**Maize Yield Prediction with Vegetation Indices and Meteorological data in Kenya**

**By**

Kedera Tom Elvis

**ENC222-0409/2017**

*Project report submitted to the Department of Geomatic Engineering and geospatial Information Systems degree of Bachelor of Science in Geospatial Information Science.*



**Department of Geomatic Engineering and Geospatial Information Systems (GEGIS)**

DECLARATION

*I declare that this project is my work and has not been submitted by anybody else in any other university for the award of any degree to the best of my knowledge.*

Sign………………………………… Date………………………………………..

Kedera Tom Elvis

ENC222-0409/2017

Department of Geomatic Engineering and Geospatial Information Systems (GEGIS)

Jomo Kenyatta University of Agriculture and Technology

CERTIFICATION

This project has been submitted for examination with my approval as the candidate’s supervisor.

Sign……………………………. Date ………………………………………………

Dr. –Ing. Benson Kenduiywo

Lecturer

GEGIS

©GEGIS 2021

# Acknowledgments

I would like to earnestly acknowledge the sincere efforts and valuable time given by my supervisor, (Dr. –Ing. Benson Kenduiywo) and lab technicians. Their valuable guidance and feedback have helped me in completing this project.

I would also like to appreciate the love and support from my mother, Nancy Kedera to whom I dedicate this work, and my family members, my aunties, uncles and cousins who have been encouraging me through my four years of study.

Last but not the least, my friends [Mercy, Bob, Iyvonne, Faith, Casper, Jones, Irene, and Lydia] who have always been there with my side. Without them, I could never have completed this task. Thanks a lot.

# Abstract

Predicting staple crop output can help to achieve sustainable agricultural, crop insurance, and government policy goals. Accurate in-field yield estimation is expensive to obtain, and estimating it through remote sensing combined with meteorological data is promising. This study forecasted county-level maize production for Kenya's 20 maize-producing counties for 2021 using satellite-derived indices from the Moderate Resolution Imaging Spectroradiometer (MODIS) and the county meteorological data from 2012 to 2018. For the execution of the study Machine Learning (ML) techniques were employed namely the Random Forest (RF) Support Vector Machine (SVM) and Linear Regression was used as the baseline. The models were trained and validated using a 5-fold cross-validation technique. The final obtained models for the training were then used for the prediction of the potential yields of the respective counties for the year 2021. The predictions done by the vegetation indices were the best with a MAPE (Mean Absolute Percentage Error) of 29.00%, LR had a MAPE of 32.46% and SVM was 31.17%. For the prediction done using the weather variables, again RF performed best with a MAPE of 33.05% followed by SVM with 33.97%. For the prediction with both the vegetation indices and weather variables, RF still performed best with a MAPE of 29.52% followed by LR with 31.79% and SVM had 32.35%. This thus shows that ML can be used to predict maize yields in the country when used together with the vegetation indices and weather variables.

# Table of contents

[Acknowledgments i](#_Toc90909848)

[Abstract iii](#_Toc90909849)

[Table of contents iii](#_Toc90909850)

[List of figures vii](#_Toc90909851)

[List of tables vii](#_Toc90909852)

[Acronyms and abbreviations vii](#_Toc90909853)

[Chapter 1 1](#_Toc90909854)

[1.0 Introduction 1](#_Toc90909855)

[1.1 Background 1](#_Toc90909856)

[1.2 Motivation and problem statement 2](#_Toc90909857)

[1.3 Research identification 3](#_Toc90909858)

[1.3.1 Research objectives 3](#_Toc90909859)

[1.4 Research outline 3](#_Toc90909860)

[Chapter 2 5](#_Toc90909861)

[2.0 Literature Review 5](#_Toc90909863)

[2.1 Introduction 5](#_Toc90909864)

[2.2 Maize 6](#_Toc90909865)

[2.3 Temperature 6](#_Toc90909866)

[2.4 Rainfall 6](#_Toc90909867)

[2.5 Wind speed 7](#_Toc90909868)

[2.6 Normalized Difference Vegetation Index (NDVI) 7](#_Toc90909869)

[2.7 Enhanced Vegetation Index (EVI) 7](#_Toc90909870)

[2.8 Evapotranspiration 7](#_Toc90909871)

[2.9 Machine learning 8](#_Toc90909872)

[2.9.1 Random Forest (RF) 9](#_Toc90909873)

[2.9.2 Linear Regression (LR) 9](#_Toc90909874)

