Strathmore UNIVERSITY

SU+ @ Strathmore University Library

Electronic Theses and Dissertations

2025

An Ensemble machine learning model for drought prediction in Kilifi County, Kenya.

Bosire, Violet Kwamboka
School of Computing and Engineering Sciences
Strathmore University

Recommended Citation

Bosire, V. K. (2025). *An Ensemble machine learning model for drought prediction in Kilifi County, Kenya* [Strathmore University]. https://hdl.handle.net/11071/16372

Follow this and additional works at: https://hdl.handle.net/11071/16372

This work is availed for free and open access by Strathmore University Library.
It has been accepted for digital distribution by an authorized administrator of SU+ @Strathmore University.
For more information, please contact library@strathmore.edu

An Ensemble Machine Learning Model for Drought Prediction in Kilifi County, Kenya

By

Violet Kwamboka Bosire

145573

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computing and Information Systems at Strathmore University

School of Computing and Engineering Sciences

Strathmore University

Nairobi, Kenya

March,2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this university or any other university. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: Violet Kwamboka Bosire

Signature: ____ V.K.B __ _____Date: __3/27/2025_____

Approval

The dissertation proposal of Violet Kwamboka Bosire has been reviewed and approved for examination by the following:

Signature: _____B.K.S_____ Date: 28/03/2025

Prof. Bernard Shibwabo

Associate Professor,

School of Computing & Engineering Sciences,

Strathmore University

Abstract

Drought prediction is critical for mitigating its adverse effects on agriculture, water resources, and livelihoods. This study developed a stacking ensemble learning model for multi-class drought prediction using machine learning techniques, including Random Forest, Support Vector Machine (SVM), XGBoost, and LightGBM. The CRISP-DM methodology guided the research, ensuring a systematic approach to data collection, pre-processing, model training, evaluation, and interpretation. Each model was assessed based on accuracy, recall, precision, and F1-score to determine its effectiveness in classifying drought severity levels. Among the base models, Random Forest achieved the highest accuracy (62.79%) and recall (62.79%), demonstrating its ability to capture complex patterns in the data while maintaining a reasonable balance between precision and recall. In contrast, SVM performed the weakest, with an accuracy of 25.58% and an F1-score of 29.91%, indicating its limited effectiveness in handling the dataset's complexity and class imbalance. The XGBoost and LightGBM models showed moderate performance, with both achieving an accuracy of 55.81% and recall of 55.81%, but their high precision (90.93% for XGBoost and 83.97% for LightGBM) suggests they are effective in minimizing false positives, though they struggle to balance precision and recall effectively. The stacking ensemble model, which integrated these classifiers with Logistic Regression as the meta-model, yielded an accuracy of 65.12% and an F1-score of 62.68%, outperforming all individual base models. To further enhance model performance, the study recommends incorporating additional environmental variables, such as land surface temperature (LST), Standardized Precipitation Evapotranspiration Index (SPEI) and the Palmer Drought Severity Index (PDSI) to significantly improve drought prediction accuracy. Future research should explore deep learning approaches, hybrid models combining physical hydrological and machine learning methods, and explainability techniques like SHAP values or LIME to strengthen trust in AI-driven drought forecasting systems.

Keywords: Drought Severity, Drought Prediction, Stacking, Ensemble Learning, Model Evaluation

Table of Contents

Declaration and Approval ii
Abstract iii
Table of Contents iv
List of Figures vii
List of Tables viii
Abbreviations/ Acronyms ix
Chapter 1: Introduction 1
1.1 Background of the Study 1
1.2 Problem Statement 3
1.3 Objective of the Study 4
1.3.1 General Objective 4
1.3.2 Specific Objectives 4
1.4 Research Questions 4
1.5 Justification of the study 4
1.6 Scope and Limitation of the Study 5
Chapter 2: Literature Review 6
2.1 Introduction 6
2.2 Theoretical Framework 6
2.2.1 Types of Drought and Its Impact 6
2.2.2 Drought Indicators 8
2.2.3 Drought Modeling Components 9
2.2.4 Ensemble Machine Learning Methods 10
2.3 Empirical Review 11
2.4 Research Gap 16
2.5 Conceptual Framework 17
2.6 Summary of Literature 18
Chapter 3: Research Methodology 23
3.1 Introduction 23
3.2 Business Understanding 23
3.3 Data Understanding 24
3.4 Data preparation 25
3.5 Modelling 27

3.6 Evaluation 27
3.7 Deployment 28
3.8 Dissemination of Findings 28
3.9 Tools and Software 28
3.10 Ethical Consideration 28
Chapter 4: System Analysis and Design 29
4.1 Introduction 29
4.2 System Requirement Analysis 29
4.2.1 Functional Requirements 29
4.2.2 Non-Functional Requirements 30
4.3 Systems Architecture 30
4.4 System Design 31
4.4.1 Use Case Diagram 31
4.4.2 Sequence Diagram 32
Chapter 5: Systems Implementation and Testing 34
5.1 Introduction 34
5.2 Data Pre-processing 34
5.2.1 Data Cleaning 34
5.2.2 Feature Engineering 35
5.2.3 Data Splitting 37
5.2.3 Feature Extraction 37
5.2.4 Balancing Dataset 41
5.3 Model Development 42
5.3.1 Selection of Base Models 42
5.3.2 Stacking Ensemble Model Development 43
5.4 Implementation 45
5.5 Model Evaluation 46
Chapter 6: Discussion 48
6.1 Research Objectives Review 48
6.1.1 Explore the key predictors that significantly influence drought occurrence 48
6.1.2 Develop ensemble machine learning model 50
6.1.3 Evaluate the Performance of the Developed Model 51
6.2 Contribution of the Study 54

6.3 Limitation of the Study 54
Chapter 7: Conclusions and Recommendations 56
7.1 Conclusions 56
7.2 Recommendations 56
7.3 Future Research Direction 58
References 59
Appendices 71

List of Figures

Figure 2.1: Cascading relationship between different types of drought, illustrating the interconnections between them 8
Figure 3.1: CRISP-DM Framework 23
Figure 3.2: Ensemble Learning Architecture for Stacking 27
Figure 4.1: System Architecture 31
Figure 4.2: Use case diagram 32
Figure 4.3: Sequence Diagram 33
Figure 5.1: Distribution of Precipitation 35
Figure 5.2: Distribution of Soil Moisture 35
Figure 5.3:Feature Engineering 36
Figure 5.4: Train and Test Data Split 37
Figure 5.5: Feature Selection using Correlation Matrix 39
Figure 5.6:Correlation Matrix for Selected Features 41
Figure 5.7: Balancing Dataset Using SMOTE 42
Figure 5.8: Base Models 43
Figure 5.9: Stacking Model Development 44
Figure 5.10: Accuracy of Base Models 47

VT OMNES VNVM SINT

List of Tables

Table 2.1: Summary of Literature and Gaps 19
Table 3.1: Description of Variables 24
Table 3.2: SPI Categories 26

Abbreviations/ Acronyms

ANN	Artificial Neuro Networks
ASAL	Arid and Semi-Arid Land
EVI	Enhanced Vegetation Index
EWS	Earlier Warning Systems
LGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
MODIS	Moderate Resolution Imaging Spectroradiometer
NDMA	National Drought Management Authority
NDVI	Normalized Difference Vegetation Index
NSE	Nash-Sutcliffe Efficiency
PCA	Principle Component Analysis
NSE	Nash-Sutcliffe Efficiency
R^2	Coefficient of Determination
RMSE	Root-Mean-Square Error
SPI	Standardized Precipitation Index
SVM	Support Vector Machine
SVR	Support Vector Regression
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

Chapter 1: Introduction

1.1 Background of the Study

Drought is a prolonged dry period in which insufficient rainfall causes water scarcity. This slowly developing disaster can significantly impact health, agriculture, the economy, energy resources, and the environment (Otieno, 2020). Drought affects approximately 55 million people worldwide yearly, significantly threatening livestock and crops. This puts livelihoods at risk, increases the risk of illness and death, and leads to large-scale migration. Water scarcity affects 40% of the world's population, and up to 700 million people could be displaced by drought by 2030 (World Health Organization [WHO], n.d). Moreover, climate change exacerbates these conditions by increasing temperatures, making dry areas drier and wetter regions. In dry areas, rising temperatures increase water evaporation, increasing drought risk and prolonging it. Over the past decade, 80-90% of all recorded natural disasters were due to floods, droughts, tropical cyclones, heat waves, and severe storms (Mirzabaev et al., 2023).

Since October 2020, Eastern Africa has experienced significant drought, marked by extended dry periods and intermittent heavy rainfall resulting in flash floods. This disaster has led to extensive crop failures, deteriorating pasture conditions, livestock fatalities, and diminished water supply, culminating in a humanitarian emergency affecting 4.35 million individuals (World Weather Attribution, 2023). The drought has resulted in substantial refugee influxes, with 180,000 individuals from Somalia and South Sudan seeking asylum in Kenya and Ethiopia. As of January 2023, assistance comprised 9,210 metric tons of food and USD 7.29 million in monetary transfers. In Kenya, 5.4 million individuals experienced acute food insecurity from March to June 2023, representing a 43% rise compared to the prior year.

Furthermore, about 970,000 children and 142,000 pregnant and lactating women had acute malnutrition (International Rescue Committee, 2023). Moreover, the drought resulted in the demise of over 2.4 million animals, exacerbated gender-based violence, and obstructed children's access to education. Therefore, drought is a severe challenge that needs to be addressed.

In research by Hayes et al. (2021) on drought monitoring, it was observed that meteorologists relied on terrestrial measurements, such as weather stations, rain gauges, and manual assessments of soil moisture and vegetation health. These techniques are vital in delivering information on rainfall, temperature, and the status of the soil, which is essential for drought assessment. However, they have various significant limitations regarding their spatial coverage

and time resolution. The ground-based systems mostly fail to acquire data effectively over vast areas with difficult-to-reach locations, leading to potential gaps in monitoring (Alahacoon & Edirisinghe, 2022). Additionally, these techniques sometimes involve labor and many resources in terms of maintenance and operations, which may not be feasible in places with weak infrastructures.

Machine learning (ML) algorithms have significantly advanced the field of drought prediction (Piri et al., 2023; Danandeh Mehr et al., 2023; Felsche and Ludwig, 2021.Various machine learning (ML) algorithms have been developed and applied for drought prediction, this algorithms include Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Extreme Learning Machines (ELM), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), M5 Trees, Multivariate Adaptive Regression Splines (MARS), and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) (Citakoglu & Coşkun, 2022; Docheshmeh Gorgij et al., 2022; Kikon & Deka, 2022; Kisi et al., 2019).

Otieno (2020) conducted a study on drought severity prediction and impact monitoring using remote sensing and socio-economic data. The research focused on four counties in Kenya—Marsabit, Mandera, Turkana, and Wajir. These areas are located in the northern region and classified as part of the country's arid and semi-arid lands (ASALs). Notably, no prior studies have been conducted to predict drought severity in Kilifi County, Kenya. This research gap is particularly concerning as climate change continues to amplify existing vulnerabilities and disparities among different population groups. These communities experience increased risks and restricted access to essential resources, making them more susceptible to climate-related hazards (International Union for Conservation of Nature, 2024). The recurrence of droughts and floods, along with their associated consequences—such as ocean pollution, degraded agricultural land, and loss of marine ecosystems—has contributed to food insecurity, displacement, fatalities, and economic setbacks within the county. The ongoing adverse effects of drought, particularly in agriculture, highlight the urgent need for a robust prediction and monitoring system.

Forecasting drought in Kilifi County is essential, as the region heavily depends on agriculture and livestock, both of which are highly vulnerable to changing weather patterns. Accurate drought forecasting provides early warnings that enable local governments, farmers, and communities to plan and implement water-saving strategies, adjust planting schedules, and secure food supplies, thus mitigating the impacts of drought on food security. Early predictions

can aid humanitarian efforts, as local and international organizations could mobilize resources to support affected populations, minimizing malnutrition, water scarcity, and economic losses. Therefore, this research can enhance the accuracy of machine learning models, as each model refines the parameters and variables most influential in predicting drought, offering insights that can inform policy, preparedness, and adaptive strategies in other vulnerable areas.

1.2 Problem Statement

In June 2024, Kilifi county underwent an extended period of drought accompanied by high winds (National Drought Management Authority [NDMA], 2024). Drought in Kilifi County, Kenya, has led to severe water scarcity, widespread food insecurity, and substantial livestock losses, particularly impacting vulnerable populations who rely heavily on agriculture and livestock for sustenance. The prolonged dry spells have reduced crop yields, leaving residents with limited food resources and forcing many to resort to foraging for wild roots, which provide minimal nutrition and can cause digestive issues (Kithi, 2024). This food shortage has led to widespread hunger and malnutrition, with children especially at risk due to inadequate dietary options.

Despite these challenges, little study has been conducted in this area. Numerous studies have been conducted in developed countries, and those done in Kenya focus on northern Kenya. Surprisingly, no study has been undertaken to predict drought severity in Kilifi County. Otieno (2020) conducted a study to predict drought severity and monitoring effects based on integrated remote sensing and socio-economic data. This study was conducted in 4 counties in the Northern part of Kenya. The study applied ANN and SVR techniques, which had 69% and 71% accuracy, respectively. Barrett et al. (2020) researched the forecasting of vegetation conditions toward improving drought early warning systems in pastoralist communities of Kenya.

The current study will address this gap by developing a stacking ensemble learning leveraging the strength of 4 machine learning models. This will combine the prediction accuracy of models like extreme gradient boosting (XGBoost), SVM, Random Trees, and Light Gradient Boosting Machine (LGBM) classifiers to predict drought occurrences. The study aims to enhance the ability to forecast drought and inform decision-making in the preparations and mitigations against drought in Kilifi County, Kenya.

1.3 Objective of the Study

1.3.1 General Objective

The primary goal of this study is to explore an ensemble machine learning model for drought severity prediction for Kilifi County using remote sensing data.

1.3.2 Specific Objectives

i. To explore the key predictors that significantly influence drought occurrence in Kilifi County, Kenya.
ii. To develop an ensemble machine learning model using selected variables to predict drought in Kilifi County, Kenya.
iii. To assess the effectiveness and performance of the developed ensemble machine learning model compared to individual models in predicting drought severity.

1.4 Research Questions

i. What are the key predictors among remote sensing variables that significantly influence drought occurrence in Kilifi County, Kenya?
ii. How can an ensemble machine learning model be developed using selected remote sensing variables to predict drought in Kilifi County, Kenya?
iii. How effective is the developed ensemble machine learning model compared to individual models in predicting drought severity?

1.5 Justification of the study

This study is significant due to its innovative use of ensemble machine learning techniques for predicting drought severity through remote sensing data. The study utilizes an ensemble approach that integrates various machine learning models to enhance the predictive performance, thereby improving the accuracy and reliability of drought forecasts. This is especially significant in areas where conventional monitoring techniques face difficulties due to limited data and variability. Integrating remote sensing data with advanced machine learning techniques enhances understanding of drought dynamics, facilitating improved forecasting and more effective drought management strategies.

The primary beneficiaries of this study include the following: Weather forecasting of droughts helps agricultural and water resource managers make the right decisions concerning crop management and water usage. Decision makers at the local level and policy makers will also benefit from better forecasts for droughts and their impact since it enables them to formulate

and implement appropriate policies to address the issues of droughts affecting society and infrastructure.

In the ASAL regions, improved drought prediction for the community members can help them have appropriate and early intervention measures that will help them deal with drought and its consequences on their livelihoods better. Additionally, humanitarian organizations based in drought-affected areas will benefit from the improved data quality to efficiently plan and implement their planned interventions.

1.6 Scope and Limitation of the Study

This study focused on developing a drought prediction model for Kilifi County, Kenya. The model was tailored to particular conditions of the selected region, considering basic remote sensing parameters such as NDVI and EVI. By utilizing these and related environmental factors, the model demonstrated reasonable effectiveness in predicting drought-associated conditions, its accuracy and reliability remain moderate. Additionally, its applicability to other regions or broader drought conditions is limited by site-specific factors and climatic differences.

VT OMNES VNVM SINT

Chapter 2: Literature Review

2.1 Introduction

The previous chapter provided an overview of the background, problem statement, research objectives, research questions, justification, and scope of the study. As part of the background discussion, the study references literature on drought prediction and machine learning. Additionally, this chapter explores various studies conducted in this research area. A comprehensive search was performed to gather relevant literature, covering multiple aspects of the topic and identifying key resources for the review.

The Google scholar search engine was utilized. Keywords include Machine learning, drought indicators, ensemble learning, and drought prediction. This literature review considers the research objectives and questions to form the basis of the discussion. The review seeks to guide the research and form a basis for identifying research gaps and, therefore, the need to conduct research in this field of study. Thus, a detailed literature review on this study's objectives and research questions is organized according to sections and subsections. The chapter concludes with a conceptual framework.

2.2 Theoretical Framework

2.2.1 Types of Drought and Its Impact

According to Yang and Liu (2020), drought is a condition characterized by an increased length of time when water is insufficient for various environmental and human needs. Droughts differ significantly from other disasters in that they develop slowly and may persist for months or years, impacting agricultural production, water supply, and various ecosystems. Drought onset and its development are widely associated with meteorological, hydrological, agricultural, and or/ socio-economic factors. It is, therefore, essential to understand the various types of drought to adequately address the monitoring, forecasting, and measures to take.

Primarily meteorological drought can be defined as a lack of rainfall for an extended period of time. This drought type often acts as the initial warning of a potential drought situation. It is usually evaluated utilizing the indices (Ndayiragije, 2022) for meteorological drought assessment, Palmer Drought Severity Index (PDSI), Percent of Normal Precipitation (PNP) and Standardized Precipitation Index (SPI) are adopted here. SPI values act in the capacity of surrogates for drought occurrences by capturing variations of regular precipitation (Okal et al., 2020).

Agricultural drought directly correlates to soil moisture deficits and can drastically affect crop production and the health of plants. This dry spell is caused by a lack of enough soil moisture to meet the needs of the crops; this, in turn, leads to lower yields and possibly significant impacts on food security (Cullen, 2023). Agricultural drought generally follows meteorological drought due to a lack of precipitation that gradually depletes the soil moisture. They are several drought indicators that are commonly used for monitoring and forecasting agricultural drought they include Crop Moisture Index (CMI), Soil Moisture Percentile (SMP), Soil Moisture Deficit Index (SMDI) and Normalized Difference Vegetation Index (NDVI) (Alahacoon & Edirisinghe, 2022). High temperatures and evapotranspiration rates worsen the soil moisture deficit, extending drought and its impacts on agricultural systems.

Hydrologic drought is defined by depletions in surface water and groundwater stocks, including reduced streamflow, reservoir storage, and groundwater replenishment (Carroll et al., 2021). This generally follows a meteorological drought, where the effects of reduced precipitation work their way gradually through the hydrologic system. Carroll et al. explain that hydrologic drought strongly impacts water supply, aquatic life, and hydroelectric power production. Indicators of hydrological drought are the runoff or streamflow percentiles, Palmer Hydrologic Drought Index (PHDI and the Standardized Runoff Index (SRI) (Zellou, El Moçayd, & Bergou, 2023). Hydrological drought occurrence is mostly influenced by climatic conditions and catchment attributes, such as soil composition and water retention capacity.

Socio-economic drought is the feedback loop between physical water scarcity and human activity in the form of supply and demand economics of commodities such as water and food. This drought type is defined by its effects on society, economy, and environment, encompassing diminished agricultural output, water supply deficits, and heightened susceptibility to wildfires (Liu et al., 2020). Socio-economic drought is difficult to measure because it has many dimensions, evolves slowly, involves complex socio-ecological interactions, and is defined differently by different people. While drought indicators such as the Social Water Stress Index (SWSI) have been developed to assess water scarcity, the literature on drought prediction and characterization of socio-economic drought is sparse. Some data like water quality, crop yield and remotely sensed vegetation stress can provide indirect information on socio-economic drought's occurrence and effects.

