use_label_encoder=False,
    eval_metric='logloss',
    random_state=42
)

clf.fit(X_train, y_train)
```

```
🔍 XGBoost Binary Classifier Performance:
Accuracy : 0.7675919948353777
Precision: 0.7014218009478673
Recall   : 0.3325842696629214
F1 Score : 0.45121951219512196
ROC AUC  : 0.7486362156000652
```

Figure 7: XGB Binary Classifier before tuning

The ROC curve representing the accuracy rate is displayed as below:

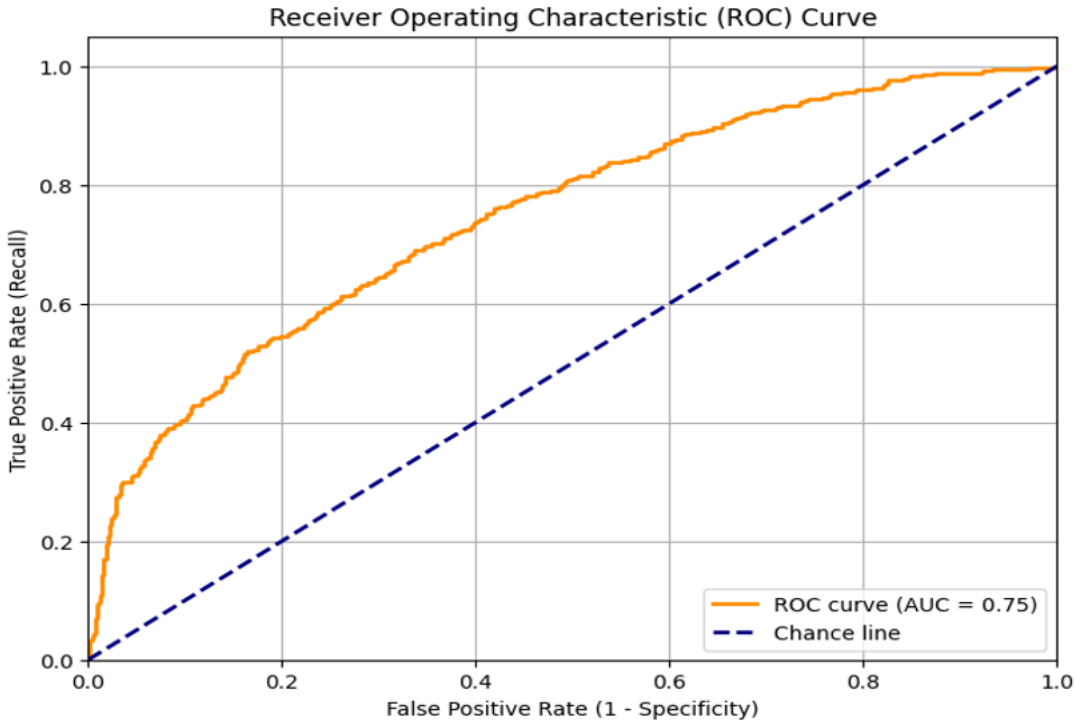

Figure 8: ROC Curve

The results show that the XGBoost model's ability to balance sensitivity correctly detecting rain events and specifically correctly identify no rain days. Achieving an overall 0.75 AUC indicates a good discriminatory performance.

4.7 Optimizing the XGBoost Model via Hyperparameter Tuning

Hyperparameter tuning was done via GridSearchCV and adjusting the parameters along with number of estimators (n=300), 0.05 learning rate, with maximum depth of 5, and a sub-sample of 0.8. After the model shifting to XGBoost binary classification, the classification was tuned by reducing the the default adjustment to 0.5 to increase the rain occurrence sensitivity and to balance the precision and recall considering the dataset imbalance. After tuning using via reduced threshold, the model attained an overall accuracy of 73%, correctly classifying nearly three-quarters of rain and no-rain cases, a value less than the initial model by 2%. The model recorded a precision and recall of 81% for the no-rain class (0), resulting in an F-1 score of 0.81 across 1,104 instances. The rain class precision for the rain class

(1) was 53% with improved recall for rain 54%, and F-1 score of 0.54 over 445 cases. The model further resulted in a 0.67 macro average for precision, recall, and F-1 score and the weighted averages of 0.73, reflecting class distribution. From the perspective of binary classification, the precision of the model precision (70.14%) and ROC AUC (0.75) suggested good discriminatory capacity. However, the critical lower recall for the rain class shows that the model missed a significant percentage of actual rainfall events, which may constrain its application effectiveness where capturing all possible rain incidences is pivotal.