[2.9.3 Support Vector Machines (SVM) 9](#_Toc90909875)

[Chapter 3 10](#_Toc90909876)

[3.0 Materials and methods 10](#_Toc90909880)

[3.1 Study area 10](#_Toc90909881)

[3.2 Data 11](#_Toc90909882)

[3.3 Methods 12](#_Toc90909883)

[3.3.1 Extracting vegetation indices 12](#_Toc90909884)

[3.3.2 Data compilation 12](#_Toc90909885)

[3.3.3 Correlation of the variables 13](#_Toc90909886)

[3.3.4 Model training 13](#_Toc90909887)

[3.3.5 Validation 13](#_Toc90909888)

[3.3.6 Yield prediction 14](#_Toc90909889)

[Chapter 4 15](#_Toc90909890)

[4.0 Results 15](#_Toc90909895)

[4.1 Correlation results 15](#_Toc90909896)

[4.2 Yield prediction results 17](#_Toc90909897)

[4.3 Prediction maps 21](#_Toc90909898)

[4.3.1 Linear regression prediction maps 23](#_Toc90909899)

[4.3.2 Random Forest Prediction maps 24](#_Toc90909900)

[4.3.3 Support Vector Machines Prediction maps 25](#_Toc90909901)

[4.4 Validation and comparison results for the respective models used 26](#_Toc90909902)

[Chapter 5 29](#_Toc90909903)

[5.0 Discussion 29](#_Toc90909906)

[Chapter 6 31](#_Toc90909907)

[6.0 Conclusion and outlook 31](#_Toc90909909)

[References 31](#_Toc90909910)

[Appendix A](#_Toc90909911)

[Combined vegetation indices and weather variables. A](#_Toc90909913)

[Vegetation indices B](#_Toc90909914)

[Weather Variables C](#_Toc90909915)

[Plots E](#_Toc90909916)

# List of figures

[Figure 1: Study Area 10](#_Toc90909354)

[Figure 2: Workflow adopted for maize yield prediction 12](#_Toc90909355)

[Figure 3: Correlation of the maize yield and the variables used for prediction 16](#_Toc90909356)

[Figure 4: Shows the prediction results done by using Linear regression. 23](file:///C:\Users\USER\Desktop\Project\ENC222-0409-2017_KEDERA%20TOM%20ELVIS.docx#_Toc90909357)

[Figure 5: Shows the prediction results done by using Random Forest. 24](file:///C:\Users\USER\Desktop\Project\ENC222-0409-2017_KEDERA%20TOM%20ELVIS.docx#_Toc90909358)

[Figure 6: Shows the prediction results done by using Support Vector Machines. 25](file:///C:\Users\USER\Desktop\Project\ENC222-0409-2017_KEDERA%20TOM%20ELVIS.docx#_Toc90909359)

# List of tables

[Table 1: Data 11](#_Toc90909426)

[Table 3: Predicted yields for the year 2021 with both the vegetation indices and the weather variables in ton/ha. 18](#_Toc90909427)

[Table 4: Predicted yields for the year 2021 with the vegetation indices in ton/ha. 19](#_Toc90909428)

[Table 5: Predicted yields for the year 2021 with the Weather Variables in ton/ha. 20](#_Toc90909429)

[Table 6: Table showing the MAPE (Mean Absolute Percentage Error) of the models used 27](#_Toc90909430)

[Table 7: Table showing the RMSE (Root Mean Square Error) of the models used 27](#_Toc90909431)

[Table 8: Table showing the R2 of the models used 27](#_Toc90909432)

# Acronyms and abbreviations

**GDP** – Gross Domestic Product

**LR** – Linear Regression

**RF** – Random Forest

**SVM** – Support Vector Machine

**ML** – Machine Learning

**NDVI** – Normalized Difference Vegetation Index

**EVI** – Enhanced Vegetation Index

**GPP** – Gross Primary Product

**RMSE** – Root Mean Square Error

**MAPE** – Mean Absolute Percentage Error

# Chapter 1

## Introduction

## Background

The Intergovernmental Panel on Climate Change (IPCC, 2014) states unequivocally that climate variability, combined with extreme climatic events (heat waves, droughts, floods, and wildfires), endangers natural and human systems and agriculture worldwide. Although distributional effects are to be expected, climate change is predicted to diminish food production and perhaps worsen food insecurity in many parts of the world. The extent to which economic players adapt their production methods to environmental changes is critical in assessing food security prospects (Ma & Maystadt, 2017).