Figure 2. 1 presents the relation of many variables involved in drought and its surrounding conditions. The illustration presents four main categories: meteorological, hydrological, agricultural, and socio-economic impacts, abbreviated as the MHAS circle. The

interconnections performed during the drought events make the apparent arrows applicable to the relationships between these categories. The climate in a geographic area relates to parameters such as precipitation and temperature, which impact the hydrological conditions in a geographical region and, hence, affect agriculture. Agriculture yields define the socio-economic features of the areas, and some of the yields affect hydrological and meteorological conditions, affecting agriculture yields. This somewhat integrated structure suggests that, while drought is the climatic condition, it is also a result of climate interaction with agriculture, water resources, and socio-economics.

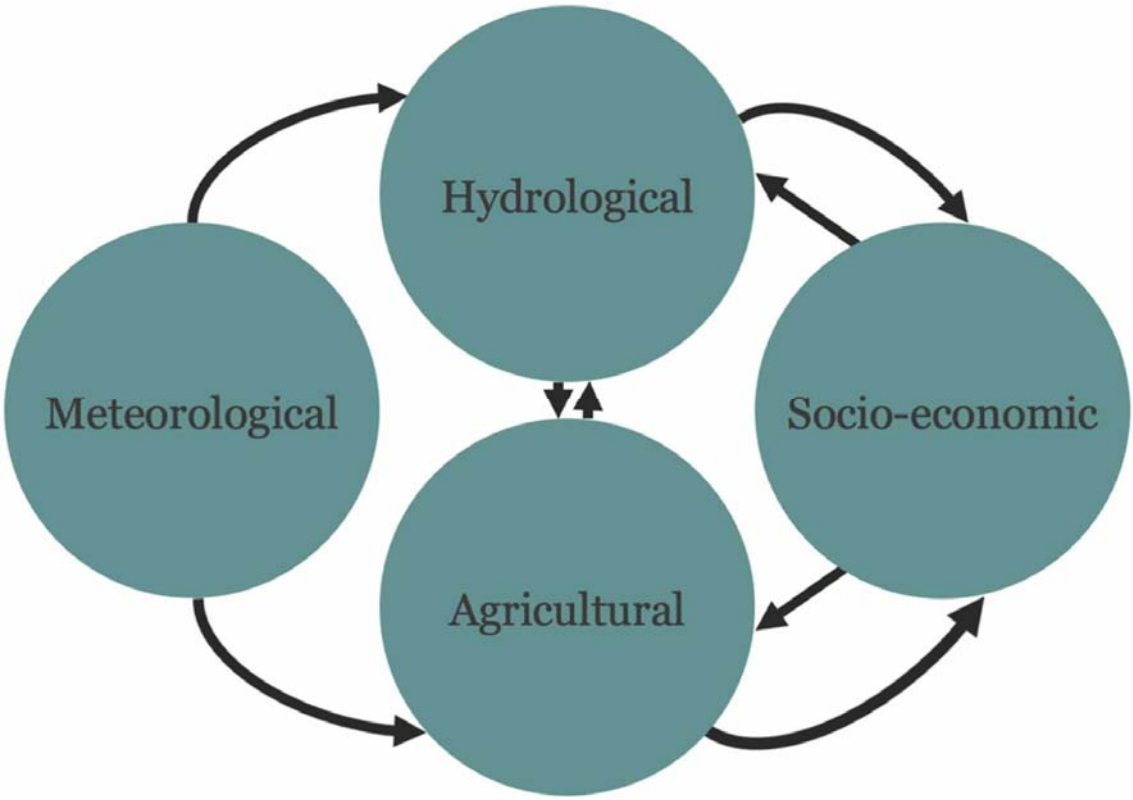


Figure 2.1: Cascading relationship between different types of drought, illustrating the interconnections between them

2.2.2 Drought Indicators

Historically, Drought monitoring relies on meteorological indicators and ground measurements (Li, Tao, & Zhang, 2022). The Palmer Drought Severity Index (PDSI) is commonly used across multiple fields to monitor drought conditions by calculating the water balance in a region, considering variables such as rainfall, temperature, soil moisture levels, and previous PDSI

readings. This comprehensive approach allows the PDSI to effectively reflect moisture availability in the area. However, a fundamental limitation of the PDSI is its response time, as it requires a 9-month calculation period, making it unsuitable for identifying droughts promptly (Han & Singh, 2023). While the index is effective for analyzing long-term trends and persistent weather anomalies—whether excessive rainfall or prolonged dryness—it lacks the sensitivity to detect short-term droughts that can quickly affect crop health during vital growth stages.

Drought assessments can leverage remote sensing data by analysing the health of vegetation and tracking changes in land cover. The Normalized Difference Vegetation Index (NDVI), a commonly used vegetation index is closely correlated with soil moisture and is instrumental in assessing drought stress by measuring vegetation health (Felegari et al., 2022). Additionally, satellite-derived surface brightness temperatures obtained from thermal channels provide valuable information on moisture levels, as thermal infrared (TIR) bands reveal surface moisture conditions. Temperature Condition Index (TCI), derived from TIR data, quantifies plant stress caused by temperature.

The Standardized Precipitation Index (SPI) is endorsed by The World Meteorological Organization (WMO) as a standard tool for tracking meteorological drought over varying timeframes, focusing on precipitation alone (Musonda et al., 2020). However, for a more comprehensive approach, the Standardized Precipitation Evapotranspiration Index (SPEI) accounts for precipitation and evapotranspiration, thus balancing water input and loss and better representing water deficits or surpluses in an area.

For agriculture-specific drought impacts, the Crop Moisture Index (CMI) combines precipitation, soil moisture, and crop evapotranspiration data to evaluate crop moisture levels and predict potential drought effects on agriculture (Sharma et al., 2022). Similarly, the Soil Moisture Deficit Index (SMDI) and Evapotranspiration Deficit Index (ETDI) target short-term agricultural drought conditions by using soil moisture and evapotranspiration data to gauge water availability and stress. These tools provide a multidimensional approach for capturing drought severity across different time scales and ecological impacts.

2.2.3 Drought Modeling Components

Drought severity and duration can be evaluated using multiple parameters, including precipitation, soil moisture, streamflow, evapotranspiration, temperature, wind speed, relative humidity, and vegetation health. These elements offer essential insights into drought conditions. Research on drought commonly relies on data sourced from climatic, remote

sensing (RS), hydrological, and atmospheric observations, collected through various platforms such as ground-based monitoring stations, earth-observing satellites, open-linked datasets, diverse sensors, and Internet of Things (IoT) devices. The gathered data is often heterogeneous, consisting of different formats—including continuous, discrete, image, text, and video—enabling a more comprehensive analysis of drought indicators (Pujar et al., 2024).

2.2.4 Ensemble Machine Learning Methods

According to Sarker (2021), machine learning is a set of techniques that allow a computer to learn patterns from raw data and give predictions without being programmed. Individual machine learning models have often been seen to lack the variety and robustness needed to cover challenging operations (Barbierato & Gatti, 2024). Ensemble learning methods address these limitations by integrating a collection of base learners' models that results in higher generalization performance than each model can provide. This works by eliminating errors that may be inherent in individual models to bring forth a result that is more accurate and reliable. It thus turns out to be ensemble learning, viewed as a collective intelligence twin in machine learning and offers a new way of looking at many different problems in the field. Most previous research justifications have established that ensemble learning is much more powerful than traditional approaches relying on a single model, particularly in hydrological modeling, which is an excellent boost for forecasting results (Zounemat-Kermani et al., 2021; Khan, Chaudhari, & Chandra, 2023). The recent rise in the application of ensemble machine learning models in hydrological modeling shows that this approach is effective and efficient for capturing complex patterns.

Popular Ensemble Learning methods include Bagging, Boosting, Stacking, and Blending, but the most favorable method is Stacking. Stacking helps to combine different models and, therefore, increases the prediction accuracy because it uses the strong points of each model and compensates for their weaknesses (Odegua, 2019). Stacking ensemble learning is a powerful method in machine learning, as it takes full advantage of various base models to achieve better overall predictive performance and generalization capability. This approach contains two phases: training base models and training a meta-model. First, divide the original dataset into two sets namely training and testing sets, on the training set apply k-fold cross-validation. This splits the training dataset into k subsets: one subset is used as a test set while the remaining k-1 subsets form the training set. The predictions are generated for every fold; hence, one gets a complete set of predictions for the training dataset. These predictions become validation folds,

combined into a new dataset on which the meta-model learns, hence the name "stacking" (Gu et al., 2022).

The contribution of base models to the stacking model is significant for the overall performance. Further, the meta-model integrates the output of the base models, very often using techniques such as multi-linear regression. RF can serve effectively as a base and meta-model because of its predictive solid power. Second, all the predictions from k-fold cross-validation are rearranged into a new dataset according to the order of the original training set. Therefore, the new dataset is used as input to train the meta-model (Lu et al., 2023). The meta-model, conceptualized as a model that draws an inference from the outputs of base models, is thereby trained on this revised dataset. The predictions of the testing set of base models are aggregated to form the testing set of the meta-model. The most important factors influencing the efficacy of stacking ensemble learning include the base model choice and the meta-model. The chosen base models are usually based on their strengths since these different methodologies in handling data will complement one another.

2.3 Empirical Review

Ghazaryan et al. (2020) discussed the challenges in drought assessment at regional scales, focusing on the temporal dynamics of crop state and drought effects using the time-series data derived from the optical and SAR satellites. Such derived indices were the Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI), Land Surface Temperature (LST), and SAR-based parameters. The logistic regression tested their drought-generated variability. The results showed that NDVI, NDMI, and SAR parameters allowed substantial accuracy in measuring agricultural dryness, and prediction accuracy ranged from 70–75 percent to 60 percent. LST acted as a leading tool for all the crops but more so for maize and sunflower crops. This model allows the decision-makers at higher levels in areas prone to drought to arrive at comprehensive assessments of crop status.

Monitoring climate change impacts on drought in semi-arid areas is a fundamental necessity, especially from a management point of view, and it is also concerned with establishing water resource policies. En-Nagre et al. (2024) presented a study on the UDB (Upper Drâa Basin), Morocco, based on data from 1980 to 2019. This work presents an application of some basic metrics describing droughts-the Standardized Precipitation Index (SPI) and the Standardized Precipitation Evapotranspiration Index (SPEI)-evaluated for various time scales: 1, 3, 9, and

12 months. The Mann-Kendall test and Sen's Slope estimator were employed for trend analysis, and appropriate statistical tests were performed.

The results indicated that the drought distribution in the UDB is non-uniform and has a definite downward trend in SPEI values, indicating deteriorating drought conditions. Out of the investigated machine learning algorithms, the algorithms of Random Forest, Voting Regressor, and AdaBoost Regressor yielded comparatively better performance, with the NSE varying between 0.74 and 0.93. In contrast, the performance of the K-Nearest Neighbors Regressor algorithm was relatively poor, with NSE between 0.44 and 0.84 (En-Nagre et al. 2024). The results confirm that machine learning models, especially ensemble approaches, perform better in predicting the drought indices and guiding appropriate measures for water management in semi-arid locations.

The SPI 6- and 9-months-ahead prediction by Lalika et al. (2024) was done using five machine-learning algorithms, namely LSTM, MARS, SVM, ELM, and M5 Tree based on rainfall data across the Wami River sub-catchment, Tanzania. The results indicated that LSTM was best for forecasting drought conditions, with the highest NSE, reaching 0.99 for all the stations and an R-value of 0.99 for most stations. Generally, ASAL drought prediction faces unique challenges, including but not limited to data scarcity, climate unpredictability, and socio-economic influences. Other critical issues Prodhan et al. (2022) identified included challenges the machine learning models faced because of deficient training datasets, which included noise, outliers, and observational bias in spatial data.

Wehbe and Temimi (2021) investigated the spatiotemporal distribution of the Arabian Peninsula water resources using remote sensing data. The approach used in this study consisted of analysing data from GRACE, TRMM, and AMSR-E to evaluate TWS, precipitation, and soil moisture by using the modified Mann-Kendall test, trend analysis, and change-point detection statistics. The results indicated a significant temporal variability and spatial distribution of TWS. The decline in soil moisture and precipitation in southern AP showed distinct trends and severe TWS depletions in areas of intense groundwater extraction in northern AP. These results have emphasized the role of climate change and groundwater extraction in regional water resource depletion.

Various models of drought prediction and technologies being used to improve the accuracy are discussed in detail by Nandgude et al. (2023). The authors identified that drought is a widespread but complex natural hazard caused principally by deficiencies in precipitation and

anomalies in temperature, bringing huge human and economic losses. The study explores the occurrence of drought all over the globe arising out of hydro-meteorological conditions and fluctuations in sea surface temperature. This covers a wide range, from conventional statistical methods to sophisticated ML and DL techniques for drought predictions. The process designed by Nandgude et al. was based on the essential analysis of the literature available to assess the efficiency of various methods. The authors pointed out that, in comparison with conventional models, both machine learning and deep learning models have considerably improved predicting droughts, primarily due to their skill to capture intricate nonlinear relationships that the huge datasets represent. Based on this, this paper reviews essential elements of remote sensing data, meteorological observations, hydrologic models, and climatic indices crucial for enhancing predictive capabilities.

Latif et al. (2023) evaluated a range of rainfall prediction models by statistically and comparatively testing machine-learning algorithms. The study used satellite images, radar data, and terrestrial observations to test the adequate performance of deep learning methodologies, such as the LSTM network. The results indicated that the LSTM models better predicted rainfall than other models. The LSTM models showed high accuracy, as portrayed in the resultant measures such as RMSE, R^2, and MAE. The study indicated that while machine learning models are good at forecasting rainfall, remote sensing and hybrid models require further research due to their potential benefit, and few studies have characterized these models.

Senhorelo et al. (2024) applied path analysis with interdependencies among six meteorological variables: air temperature, net solar radiation, reference evapotranspiration (ET0), precipitation, relative humidity, and water deficiency. The research investigated the impacts of these elements on NDVI and EVI in the grasslands of Espírito Santo, Brazil, at different time lags. It was observed that climatic variables explained more than 50% of the total variance of both indices, and this influence is due to the direct effect of temperature and relative humidity, while the impact of precipitation is indirect. Indeed, this study showed that ET0 contributes to changing both indices mainly after 32 days, but the response of solar radiation and water scarcity is immediate within 32 days, showing the fast response of vegetation to these variables.

The study by Elbeltagi et al. (2024) focused on developing and testing hybrid machine-learning methods to estimate drought severity for eight governorates in Upper Egypt. Different combinations of CNN with LSTM, RF, SVR, and XGB were considered when developing a Palmer Drought Severity Index forecast. Generally, the CNN-LSTM model was the best performing model among all models, with higher performances for the statistical metrics such

as the NSE and R^2. However, the CNN-SVR model performed outstandingly during all the tests. For this reason, the study recommended the CNN-LSTM model for future works because it showed outstanding effectiveness in PDSI value forecasting.

Taiwo et al. (2023) studied the land use and land cover LULC changes around FUTA in Nigeria from 1991-2031 and its effect on agricultural parameters. It was interpreted from the analysis of the Landsat image and various vegetation indices that both built-up and unused land areas increased while the lightly vegetated area decreased. It also projects further vegetation reduction and expansion in built-up areas by 2031 at a pace that will significantly impact drought severity and the health of crops. The results revealed that sustainable land management reduces water stress and drought conditions.

The hybrid models were utilized by Altunkaynak and Jalilzadnezamabad (2021) to investigate the methods that could improve forecast accuracy and augment the lead time of the Palmer Drought Severity Index-PDSI. They introduced new W-Fuzzy, W-kNN, and W-SVM models, incorporating the DWT into Fuzzy, kNN, and SVM methodologies. These hybrid models showed better performance for a lead time up to six months earlier than their solo counterparts in the Marmara region, Turkey. It is indicated by the W-Fuzzy model in terms of MSE, CE, and R2.

Nguyen and Ly (2023) formulated and assessed three machine learning models—Decision Tree, Light Gradient Boosting Machine, and Extreme Gradient Boosting (XGBoost)—to forecast the compressive strength (CS) of fiber-reinforced self-compacting concrete (FRSCC). They evaluated model performance using 387 data samples and 17 input parameters by Monte Carlo and K-fold cross-validation. Results showed that XGBoost outperformed the other models regarding accuracy and stability, with the highest R^2 of 0.992 and the lowest RMSE and MAE of 1.892 and 1.438 MPa, respectively. The sensitivity study indicated that the most influential variables on the compressive strength of FRSCC are cement, aggregate, water, and sample age, although their impact is at different degrees.

Varela and Hernández (2024) investigated the application of machine learning in the prediction of hydrological drought in Chihuahua, Mexico, based on the Standardized Precipitation Evapotranspiration Index. Since this region often has to bear the drought, the people in this area also suffer from such a calamity without adequate warnings. The authors, in this paper, develop predictive models using ANN, LSTM, and SVR to estimate SPEI at 12-month scales (SPEI 12) and 24-month scales (SPEI 24) with a sample period ranging from 1901 to 2020,

applying the capability of machine learning for time series forecasting. Drought cycle simulation is an attempt at improving predictability regarding drought impacts. It was a methodology of research that encompassed 956 experiments, using three different machine learning techniques and changing parameters such as the number of neurons, kernel type, and polynomial degree. The models have been evaluated by different performance indices such as MSE, MAE, MBE, the coefficient of determination R^2, and the Kendall coefficient (Varela & Hernández, 2024). After this, two of the best models for each approach were selected for further analysis. The average performance showed that for SPEI 12, the best models had an MSE of 0.0051, MAE of 0.0537, MBE of 0.0218, R^2 of 0.8495, and a Kendall coefficient of 0.7592. For SPEI 24, the respective performances were an MSE of 0.0024, MAE of 0.0375, MBE of 0.0162, R^2 of 0.9218, and a Kendall coefficient of 0.8558. These results confirm that machine learning systems, including ANN, LSTM, and SVR, can potentially succeed in estimating SPEI values for the successful prediction of hydrological drought, especially at an extended temporal scale like SPEI 24. This study highlights the recent emphasis on sophisticated techniques for modelling drought dynamics to enhance the reliability of early warning systems and preparedness.

Understanding the processes of extreme drought events is of enormous scientific interest and is highly relevant for developing better risk management. Using artificial neural networks, Felsche and Ludwig (2021) predicted the occurrence of droughts in Munich and Lisbon, two divergent geographical regions in Europe, one month in advance. The methodology used a single-model initial-condition large ensemble CRCM5-LE and required 28 atmospheric and soil variable inputs. These data have been produced at Ouranos within the frame of the ClimEx project by driving the Canadian Regional Climate Model CRCM5 with 50 members of the Canadian Earth System Model CanESM2. Drought events have been selected using the standardized precipitation index, and the most efficient machine learning algorithms classified - drought or no drought - correctly around 55%-57% of the events in each domain. Explainable AI techniques, such as SHapley Additive exPlanations (SHAP), were employed to elucidate the trained models, demonstrating that variables including the North Atlantic Oscillation index and air pressure one month prior to the event were essential for precise predictions (Varela & Hernández, 2024). The study indicated that seasonality significantly impacted prediction performance, especially in Lisbon, where seasonal variations created challenges for sustaining prediction accuracy. This study underlines possible gaps in prediction accuracy and encourages further study with feature selection optimized to better understand complex interactions within

various climatic contexts. Therefore, the research calls for additional model performance testing for a longer lead time through varied seasons to enhance reliability.