	precision	recall	f1-score	support
0	0.81	0.81	0.81	1104
1	0.53	0.54	0.54	445
accuracy			0.73	1549
macro avg	0.67	0.67	0.67	1549
weighted avg	0.73	0.73	0.73	1549

Figure 9: Refined Model

Interpretation The tuned XGBoost model (threshold 0.3) achieved a balance trade-off between precision recall after tuning. Despite a slight drop in overall accuracy, the model became better in identifying rain events, indicating improved sensitivity and prediction fairness.

4.7.1 Benchmark Results

The tuned performance of the XGBoost Binary Classifier, ARIMA, and Ridge models was compared using key classification metrics, as visualized in Figure **??**. The XGBoost model achieved a recall of 0.5573, outperforming ARIMA and supporting enhanced rain detection for early warning systems.

The XGBoost model achieved a recall of 0.5573, outperforming ARIMA and supporting enhanced rain detection for early warning systems.

Table 3: Benchmark Model Results with Thresholds

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
XGBoost Binary	0.7308	0.5299	0.5573	0.5433	0.7549
ARIMA	0.6979	0.3175	0.0449	0.0787	0.4552
Ridge	0.7082	0.4941	0.6629	0.5662	0.7610

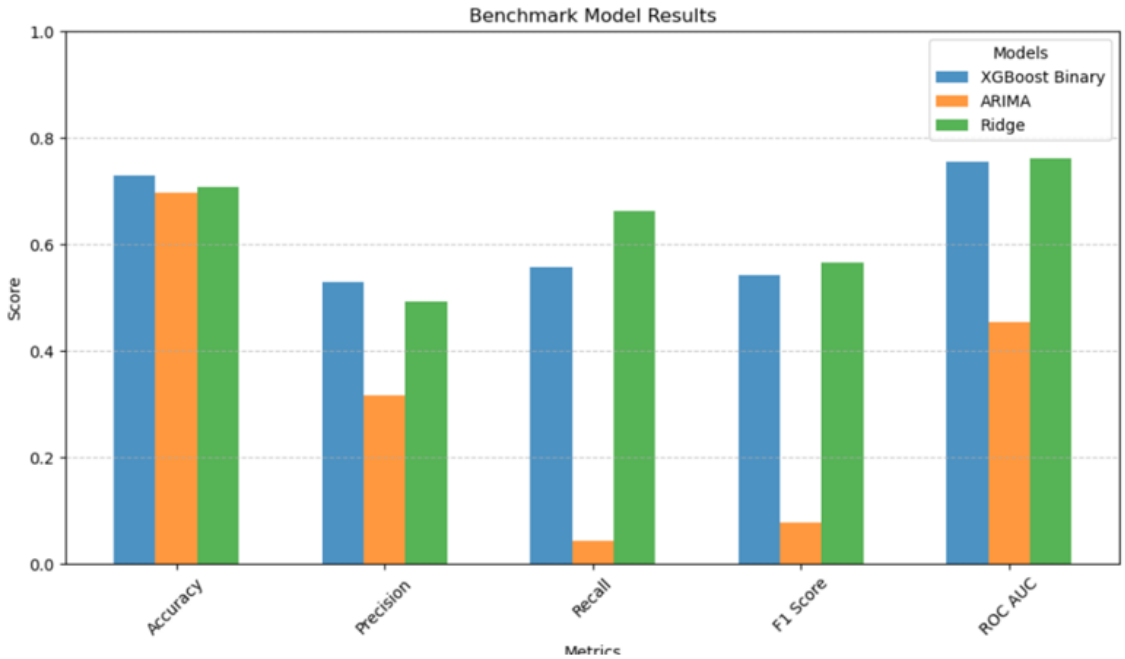

Figure 10: Histogram Benchmark Model Results

4.8 Assessment of the Implications of the enhanced forecasts in agriculture, water control management,and catastrophe preparedness in Kenya

With 0.73 accuracy and 0.54 recall for rain, the optimized XGBoost Binary Classification Model supports the practical applications across various sectors: In Agriculture, accurate predictions of rain or no-rain events, for instance,81% no rain precision aid farmers in planning for planting schedules during Kenya's bimodal patterns for March-May and October-December rain. For water resource management, the predictions inform reservoir efficiency hence reducing flood likelihood due to reliable no-rain forecasting. The recall

of 0.54 show rather successful rain event detection thus providing early warning for disaster preparedness, although there is a need for further improvement of extreme events improvement.