Crop cultivation is an important contributor to food security and employment in Kenya. Directly, the sector accounts for around 26% of GDP, with another 25% coming from indirect sources. Kenya's primary food source is maize. Kenya has 2.1 million hectares of maize, accounting for more than 40% of the country's total cultivated land. Droughts and pests alter maize yields, which makes them unpredictable. Better investments, more efficient markets, and better policymaking can all benefit from a quantitative and spatially detailed understanding of maize yield variations. If yield projections are accurate, they can be utilized to prevent food shortages by implementing suitable measures such as imports (Kenduiywo et al., 2020).

Maize productivity is highly influenced by local weather conditions. Research on the effects of climate change on maize yield and yield gap has been undertaken. In different places, the primary meteorological factors limiting climate-induced yield vary (H. Sun et al., 2016).

Rainfall and temperature are the key meteorological parameters that largely contribute to the fluctuation in maize output in Kenya. For proper growth and development, maize requires an average daily temperature of at least 20°C. Temperatures around 30°C are ideal for optimum harvests. Photoperiod and temperature have an impact on flowering time. Maize is regarded as a high-yielding, short-day crop. It also requires ideal moisture levels at the time of planting. In the tropics, 600-900mm of rain during the growing season is ideal (Amudavi et al., 2017).

To create effective adaptation methods, knowledge of the projected implications of climate change and different weather conditions on agriculture is essential. In addition to long-term climate change information, information on near-term production hazards is required to provide advice on how to reduce weather risks through short-term adaptation planning and how to deal with already altered climate circumstances (Laudien et al., 2020).

## Motivation and problem statement

For many years, farmers have been predicting crop production based on their past experiences and historical data, and then make critical cropping decisions based on that projection. (Shahhosseini et al., 2020). However, new technologies such as simulation crop models and machine learning, as well as the ability to analyze huge data with high-performance computing, have resulted in more accurate yield predictions in recent years.

Infield prediction of maize yield is an expensive undertaking in Kenya. With the continuous advancement in remote sensing technology, and meteorological data collection and analysis of the different weather parameters. The derived vegetation indices can be utilized for less costly prediction purposes for the yields e.g. maize. Also, the meteorological data can be used as well and a combination of the two can be used to yield more accurate yield estimate results as a study done by Bouras et al., 2021 in Morocco on the prediction of cereals in the country revealed that combining the remote sensing indices with the meteorological data and using machine learning models. The results of the study were promising which shows that a similar study can be implemented in Kenya.

In Kenya's studies have been done using several machine learning (ML) techniques to try and do prediction of the yields, these ML techniques include Support Vector Regression and Random Forest regression models that have been used to forecast maize yields using MODIS vegetation indices (Kenduiywo et al., 2020). Other studies like (Mupangwa et al., 2020) have applied Machine Learning approaches to estimating maize yields from various Conservation Agriculture-based cropping systems in Eastern and Southern Africa's highland and lowland regions.

These studies predict agricultural yields, especially maize, using a combination of multi-source datasets and machine learning techniques. However, the use of remote sensing vegetation indices in conjunction with weather data as predictors of cereal yield has yet to be evaluated. In this context, the goal of this research is to look at the feasibility of employing machine learning to develop dynamic decision support systems for maize production in Kenya by merging satellite-based vegetation indices and weather data.

Maize yield prediction early in the season will provide information and insights that will help in:

* Improving crop management.
* Economic trading.
* Food production monitoring.
* Ensure food security.

For this study, the models adopted are the support vector machine which from the study done by Kenduiywo et al., 2020, its performance was similar to that of the random forest model which was both being used for the prediction. The second model is the random forest which from the study by Jeong et al., 2016 performed comparably better than the multilinear regression was used as the baseline, it is also used since one of its advantages is its ability to avoid overfitting of the model. The final model is the linear regression which for this study will be used as a baseline to check the performance of the other models being used.

## Research identification

The main objective of the study is to predict maize yields in Kenya using machine learning models Random Forest, Support Vector Machine, and Linear Regression.

### Research objectives

The specific objectives for the study are;

1. To establish the correlation between maize yields and the variables.
2. To train prediction models and predict the yields.
3. To validate and compare the models used for prediction.