Charles (2020) investigated approaches for predicting drought severity and its impacts based on remote sensing and socio-economic data. This research developed methods for quantifying drought severity and effects through the Vegetation Condition Index and Mid-Upper Arm Circumference, abbreviated VCI3M and MUAC, respectively. Homogeneous and heterogeneous model ensembles were developed using ANN and SVR following the strategy "over-produce then select" to get reliable predictions. In all, 244 different models were created from comprehensive data, of which 111 models were selected for developing model ensembles. The heterogeneous stacked ensemble achieved an R^2 of 0.94 with 80% classification accuracy, considerably outperforming the performance of the individual ANN and SVR models with R^2 values of 0.83 and 0.78, respectively. Some gaps identified from this study are evaluating other techniques, constructing heterogeneous ensembles, and forecasting periods as long as six months to develop improved prediction accuracy and drought management strategies.

Barrett et al. (2020) sought to develop a suite of state-of-the-art forecasting methodologies that explicitly targeted enhancing EWS for pastoralist communities in Kenya, with extreme dependence upon satellite-based indicators, including the 3-month Vegetation Condition Index-VCI3M. Employing linear autoregression and Gaussian process modelling of the MODIS and Landsat data, the authors generated vegetation condition forecasts several weeks into the future. These models have a high hit rate of 89% four weeks in advance and 81% six weeks in advance, with respective false alarm rates of 4% and 6%. These results prove the efficiency of these models in depicting vegetation decline in good time for timely intervention. Future studies should incorporate more modelling approaches, expand the ranges of environmental and socio-economic data, and assess practical applications of those forecasting methods in real-world early warning systems for scalability and adaptability.

2.4 Research Gap

While existing studies have explored drought severity and prediction using SPI and other indices, many rely on single or limited data sources and often do not combine the full range of meteorological and remote sensing variables. Additionally, no study specifically focused on drought prediction for Kilifi County, a region prone to droughts that significantly impact local agriculture and water resources. This study aims to bridge this gap by employing a comprehensive set of predictive variables through an ensemble machine-learning approach that

incorporates both local meteorological data and remote sensing data, offering a more precise and regionally relevant model for drought prediction. This study also contributes by applying a novel machine learning stacking approach to improve predictive accuracy, significantly enhancing traditional statistical models.

2.5 Conceptual Framework

The conceptual framework illustrated in Figure 2.2 represents a stacking ensemble approach for drought severity prediction. The process begins with the input of features such as temperature, precipitation, wind speed, and NDVI, which undergoes data pre-processing to handle missing values, normalize data, and prepare it for further analysis. Simultaneously, feature engineering is performed to extract Standard Precipitation Index attribute that enhance predictive accuracy. The processed data is then split into two -training and testing sets- to train and evaluate the model.

The stacking ensemble method is built using multiple base classifiers: Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). Each of these models independently learns patterns from the training set and generates predictions. These predictions are then passed to a meta-model, which in this case is Logistic Regression. The meta-model integrates the outputs of the base models to make the final decision, leveraging the strengths of all classifiers while minimizing individual weaknesses. Finally, the trained stacking model produces the drought severity prediction output, providing a more robust and accurate prediction compared to using individual models. This ensemble approach improves prediction performance by combining multiple machine learning algorithms into a single, optimized predictive framework.

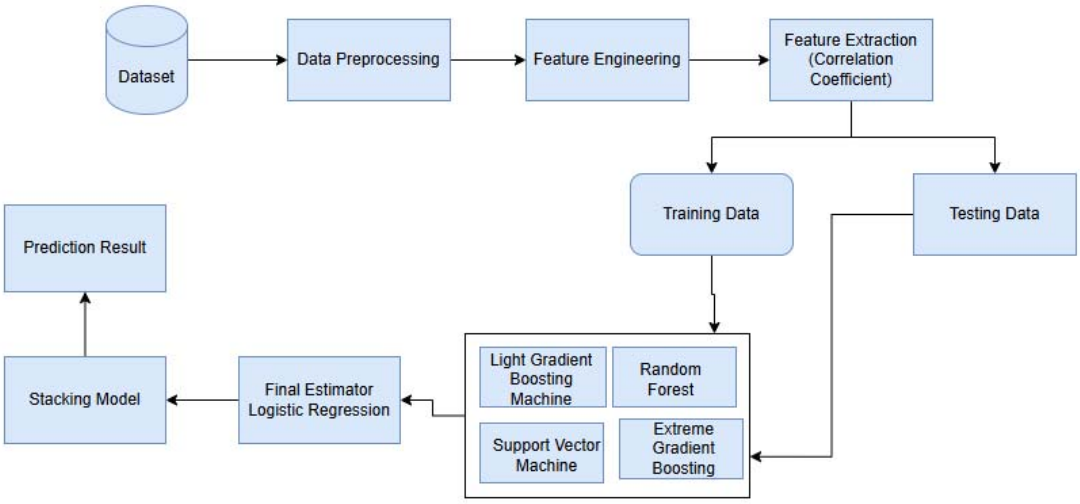


Figure 2.2: Conceptual Framework

2.6 Summary of Literature

Table 2.1 summarizes some recent studies pertinent to drought prediction and early warning systems and the impacts of drought on various environmental and socio-economic factors. The methodologies applied vary between traditional statistical techniques, advanced machine learning, and hybrid models and thus reflect the diversity of approaches characteristic of this developing field of study. The results also show the key improvements in forecast precision, the input of remote sensing data, and new models that comprise LSTM, CNN, and ensemble techniques. The table also indicates the gaps in each study by mentioning the limitations within data collection, more comprehensive geographical validation, and capturing long-term drought trends. This paper reviewed research that can shed light on existing methodologies and suggest further areas of study and improvement in drought monitoring and forecasting.

Table 2.1: Summary of Literature and Gaps

Study	Author(s)	Technique/methodology	Findings	Gaps
Forecasting 3-month Vegetation Condition Index (VCI3M) for Drought Early Warning	Barrett et al. (2020)	Linear autoregression, Gaussian process modeling, MODIS, Landsat data	Achieved 89% hit rate and 4% false alarm rate for identifying drought conditions four weeks in advance. Enhances early action capabilities for vulnerable communities.	Limited to 3-month VCI3M; may not capture longer-term drought trends.
Drought Prediction in Upper Drâa Basin, Morocco	En-Nagre et al. (2024)	SPI, SPEI, Mann-Kendall test, Sen's Slope estimator, Random Forest, Voting Regressor, AdaBoost Regressor, KNN Regressor	Revealed heterogeneous drought distribution with a decreasing trend in SPEI values; Random Forest, Voting Regressor, and AdaBoost Regressor were most effective. NSE values ranged from 0.74 to 0.93.	Enhanced data collection and spatial distribution of measurement sites are needed to improve data representativeness.
Predicting SPI in Wami River Sub-catchment, Tanzania	Lalika et al. (2024)	LSTM, MARS, SVM, ELM, M5 Tree	LSTM performed best with an NSE of 0.99 and Pearson's correlation coefficient (R) of 0.99, showing high accuracy in drought prediction.	Challenges include data scarcity and variability in climatic conditions.
Advanced stacked integration method for forecasting long-term drought severity: CNN	Elbeltagi et al. (2024)	Convolution Neural Networks (CNN)-Long Short-Term Memory (LSTM), CNN-Random	CNN-LSTM model showed the highest accuracy with high NSE and R^2; the CNN-SVR	There is a need for further testing of hybrid models in different regions and conditions.

with machine learning models.		Forest (RF), CNN-Support Vector Machine (SVR), and CNN-Extreme Gradient Boosting (XGB).	model excelled during testing.	
Monitoring and predicting the influences of land use/land cover change on cropland characteristics and drought severity using remote sensing techniques.	Taiwo et al. (2023)	Landsat imagery	Increase in built-up and bare land areas; decrease in light vegetation; predicted declines in vegetation and increases in built-up areas by 2031, impacting drought severity and crop health.	Need for sustainable land use management strategies to address water stress and drought.
Extended lead time accurate palmer drought severity index forecasting using hybrid wavelet-fuzzy and machine learning techniques.	Altunkaynak & Jalilzadnezamabad (2021)	Discrete Wavelet Transform (DWT) with Fuzzy, k-Nearest Neighbor (kNN), and Support Vector Machine (SVM)	Hybrid models (W-Fuzzy, W-kNN, W-SVM) demonstrated superior performance; W-Fuzzy showed the best overall performance.	There is a need for further validation in different regions and extended lead times.
Review of Drought Prediction Models and Technologies	Nandgude et al. (2023)	Survey of statistical, machine learning (ML), and deep learning (DL) models; integration of remote sensing and climate indices	ML and DL models showed improved prediction accuracy and significant advancements in technology for drought prediction.	Integration of diverse data sources requires high-quality data, computational resources, and interpretation challenges of DL models.
Comparison of Rainfall Prediction Models	Latif et al. (2023)	Satellite imagery, radar data, ground-based observations, deep learning (LSTM)	LSTM models outperformed others in rainfall prediction with high accuracy	Need for further exploration of remote sensing and hybrid models.

			metrics (RMSE, R^2, MAE).	
Relationships Among Meteorological Variables and Vegetation Indices in Brazil	Senhorelo et al. (2024)	Path analysis, NDVI, EVI	Meteorological variables explained over 50% of the variability in NDVI and EVI; temperature and relative humidity had direct effects, while precipitation had indirect effects.	Seasonal variations and delays in impact need more attention for accurate predictions.
Spatiotemporal Distribution of Water Resources in the Arabian Peninsula	Wehbe and Temimi (2021)	GRACE, TRMM, AMSR-E, trend analysis, modified Mann-Kendall test, change point detection	Significant temporal variability and spatial distribution of TWS; decreasing precipitation and soil moisture in the southern AP, severe TWS depletion in northern AP.	Focus on groundwater extraction impacts; may need broader regional data for comprehensive understanding.
Predicting Drought Severity and Effects using Remote Sensing and Socio-Economic Data	Charles (2020)	ANN, SVR, VCI3M, MUAC, heterogeneous and homogeneous model ensembles, "over-produce then select" strategy	Heterogeneous stacked ensembles outperformed homogeneous ensembles and individual models and achieved R^2 of 0.94 and 80% classification accuracy—improved	We need to evaluate additional techniques, build more diverse ensembles, and extend forecasting periods up to six months.

			performance in outlier classes.	
Enhancing Drought Early Warning Systems for Pastoral Communities in Kenya	Barrett et al. (2020)	Linear autoregression, Gaussian process modeling, VCI3M, MODIS, Landsat data	Forecasting models achieved high accuracy: 89% hit rate four weeks ahead and 81% six weeks ahead with 4% and 6% false alarm rates. Effective in anticipating vegetation conditions in advance.	Explore additional modeling techniques, integrate more data sources, and evaluate practical implementation in real-world EWS.

VT OMNES VNVM SINT

Chapter 3: Research Methodology

3.1 Introduction

This chapter explains the research methodology which is structured around the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, as illustrated in Figure 3.1. The CRISP-DM framework is a widely recognized and standardized approach for data mining projects, comprising six key phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment**.** Its structured yet adaptable approach has earned it widespread recognition, particularly for advancing data-driven solutions in drought prediction (Vanegas et al., 2023). Despite being introduced over two decades ago, CRISP-DM remains the industry standard, valued for its dependability, flexibility, and capability to produce explainable insights and robust predictive models (Martínez-Plumed et al., 2019).

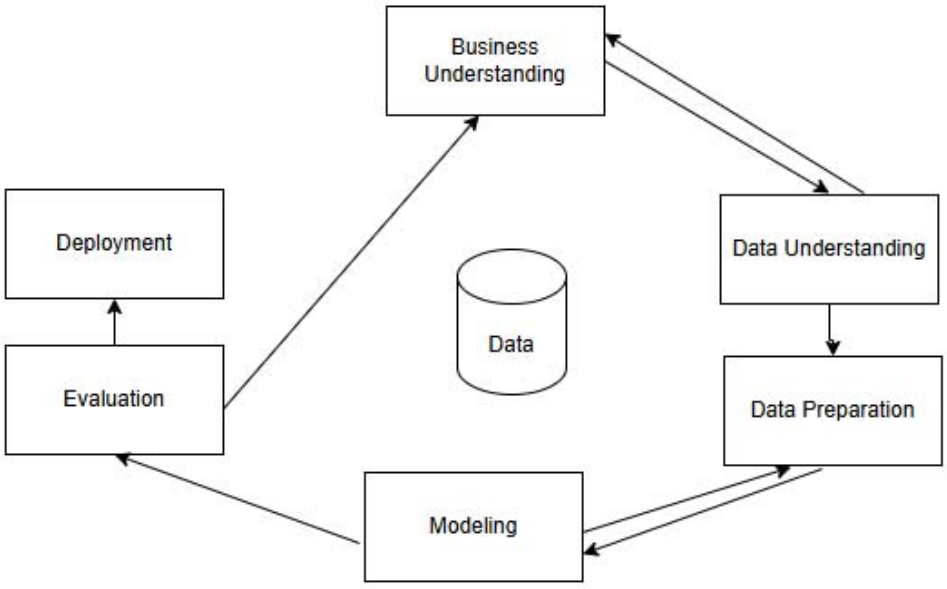


Figure 3.1: CRISP-DM Framework Framework (Adapted from Niaksu, 2015*)*

3.2 Business Understanding

The first step in the CRISP-DM framework is focused on understanding the objectives of the study and requirements from a business perspective. This involved defining the problem, setting specific goals, and evaluating the current situation to establish the project's scope. This study is determined to address severe drought occurrence in Kilifi County, Kenya, which

significantly impacts agriculture, water availability, and livelihoods. Droughts disrupt food production, cause water scarcity, and exacerbate poverty and malnutrition in affected communities. Predicting drought severity accurately is crucial for effective resource allocation, early warning systems, and disaster preparedness. The primary objective in this study was to develop an ensemble machine learning model to predict drought severity using selected variables that influence drought occurrence.

Traditional methods of drought prediction often fail to account for the complex interactions between climatic variables such as temperature, precipitation, humidity, and vegetation indices like NDVI and EVI. These methods also lack the precision needed for localized forecasting, limiting their effectiveness in addressing the specific needs of vulnerable regions like Kilifi County. By addressing these limitations, this study aimed to develop a robust, data-driven predictive model that offers enhanced accuracy. The insights generated can empower government agencies, humanitarian organizations, farmers, and other stakeholders to make informed decisions, mitigate drought impacts, and promote sustainable resource use.

3.3 Data Understanding

The Data Understanding phase of this study involved gathering relevant data, familiarizing ourselves with it, and identifying key insights to inform the development of a predictive model for drought forecasting in Kilifi County, Kenya. Purposive sampling method was employed in this study to select key features for drought prediction in Kilifi County. The dataset consisted of monthly data recorded from 2000 to 2022, which covered a significant period for identifying patterns in climate variables and their relation to drought occurrences. The sample size included a total number of 276 records for these years across the selected variables.

The dataset was obtained from Google Earth Engine and NASA POWER since they were considered to be reputable sources and the data was downloaded in CSV format suitable for analysis. Key variables include vegetation indices like EVI (Enhanced Vegetation Index) and NDVI (Normalized Difference Vegetation Index) which provided valuable insights into vegetation health and drought conditions. Metrological data like Evaporation Land, Evatranspiration, Soil Moisture, Soil Wetness temperature, dew point, precipitation, humidity, wind speed, and surface pressure, were also included. These variables are further described in Table 3.1.

Table 3.1: Description of Variables

Variable	*Description*
EVI	Indicates vegetation stress due to water scarcity
NDVI	Tracks vegetation health, with lower values signaling drought effects
Temperature	Atmospheric heat affecting evaporation
Dew point	The temperature at which air becomes saturated with moisture
Wind speed	Rate of air movement
Precipitation	Water falling from the atmosphere
Surface pressure	The force exerted by air above a surface
Evaporation Land	Represents the process where water is converted from liquid to vapor on land surfaces
Evatranspiration	shows the total water loss from the soil and vegetation to the atmosphere
Soil Moisture	Measures the water content in the soil
Soil Wetness	Indicates the degree of saturation in the soil
Humidity	Concentration of water vapor present in the air

3.4 Data preparation

Data Preparation involved cleaning and pre-processing the dataset, and transforming variables where necessary to create more informative features. Feature engineering included generating Standardized Precipitation index (SPI) as a new column. In this study, the SPI was selected as the primary drought index due to its practicality and endorsement by WMO as a key metric for global drought monitoring (Oruc et al., 2024). The SPI's accessibility, computational simplicity, and ease of analysis made it particularly effective for assessing drought conditions across diverse climates and timeframes. SPI values were derived from precipitation data, providing long-term precipitation data to capture regional drought characteristics across multiple timescales (Mukhawana et al., 2023).

To compute SPI, long-term precipitation records were adapted to a probability distribution that fits climatic precipitation data; in this study, the gamma distribution was chosen due to its appropriateness for modelling precipitation and broad acceptance in drought research. The precipitation data was transformed to approximate a normal distribution, with an SPI mean of

"0" and a standard deviation of "1," enabling standardized comparisons across wet and dry conditions. Positive SPI values were indicated wetter-than-average conditions, while negative values were indicated drier-than-average conditions, with increasingly negative values signifying more severe droughts. This normalization feature made SPI a highly effective and standardized measure for quantifying both drought severity and wet periods in Kilifi County. Table 3.2 summarizes commonly used SPI threshold values for defining drought severity.

Exploratory Data Analysis (EDA) involved data exploration and identifying patterns and relationships was also conducted. This included creating visualizations such as histograms and scatter plots, to explore trends, distributions, and correlations between various variables. A correlation matrix was constructed to help in feature selection to be used in model training. Features that were highly correlated were removed and only importance features which influence drought severity were retained and used to train and test the model. Also, data points with inconsistencies, such as outliers that cannot be resolved through standard data cleaning techniques were eliminated.

Additionally, there was splitting of data into training and testing sets for evaluation of the model's performance in the subsequent phases. Data splitting is considered one of the essential steps in ensemble machine learning since it significantly impacts the model's ability to generalize newer data.

Table 3.2: SPI Categories

SPI	Category
Above 1.50)	Very wet
(1.0)–(1.49)	Moderately wet
(-1.0)–(0.99)	Normal
(−1.5)–(−1.0)	Mild Drought
(−2.0)–(−1.5)	Moderate Drought
−2.0 and less	Severe Drought

3.5 Modelling

During Modelling phase, various machine learning algorithms were trained to predict drought severity. Some of the models used include Random Forest, Light Gradient Boosting Machine (LGBM), Support Vector Machine (SVM), and XGBoost. The study implemented a stacking ensemble method to combine the strengths of these individual models to enhance prediction accuracy. Stacking is an example of an ensemble learning method that embeds the strengths of many base models to form a meta-model, allowing for enhanced predictive accuracy and overcoming the identified methodological deficiencies in earlier studies. This involved taking four machine learning models, namely XGBoost, Support Vector Machine, Random Forest, and LGBM as base models. Each base model was trained and these predictions became, in turn, new input features to the next layer. They were then used to train a meta-model, usually a linear regression model, to find an optimal combination of base model outputs. Figure 3.1 shows the process of stacking.