5 CHAPTER FIVE: DISCUSSION, CONCLUSION, AND RECOMMENDATION

5.1 Introduction

This chapter presents the interpretation and discussion of the study findings in relation to the stated objectives and the broader body of knowledge. The analysis moves beyond the numerical outputs provided in the results section to explain the meaning, implications, and potential applications of the findings. It combines the results with insights from existing literature, highlighting areas of agreement and divergence, and considers the underlying factors that may have influenced the outcomes. Furthermore, the discussion explores the extent to which these findings can be generalized to similar contexts, while acknowledging the scope and limitations of the study. This structured approach ensures that the results are critically examined and positioned within the wider academic and practical discourse.

5.2 Discussion

The initial XGBoost model using regression method focused on predicting continual precipitation amounts which reflected poor performance with low R^2 (0.066) and a high RMSE (1.06). The shift to XGBoost Binary Classification modeling to predict the occurrence of rain or no rain next-day precipitation prediction exhibited more effectively; from the results, it was found that the model learned to distinguish dry and wet conditions consistently, showing reliable performance when identifying non-rain days while taking a more conservative stance when predicting rain events. Qualitatively, this pattern indicates that the model successfully captured dominant signal structures in the input variables particularly recent precipitation patterns, seasonal cycles, and humidity–temperature interactions yet found finer distinctions that separate marginal wet cases from dry cases more challenging to resolve. The model's behaviour therefore reflects a sound capacity to map meteorological predictors

to binary rainfall outcomes while exhibiting a trade-off in sensitivity that is common when minority events are less frequent or less distinct in predictor space. Interpreting these outcomes to the study objectives, the first objective (evaluating model performance using key meteorological variables) is supported by evidence that engineered lag features and seasonally smoothed temporal descriptors contributed materially to the classifier's decision rules. The prominence of lagged precipitation and humidity–temperature interactions among the model's influential features aligns with meteorological theory concerning convective processes and persistence effects, suggesting the model's learned patterns are physically plausible. Relative stability of performance across held-out data indicates that the modeling pipeline feature engineering, temporal split, and algorithmic choice provided a robust framework for translating observed weather dynamics into actionable classifications, meeting the study's aim of producing operationally relevant short-term forecasts.

Comparison with existing literature shows substantial convergence in both method and outcome. The adoption of gradient boosting for precipitation or event classification is well represented in regional studies where ensemble tree methods commonly outperform simpler parametric models on structured meteorological data (Papacharalampous et al., 2023). Several regional analyses have reported behaviour comparable to the present findings: ensemble methods reliably identify the prevailing (more frequent) conditions while showing reduced sensitivity to less frequent extremes, unless additional predictors or resampling strategies are introduced. Work that focused on East and West African contexts likewise emphasizes the utility of lagged rainfall and humidity variables as core predictors (Noorbakhsh, 2022) which corroborates the present model's feature importance profile.