## Research outline

This study consists of six chapters. Chapter one sets the research background, motivation and problem, and objectives. A review of the theoretical concepts on machine learning and yield prediction is discussed in chapter two. Chapter three describes the study area chosen, the data used, and the overall execution framework of the study. The results are presented in the fourth chapter and followed by the discussion in chapter five. The final chapter concludes the study and gives recommendations for future studies.

# Chapter 2



## Literature Review

This chapter provides a theoretical background to the research content. The use of machine learning techniques in yield prediction as well as the best machine learning models for the prediction of the yields have been discussed. Key weather variables that affect the yields together with vegetation indices acquired from remote sensing techniques.

## Introduction

Crop growth is highly dependent on the weather. Some crops require specific temperatures, either high or low, to initiate germination and continue development. Temperature and humidity, on the other hand, are frequently utilized to predict the occurrence of many insect pests and diseases. Farmers can schedule planting, protection, harvesting, and other field activities based on this data to avoid negative weather effects and production losses.

Weather monitoring provides a variety of measuring choices for providing relevant information about soil and crop conditions, including air temperature, relative humidity, rainfall, wind speed/direction, and evapotranspiration (Folnovic, 2021). Weather forecasts are critical because they determine when crops are planted, harvested, irrigated, protected, fertilized, and sprayed.

Yin et al., 2021 investigated the impact of weather on maize yields in three provinces along China's northeast coast (Heilongjiang, Jilin, and Liaoning); they discovered that minimum temperature deviations for May and September had a strong positive effect on maize yield, where yield in a given year was also measured as differences from average yield over the 44 years from 1965. Multiple regressions were estimated for each province, as well as for the three provinces combined, with intercept terms included only in regressions for Heilongjiang and Liaoning provinces, but not in the aggregated regression or for Jilin province.

The key variable of relevance is temperature, while precipitation may be significant at certain periods of the year. For example, excessive rainfall in September may cause harvests to be delayed, grain quality to suffer, and overall production to be reduced. Similarly, if fields are too wet in the autumn, harvesting may be delayed, and yields may decline due to decay; or if fields are too wet in the spring, planting may be delayed, resulting in less exposure to heat, or exposure to heat at the wrong time in the growing cycle, and thus lower yields (B. J. Sun & Van Kooten, 2014).

Weather parameters such as radiation, daily maximum temperature, daily minimum temperature, and daily average temperature during the grain-filling stage all have an impact on maize production, resulting in greater seasonal yield fluctuation. Optimizing the sowing date and plant density could reduce seasonal yield fluctuation greatly (H. Sun et al., 2016).

## Maize

Maize, *Zea mays* *L*. (corn), is the world's most abundant cereal crop. Except for Antarctica, it is grown on every continent. There are approximately 50 species that vary in color, texture, grain shape and size, and grain form and size. The most commonly cultivated maize varieties are white, yellow, and red. Depending on where you live, most people prefer the white and yellow types. Maize, which was developed in central Mexico approximately 1500 BC, was brought into Africa around 1500 AD and quickly expanded to every corner of the continent, becoming Africa's most significant grain crop (Amudavi et al., 2017).

## Temperature

Crops are vulnerable to climate change, particularly temperature fluctuations. Temperature increases are the most likely to have a negative influence on agricultural yields, and regional temperature changes can be anticipated with greater accuracy from climate models than precipitation. Meteorological records suggest that mean annual temperatures in wheat, rice, maize, and soybean growing areas have risen by 1°C during the last century. To estimate the danger to global food security and subsequently establish targeted adaptive solutions to feed the world population, it is required to quantify the impact of temperature increases on global crop yields, including any geographical differences (Piao et al., 2017).

## Rainfall

Kenya's rainfall peaks twice a year, between March-May and September-November. The major growing season is supported in most regions by the March (long) rains. This is visible in the wetter mid-altitude and transitional zones, as well as in the highland and lowland tropics, where the vast majority of farmers see the March rains as their main maize season. Farmers, on the other hand, value both seasons at least equally. In the semi-arid and dry transition zones, there is very little rainfall between the two seasonal maxima, indicating a bimodal rainfall pattern. The increased amount of rainfall between the two main peaks in the somewhat wetter zones, on the other hand, supports a prolonged single cropping season. This is especially true in the highlands, where the average total precipitation during the three months between the two rain peaks nearly equals the amount of precipitation during the six months of short rains (September-February). Rainfall patterns, in conjunction with population density, can thus explain the regional variance in the intensity of maize cultivation in Kenya (Hassan, 1996).