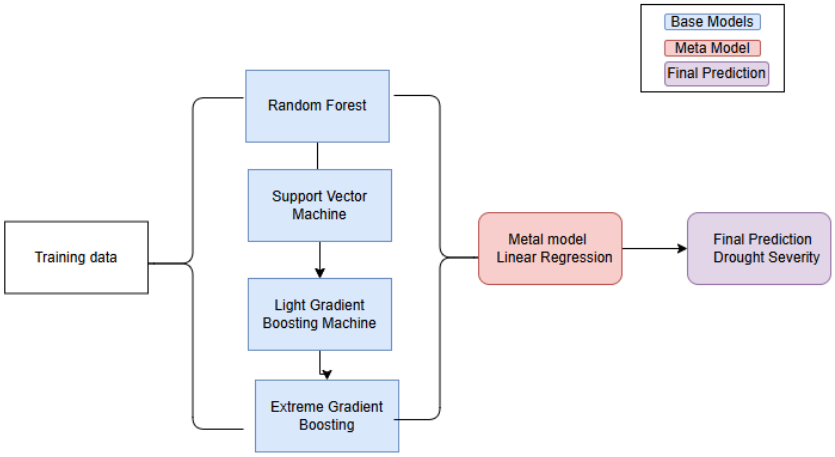


Figure 3.2: Ensemble Learning Architecture for Stacking

3.6 Evaluation

To validate the accuracy of the model, appropriate evaluation metrics were employed. This included measures like accuracy, precision, recall, and F1-score. These gave a comprehensive evaluation of the model's effectiveness in correctly identifying drought instances while minimizing false predictions. Accuracy evaluated the overall correctness of predictions, while precision and recall offered insights into the balance between false positives and false negatives, helping to ensure the model's reliability and stability. The F1-score, a harmonic

mean of precision and recall, further ensured that both false alarms and missed drought predictions were minimized, making the model suitable for real-world applications.

3.7 Deployment

The deployment phase involves implementing the predictive model in a real-world setting. Although deployment was not the primary focus of this study, insights generated from the ensemble model can inform future operational drought monitoring systems. Key outcomes include recommendations for integrating the model into drought early warning frameworks.

3.8 Dissemination of Findings

The findings of this study on drought prediction in Kilifi County, Kenya, will be disseminated through a well-structured plan to ensure the results reach all relevant stakeholders and have a meaningful impact. The research findings will be posted in Strathmore University repository. This will ensure that the study contributes to the broader scientific community and supports further drought prediction and climate resilience research. It will be published in the university repository and easily accessed by anyone.

3.9 Tools and Software

The current research used Python 3.0 to analyze the data. All the processes were done using Jupyter Notebook. Some of the libraries which were utilized include Seaborn and Matplotlib, and Scikit-Learn was used to visualize data.

3.10 Ethical Consideration

Data collection for this study was strictly carried out via online sources. The data was fetched from publicly available platforms or databases that permit research use, ensuring compliance with legal and privacy regulations. The data was securely stored and used solely for the objectives of this study. Additionally, transparency was maintained regarding the statement of purpose for data collection and its compilation, in alignment with ethical research guidelines. Proper citation of data sources, adherence to data usage agreements, and strict measures to prevent violations of privacy or proprietary rights were ensured. Furthermore, the study obtained ethical clearance from the Strathmore University Institutional Scientific and Ethical Review Committee (SU-ISERC).

Chapter 4: System Analysis and Design

4.1 Introduction

This chapter presents the system analysis and design of the proposed model for predicting drought severity using climate and vegetation cover data. The chapter outlines the functional and non-functional requirements, system architecture, data flow, and modelling techniques employed in developing the prediction system. The system analysis phase involves defining the problem, identifying user requirements, and determining the appropriate methodologies for implementing the system. The design phase focuses on structuring the machine learning framework, selecting suitable algorithms, and integrating relevant datasets to enhance model accuracy and reliability.

A comprehensive system design ensures that the developed model effectively processes remote sensing variables such as temperature, precipitation, vegetation indices, and other climatic factors to provide accurate drought predictions. Hence, this chapter discusses the logical and physical design of the system, highlighting key components such as data pre-processing, feature selection, and model evaluation.

4.2 System Requirement Analysis

System requirement analysis is a critical phase in the development of the proposed drought severity prediction model, as it defines the functional and non-functional requirements necessary for the system's effective implementation. This section outlines the key specifications needed to ensure the model's accuracy, efficiency, and usability. The system requirements are categorized into functional requirements, which describe the essential operations the system must perform, and non-functional requirements, which define performance constraints such as scalability, reliability, and security (Vance et al., 2021).

Additionally, the hardware and software requirements necessary for processing large volumes of remote sensing data and running machine learning algorithms are identified. By clearly defining these requirements, this section establishes a foundation for designing and implementing a robust and efficient drought prediction system, ensuring that it meets the needs of researchers, policymakers, and stakeholders involved in climate monitoring and disaster management.

4.2.1 Functional Requirements

The system must perform the following key functions:

i. Data Input: The system should allow input of metrological and remote sensing data, including temperature, precipitation, soil moisture, and vegetation indices.
ii. Data Processing**:** Clean and preprocess the collected data to enhance accuracy and consistency.
iii. Model Training: Apply machine learning algorithms to historical drought data to train models for predicting drought severity
iv. Prediction Generation: Produce timely and accurate drought severity forecasts.

4.2.2 Non-Functional Requirements

The system should adhere to the following non-functional criteria:

i. Performance: Ensure rapid data processing and prediction capabilities.
ii. Scalability: Accommodate increasing data volumes and user demands without degradation in performance.
iii. Reliability: Maintain high system uptime and robustness against failures.
iv. Usability: Provide an intuitive interface for users with varying levels of technical expertise.

4.3 Systems Architecture

The system architecture for the drought severity prediction model is designed to ensure scalability, efficiency, and reliability while processing large datasets and running machine learning algorithms. The architecture is composed of several key components, including data acquisition, data processing, model development, and user interface layers. At the data acquisition layer, users can input various datasets including remote sensing data, weather station data, and satellite imagery through a web interface. The data processing layer involves cleansing and changing raw data into structured formats suitable for analysis. The model development layer integrates machine learning algorithms, such as XGBoost and Random Forest, to generate drought predictions based on historical. Finally, the user interface layer presents the predictions and visualizations through an interactive dashboard, enabling decision-makers to assess drought severity and take appropriate action. Figure 4.1 shows the drought prediction Systems Architecture.

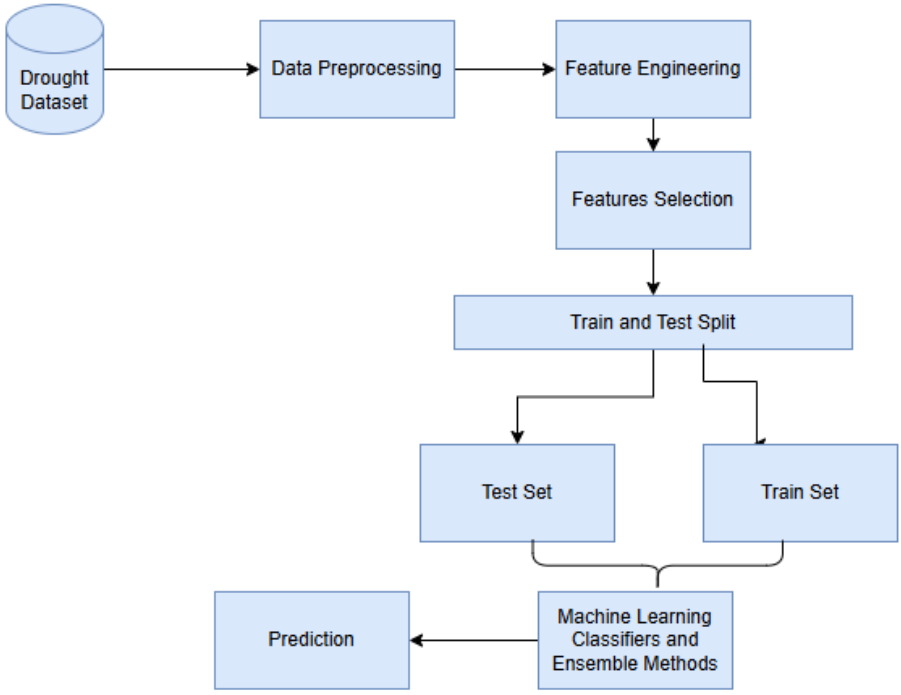


Figure 4.1: System Architecture

4.4 System Design

The system design for the drought severity prediction model is structured to ensure efficiency, scalability, and ease of use. It includes a modular architecture, with key components such as data acquisition, processing, model development, and visualization. Users will input datasets through an intuitive web interface, where the data is securely stored in a cloud-based database. The system is designed to process and cleanse the input data, making it suitable for machine learning models. Machine learning algorithms like XGBoost and Random Forest are used to generate accurate drought severity predictions. The final output, including predictions which are presented via a user-friendly dashboard, allowing decision-makers to easily interpret the results. The design ensures that the system can scale with increasing data volume and is flexible enough to incorporate additional features or data sources in the future.

4.4.1 Use Case Diagram

A Use Case Diagram visually represents the system's functionality from the perspective of the users and how they interact with the system. It provides a high-level overview of the system's key operations and identifies the primary use cases that the system supports. A Use Case Diagram for the drought prediction model in Kilifi County illustrates the interactions between

key actors and the system's core functionalities. The primary actors that is the residents of Kilifi access drought forecasts to make informed agricultural decisions. The system consists of several use cases, including data ingestion, data pre-processing, machine learning model training, real-time prediction generation, and prediction output as shown in Figure 4.2 Users interact with the system through a dashboard interface, where they can access predictive insights.

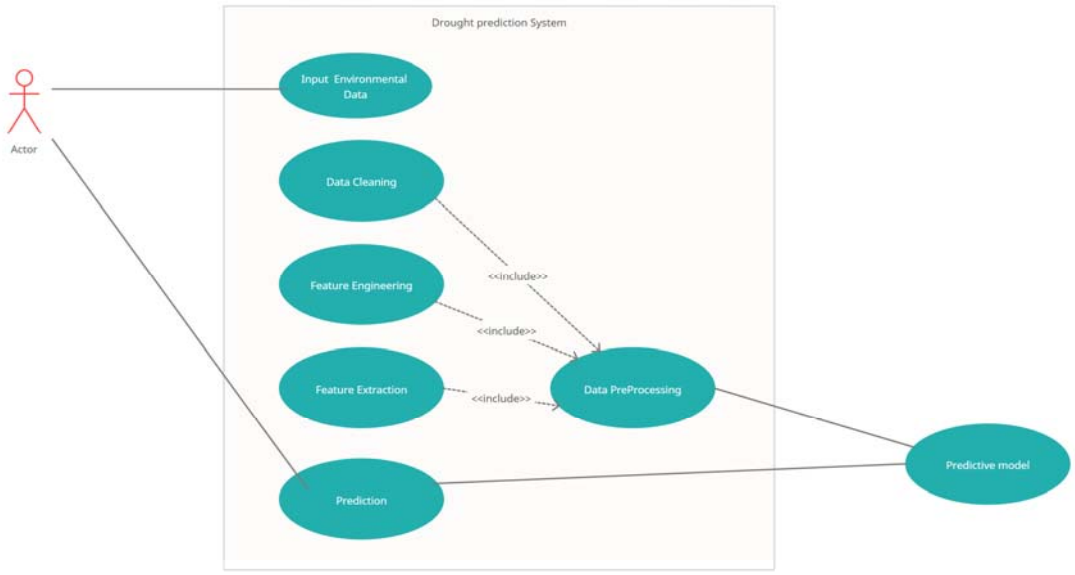


Figure 4.2: Use case diagram

4.4.2 Sequence Diagram

A Sequence Diagram is a Unified Modelling Language (UML) diagram that represents the flow of interactions between system components and users in a time-ordered manner. It visually depicts how different elements of the system communicate to achieve specific tasks. In the drought severity prediction model, the Sequence Diagram outlines the step-by-step process from the user inputting the data to the final output. The sequence begins when a user inputs dataset(s). The system then validates and preprocesses the data, ensuring it is cleaned and structured correctly. Once processed, the machine learning model is triggered to analyze the data and generate drought severity predictions. The results are then stored in the system database, and the user can access and visualize the predictions through an interactive dashboard. If needed, users can also export the predictions for further analysis or reporting.

This diagram helps in understanding the logical flow of operations within the system, ensuring a well-structured design that meets user expectations efficiently as shown in Figure 4.3.

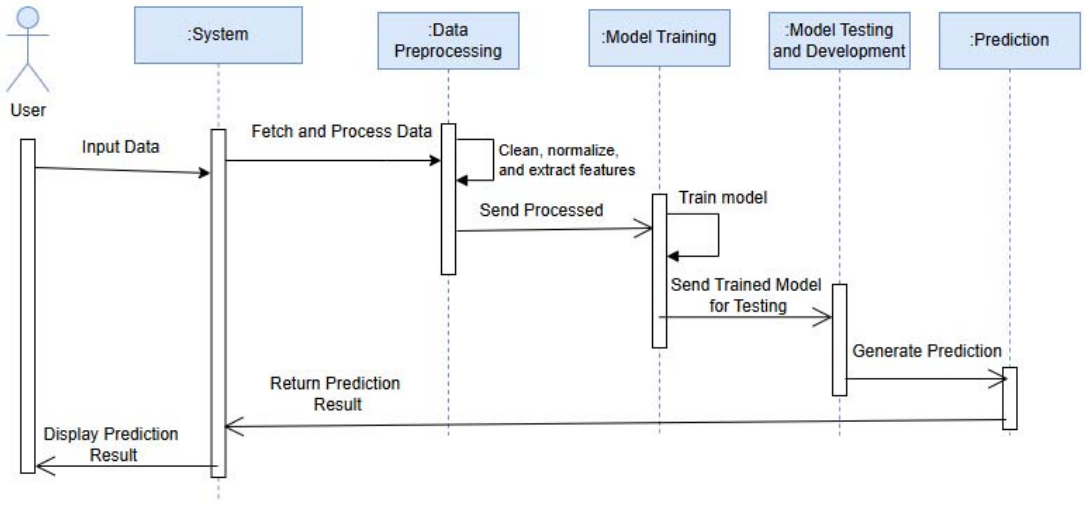


Figure 4.3: Sequence Diagram

VT OMNES VNVM SINT

Chapter 5: Systems Implementation and Testing

5.1 Introduction

Drought prediction is a critical component of climate risk management, particularly in regions like Kilifi County, Kenya, where agriculture is a primary livelihood and highly dependent on rainfall. Accurate and timely prediction of drought events can significantly enhance preparedness, mitigate adverse impacts, and support decision-making processes for farmers, policymakers, and other stakeholders. In this chapter, we delve into the implementation and testing of a robust predictive model designed to forecast drought occurrences in Kilifi County. The primary objective of this study was to develop a stacking ensemble model that leverages the combined strengths of various machine learning algorithms to enhance predictive accuracy and reliability.

This chapter outlines the systematic process of implementing and testing the stacking ensemble model. It begins with a detailed description of the data pre-processing steps, including collection of data, cleaning the data, and feature engineering, which are necessary for maintaining the relevance and quality of the input data. Subsequently, the chapter discusses the selection and configuration of base models, the training process, and the development of the meta-model. Rigorous testing and validation procedures are then employed to check the model's performance.

Furthermore, the chapter explores the problems faced during the implementation phase and the methods adopted to resolve them. These challenges include handling imbalanced datasets, managing computational complexity, and ensuring the model's generalizability to unseen data. The testing phase also involves comparing the stacking ensemble model's performance with that of individual base models and other traditional ensemble methods, such as Random Forest, Support Vector Machine, Extreme Gradient Boosting, and Light Gradient Boosting Machine to demonstrate its superiority.

5.2 Data Pre-processing

5.2.1 Data Cleaning

Data cleaning was an essential process in preparing the dataset for model development. Null values, which could introduce bias or errors in the analysis, were identified and dropped to ensure the integrity of the dataset. Additionally, outliers were examined to determine their influence on the model's performance. Box plots were used for visual representation of the

distribution of continuous variables, such as precipitation and soil moisture, and to identify extreme values that fell outside the interquartile range (IQR). As shown in Figure 5.1, the box plot for precipitation revealed several outliers, particularly in periods of unusually high or low rainfall. Similarly, Figure 5.2 illustrates the presence of outliers in soil moisture data, which could skew the model's predictions. To maintain the robustness of the dataset, these outliers were carefully analyzed and subsequently dropped. This step ensured that the dataset was free from anomalies that could negatively affect the training and performance of the stacking ensemble model.

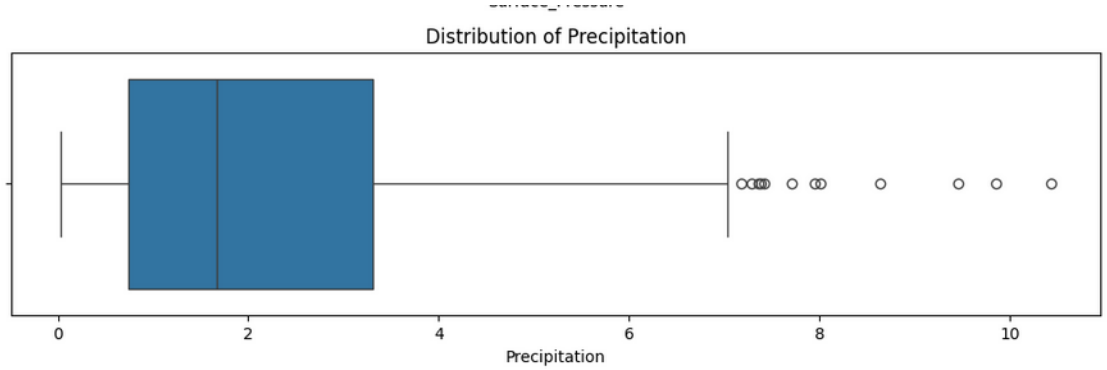


Figure 5.1: Distribution of Precipitation

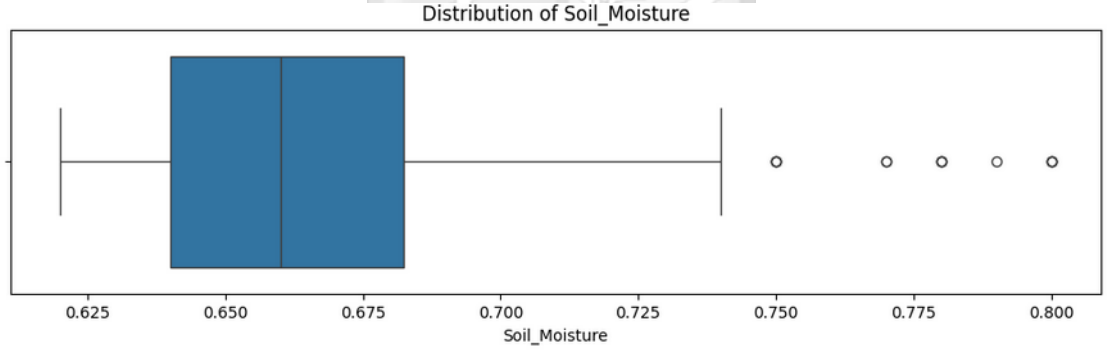


Figure 5.2: Distribution of Soil Moisture

5.2.2 Feature Engineering

This is the process of converting raw data into relevant features so as to improve the effectiveness of machine learning models. Feature engineering was applied to derive the Standardized Precipitation Index (SPI), a crucial variable for evaluating drought severity. Figure 5.3 illustrates the calculation process of SPI using precipitation data to assess drought conditions. Zero precipitation values were replaced with a small number (0.01) to avoid issues when fitting a Gamma distribution, which is commonly used to model precipitation. A rolling

sum of precipitation was computed over a specified scale to smooth short-term fluctuations. The Gamma distribution was fitted to the rolling precipitation data, and its Cumulative Distribution Function (CDF) was computed. The CDF values were then converted into Z-scores (SPI values) using the inverse normal distribution function, ensuring data standardization. The SPI values were integrated back into the dataset and classified into drought severity categories based on predefined thresholds, ranging from Severe Drought (SPI < -2) to Very Wet (SPI > 1.5). Finally, the updated dataset was saved as a CSV file for further analysis and visualization. The SPI metric is particularly valuable for drought monitoring because it accounts for long-term climate variability, allowing for more accurate comparisons across different regions and time periods. By engineering this feature, the model gains a standardized and interpretable measure of precipitation anomalies, aiding in effective drought prediction and decision-making for agricultural and water resource management.