There are also studies that both support and nuance these parallels. Research that incorporates satellite-derived cloud and moisture indices, atmospheric pressure fields, or high-resolution reanalysis data tends to report improved detection of rare or extreme events relative to models restricted to surface station inputs. In those comparisons, gradient boosting remains competitive but the inclusion of richer spatial or vertical atmospheric information

frequently narrows or reverses the sensitivity gap for positive events. Conversely, where datasets are purely tabular and station-based, gradient boosting often matches or exceeds the performance of more complex neural architectures, owing to its efficiency with engineered features and its implicit handling of nonlinear interactions. These contrasting findings indicate that algorithmic choice and data richness jointly determine whether boosting methods will be optimal or whether hybrid/deep approaches are warranted. Similar findings in the literature reinforce the argument that careful feature engineering particularly the use of temporal lags and seasonally aware encoding can deliver strong predictive power for short-horizon precipitation classification. Multiple studies have reported that models which definately encode seasonality and recent persistence produce more interpretable and transferable decision rules than those that rely solely on raw temporal indices. At the same time, contrasting literature suggests that where the objective is to prioritize detection of infrequent events, additional input modalities (satellite, radar, pressure tendencies) or targeted modeling strategies (resampling, cost-sensitive learning, ensemble stacking) can materially change the balance between sensitivity and specificity. To generalization, the study's qualitative alignment with regional and cross-domain findings implies that the modeling framework is well suited for replication across climatologically similar areas, particularly where the same core predictor set is available and station density is adequate. The underlying meteorological drivers captured by lagged precipitation and humidity–temperature interactions are broadly relevant across many parts of East Africa, which supports potential portability. That said, practical generalization requires local calibration: transfer to distinct micro-climates or regions with different rainfall generation mechanisms should be accompanied by site-specific model tuning and validation against local observations. Where richer data sources exist, incorporation of additional atmospheric predictors will likely improve sensitivity to positive events and extend the model's operational utility.

The qualitative results substantiate that an XGBoost-based approach with thoughtfully

engineered temporal and interaction features provides a competent and interpretable solution for next-day precipitation classification. The findings are consistent with a body of literature that recognizes gradient boosting as a powerful method for structured meteorological data, while also echoing the wider evidence that data richness and task framing determine whether boosting alone suffices or must be augmented with complementary data and methods. The study thus contributes to a growing consensus that practical precipitation forecasting benefits from the combined application of sound feature design and robust ensemble learning, with local validation guiding broader generalization.

Limitation While the study achieved its objectives and provided meaningful insights, certain limitations must be acknowledged. First, the analysis relied on a single dataset, which, although comprehensive, may not fully capture the variability present in broader contexts. As such, the generalizability of the findings to other populations or settings may be limited. Additionally, the study's design was observational, meaning that causal relationships cannot be conclusively established; the results should be interpreted as indicative rather than definitive. Another limitation lies in the potential influence of unmeasured variables. Although the model incorporated relevant predictors, factors not included in the dataset could have affected the results, leading to residual confounding. Moreover, the binary classification framework employed may have oversimplified complex relationships within the data, potentially affecting the precision of the findings. Finally, while the methods and algorithms applied are robust, they are also sensitive to hyperparameter tuning and data preprocessing choices. Different parameter configurations or preprocessing strategies might yield slightly different results. Recognizing these constraints, future research should employ larger, more diverse datasets, consider alternative modelling approaches, and incorporate additional variables to enhance the robustness and applicability of the findings.

5.3 Conclusions and Recommendations

Conclusion

The effectiveness of the machine learning model was evaluated, analyzing and forecasting patterns within the weather dataset with a goal to support decision making and to improve operational efficiency. The model successfully captured interactions within the key variables, aligned with the research objectives. The results showed the capability of using advanced analytics to allow data-driven and practical insights that can inform measures. The structure and the quality of the input variables showed an important role in determining the model's performance, thus emphasizing the need for thorough preparation of data in predictive analytics.