## Wind speed

Strong winds can cause newly emerged crops to be buried, rip immature seedlings from the soil, or partially expose their roots. Inadequate canopy cover permits soil particles to be mobilized by the shearing force of the wind on the soil surface. Seedlings will be roughened and damaged by these blowing particles. Wind-induced movements can cause leaves to rub together as the canopy grows, extending the abrasion process. As the crop grows taller, especially as the grain head fills, the force of the wind can lodge the crop by breaking stems or causing soil and root collapse. Fallen crops are difficult to harvest, and grain recovery may be prohibitively expensive. (Cleugh et al., 1998)

## Normalized Difference Vegetation Index (NDVI)

NDVI measures the difference between near-infrared, which vegetation strongly reflects, and red light to quantify vegetation, which vegetation absorbs. The NDVI scale is always between -1 and +1. However, each form of land cover does not have a unique limit. When you have negative readings, for example, it's likely water. On the other side, if your NDVI number is near +1, you're likely looking at dense green foliage. However, when the NDVI is close to zero, there are no green leaves and the area may be urbanized. The most popular index used by remote sensing experts is the NDVI.

## Enhanced Vegetation Index (EVI)

EVI is an 'optimal' vegetation index meant to improve vegetation monitoring by isolating the canopy background signal and reducing atmospheric impacts.

## Evapotranspiration

Crop water consumption, also known as evapotranspiration, is the amount of water that a crop uses for growth and cooling. This water is removed from the soil root zone by the root system, which symbolizes transpiration, and is no longer available in the soil as stored water. The extraction of soil water gets more difficult as the water content of the soil declines. The roots extract 99.9% of the water consumed by an irrigated crop, while the leaves transpire the rest. (Onainor, 2019)

## Machine learning

ML techniques have emerged as a possible alternative to and complement traditional crop production modelling. When compared to other study fields in agriculture, such as soil and water management and livestock production, ML techniques are increasingly being used in crop production research. In crop production, machine learning tools have mostly been used to estimate yield and detect disease in cropping systems. At multiple scales, such as local, regional, and national levels, machine learning techniques can be used to estimate crop yields (Mupangwa et al., 2020).

A study was done by Mupangwa et al., 2020, on evaluating machine learning algorithms for predicting maize yields under conservational agriculture in eastern and southern Africa. He used several models both linear and machine learning-based, i.e. Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbour, Gaussian Naïve Bayes, and the Support Vector Machine. For this study, the type of prediction that was done was classification to fine the area/ agroecological zones with the highest yields. It was concluded the linear discriminant analysis and the Naïve Bayes models were the best performers in the prediction.

Although simulation crop modelling has a fair forecast accuracy, it is not as easily applicable as machine learning models due to user skill, data calibration requirements, long runtimes, and data storage limits. ML, on the other hand, has found a wide range of applications in a variety of situations due to its capacity to cope with linear and nonlinear relationships, non-normal data, and the quality of outputs while having substantially shorter runtimes. Based on the type of response variables, supervised learning is divided into regression and classification tasks. Many studies have attempted to handle regression problems using machine learning (Shahhosseini et al., 2020).

This study uses weather variables together with satellite-derived vegetation indices to predict maize yields together with machine learning models (RF, SVM) and linear-based models (LR) to predict the yields.

### Random Forest (RF)

Jeong et al., 2016, did a study that assessed the efficacy of RF regression on global and regional scales, using MLR as a baseline to estimate complex yield responses of wheat, grain maize, potato, and silage maize. Although the RF method had several advantages for regressing complicated crop systems, it is not commonly used in this field. We demonstrated that RF outperforms all crops and areas examined in terms of yield prediction. The findings of the study indicated a high potential for implementing an RF algorithm as an alternative statistical modelling tool for crop yield predictions. It should be emphasized that RF runs the risk of overfitting data in situations where training data is concentrated. It should be noted that RF carries the danger of overfitting data in situations when the training data is dense, whilst its accuracy can suffer in situations where the training data is sparse.

### Linear Regression (LR)

In Linear Regression (LR), the dependent variable y is linearly related to multiple independent variables x*i* = 1. . . n as:

Where for equation (1), y in this study is the predicted yield, x*i* (*i* = 0 . . . n) are the satellite-based vegetation indices, the weather data, and a*i* (*i* = 0 . . . n) are the regression coefficients.