[7]:
```python
#creating SPI
def calculate_spi(precipitation, scale=9):
    """
    Calculate the Standardized Precipitation Index (SPI) for a given precipitation time series.

    :param precipitation: Pandas Series of monthly precipitation values
    :param scale: SPI time scale (default is 3-month SPI)
    :return: Pandas Series of SPI values
    """
    precipitation = precipitation.replace(0, 0.01)  # Avoid zero values for Gamma distribution
    rolling_precip = precipitation.rolling(scale).sum().dropna()  # Compute rolling sum

    # Fit Gamma distribution
    shape, loc, scale_param = stats.gamma.fit(rolling_precip, floc=0)
    cdf_values = stats.gamma.cdf(rolling_precip, shape, loc=loc, scale=scale_param)  # Compute CDF

    # Convert CDF to SPI (Z-score)
    spi_values = stats.norm.ppf(cdf_values)

    return pd.Series(spi_values, index=rolling_precip.index)  # Return SPI values with correct index
# Compute SPI and integrate it back
data["SPI"] = calculate_spi(data["Precipitation"])
# Reset index and save the updated dataset
data.reset_index(inplace=True)
data.to_csv("updated_dataset_with_spi.csv", index=False)
```

[9]:
```python
# Classify drought severity based on SPI thresholds
data['Drought_Severity'] = pd.cut(
    data['SPI'],
    bins=[-float('inf'), -2, -1.5, -1, 1, 1.5, float('inf')],
    labels=['Severe Drought', 'Moderate Drought', 'Mild Drought',
            'Normal', 'Moderately Wet', 'Very Wet']
```

Figure 5.3:Feature Engineering

5.2.3 Data Splitting

The data was split using TimeSeriesSplit from the sklearn.model_selection module, this was done so as to maintain the temporal order of the data. This approach was well-suited for our study since we wanted to preserves the chronological order of observations, prevent data leakage and ensure the model is trained on past data while being tested on future data. The dataset was partitioned into five sequential splits (n_splits=5), forming a rolling window of training and testing sets. For each split, training and testing indices were generated, with the corresponding features (X) and target variable (y) extracted accordingly. Specifically, X_train and y_train were constructed using the training indices, while X_test and y_test were derived from the testing indices, as illustrated in Figure 5.4. This method enabled the model to be trained and validated across multiple time periods, enhancing its robustness and reliability in capturing temporal patterns within the data.

```
[61]: from sklearn.model_selection import TimeSeriesSplit

tscv = TimeSeriesSplit(n_splits=5)  # Number of splits

for train_index, test_index in tscv.split(X):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
```

Figure 5.4: Train and Test Data Split

5.2.3 Feature Extraction

To determine the most relevant features for predicting drought occurrence, a correlation matrix was created to examine relationships among the dataset variables. This matrix quantifies the linear association between feature pairs, with values ranging from -1 to 1. A value near 1 signifies a strong positive correlation, a value near -1 indicates a strong negative correlation, and a value close to 0 suggests minimal or no linear relationship.

To mitigate the issue of multicollinearity, which arises when two or more features in a dataset exhibit high correlation, potentially causing redundancy and instability in model performance, several variables were dropped based on their high correlation with other features, as indicated by the correlation matrix in Figure 5.5. Temporal variables such as Month and Year were removed because they showed weak correlations with most features. For instance, Year had correlations close to 0 with most variables, and Month had moderate correlations with some features like Evaporation Land and Humidity. Since temporal variables often do not contribute directly to the predictive power of the model and can introduce noise, they were excluded.

Wind Speed was also dropped due to its weak correlations with most features, except for a strong negative correlation with Dew Point (-0.69) and Temperature (-0.55). However, its overall contribution to explaining drought severity was minimal, as indicated by its near-zero correlation with Drought_Severity (-0.01).

Variables such as Soil Wetness, Humidity, and Evatranspiration were removed because they exhibited extremely high correlations with other moisture-related features. For example, Soil Wetness had strong correlations with Evaporation Land (0.96), Soil Moisture (0.70), and Humidity (0.97), while Humidity was highly correlated with Evaporation Land (0.92) and Soil Wetness (0.97). Similarly, Evatranspiration showed very high correlations with Evaporation Land (0.92) and Soil Wetness (0.84). Retaining these variables would have introduced redundancy, as their information was already captured by other strongly correlated features. Precipitation was also dropped because, despite having moderate correlations with Evaporation Land (0.78) and Soil Moisture (0.33), its correlation with Drought Severity was weak (0.18), and its information overlapped with other moisture-related features.

Additionally, Surface Pressure was excluded due to its strong negative correlations with Temperature (-0.80) and Dew Point (-0.70), despite its negligible correlation with Drought Severity (-0.00). Its inclusion would have added little predictive value while increasing model complexity. Finally, EVI (Enhanced Vegetation Index) was removed because it was highly correlated with NDVI (0.97), another vegetation index in the dataset. Since NDVI had a stronger correlation with Drought Severity (0.27) compared to EVI (0.25), EVI was dropped to avoid redundancy. By removing these variables, the dataset was streamlined to include only the most relevant and non-redundant features, such as Evaporation Land, Soil Moisture, Dew Point, Temperature, and NDVI. This step ensured that the model could focus on the most impactful predictors of drought severity while maintaining stability and interpretability.

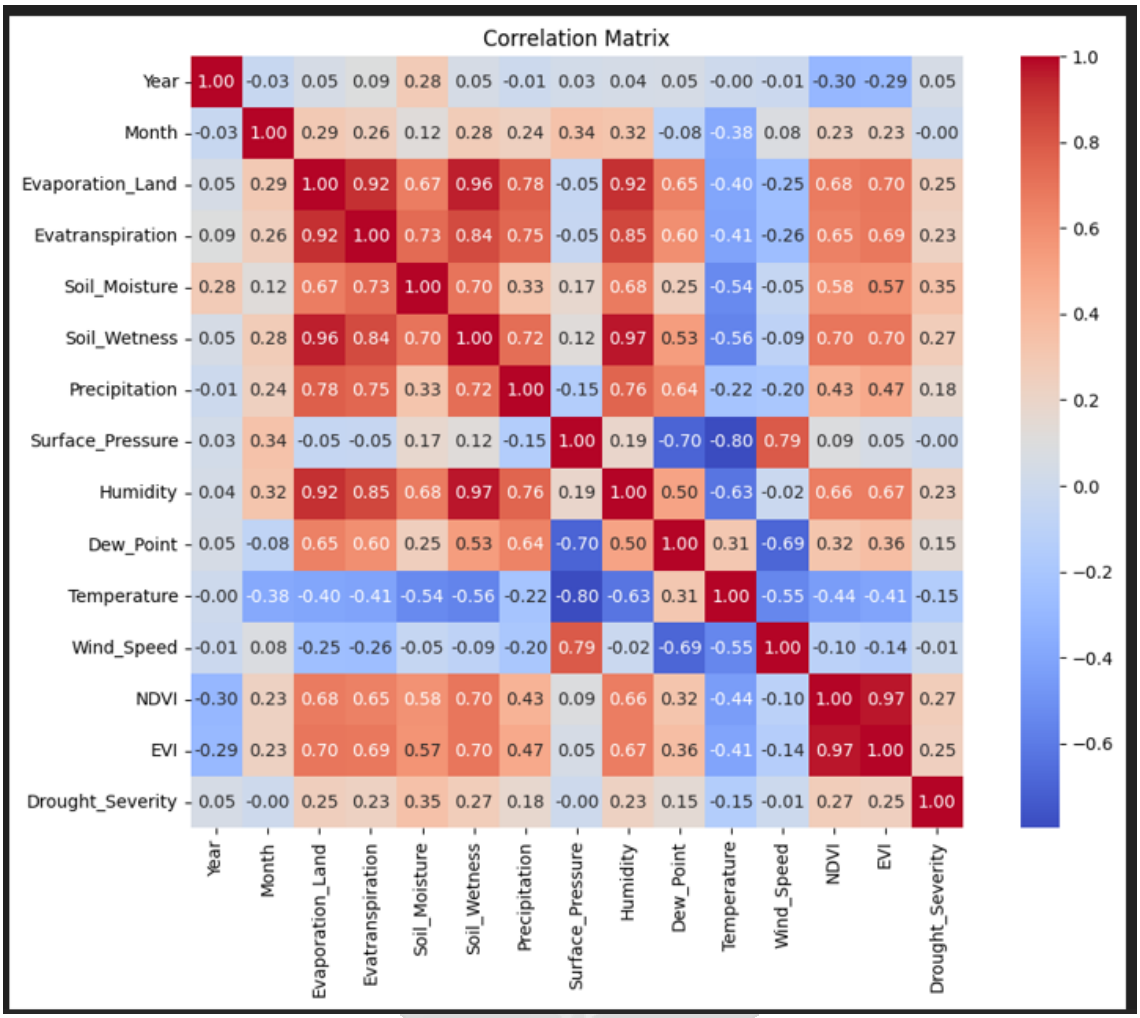


Figure 5.5: Feature Selection using Correlation Matrix

Figure 5.6 revealed several key insights about the relationships between the features which were selected. For instance, Evaporation Land showed a strong positive correlation with NDVI (0.68) and a moderate positive correlation with Soil Moisture (0.67) and Dew Point (0.65). This suggests that higher evaporation rates are associated with increased vegetation health (NDVI), soil moisture levels, and dew point temperatures. On the other hand, Evaporation Land exhibited a moderate negative correlation with Temperature (-0.40), indicating that higher temperatures may reduce evaporation rates in certain conditions.

Soil Moisture also demonstrated a strong negative correlation with Temperature (-0.54), implying that higher temperatures tend to reduce soil moisture levels. Additionally, Soil Moisture had a moderate positive correlation with NDVI (0.58), reinforcing the relationship between soil moisture and vegetation health. Dew Point, while positively correlated

with Evaporation Land (0.65) and NDVI (0.32), showed a weaker relationship with other variables, such as Temperature (0.31) and Drought Severity (0.15).

Temperature stood out as a feature with notable negative correlations, particularly with Soil Moisture (-0.54) and Evaporation Land (-0.40), highlighting its role in influencing moisture-related variables. NDVI, a measure of vegetation health, showed strong positive correlations with Evaporation Land (0.68) and Soil Moisture (0.58), as well as a moderate positive correlation with Drought Severity (0.27). This suggests that healthier vegetation is associated with higher evaporation rates, better soil moisture retention, and potentially lower drought severity.

Finally, Drought Severity exhibited weak to moderate correlations with most features, with the strongest positive correlation observed with Soil Moisture (0.35). This indicates that soil moisture levels are a key indicator of drought conditions, while other features like NDVI (0.27) and Evaporation Land (0.25) also contribute to understanding drought severity. Based on the correlation matrix, features such as Evaporation Land, Soil Moisture, NDVI, and Temperature were identified as the most influential predictors of drought occurrence due to their strong relationships with other variables and their potential impact on drought severity. These features were selected for inclusion in the model to ensure that it captures the most relevant patterns and relationships in the data.

VT OMNES VNVM SINT

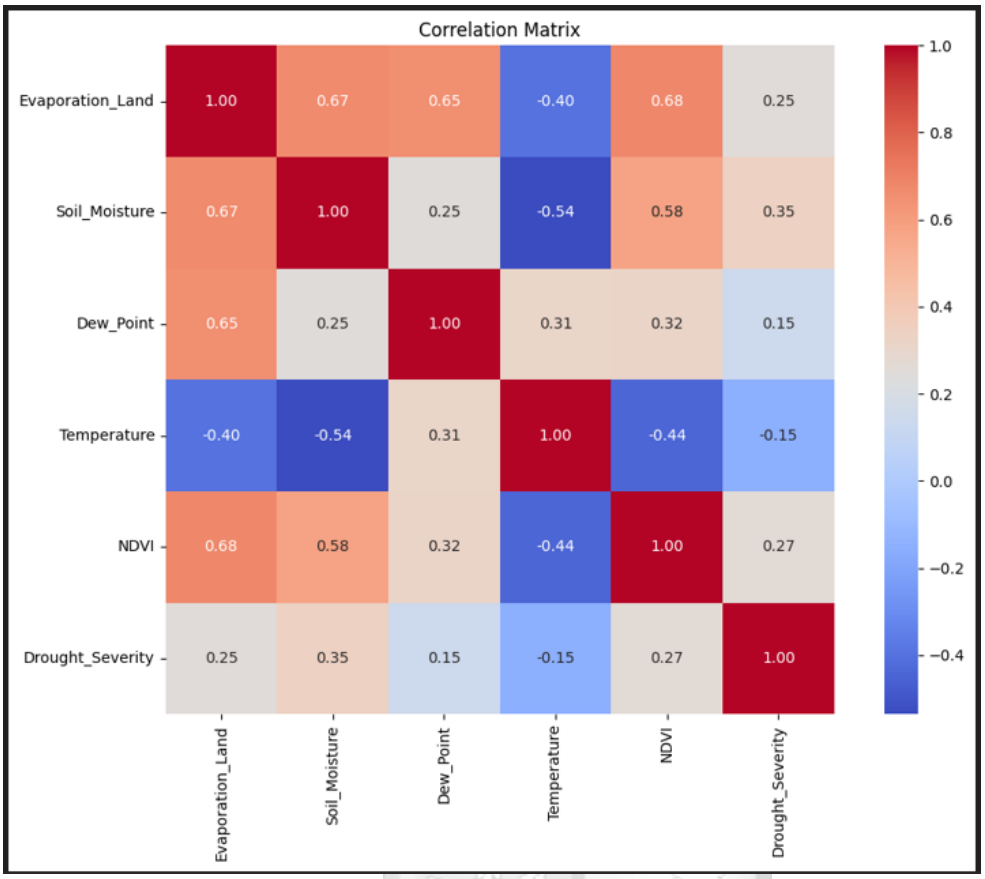


Figure 5.6: Correlation Matrix for Selected Features

5.2.4 Balancing Dataset

The dataset displayed a notable class imbalance, which could result in biased model performance, especially when predicting minority classes. To mitigate this, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE is an oversampling approach that enhances class balance by generating synthetic samples for minority classes through interpolation between existing instances. Before applying SMOTE, the training dataset consisted of 215 samples with 5 features as show in Figure 5.6. The class distribution was highly skewed, with class 3 being the majority (160 instances) and classes 0, 1, 2, and 4 having significantly fewer instances (18, 9, 15, and 7, respectively). This indicated that some classes, such as class 3, were overrepresented, while others, like classes 1, 4, and 5, had very few instances. To mitigate this imbalance, SMOTE was applied with a random_state of 5 and k_neighbors set to 2, ensuring reproducibility and controlling the number of nearest neighbors used to create synthetic samples. After applying SMOTE, the dataset was balanced, resulting in 960 samples with an equal number of instances across all classes. Specifically, the counts for each class were balanced to 160 instances, as demonstrated below. This step was crucial in

improving the model's ability to learn patterns from minority classes and enhancing its overall predictive performance.

```
[65]: #Fixing class imbalance
      #Upsampling using SMOTE
      sm = SMOTE(random_state = 5, k_neighbors=2)
      X_train_ures_SMOTE, y_train_ures_SMOTE = sm.fit_resample(X_train, y_train.ravel())

[67]: print('Before OverSampling, the shape of train_X: {}'.format(X_train.shape))
      print('Before OverSampling, the shape of train_y: {} \n'.format(y_train.shape))

      print('After OverSampling, the shape of train_X: {}'.format(X_train_ures_SMOTE.shape))
      print('After OverSampling, the shape of train_y: {} \n'.format(y_train_ures_SMOTE.shape))

      print("Counts of label '0' - Before Oversampling:{}, After OverSampling: {}".format(sum(y_train == 0),sum(y_train_ures_SMOTE == 0)))
      print("Counts of label '1' - Before Oversampling:{}, After OverSampling: {}".format(sum(y_train == 1),sum(y_train_ures_SMOTE == 1)))
      print("Counts of label '2' - Before Oversampling:{}, After OverSampling: {}".format(sum(y_train == 2),sum(y_train_ures_SMOTE == 2)))
      print("Counts of label '3' - Before Oversampling:{}, After OverSampling: {}".format(sum(y_train == 3),sum(y_train_ures_SMOTE == 3)))
      print("Counts of label '4' - Before Oversampling:{}, After OverSampling: {}".format(sum(y_train == 4),sum(y_train_ures_SMOTE == 4)))

      Before OverSampling, the shape of train_X: (215, 5)
      Before OverSampling, the shape of train_y: (215,)

      After OverSampling, the shape of train_X: (960, 5)
      After OverSampling, the shape of train_y: (960,)

      Counts of label '0' - Before Oversampling:18, After OverSampling: 160
      Counts of label '1' - Before Oversampling:9, After OverSampling: 160
      Counts of label '2' - Before Oversampling:15, After OverSampling: 160
      Counts of label '3' - Before Oversampling:160, After OverSampling: 160
      Counts of label '4' - Before Oversampling:7, After OverSampling: 160
```

Figure 5.7: Balancing Dataset Using SMOTE

5.3 Model Development

5.3.1 Selection of Base Models

The stacking ensemble model enhances predictive performance by combining multiple base models, leveraging their individual strengths. In this study, four diverse and robust machine learning algorithms were selected as base models: Random Forest, Support Vector Machine (SVM), XGBoost, and LightGBM. Each model was chosen for its distinct characteristics and ability to capture different patterns within the data.

The Random Forest model was initialized with 100 estimators and a maximum depth of 70, along with additional hyperparameters such as minimum samples split of 5 and minimum samples leaf of 2 to control overfitting. Random Forest is an ensemble method that builds multiple decision trees and aggregates their predictions, making it highly effective for handling complex, non-linear relationships in the data. Its robustness to outliers and ability to handle high-dimensional data made it a strong candidate for the ensemble

The Support Vector Machine (SVM) was configured with a polynomial kernel and a regularization parameter (C=1). SVMs are effective in identifying optimal hyperplanes for class separation in high-dimensional spaces, making them well-suited for classification tasks. The polynomial kernel enhances the model’s ability to capture non-linear relationships, while

setting probability=True enables the model to generate probability estimates, which are crucial for the stacking ensemble.

XGBoost (Extreme Gradient Boosting) was selected for its exceptional performance in structured data tasks. It was initialized with 200 estimators, a maximum depth of 70, and a learning rate of 0.05. Additional hyperparameters such as subsample=0.8, colsample_bytree=0.8, gamma=0.2, reg_lambda=10, and reg_alpha=5 were tuned to control overfitting and enhance generalization. XGBoost is a gradient-boosting algorithm that builds trees sequentially, correcting errors from previous trees, making it highly accurate and efficient.

Similarly, LightGBM (Light Gradient Boosting Machine) was included for its speed and efficiency, especially with large datasets. It was configured with 200 estimators, a maximum depth of 70, and a learning rate of 0.05. Parameters such as subsample=0.8, colsample_bytree=0.8, reg_lambda=10, and reg_alpha=5 were used to regularize the model and prevent overfitting. LightGBM uses a leaf-wise tree growth strategy, which often results in faster training times and better performance compared to traditional depth-wise growth.

The stacking ensemble integrates multiple base models, leveraging their individual strengths—such as Random Forest's robustness, SVM's capability to handle non-linear boundaries, XGBoost's high accuracy, and LightGBM's computational efficiency, as illustrated in Figure 5.8. This diverse combination enables the ensemble to capture a broad spectrum of patterns in the data, enhancing overall predictive performance and generalization ability.