Recommendation

Future research should explore hybrid approaches for modeling that integrate the strengths of multiple algorithms to improve prediction accuracy and robustness. The Kenya Meteorological Department to integrate the machine learning models as complementary tools with existing systems for accuracy and reliability enhancement. Integrate domain-specific knowledge into the modelling process to ensure findings remain in practical use and are interpretable for operational purposes. To adapt to changing conditions, ensure regular model training and monitoring of the models' performance and predictive accuracy maintenance periodically.

REFERENCES

References

Abisha, R., Krishnani, K. K., Sukhdhane, K., Verma, A., Brahmane, M., and Chadha, N. (2022). Sustainable development of climate-resilient aquaculture and culture-based fisheries through adaptation of abiotic stresses: a review. *Journal of Water and Climate Change*, 13(7):2671–2689.

Affoh, R., Zheng, H., Dangui, K., and Dissani, B. M. (2022). The impact of climate variability and change on food security in sub-saharan africa: Perspective from panel data analysis. *Sustainability*, 14(2):759.

Anwar, M., Winarno, E., Hadikurniawati, W., and Novita, M. (2021). Rainfall prediction using extreme gradient boosting. In *Journal of Physics: Conference Series*, volume 1869, page 012078. IOP Publishing.

Awad, M. and Fraihat, S. (2023). Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. *Journal of Sensor and Actuator Networks*, 12(5):67.

Babu Nuthalapati, S., Nuthalapati, A., et al. (2024). Accurate weather forecasting with gradient boosting using machine learning. *Int. J. Sci. Res. Arch*, 12(2):408–422.

Deo, A., Karmakar, S., and Arora, A. (2022). Rainwater harvesting and water balance simulation-optimization scheme to plan sustainable second crop in small rain-fed systems. *Journal of Environmental Management*, 323:116135.

Gleick, P. H. and Cooley, H. (2021). Freshwater scarcity. *Annual Review of Environment and Resources*, 46(1):319–348.

Habib-ur Rahman, M., Ahmad, A., Raza, A., Hasnain, M. U., Alharby, H. F., Alzahrani, Y. M., Bamagoos, A. A., Hakeem, K. R., Ahmad, S., Nasim, W., et al. (2022). Impact of climate change on agricultural production; issues, challenges, and opportunities in asia. *Frontiers in Plant Science*, 13:925548.

IPCC (2022). *Climate change 2022: Impacts, adaptation, and vulnerability. Working Group II contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Khodakhah, H., Aghelpour, P., and Hamedi, Z. (2022). Comparing linear and non-linear data-driven approaches in monthly river flow prediction, based on the models sarima, lssvm, anfis, and gmdh. *Environmental Science and Pollution Research*, 29(15):21935–21954.

Kilonzo, S. M. (2022). Women, indigenous knowledge systems, and climate change in kenya. In *African Perspectives on Religion and Climate Change*, pages 79–90. Routledge.

KMD (2023). State of the climate in kenya 2023. Technical report, Kenya Meteorological Department.

Kogo, B. K., Kumar, L., and Koech, R. (2021). Climate change and variability in kenya: a review of impacts on agriculture and food security. *Environment, development and sustainability*, 23(1):23–43.

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., and Matsopoulos, G. K. (2023). A review of arima vs. machine learning approaches for time series forecasting in data-driven networks. *Future Internet*, 15(8):255.

Mishra, P., Al-Khatib, A. M. G., Yadav, S., Ray, S., Lama, A., Kumari, B., Sharma, D., and Yadav, R. (2024). Modeling and forecasting rainfall patterns in india: a time series analysis with xgboost algorithm. *Environmental Earth Sciences*, 83(6):163.

Noorbakhsh, M. (2022). *Improving drought predictability in Africa by data-driven models*. PhD thesis, University of Warwick.

Oino, P. G. and Musau, E. (2024). Community engagement in climate change and adaptation in kenya: A socio-anthropological and linguistic perspective. *African Journal of Climate Change and Resource Sustainability*, 3(1):387–404.