### Support Vector Machines (SVM)

SVM (Support Vector Machine) is a supervised machine learning technique that is used to solve binary classification issues. The goal of this approach is to draw a hyperplane in an N-dimensional space, where N is the number of characteristics in a dataset that will be used to categorize the data points. There can be any number of hyperplanes drawn, but the algorithm's main goal is to locate the plane with the greatest margin, or the greatest distance between the data points of the characteristics being plotted. The greater the distance, the more accurate the classification (Suresh et al., 2020).

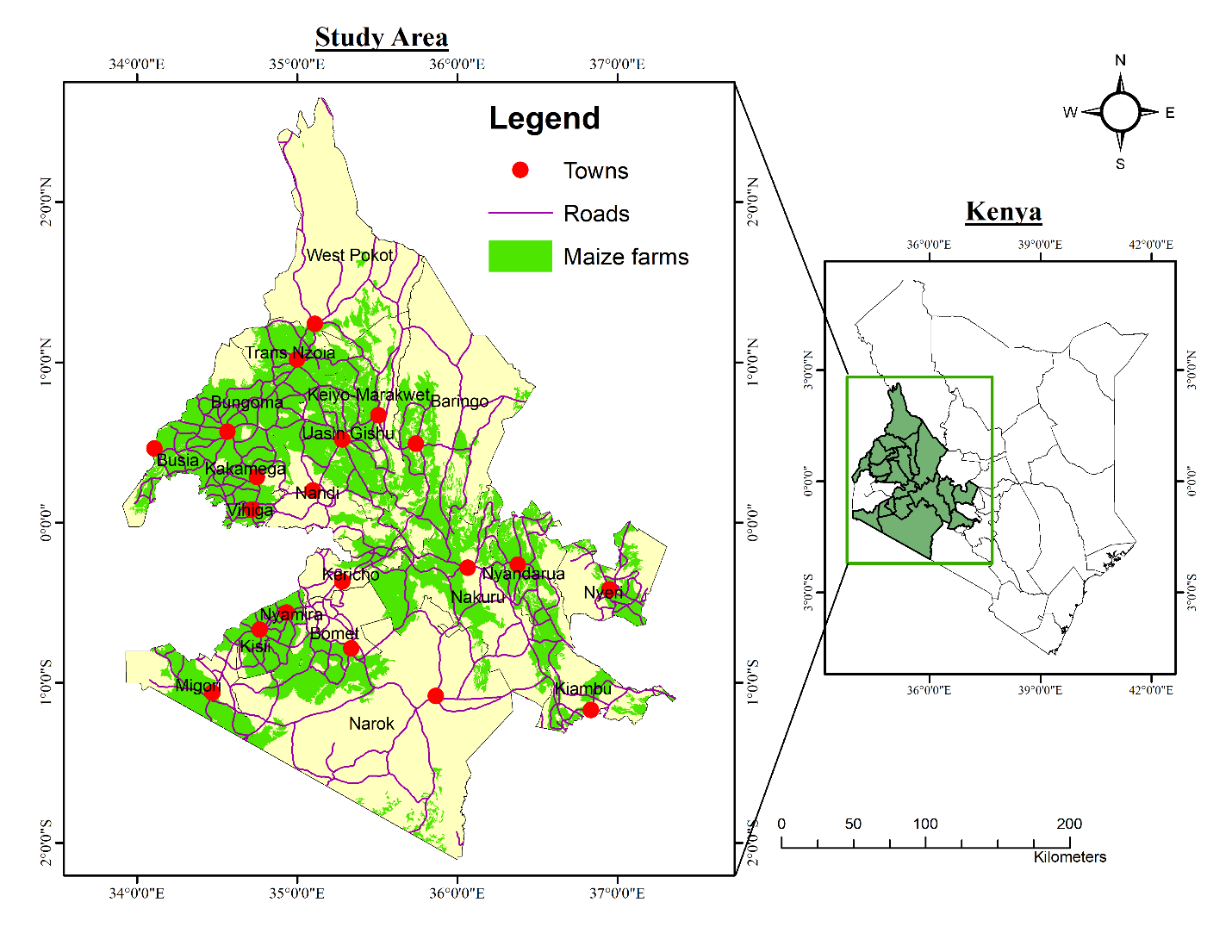
# Chapter 3



## Materials and methods

This chapter describes the selected study area, data used for the prediction of the yields.