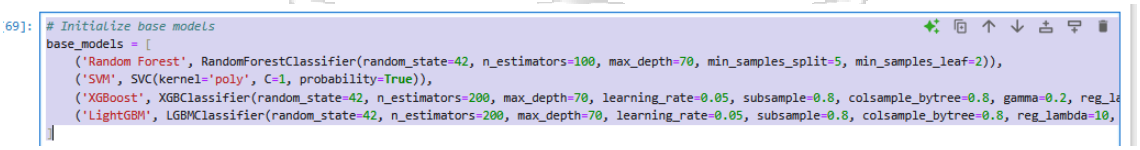


Figure 5.8: Base Models

5.3.2 Stacking Ensemble Model Development

The stacking ensemble model was created and trained using the StackingClassifier from the sklearn ensemble module. This approach combines the predictions of multiple base models (Random Forest, SVM, XGBoost, and LightGBM) into a single, more robust model by leveraging a meta-model, which in this case is Logistic Regression. The base models were selected for their diverse strengths and ability to capture different patterns in the data, while Logistic Regression was chosen as the meta-model due to its simplicity, interpretability, and effectiveness in combining predictions for multi-class classification tasks. The multi parameter

was used to handle the multiple drought severity classes, and the solver was selected for optimization, as it is well-suited for small to medium-sized datasets. To ensure convergence, the maximum iteration was set at1000, and a random state=42 was used for reproducibility.

The training process involved several key steps. First, the base models were trained on the SMOTE-balanced training dataset, which was pre-processed to address class imbalance. After training, each base model generated predictions (class probabilities) for the training dataset. These predictions were then combined into a new dataset, where each column represented the predictions from one base model. This new dataset served as the input features for the meta-model, while the original target variable (y_train_ures_SMOTE) remained unchanged. The meta-model (Logistic Regression) was trained on this dataset to learn the optimal way to combine the base models' predictions. To ensure robustness and reduce the risk of overfitting, 5-fold cross-validation was used during training. This means that the base models' predictions were generated using out-of-fold samples, ensuring that the meta-model was trained on predictions that the base models had not seen during their own training. The n_jobs=-1 parameter was used to enable parallel processing, speeding up the training process by utilizing all available CPU cores.

By combining the predictions of the base models using Logistic Regression as the meta-model, the stacking ensemble leverages the strengths of each individual model while mitigating their weaknesses. This approach ensures that the final model is robust, accurate, and well-suited for the complex task of predicting drought occurrence in Kilifi County. The use of cross-validation and parallel processing further enhances the model's reliability and efficiency, making it a powerful tool for drought prediction and risk management. This stacking process is illustrated in Figure 5.9.

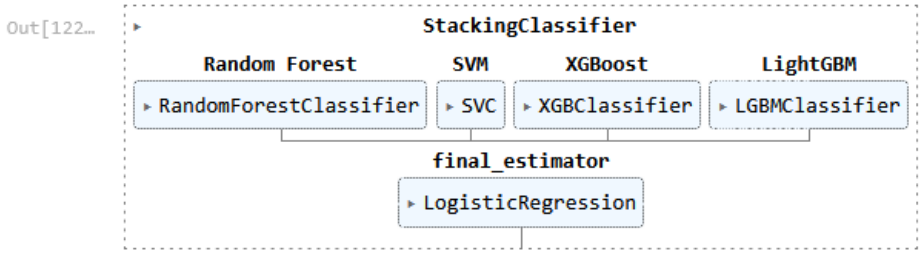


Figure 5.9: Stacking Model Development

5.4 Implementation

The implementation of the drought prediction model was carried out using Python as the primary programming language, chosen for its versatility, extensive libraries, and strong support for machine learning and data analysis tasks. The development environment used was Jupyter Notebook, a powerful and interactive tool that provides excellent support for Python development, including debugging, code execution, and visualization of results. Jupyter Notebook is particularly well-suited for data science and machine learning tasks, as it allows for the creation of documents that combine live code, equations, visualizations, and narrative text in a single interface. Its cell-based structure enables incremental development and testing, making it easier to experiment with code and visualize intermediate results. Key Python libraries and frameworks were utilized for various stages of the project, including Pandas and NumPy for data manipulation and numerical computations, Scipy for scientific computations such as calculating the Standardized Precipitation Index (SPI), Scikit-learn for implementing machine learning algorithms and evaluation metrics, XGBoost and LightGBM for advanced ensemble modelling, and Matplotlib and Seaborn for data visualization. Additionally, the Imbalanced-learn library was used to address class imbalance through techniques like SMOTE (Synthetic Minority Oversampling Technique). These tools collectively enabled efficient data pre-processing, model development, and evaluation.

Despite the availability of powerful tools and libraries, several technical and practical challenges were encountered during the implementation of the drought prediction model. One of the key challenges was feature engineering to generate the Standardized Precipitation Index (SPI), a critical feature for drought prediction. Calculating SPI involves fitting historical precipitation data to a probability distribution, typically the gamma distribution, and transforming it into a standardized index. This process required handling large time-series datasets and ensuring the accuracy of statistical computations. To address this, the Scipy library was used to fit the gamma distribution and compute the SPI, while a custom function was written to automate the calculation of SPI for 9-month time scale SPI.

Another significant challenge was the imbalanced dataset, where some drought severity classes had very few instances compared to others. This imbalance could lead to biased model performance, as the model might prioritize the majority class and ignore the minority classes. To address this issue, the SMOTE (Synthetic Minority Oversampling Technique) from the imbalanced-learn library was applied.

Hyperparameter tuning was another challenge, as selecting the optimal hyperparameters for the base models and the meta-model required extensive experimentation and validation. A combination of grid search and cross-validation was used to tune the hyperparameters, ensuring that the models were optimized for performance while avoiding overfitting.

5.5 Model Evaluation

During the testing and evaluation phase, the performance of each individual base model, along with the stacking ensemble, was assessed using key metrics such as accuracy, recall, precision, and F1-score. The Random Forest model achieved an accuracy of 0.6279, with a recall of 0.6279, precision of 0.7003, and an F1-score of 0.6071. This indicates its ability to capture complex patterns in the data while maintaining a reasonable balance between precision and recall.

In contrast, the Support Vector Machine (SVM) model exhibited lower performance, with an accuracy and recall of 0.2558, despite achieving a relatively high precision of 0.7557. The disparity between precision and recall resulted in a low F1-score of 0.2991, suggesting that the SVM model struggled to effectively handle the dataset's complexity.

The XGBoost model demonstrated moderate performance, attaining an accuracy and recall of 0.5581, along with a high precision of 0.9093, leading to an F1-score of 0.6036. Similarly, the LightGBM model produced comparable results, with an accuracy and recall of 0.5581, precision of 0.8397, and an F1-score of 0.5939. While both XGBoost and LightGBM exhibited strong precision, their relatively lower recall values indicate a trade-off between identifying true positives and minimizing false positives

The stacking ensemble model, which integrated the predictions of the base models using Logistic Regression as the meta-model, achieved superior performance compared to the individual models. It attained an accuracy of 0.6512, a recall of 0.6512, and a precision of 0.7632. Additionally, its F1-score of 0.6268 highlighted its effectiveness in maintaining a balance between precision and recall, establishing it as the best-performing model overall.

The enhanced performance of the stacking ensemble can be attributed to its ability to capitalize on the strengths of each base model while compensating for their individual limitations. By combining diverse algorithms, the ensemble model demonstrated improved generalization and robustness, making it particularly suitable for the complex task of drought prediction. These findings underscore the efficacy of the stacking ensemble approach in enhancing predictive

accuracy and highlight its potential for real-world applications in drought risk management. Figure 5.10 illustrates the accuracy of each model.

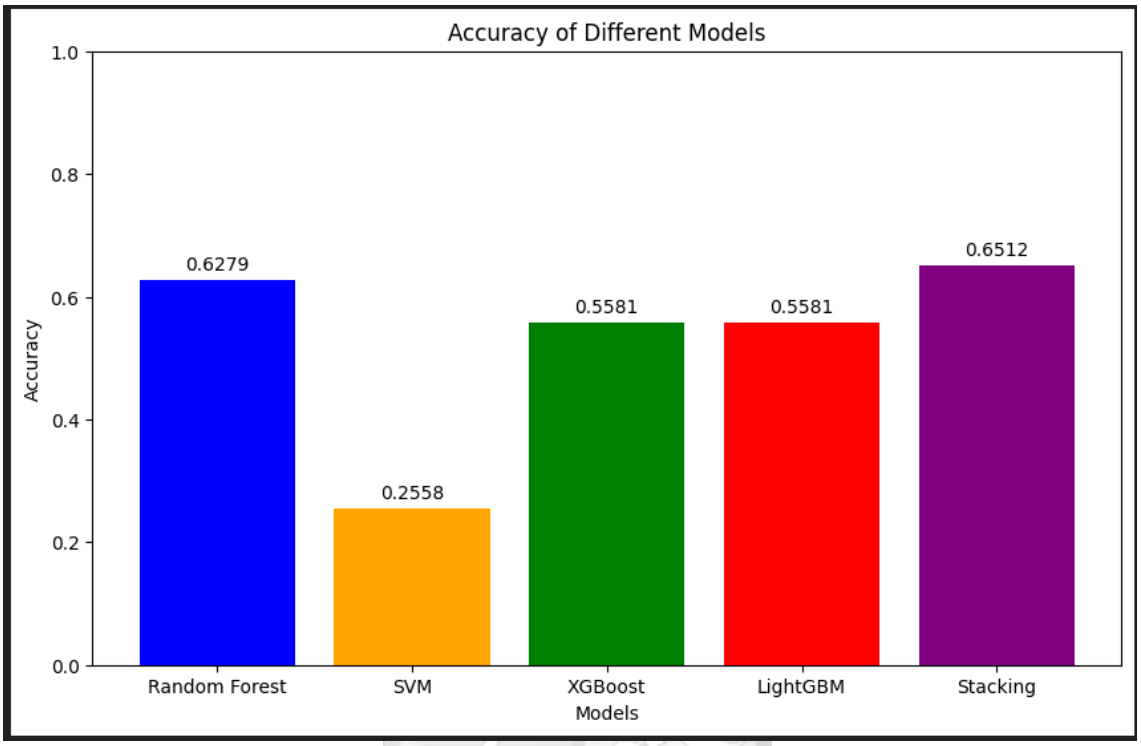


Figure 5.10: Accuracy of Base Models

VT OMNES VNVM SINT

Chapter 6: Discussion

6.1 Research Objectives Review

In the development of a stacking ensemble learning model, 4 base learners' model were trained including Random Forest, Support Vector Machine, Light Gradient Boosting Machine, and Extreme Gradient Boost. These models were trained on a time series dataset which was split using TimeSeriesSplit to maintain its chronological order. Unlike random splitting methods, TimeSeriesSplit ensures that training data always comes before test data to prevent data leakage. The dataset was divided into five sequential train-test sets, where each split expands the training set by including more past data while shifting the test set forward in time. This approach is crucial for time-dependent datasets, such as drought prediction, as it ensures that models are trained on past observations and tested on future ones, mimicking real-world predictive scenarios. This chapter explores the findings of the study, which focused on leveraging stacking ensemble learning model to predict drought severity in Kilifi county.

6.1.1 Explore the key predictors that significantly influence drought occurrence

The present study identified evaporation land, soil moisture, dew point, NDVI (Normalized Difference Vegetation Index), and temperature as the most influential factors affecting drought severity. These findings align with previous studies, which have consistently emphasized the importance of meteorological and remote sensing variables in drought assessment. For instance, soil moisture and temperature are widely recognized as critical indicators of drought conditions due to their direct impact on water availability and evapotranspiration rates. Similarly, NDVI, a measure of vegetation health derived from remote sensing data, has been shown to correlate strongly with drought severity, as it reflects the stress on vegetation caused by water scarcity. These findings align with previous studies but also highlight key differences in the role of meteorological and remote sensing variables in drought assessment.

Elbeltagi et al. (2024) primarily focused on using meteorological variables to develop a Palmer Drought Severity Index (PDSI) forecast. While their study did not explicitly outline the impact of individual features, the use of CNN-LSTM models suggests that meteorological factors such as temperature and precipitation played a critical role in model performance. In contrast, the current study expands on this approach by incorporating both meteorological and remote sensing variables, identifying evaporation land, soil moisture, dew point, NDVI (Normalized Difference Vegetation Index), and temperature as the most influential factors affecting drought severity. Unlike Elbeltagi et al. (2024), which relied heavily on meteorological data, this study

highlights the importance of integrating remote sensing data, particularly NDVI, to capture vegetation health and its relationship with drought conditions. Additionally, while Elbeltagi et al. (2024) utilized advanced deep learning models like CNN-LSTM, the current study employs a stacking ensemble model that combines multiple machine learning algorithms to improve predictive accuracy. This approach not only leverages the strengths of diverse models but also provides a more comprehensive understanding of the key predictors of drought severity. By comparing the two studies, it is evident that while meteorological variables remain crucial, the inclusion of remote sensing data and the use of ensemble modelling techniques offer a more holistic and accurate framework for drought prediction. This underscores the importance of integrating multiple data sources and methodologies to enhance drought assessment and mitigation strategies.

Ghazaryan et al. (2020) examined remote sensing indicators, including NDVI (Normalized Difference Vegetation Index), NDMI (Normalized Difference Moisture Index), Land Surface Temperature (LST), and SAR-based parameters, to assess the effects of drought on vegetation. Their study found that NDVI and NDMI provided substantial accuracy in measuring agricultural dryness, with NDVI being particularly effective in capturing vegetation health and stress. Additionally, Ghazaryan et al. (2020) identified LST as a strong predictor of drought stress in crops, as higher surface temperatures are often associated with reduced soil moisture and increased evapotranspiration, exacerbating drought conditions. The current study also highlights the importance of NDVI as a key predictor of drought severity, with a positive correlation of 0.27, reinforcing its role in assessing vegetation health and water stress. However, while Ghazaryan et al. (2020) focused primarily on remote sensing indicators, the current study integrates both meteorological and remote sensing variables, such as evaporation land, soil moisture, dew point, and temperature, to provide a more comprehensive understanding of drought dynamics. For instance, temperature in the current study showed a negative correlation (-0.15) with drought severity, aligning with the findings of Ghazaryan et al. (2020) regarding the impact of temperature on drought stress. Furthermore, the current study employs a stacking ensemble model that combines multiple machine learning algorithms, offering a more robust and accurate framework for drought prediction compared to the individual remote sensing indicators used by Ghazaryan et al. (2020). By integrating diverse data sources and methodologies, the current study not only corroborates the importance of remote sensing indicators like NDVI but also expands on their findings by incorporating additional meteorological factors and advanced modelling techniques. This approach enhances

the ability to predict and mitigate drought impacts, particularly in regions like Kilifi County, where both environmental and climatic factors play a significant role in drought occurrence.

6.1.2 Develop ensemble machine learning model

The development of ensemble models for drought prediction has been explored extensively in recent studies, each employing unique methodologies and classifiers. En-Nagre et al. (2024) applied ensemble learning techniques, including Random Forest and ensemble-based approaches like Voting Regressor and AdaBoost Regressor, to predict drought indices such as SPI and SPEI in the Upper Drâa Basin, Morocco. Their reliance on SPI and SPEI closely aligns with the current study, which also uses SPI as the primary drought index. However, while En-Nagre et al. (2024) focused on regression-based ensemble methods, the current study employs a stacking ensemble model that combines diverse classifiers such as Random Forest, SVM, XGBoost, and LightGBM, with Logistic Regression as the meta-model. This approach leverages the strengths of multiple algorithms to improve predictive accuracy, offering a more robust framework compared to the Voting and AdaBoost Regressors used by En-Nagre et al. (2024).

Lalika et al. (2024) employed five machine learning models; Long Short-Term Memory (LSTM), Multivariate Adaptive Regression Splines (MARS), Support Vector Machines (SVM), Extreme Learning Machine (ELM), and M5 Tree for drought prediction in Tanzania based on rainfall data. While their study incorporated SVM, similar to the current study, their emphasis on deep learning through LSTM differs significantly from the stacking ensemble approach used here. The current study avoids the computational complexity of deep learning models like LSTM and instead focuses on combining simpler, interpretable models to achieve comparable or superior performance. This makes the current approach more accessible and computationally efficient, particularly for regions with limited resources.

Varela and Hernández (2024) explored the use of Artificial Neural Networks (ANN), LSTM, and Support Vector Regression (SVR) to predict drought in Chihuahua, Mexico, utilizing SPEI at different temporal scales. Their methodology involved extensive hyperparameter tuning to optimize model performance, a practice also adopted in the current study. However, while Varela and Hernández (2024) relied heavily on deep learning models like ANN and LSTM, the current study employs a stacking ensemble that combines traditional machine learning models, offering a balance between performance and interpretability. The use of Logistic

Regression as the meta-model in the current study further enhances interpretability, making it easier to understand how predictions are derived.

Felsche and Ludwig (2021) developed a drought classification model using neural networks based on atmospheric and soil variables. Their approach involved a large ensemble of climate model outputs and explainable AI techniques to interpret predictions. While their study differs in its use of ANN and large-scale climate data, it shares similarities with the current study in integrating multiple inputs to improve drought forecasting. The current study also emphasizes the integration of diverse data sources, including meteorological and remote sensing variables, to enhance model performance. However, instead of relying on neural networks, the current study uses a stacking ensemble model, which is less computationally intensive and more interpretable.

Charles (2020) constructed homogeneous and heterogeneous model ensembles using ANN and SVR for drought prediction based on remote sensing and socio-economic data. Their approach followed an "over-produce then select" strategy, generating multiple models and selecting the best-performing ones. While this strategy is effective, the current study adopts a more streamlined approach by directly combining the predictions of multiple base models using a meta-model. This eliminates the need for extensive model selection and simplifies the ensemble development process.

While the studies by En-Nagre et al. (2024), Lalika et al. (2024), Varela and Hernández (2024), Felsche and Ludwig (2021), and Charles (2020) each contribute valuable insights into ensemble model development for drought prediction, the current study distinguishes itself by employing a stacking ensemble model that combines diverse classifiers (Random Forest, SVM, XGBoost, LightGBM) with Logistic Regression as the meta-model. This approach balances performance, interpretability, and computational efficiency, making it particularly suitable for drought prediction in resource-constrained regions like Kilifi County. By integrating multiple data sources and leveraging the strengths of various machine learning algorithms, the current study advances the field of drought prediction and offers a practical framework for real-world applications.

6.1.3 Evaluate the Performance of the Developed Model

The evaluation metrics of the models developed in this study provide valuable insights into their performance and highlight the effectiveness of the stacking ensemble approach compared

to individual base models. The Random Forest model achieved an accuracy of 62.79%, with a recall of 62.79%, precision of 70.03%, and an F1-score of 60.71%. These results indicate that Random Forest performs reasonably well in capturing the patterns in the data, with a relatively high precision suggesting its ability to minimize false positives. However, the recall and F1-score, while moderate, indicate room for improvement in correctly identifying true positives and balancing precision and recall. In comparison, the Support Vector Machine (SVM) model struggled significantly, with an accuracy of 25.58%, recall of 25.58%, and an F1-score of 29.91%, despite a high precision of 75.57%. This discrepancy between precision and recall suggests that the SVM model is less effective in handling the dataset's complexity and class imbalance, resulting in poor overall performance.

The XGBoost and LightGBM models demonstrated moderate performance, with both achieving an accuracy of 55.81% and recall of 55.81%. XGBoost showed a high precision of 90.93% and an F1-score of 60.36%, while LightGBM achieved a precision of 83.97% and an F1-score of 59.39%. These results indicate that both gradient boosting models are effective in minimizing false positives, as reflected in their high precision, but they struggle to achieve a balanced performance in terms of recall and F1-score. This suggests that while these models are strong in certain aspects, they may not fully capture the complexity of the dataset, particularly in identifying minority classes.

In contrast, the stacking ensemble model outperformed the individual base models, achieving an accuracy of 65.12%, recall of 65.12%, precision of 76.32%, and an F1-score of 62.68%. The stacking model's superior performance can be attributed to its ability to leverage the strengths of multiple base models while mitigating their individual weaknesses. By combining the predictions of Random Forest, SVM, XGBoost, and LightGBM using Logistic Regression as the meta-model, the stacking ensemble achieved a more balanced and robust performance across all evaluation metrics. This is particularly evident in the F1-score, which reflects a better balance between precision and recall compared to the individual models.