Okedele, P. O., Aziza, O. R., Oduro, P., and Ishola, A. O. (2024). Integrating indigenous knowledge systems into global climate adaptation policies. *Int J Eng Res Dev*, 20(12):223–31.

Owino, D. O. (2022). Responding to impacts of climate change: A case study of kenya. Master's thesis, Oslo Metropolitan University.

Papacharalampous, G., Tyralis, H., Doulamis, A., and Doulamis, N. (2023). Comparison of tree-based ensemble algorithms for merging satellite and earth-observed precipitation data at the daily time scale. *Hydrology*, 10(2):50.

Parmesan, C., Morecroft, M. D., and Trisurat, Y. (2022). *Climate change 2022: Impacts, adaptation, and vulnerability*. PhD thesis, GIEC.

Pello, K., Okinda, C., Liu, A., and Njagi, T. (2021). Factors affecting adaptation to climate change through agroforestry in kenya. *Land*, 10(4):371.

Rojas-Campos, A., Langguth, M., Wittenbrink, M., and Pipa, G. (2023). Deep learning models for generation of precipitation maps based on numerical weather prediction. *Geoscientific Model Development*, 16(5):1467–1480.

Sagi, O. and Rokach, L. (2021). Approximating xgboost with an interpretable decision tree. *Information sciences*, 572:522–542.

SAMWEL, M. P. (2021). *CLIMATE VARIABILITY AND FOOD SECURITY IN KISII COUNTY; KENYA*. PhD thesis.

Sarma, H. H., Paul, A., Kakoti, M., Talukdar, N., and Hazarika, P. (2024). Climate resilient agricultural strategies for enhanced sustainability and food security: A review. *Plant Archives*, 24(1):787–792.

Appendix A: Research Permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR
SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **129892**

Date of Issue: **28/May/2025**

RESEARCH LICENSE

This is to Certify that Ms.. Damaris Mulinge of The Cooperative University of Kenya, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: DEVELOPMENT OF A MACHINE LEARNING MODEL FOR PRECIPITATION FORECASTING IN KENYA for the period ending : 28/May/2026.

License No: **NACOSTI/P/25/4174459**

129892

Applicant Identification Number

Deputy Director
NATIONAL COMMISSION FOR
SCIENCE,TECHNOLOGY &
INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document,
Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix B: Publication

Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(RESEARCH ARTICLE)

Check for updates

Development of a machine learning model for precipitation forecasting in Kenya

Damaris Muthoki Mulinge [1,*], John M. Kihoro [2], Shadrack S. Madila [3] and Shem Mbandu Angolo [4]

[1] *Department of Information and Communication Technology, The Cooperative University of Kenya, Karen, Nairobi, Kenya.*
[2] *Department of Mathematical Sciences, The Cooperative University of Kenya, Karen, Nairobi, Kenya.*
[3] *Department of Information and Communication Technology, Moshi Cooperative University, Moshi, Tanzania.*
[4] *Department of Information and Communication Technology, The Cooperative University of Kenya, Karen, Nairobi, Kenya.*

Global Journal of Engineering and Technology Advances, 2025, 24(03), 043-050

Publication history: Received on 26 July 2025; revised on 29 August; accepted on 01 September 2025

Article DOI: https://doi.org/10.30574/gjeta.2025.24.3.0261

Abstract

Accurate precipitation forecasting is important for mitigating the impacts of climate variability in Kenya, where erratic rainfall events considerably affect agriculture, water control and disaster preparedness. Traditional methods such as ARIMA (Autoregressive Integrated Moving Average) and NWP (Numerical Weather Prediction) have shown to struggle with complex weather patterns due to linearity assumptions, high computational demands and limited spatial resolution. This research develops and evaluates an XGBoost-based machine learning model to enhance precipitation predictions both long-term and short-term. Utilizing a a 20-year weather dataset (2004 - 2024) with 7300 daily data records sourced from online Visual Crossing Weather Data, key features include temperature, humidity, wind speed, lagged precipitation (1-7), rolling means and seasonal encoding to capture bimodal rainfall patterns of the months of march-May, and October-December. Data processing involved min-max normalization of 0-1 range, feature selection, sin/cosine transformations for seasonal patterns and temperature-humidity interactions for connective modelling processes. The dataset used was split with 80% for training and 20% for testing and a temporal split ≤ 2020 for training and > 2020 for testing maintaining the chronological data order. The initial attempts exhibited poor performance with low R^2 = 0.066 and a high RMSE=1.06 hence leading to XGBoost binary classification shift to predict the likelihood of rain/no-rain tomorrow. Bayesian optimization and GridSearchCV hyperparameter tuning was applied with default 0.5 threshold adjustment for improved rain class sensitivity using classification metrics and resulted 76.76% accuracy, 70.14% precision, 33.36% recall, 45.12% F1-Score and ROC-AUC 0.75. Post-tuning accuracy by reducing the threshold to 0.3 to capture missed rainfall events: 73% accuracy, no-rain precision and recall 81%, 53% rain precision, 54% recall, F1-Score 54%. Temperature-humidity interaction as the top predictor in feature importance. The results contribute to improved precipitation prediction accuracy hence supporting decision making in agriculture, water resource management and early disaster preparedness in Kenya's climate vulnerable regions.

Keywords: Machine Learning; Climate variability; Precipitation forecasting; XGBoost; Binary Classification

1. Introduction

Accurate precipitation forecasting is critical for mitigating the impacts of climate change, especially in Kenya, which is vulnerable to extreme weather events. Many areas face challenges such as food insecurity and water scarcity due to unpredictable rainfall patterns. Precipitation variability poses a significant effect on the local communities dependent on natural resources, agricultural practices, and the region's socio-economic stability (IPCC, 2022) (KMD, 2023). Kenya's Climatic and weather conditions extremely contribute to food insecurity and water scarcity in about 75% of the country, and with erratic rainfall leading to droughts and floods that disrupt livelihoods (Affoh et al., 2022). Traditional approaches to weather forecasting though valuable, struggle with non-linear dynamics due to historical reliance on statistical correlations and have shown to struggle with non-linear dynamics, recent approaches such as Genetic

* Corresponding author: Damaris Muthoki Mulinge

Appendix C: AI Report

Damaris Mulinge

PROJECT_FOR_MACHINE_LEARNING_MODEL_FOR_PRECIP...

Final Thesis/Project Submission

MSC_March_2025_class

The Cooperative University of Kenya

Document Details

Submission ID
trn:oid:::1:3350441413

Submission Date
Sep 24, 2025, 10:11 PM GMT+3

Download Date
Sep 25, 2025, 9:41 AM GMT+3

File Name
PROJECT_FOR_MACHINE_LEARNING_MODEL_FOR_PRECIPITATION_FORECASTING_IN_KENYA.pdf

File Size
860.1 KB

51 Pages

11,362 Words

68,157 Characters

turnitin Page 1 of 53 - Cover Page Submission ID trn:oid:::1:3350441413

AI Report

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Appendix D: Similarity

Damaris Mulinge

PROJECT_FOR_MACHINE_LEARNING_MODEL_FOR_PRECIP...

Final Thesis/Project Submission

MSC_March_2025_class

The Cooperative University of Kenya

Document Details

Submission ID
trn:oid:::1:3350441413

Submission Date
Sep 24, 2025, 10:11 PM GMT+3

Download Date
Sep 25, 2025, 9:41 AM GMT+3

File Name
PROJECT_FOR_MACHINE_LEARNING_MODEL_FOR_PRECIPITATION_FORECASTING_IN_KENYA.pdf

File Size
860.1 KB

51 Pages

11,362 Words

68,157 Characters

turnitin Page 1 of 59 - Cover Page Submission ID trn:oid:::1:3350441413

Similarity

11% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

104 Not Cited or Quoted 10%
Matches with neither in-text citation nor quotation marks

11 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

9% Internet sources
8% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.