## Study area



**Counties**

Figure 1: Study Area

The study area for this study includes the counties within the food basket of Kenya which produce maize in large quantities. Most of these counties are located in the western and central regions of the country since, in these regions, the soils located are well fertile and are well suited for the planting of maize. They are also located within the country’s tropical climatic region which is a favorable climatic condition for the cultivation of maize.

These counties include Trans-Nzoia which is the leading maize producing county in the country, it’s followed by Uasin Gishu county, the other counties are, West Pokot, Kakamega, Bungoma, Elgeyo-Marakwet, Kericho, Nakuru, Narok, Bomet, Migori, Kisii, Kiambu, Nyandarua, Baringo, Nandi, Busia, Nyeri, and Vihiga.

## Data

Table 1: Data

|  |  |  |  |
| --- | --- | --- | --- |
| **Data** | **Duration** | **Source** | **Link** |
| Maize yield data | 2012 - 2018 | Ministry of agriculture, livestock, and fisheries | (MOALFI, 2020) |
| Weather data | 2012 – 2021  (Monthly) | Climate Data Website | (Climate, 2021) |
| Evapotranspiration | 2012 – 2021  (Monthly) | FAO Wapor | (FAO, 2021) |
| NDVI, EVI GPP | 2012 – 2021  (Monthly) | MODIS | (MODIS, 2021) |

The data used was obtained from the Ministry of Agriculture Livestock and Fisheries website (MOALFI, 2020) and website the only maize yield data available was from 2012 to 2018 and this is the data that was obtained and which will later be used to both train and validate the models that will be used. The meteorological data used was also obtained online from the climate data organization’s website. Together with this data on weather also data on the evapotranspiration data obtained by the Food and Agriculture Organisation (FAO) for the same month of April. Also added to this data used are the satellite-derived indices namely NDVI from the MOD13Q1 product, EVI from the MOD13Q1 product too and GPP from the MOD17A2H product. These were obtained from MODIS (Moderate Resolution Imaging Spectroradiometer).

The data obtained was for April since in Kenya the planting season is in March, this study aims to see and understand if the prevailing conditions post-harvest have an impact on the yield at the end of the season during the harvesting period. These data were analyzed and combined to create one excel spreadsheet to be used for the study.

## Methods

The methodology utilized for this study is summarized in the flow shat below;