When compared to other studies, the stacking ensemble's performance is competitive and aligns with the findings of several recent works. For instance, Ghazaryan et al. (2020) achieved prediction accuracies ranging from 60% to 75% using remote sensing indicators like NDVI and NDMI, which is comparable to the stacking model's accuracy of 65.12.%. Similarly, En-Nagre et al. (2024) reported Nash-Sutcliffe Efficiency (NSE) values between 0.74 and 0.93 for ensemble models like Random Forest and AdaBoost Regressor, demonstrating the effectiveness of ensemble approaches in drought prediction. The stacking ensemble's

performance is also consistent with Lalika et al. (2024), who demonstrated the effectiveness of LSTM models with NSE values reaching 0.99, although their focus on deep learning differs from the current study's ensemble-based approach.

The stacking ensemble's performance is further supported by studies such as Varela and Hernández (2024), who explored the use of ANN, LSTM, and SVR to predict drought in Chihuahua, Mexico, achieving high R^2 values (0.8495 for SPEI 12 and 0.9218 for SPEI 24). While their study utilized deep learning models, the stacking ensemble's accuracy of 65.12% and F1-score of 62.68% demonstrate that simpler ensemble methods can also achieve competitive results. Additionally, Felsche and Ludwig (2021) developed a drought classification model using neural networks and explainable AI techniques, achieving prediction accuracies of 55%-57%. Although their approach differs from the current study, the stacking ensemble's higher accuracy highlights the potential of combining multiple models to improve predictive performance.

The findings of Charles (2020), who constructed homogeneous and heterogeneous model ensembles using ANN and SVR, further validate the effectiveness of ensemble methods. Their heterogeneous stacked ensemble achieved an R^2 of 0.94 and 80% classification accuracy, outperforming individual models. Similarly, Elbeltagi et al. (2024) demonstrated the effectiveness of hybrid machine-learning methods, with CNN-LSTM models achieving high NSE and R^2 values for PDSI forecasting. These studies collectively underscore the importance of ensemble approaches in improving drought prediction accuracy, as demonstrated by the stacking ensemble's performance in this study.

The stacking ensemble model developed in this study demonstrates superior performance compared to the individual base models, achieving a balanced and robust prediction capability. While the base models like Random Forest, XGBoost, and LightGBM showed moderate performance, the stacking approach effectively combined their strengths to improve accuracy, recall, precision, and F1-score. This aligns with the findings of other studies, such as Ghazaryan et al. (2020), En-Nagre et al. (2024), Lalika et al. (2024), and Varela and Hernández (2024), which emphasize the effectiveness of ensemble methods in drought prediction, particularly when integrating diverse data sources and methodologies. The stacking ensemble's performance underscores its potential as a reliable tool for drought prediction and highlights the importance of leveraging multiple models to enhance predictive accuracy and robustness.

6.2 Contribution of the Study

This study contributes to the existing body of literature on drought prediction by demonstrating the effectiveness of ensemble machine learning models in classifying drought severity using Standardized Precipitation Index (SPI). While previous studies have primarily relied on single machine learning models, such as Decision Trees or Support Vector Machines, this research highlights the advantages of integrating multiple classifiers, including Random Forest, XGBoost, and LightGBM, into a stacking ensemble framework. By evaluating the performance of these models based on multiple classification metrics, the study provides a comparative analysis that enhances understanding of how different algorithms respond to drought severity classification. Additionally, this research aligns with prior findings that boosting techniques are effective in handling imbalanced datasets, further validating their application in environmental data science. The study, therefore, serves as a reference point for future research seeking to refine drought prediction techniques using advanced machine learning approaches.

In the context of drought mitigation, this study provides valuable insights that can inform early warning systems and decision-making processes. Accurate classification of drought severity is crucial for policymakers, water resource managers, and agricultural stakeholders who rely on timely and reliable information to implement mitigation strategies. By leveraging ensemble learning techniques, the study presents a robust predictive framework that enhances drought monitoring capabilities, allowing for proactive measures to be taken before severe drought conditions escalate. Furthermore, the findings emphasize the importance of utilizing SPI-based models for drought prediction, reinforcing the relevance of precipitation-driven indicators in climate risk assessment. The application of machine learning in this domain not only improves forecasting accuracy but also contributes to the development of data-driven policies aimed at enhancing resilience against drought-related impacts.

6.3 Limitation of the Study

The current study, while demonstrating the effectiveness of a stacking ensemble model for drought prediction, has several limitations that warrant consideration. Firstly, the study relies heavily on historical meteorological and remote sensing data, which may not fully capture the dynamic and complex nature of drought conditions. Droughts are influenced by a multitude of factors, including anthropogenic activities, land-use changes, and socio-economic factors, which were not explicitly incorporated into the model. This omission may limit the model's

ability to account for human-induced changes in water availability and land management practices, which are increasingly significant in drought-prone regions like Kilifi County.

Secondly, the study's reliance on the Standardized Precipitation Index (SPI) as the primary drought indicator, while widely used, may not fully represent the multifaceted nature of drought. SPI is primarily based on precipitation data and does not account for other critical factors such as evapotranspiration, soil moisture, and vegetation health, which are essential for a comprehensive drought assessment. Although the study incorporates variables like NDVI and soil moisture, the integration of additional indices such as the Standardized Precipitation Evapotranspiration Index (SPEI) or the Palmer Drought Severity Index (PDSI) could provide a more holistic understanding of drought conditions.

Chapter 7: Conclusions and Recommendations

7.1 Conclusions

This study underscores the effectiveness of machine learning models in drought classification, demonstrating that ensemble approaches can enhance predictive performance. While individual models such as Random Forest and XGBoost performed well, the stacking ensemble model provided a balanced trade-off between precision and recall. The results indicate that integrating multiple climate indicators, including precipitation, temperature, NDVI, pressure, and wind speed, contributes to a more comprehensive drought assessment. These findings reinforce the significance of using advanced data-driven techniques to improve early warning systems, allowing for timely and informed decision-making in drought-prone regions.

Furthermore, the study highlights the challenges of multi-class drought classification, particularly in dealing with class imbalances and feature interactions. Despite its effectiveness, the stacking ensemble model did not yield substantial improvements over the strongest individual models, suggesting that further optimization and additional predictive variables could enhance model performance. The insights gained from this study provide a foundation for future research, encouraging the integration of more diverse datasets and the exploration of deep learning techniques to refine drought forecasting models.

7.2 Recommendations

Based on the limitations identified in the current study, several recommendations and future research directions can be proposed to enhance the robustness, generalizability, and applicability of drought prediction models. These recommendations aim to address the gaps and challenges encountered in this study while building on its strengths.

Future research should explore the integration of additional drought indices, such as the Standardized Precipitation Evapotranspiration Index (SPEI) and the Palmer Drought Severity Index (PDSI), to provide a more comprehensive assessment of drought conditions. These indices account for factors like evapotranspiration and soil moisture, which are critical for understanding the full spectrum of drought impacts. Additionally, incorporating socio-economic data, land-use changes, and anthropogenic factors (e.g., irrigation practices, groundwater extraction) could improve the model's ability to capture human-induced influences on drought dynamics.

While the stacking ensemble model demonstrated strong performance, future research could explore the use of advanced machine learning and deep learning techniques, such as Long

Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and hybrid models like CNN-LSTM. These models are particularly effective for capturing temporal and spatial dependencies in time-series data, which are crucial for drought prediction. Additionally, explainable AI techniques, such as SHapley Additive exPlanations (SHAP), could be used to interpret model predictions and identify the most influential variables.

To make the model more accessible and practical for real-world applications, future research should focus on optimizing computational efficiency. Techniques such as model pruning, quantization, and distributed computing could be employed to reduce the computational burden of training and deploying complex ensemble models. Additionally, exploring lightweight models or transfer learning approaches could enable the deployment of drought prediction systems in resource-constrained settings.

The current study focused on Kilifi County, which may limit the generalizability of the findings. Future research should validate the model in diverse geographic regions with varying climatic and environmental conditions. This would help assess the model's adaptability and robustness across different contexts, ensuring its applicability to a wider range of drought-prone areas.

To enhance the practical utility of the model, future research should explore the integration of real-time data streams from satellite sensors, weather stations, and IoT devices. This would enable the development of dynamic, real-time drought monitoring and early warning systems. Such systems could provide timely alerts to stakeholders, enabling proactive drought mitigation and resource management.

To ensure the model's practical implementation, future research should involve collaboration with stakeholders, including government agencies, agricultural organizations, and local communities. Engaging stakeholders in the model development process would help align the model's outputs with the needs of end-users and facilitate its adoption in real-world drought management practices.

The current study provides a strong foundation for drought prediction using a stacking ensemble model, but there is significant potential for further improvement. By incorporating additional drought indices, expanding datasets, exploring advanced machine learning techniques, and addressing computational challenges, future research can enhance the model's accuracy, generalizability, and applicability. Additionally, integrating real-time data, validating the model in diverse regions, and collaborating with stakeholders will ensure that

the model meets the needs of real-world drought management and early warning systems. These future research directions will contribute to more effective drought prediction and mitigation strategies, ultimately reducing the socio-economic and environmental impacts of droughts.

7.3 Future Research Direction

Future research should explore deep learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), to capture spatiotemporal dependencies in drought patterns. Investigating hybrid models that combine physical-based hydrological models with machine learning approaches could provide more robust predictions. Furthermore, expanding the study area to include multiple geographical regions with diverse climatic conditions would enhance the generalizability of the findings. Lastly, future studies should focus on explainability methods, such as SHAP values or LIME, to better understand model decisions and improve trust in AI-driven drought forecasting systems.

VT OMNES VNVM SINT

References

A Alshahrani, M., Laiq, M., Noor-ul-Amin, M., Yasmeen, U., & Nabi, M. (2024). A support vector machine-based drought index for regional drought analysis. Scientific Reports, 14(1), 9849.

Alahacoon, N., & Edirisinghe, M. (2022). A comprehensive assessment of remote sensing and traditional-based drought monitoring indices at global and regional scale. *Geomatics, Natural Hazards and Risk*, *13*(1), 762-799.

Alawsi, M. A., Zubaidi, S. L., Al-Bdairi, N. S. S., Al-Ansari, N., & Hashim, K. (2022). Drought forecasting: a review and assessment of the hybrid techniques and data pre-processing. *Hydrology*, *9*(7), 115.

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification a review. Qubahan Academic Journal, 1(2), 81-90.

Agarwal, S. (2020). Trade-offs between fairness, interpretability, and privacy in machine learning (Master's thesis, University of Waterloo).

Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52.

Aldrees, A., Awan, H. H., Javed, M. F., & Mohamed, A. M. (2022). Prediction of water quality indexes with ensemble learners: Bagging and boosting. Process Safety and Environmental Protection, 168, 344-361.

Ali, Y., Hussain, F., & Haque, M. M. (2024). Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. Accident Analysis & Prevention, 194, 107378.

Altunkaynak, A., & Jalilzadnezamabad, A. (2021). Extended lead time accurate forecasting of palmer drought severity index using hybrid wavelet-fuzzy and machine learning techniques. Journal of Hydrology, 601, 126619.

Alzamzami, F., Hoda, M., & El Saddik, A. (2020). Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation. IEEE access, 8, 101840-101858.

Bagheri, R., Ranjbar-Fordoei, A., Mousavi, S. H., & Tahmasebi, P. (2021). Assessment of MODIS-Derived NDVI and EVI for Different Vegetation Types in Arid Region: A Study in Sirjan Plain Catchment of Kerman province, Iran. Journal of Rangeland Science, 11(1), 54.

Bakır, R., Orak, C., & Yüksel, A. (2024). Optimizing hydrogen evolution prediction: A unified approach using random forests, lightGBM, and Bagging Regressor ensemble model. International Journal of Hydrogen Energy, 67, 101-110.

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal, 3, 100071.

Barbierato, E., & Gatti, A. (2024). The challenges of machine learning: A critical review. *Electronics*, *13*(2), 416.

Barrett, A. B., Duivenvoorden, S., Salakpi, E. E., Muthoka, J. M., Mwangi, J., Oliver, S., & Rowhani, P. (2020). Forecasting vegetation condition for drought early warning systems in pastoral communities in Kenya. *Remote Sensing of Environment*, *248*, 111886.

Cabitza, F., Campagner, A., Ferrari, D., Di Resta, C., Ceriotti, D., Sabetta, E., ... & Carobene, A. (2021). Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. Clinical Chemistry and Laboratory Medicine (CCLM), 59(2), 421-431.

Cao, M., Chen, M., Liu, J., & Liu, Y. (2022). Assessing the performance of satellite soil moisture on agricultural drought monitoring in the North China Plain. *Agricultural Water Management*, *263*, 107450.

Carroll, C. J., Slette, I. J., Griffin-Nolan, R. J., Baur, L. E., Hoffman, A. M., Denton, E. M., ... & Knapp, A. K. (2021). Is a drought a drought in grasslands? Productivity responses to different types of drought. *Oecologia*, 1-10.

Citakoglu, H., & Coşkun, Ö. (2022). Comparison of hybrid machine learning methods for the prediction of short-term meteorological droughts of Sakarya Meteorological Station in Turkey. *Environmental Science and Pollution Research*, *29*(50), 75487-75511.

Cullen, K. A. (2023). A review of applications of remote sensing for drought studies in the Andes region. *Journal of Hydrology: Regional Studies*, *49*, 101483.

Danandeh Mehr, A., Rikhtehgar Ghiasi, A., Yaseen, Z. M., Sorman, A. U., & Abualigah, L. (2023). A novel intelligent deep learning predictive model for meteorological drought forecasting. *Journal of Ambient Intelligence and Humanized Computing*, *14*(8), 10441-10455.

Docheshmeh Gorgij, A., Alizamir, M., Kisi, O., & Elshafie, A. (2022). Drought modelling by standard precipitation index (SPI) in a semi-arid climate using deep learning method: long short-term memory. *Neural Computing and Applications*, 1-18.

Elbeltagi, A., Srivastava, A., Ehsan, M., Sharma, G., Yu, J., Khadke, L., ... & Jinsong, D. (2024). Advanced stacked integration method for forecasting long-term drought severity: CNN with machine learning models. Journal of Hydrology: Regional Studies, 53, 101759.

En-Nagre, K., Aqnouy, M., Ouarka, A., Naqvi, S. A. A., Bouizrou, I., El Messari, J. E. S., ... & El-Askary, H. (2024). Assessment and prediction of meteorological drought using machine learning algorithms and climate data. *Climate Risk Management*, *45*, 100630.

Felegari, S., Sharifi, A., Moravej, K., Golchin, A., & Tariq, A. (2022). Investigation of the relationship between ndvi index, soil moisture, and precipitation data using satellite images. *Sustainable Agriculture Systems and Technologies*, 314-325.

Felsche, E., & Ludwig, R. (2021). Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations. *Natural Hazards and Earth System Sciences*, *21*(12), 3679-3691.

Feng, P., Wang, B., Luo, J. J., Li Liu, D., Waters, C., Ji, F., ... & Yu, Q. (2020). Using large-scale climate drivers to forecast meteorological drought condition in growing season across the Australian wheatbelt. *Science of the Total Environment*, *724*, 138162.

Ghazaryan, G., Dubovyk, O., Graw, V., Kussul, N., & Schellberg, J. (2020). Local-scale agricultural drought monitoring with satellite-based multi-sensor time-series. GIScience & Remote Sensing, 57(5), 704-718.

GladShiya, V. B., & Sharmila, K. K. (2024). Using Ensemble Learning and Random Forest Techniques to Solve Complex Problems. In Machine Learning Algorithms Using Scikit and TensorFlow Environments (pp. 388-407). IGI Global.

VT OMNES VNVM SINT

Glantz, M. H., & Ramirez, I. J. (2020). Reviewing the Oceanic Niño Index (ONI) to enhance societal readiness for El Niño's impacts. *International Journal of Disaster Risk Science*, *11*, 394-403.

Gu, J., Liu, S., Zhou, Z., Chalov, S. R., & Zhuang, Q. (2022). A stacking ensemble learning model for monthly rainfall prediction in the Taihu Basin, China. *Water*, *14*(3), 492.

Gyaneshwar, A., Mishra, A., Chadha, U., Raj Vincent, P. D., Rajinikanth, V., Pattukandan Ganapathy, G., & Srinivasan, K. (2023). A contemporary review on deep learning models for drought prediction. Sustainability, 15(7), 6160.

Han, J., & Singh, V. P. (2023). A review of widely used drought indices and the challenges of drought assessment under climate change. *Environmental Monitoring and Assessment*, *195*(12), 1438.

Hateren, T. C. V., Sutanto, S. J., & Lanen, H. A. J. V. (2019). Evaluating skill and robustness of seasonal meteorological and hydrological drought forecasts at the catchment scale-case Catalonia (Spain). *Environment International*, **133**, 105206. **https://doi.org/10.1016/j.envint.2019.105206**

Hayes, M. J., Svoboda, M. D., Wardlow, B. D., Anderson, M. C., & Kogan, F. (2012). Drought monitoring: Historical and current perspectives.

Hu, T., Renzullo, L. J., van Dijk, A. I., He, J., Tian, S., Xu, Z., ... & Liu, Q. (2020). Monitoring agricultural drought in Australia using MTSAT-2 land surface temperature retrievals. Remote Sensing of Environment, 236, 111419.

Huynh-Cam, T. T., Chen, L. S., & Le, H. (2021). Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance. Algorithms, 14(11), 318.

International Rescue Committee. (2023, February 21). *Severe drought is projected to leave about 5.4 million people in Kenya without adequate access to food and water between March and June 2023*. https://www.rescue.org/press-release/irc-severe-drought-projected-leave-about-54-million-people-kenya-without-adequate

International Union for Conservation of Nature. (2024, October 11). *Empowering communities in inclusive climate action in Kilifi seascape*. IUCN. https://iucn.org/story/202410/empowering-communities-inclusive-climate-action-kilifi-seascape

Jiang, X., Li, X., Wang, Q., Song, Q., Liu, J., & Zhu, Z. (2024). Multi-sensor data fusion-enabled semi-supervised optimal temperature-guided PCL framework for machinery fault diagnosis. Information Fusion, 101, 102005.

Jiang, Z., Hsu, Y. J., Yuan, L., Cheng, S., Feng, W., Tang, M., & Yang, X. (2022). Insights into hydrological drought characteristics using GNSS-inferred large-scale terrestrial water storage deficits. Earth and Planetary Science Letters, 578, 117294.

Kallis, G. (2008). Droughts. *Annual review of environment and resources*, *33*(1), 85-118.

Khan, N., Sachindra, D. A., Shahid, S., Ahmed, K., Shiru, M. S., & Nawaz, N. (2020). Prediction of droughts over Pakistan using machine learning algorithms. Advances in Water Resources, 139, 103562.

Khan, A. A., Chaudhari, O., & Chandra, R. (2023). A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. *Expert Systems with Applications*, 122778.

Kiangala, S. K., & Wang, Z. (2021). An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. Machine Learning with Applications, 4, 100024.

Kikon, A., & Deka, P. C. (2022). Forecasting of Meteorological Drought Using Machine Learning Algorithm. In *Advanced Modelling and Innovations in Water Resources Engineering: Select Proceedings of AMIWRE 2021* (pp. 43-52). Springer Singapore.

Kisi, O., Gorgij, A. D., Zounemat-Kermani, M., Mahdavi-Meymand, A., & Kim, S. (2019). Drought forecasting using novel heuristic methods in a semi-arid environment. *Journal of Hydrology*, *578*, 124053.

Kithi, M. (2024, October 21). *Thousands face starvation in Kilifi as drought situation worsens*. The Standard. https://www.standardmedia.co.ke/coast/article/2001504925/thousands-face-starvation-in-kilifi-as-drought-situation-worsens

Konapala, G., Mishra, A. K., Wada, Y., & Mann, M. E. (2020). Climate change will affect global water availability through compounding changes in seasonal precipitation and evaporation. *Nature communications*, *11*(1), 3044.

Lalika, C., Mujahid, A. U. H., James, M., & Lalika, M. C. (2024). Machine learning algorithms for the prediction of drought conditions in the Wami River sub-catchment, Tanzania. Journal of Hydrology: Regional Studies, 53, 101794.

Latif, S. D., Hazrin, N. A. B., Koo, C. H., Ng, J. L., Chaplot, B., Huang, Y. F., ... & Ahmed, A. N. (2023). Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches. Alexandria Engineering Journal, 82, 16-25.

Leng, Z., Chen, L., Yang, B., Li, S., & Yi, B. (2024). An extreme forecast index-driven runoff prediction approach using stacking ensemble learning. *Geomatics, Natural Hazards and Risk*, *15*(1), 2353144.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. International Journal of Applied Earth Observation and Geoinformation, 112, 102926.

Liu, C., Yang, C., Yang, Q., & Wang, J. (2021). Spatiotemporal drought analysis by the standardized precipitation index (SPI) and standardized precipitation evapotranspiration index (SPEI) in Sichuan Province, China. Scientific reports, 11(1), 1280.

Liu, Q., Zhang, S., Zhang, H., Bai, Y., & Zhang, J. (2020). Monitoring drought using composite drought indices based on remote sensing. *Science of the total environment*, *711*, 134585.

Lu, M., Hou, Q., Qin, S., Zhou, L., Hua, D., Wang, X., & Cheng, L. (2023). A stacking ensemble model of various machine learning models for daily runoff forecasting. *Water*, *15*(7), 1265.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., ... & Flach, P. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, *33*(8), 3048-3061.

Mathende, M., Ndlovu, B., Dube, S., Muduva, M., & Kiwa, F. (2024, May). AI-based DROUGHT FORECASTING FOR PARAMETRIC INSURANCE. In *5th South American Industrial Engineering and Operations Management Conference*, https://doi.org/10.46254/SA05.20240194.

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. IEEE Access.

Mirzabaev, A., Kerr, R. B., Hasegawa, T., Pradhan, P., Wreford, A., von der Pahlen, M. C. T., & Gurney-Smith, H. (2023). Severe climate change risks to food security and nutrition. *Climate Risk Management*, *39*, 100473.

Mohammadrezaei, M., Soltani, S., & Modarres, R. (2020). Evaluating the effect of ocean-atmospheric indices on drought in Iran. *Theoretical and Applied Climatology*, *140*, 219-230.

Mokhtar, A., Jalali, M., He, H., Al-Ansari, N., Elbeltagi, A., Alsafadi, K., ... & Rodrigo-Comino, J. (2021). Estimation of SPEI meteorological drought using machine learning algorithms. *IEEe Access*, *9*, 65503-65523.

Mukhawana, M. B., Kanyerere, T., & Kahler, D. (2023). Review of in-Situ and remote sensing-based indices and their Applicability for integrated drought monitoring in South Africa. *Water*, *15*(2), 240.

Musonda, B., Jing, Y., Iyakaremye, V., & Ojara, M. (2020). Analysis of long-term variations of drought characteristics using standardized precipitation index over Zambia. *Atmosphere*, *11*(12), 1268.

Myronidis, D., Ioannou, K., Fotakis, D., & Dörflinger, G. (2018). Streamflow and hydrological drought trend analysis and forecasting in Cyprus. *Water resources management*, *32*, 1759-1776.

Nabavi-Pelesaraei, A., Saber, Z., Mostashari-Rad, F., Ghasemi-Mobtaker, H., & Chau, K. W. (2021). Coupled life cycle assessment and data envelopment analysis to optimize energy consumption and mitigate environmental impacts in agricultural production. In *Methods in sustainability science* (pp. 227-264). Elsevier.

Nandgude, N., Singh, T. P., Nandgude, S., & Tiwari, M. (2023). Drought prediction: a comprehensive review of different drought prediction models and adopted technologies. *Sustainability*, *15*(15), 11684.

National Drought Management Authority. (2024, July 16). *Kilifi County: Drought early warning bulletin for June 2024*. ReliefWeb. https://reliefweb.int/report/kenya/kilifi-county-drought-early-warning-bulletin-june-2024

Ndayiragije, J. M., & Li, F. (2022). Effectiveness of drought indices in the assessment of different types of droughts, managing and mitigating their effects. *Climate*, *10*(9), 125.

Nguyen, M. H., & Ly, H. B. (2023). Development of machine learning methods to predict the compressive strength of fiber-reinforced self-compacting concrete and sensitivity analysis. Construction and Building Materials, 367, 130339.

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., ... & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*(1), 4832864.

Odegua, R. (2019, August). An empirical study of ensemble techniques (bagging, boosting and stacking). In *Proc. Conf.: Deep Learn. IndabaXAt*.

Okal, H. A., Ngetich, F. K., & Okeyo, J. M. (2020). Spatio-temporal characterisation of droughts using selected indices in Upper Tana River watershed, Kenya. *Scientific African*, *7*, e00275.

Oruc, S., Tugrul, T., & Hinis, M. A. (2024). Beyond Traditional Metrics: Exploring the Potential of Hybrid Algorithms for Drought Characterization and Prediction in the Tromso Region, Norway. *Applied Sciences*, *14*(17), 7813.

Otieno, C. (2020). *Model ensembles for Predictive drought severity and drought Effects monitoring using remote sensing & Socio-economic data* (Doctoral dissertation, University of Nairobi).

Panigrahi, B., Kathala, K. C. R., & Sujatha, M. (2023). A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. Procedia Computer Science, 218, 2684-2693.

Piri, J., Abdolahipour, M., & Keshtegar, B. (2023). Advanced machine learning model for prediction of drought indices using hybrid SVR-RSM. *Water Resources Management*, *37*(2), 683-712.

Prodhan, F. A., Zhang, J., Hasan, S. S., Sharma, T. P. P., & Mohana, H. P. (2022). A review of machine learning methods for drought hazard monitoring and forecasting: Current research trends, challenges, and future research directions. Environmental modelling & software, 149, 105327.

Pujar, G. S., Taori, A., Chakraborty, A., & Mitran, T. (2024). Sensing climate change through earth observations: perspectives at global and National Level. In *Digital agriculture: A solution for sustainable food and nutritional security* (pp. 225-280). Cham: Springer International Publishing.

Rafiei Sardooi, E., Azareh, A., Eskandari Damaneh, H., & Skandari Damaneh, H. (2021). Drought Monitoring Using MODIS Land Surface Temperature and Normalized Difference Vegetation Index Products in Semi-Arid Areas of Iran. Journal of Rangeland Science, 11(4), 402-418.

Rahmatinejad, Z., Reihani, H., Tohidinezhad, F., Rahmatinejad, F., Peyravi, S., Pourmand, A., ... & Eslami, S. (2019). Predictive performance of the SOFA and mSOFA scoring systems for predicting in-hospital mortality in the emergency department. *The American journal of emergency medicine*, *37*(7), 1237-1241.

Rezaiy, R., & Shabri, A. (2023). Drought forecasting using W-ARIMA model with standardized precipitation index. *Journal of Water and Climate Change*, *14*(9), 3345-3367.

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences, 2(7), 1308.

Sajid, S. W., Hasan, M., Rabbi, M. F., & Abedin, M. Z. (2023). An ensemble LGBM (light gradient boosting machine) approach for crude oil price prediction. In Novel Financial Applications of Machine Learning and Deep Learning: Algorithms, Product Modelling, and Applications (pp. 153-165). Cham: Springer International Publishing.

Saha, K., Guha, A., & Banik, T. (2021). Indian summer monsoon variability over North-East India: Impact of ENSO and IOD. *Journal of Atmospheric and Solar-Terrestrial Physics*, *221*, 105705.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

Senhorelo, A. P., Sousa, E. F. D., Santos, A. R. D., Ferrari, J. L., Peluzio, J. B. E., Carvalho, R. D. C. F., ... & Moreira, T. R. (2024). Application of Path Analysis and Remote Sensing to Assess the Interrelationships between Meteorological Variables and

Vegetation Indices in the State of Espírito Santo, Southeastern Brazil. Diversity, 16(2), 90.

Sharafati, A., Haji Seyed Asadollah, S. B., Motta, D., & Yaseen, Z. M. (2020). Application of newly developed ensemble machine learning models for daily suspended sediment load prediction and related uncertainty analysis. *Hydrological Sciences Journal*, *65*(12), 2022-2042.

Sharma, N. N., Tapas, M. R., & Kumar, A. U. (2022). Drought Monitoring Indices. *Meteorology and Climatology*, *53*.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.

Sundararajan, K., Garg, L., Srinivasan, K., Bashir, A. K., Kaliappan, J., Ganapathy, G. P., ... & Meena, T. (2021). A contemporary review on drought modeling using machine learning approaches. Computer Modeling in Engineering & Sciences, 128(2), 447-487.

Taiwo, B. E., Kafy, A. A., Samuel, A. A., Rahaman, Z. A., Ayowole, O. E., Shahrier, M., ... & Abosede, O. O. (2023). Monitoring and predicting the influences of land use/land cover change on cropland characteristics and drought severity using remote sensing techniques. Environmental and Sustainability Indicators, 18, 100248.

Vadhnani, S., & Singh, N. (2022). Brain tumor segmentation and classification in MRI using SVM and its variants: a survey. Multimedia Tools and Applications, 81(22), 31631-31656.

Vance, L., Thangavelautham, J., & Fernández, J. M. (2021, March). Top level systems requirements analysis for a ground cooperative orbital debris avoidance system. In *2021 IEEE Aerospace Conference (50100)* (pp. 1-6). IEEE.

Vanegas, C. E. D., Mejía, J. C. G., Agudelo, F. A. V., & Duran, D. E. S. (2023). A Representation Based on Essence for the CRISP-DM Methodology. *Computación y Sistemas (CyS)*, *27*(3).

Varela, J. A. M., & Hernández, J. I. R. (2024). Prediction of hydrological drought by the Standardized Precipitation Evapotranspiration Index in Chihuahua, Mexico, using machine learning algorithms. *Atmósfera*, *38*, 687-697.

Vreugdenhil, M., Greimeister-Pfeil, I., Preimesberger, W., Camici, S., Dorigo, W., Enenkel, M., ... & Wagner, W. (2022). Microwave remote sensing for agricultural drought monitoring: Recent developments and challenges. *Frontiers in Water*, *4*, 1045451.

World Health Organization. (n.d.). Drought. Retrieved July 25, 2024, from https://www.who.int/health-topics/drought#tab=tab_1

World Weather Attribution. (2023, April 27). *Human-induced climate change increased drought severity in Horn of Africa*. https://www.worldweatherattribution.org/human-induced-climate-change-increased-drought-severity-in-southern-horn-of-africa/

World Vision International. (2024, April 2). *Women groups in rural villages of Kilifi County saving for the future*. https://www.wvi.org/stories/global-hunger-crisis/women-groups-rural-villages-kilifi-county-saving-future

Wehbe, Y., & Temimi, M. (2021). A remote sensing-based assessment of water resources in the Arabian Peninsula. Remote Sensing, 13(2), 247.

Xie, Y., Li, C., Li, M., Liu, F., & Taukenova, M. (2023). An overview of deterministic and probabilistic forecasting methods of wind energy. *Iscience*, *26*(1).

Xu, X., Chen, F., Wang, B., Harrison, M. T., Chen, Y., Liu, K., ... & Hu, K. (2024). Unleashing the power of machine learning and remote sensing for robust seasonal drought monitoring: A stacking ensemble approach. *Journal of Hydrology*, *634*, 131102.

Yang, T. H., & Liu, W. C. (2020). A general overview of the risk-reduction strategies for floods and droughts. *Sustainability*, *12*(7), 2687.

Zellou, B., El Moçayd, N., & Bergou, E. H. (2023). Towards improved drought prediction in the Mediterranean region–modeling approaches and future directions. *Natural Hazards and Earth System Sciences*, *23*(11), 3543-3583.

Zheng, X., Sarwar, A., Islam, F., Majid, A., Tariq, A., Ali, M., ... & Soufan, W. (2023). Rainwater harvesting for agriculture development using multi-influence factor and fuzzy overlay techniques. *Environmental Research*, *238*, 117189.

Zhou, S., Williams, A. P., Lintner, B. R., Berg, A. M., Zhang, Y., Keenan, T. F., ... & Gentine, P. (2021). Soil moisture–atmosphere feedbacks mitigate declining water availability in drylands. Nature Climate Change, 11(1), 38-44.

Zhou, Y., Zhou, P., Jin, J., Wu, C., Cui, Y., Zhang, Y., & Tong, F. (2022). Drought identification based on Palmer drought severity index and return period analysis of drought characteristics in Huaibei Plain China. Environmental Research, 212, 113163.

Zhou, P., Liu, Z., & Cheng, L. (2020). An alternative approach for quantitatively estimating climate variability over China under the effects of ENSO events. *Atmospheric Research*, *238*, 104897.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*, *598*, 126266.

Appendices

Appendix A: Research Schedule

	Aug 2024	Sep 2024	Oct 2024	Nov 2024	Dec 2024	Jan 2025	Feb 2025	March 2025	April 2025	May 2025	Jun 2025
Literature Review											
Data Collection											
Research Methodology											
Presentation											
Addressing Feedback											
Final Proposal Submission											
Data Analysis											
Model Deployment											
Writing chapter 4 & 5											
Finishing on First Draft											
Addressing Feedback											
Final Submission											

Appendix B: Ethical Clearance Letter

Strathmore UNIVERSITY

5th December 2024

Mrs Bosire Violet,
violet.bosire@strathmore.edu

Dear Mrs Bosire,

RE: Ensemble Machine Learning Model for Drought Prediction in Kilifi County, Kenya

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2439/24.** The approval period is from **5th December 2024 to 4th December 2025.**

This approval is subject to compliance with the following requirements:

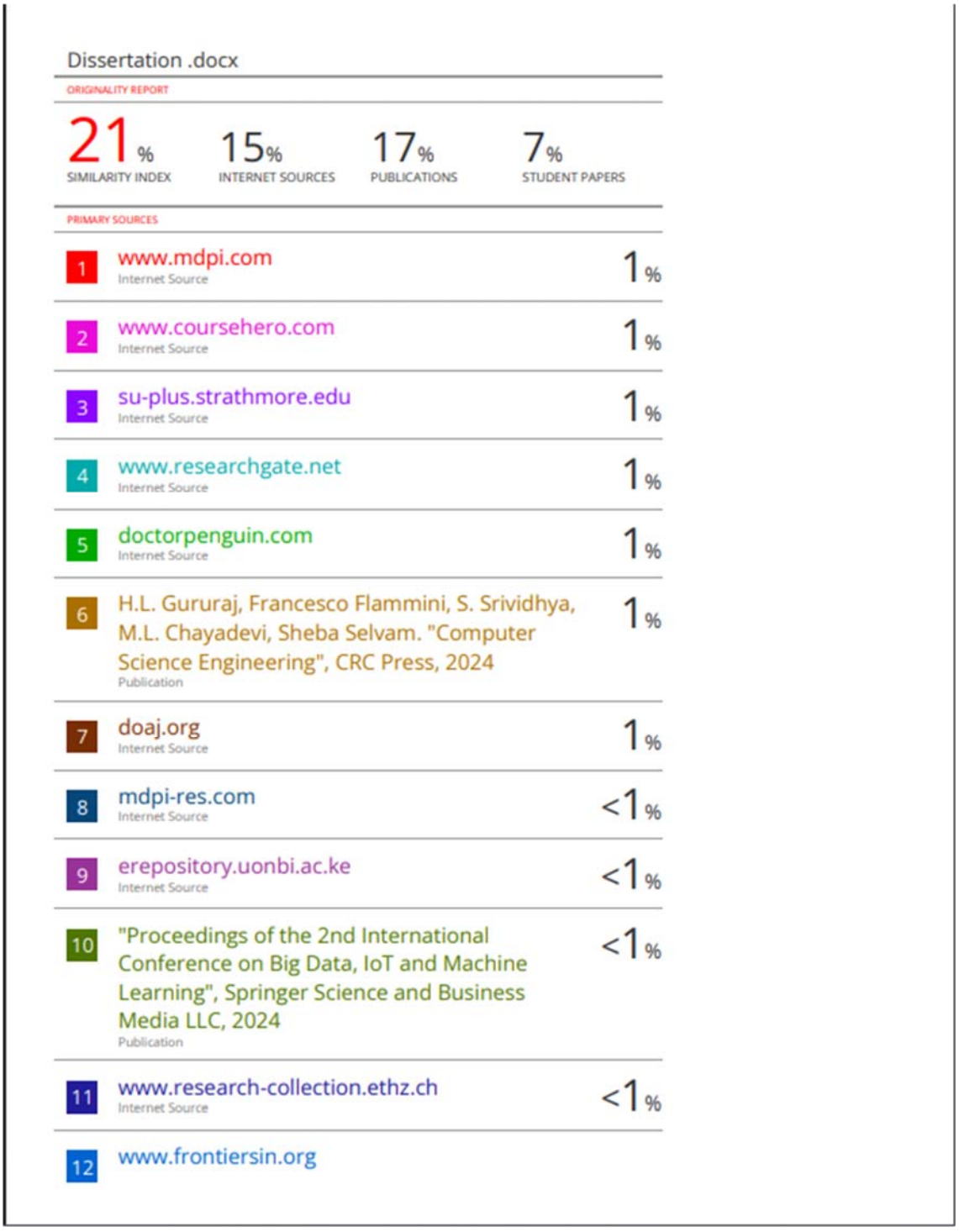
i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Appendix C: Similarity Report



Dissertation .docx

ORIGINALITY REPORT

21%	15%	17%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	www.coursehero.com Internet Source	1%
3	su-plus.strathmore.edu Internet Source	1%
4	www.researchgate.net Internet Source	1%
5	doctorpenguin.com Internet Source	1%
6	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	1%
7	doaj.org Internet Source	1%
8	mdpi-res.com Internet Source	<1%
9	erepository.uonbi.ac.ke Internet Source	<1%
10	"Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning", Springer Science and Business Media LLC, 2024 Publication	<1%
11	www.research-collection.ethz.ch Internet Source	<1%
12	www.frontiersin.org	

	Internet Source	<1%
13	www2.mdpi.com Internet Source	<1%
14	Khalid En-nagre, Mourad Aqnouy, Ayoub Ouarka, Syed Ali Asad Naqvi et al. "Assessment and prediction of meteorological drought using machine learning algorithms and climate data", Climate Risk Management, 2024 Publication	<1%
15	Submitted to Liverpool John Moores University Student Paper	<1%
16	Submitted to University of Sunderland Student Paper	<1%
17	www.uyik.org Internet Source	<1%
18	Birhan Getachew Tikuye, Ram L. Ray, Busnur Manjunatha, Gebrekidan Worku Tefera, Sanjita Gurau. "Drought Monitoring Using the Climate Hazards InfraRed Precipitation with Stations (CHIRPS) in Ethiopia", Natural Hazards Research, 2024 Publication	<1%
19	Sai Kiran Oruganti, Dimitrios A Karras, Srinesh Singh Thakur, Janapati Krishna Chaithanya, Sukanya Metta, Amit Lathigara. "Digital Transformation and Sustainability of Business", CRC Press, 2025 Publication	<1%
20	journals.plos.org Internet Source	<1%
21	Rajib Maity, Aman Srivastava, Subharthi Sarkar, Mohd Imran Khan. "Revolutionizing the future of hydrological science: Impact of	<1%