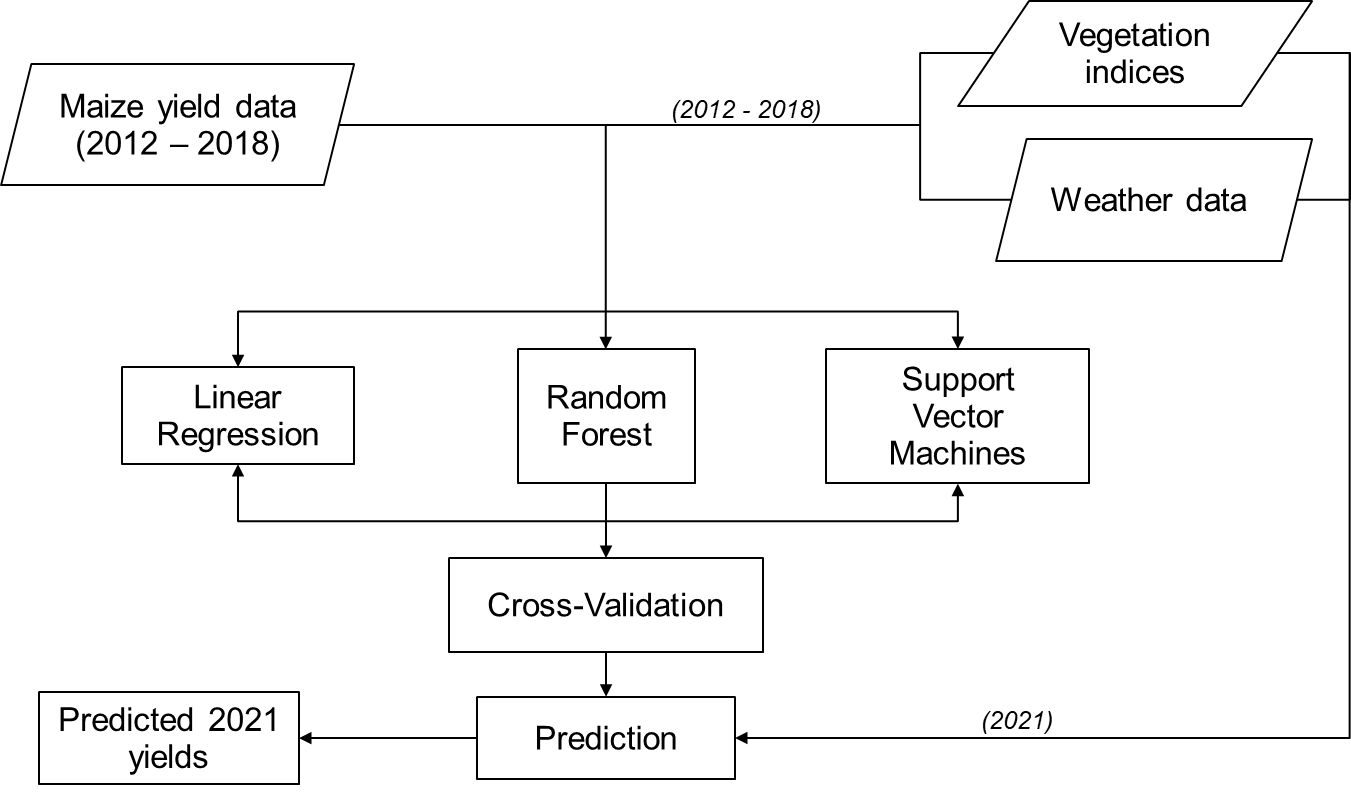


Figure 2: Workflow adopted for maize yield prediction

### Extracting vegetation indices

The vegetation indices were obtained by first downloading the images for the respective years, then clipping them to the maize farm shapefile for the countries in the study area. Random points were then generated with at least 500 points per county. These points were generated with the extent of the maize cropland shapefile data that had been clipped to only that which was in the study area. The points were then overlaid on the clipped MODIS images for sampling the index values. The extracted index data was exported from the attribute table to a .csv format. The averages for the indices were then derived which would be used in the training of the models and prediction of the maize yields.

### Data compilation

The weather data and the vegetation indices for the corresponding yield data available were matched for the respective counties with the years as the reference for merging the data. This was done for the years 2012 to 2018. These combined data were processed in R-studio for the training of the three models to be used for the prediction of the yields. The data was split into three further categories i.e. the combined yields, weather variables, and the vegetation indices, the second was that which contained only the yields and the vegetation indices and the last category was that of the yields and the weather variables.

The remainder of the independent variables i.e. weather variables and the vegetation indices which are for the year 2021 are used for the prediction of the yields for the year 2021. This is done after the validation of the models.

### Correlation of the variables

For the correlation, we were checking the kind of relationship that existed between the yields and the variables proposed for the prediction. The Pearson’s correlation technique was implemented and the values were obtained. The following thresholds for the absolute value of the correlation coefficient were used to describe the strength of the relationship between variables:

±0.0 < |*r*| ±0.3 indicated a weak relationship

±0.3 < |*r*| ±0.7 indicated a moderate relationship

±0.7 < |*r*| ±1.0 indicated a strong relationship

### Model training

The proposed models to be used for the prediction were linear regression, random forest, and support vector machines. These models were trained using the *k*-fold cross-validation technique where *k* is the number of classes or groups in which the data set is split and one of the groups is used for training and the rest is used for validation. For our study, the 5-fold cross-validation technique was employed, whereby the data was split into 5 equal classes/groups, the first group was used for training and the remaining four were used for validation, for the second iteration, the second group was used for training while the first and the last three were used for validation. This process is repeated until the last, which is the fifth group is used for training and the first four were used for validation.

### Validation

Validation is always done to check how accurate the trained model works before being used for prediction or classification purposes. Under this, the amount of error the model gives is used to know how well the model is performing. For our study, three validation methods were employed i.e. MAPE, RMSE, and the R2 value which explains the proportion of variance in the dependent variable that is explained by the independent variable.

Wherefrom equations (2, 3, and 4), **y*i***denotes the actual values, **ŷ*i***denoted the predictions, and ***n*** denotes the number of observations. These parameters are then tabulated for the three prediction scenarios and summarized in the chapter below.

### Yield prediction

After the training and validation of the models, the 2021 data of the independent variables were the fit to the models for the different scenarios i.e. the combined weather and vegetation indices, the weather variables only, and the vegetation indices. All the three gave out different results given in the following chapter

# Chapter 4



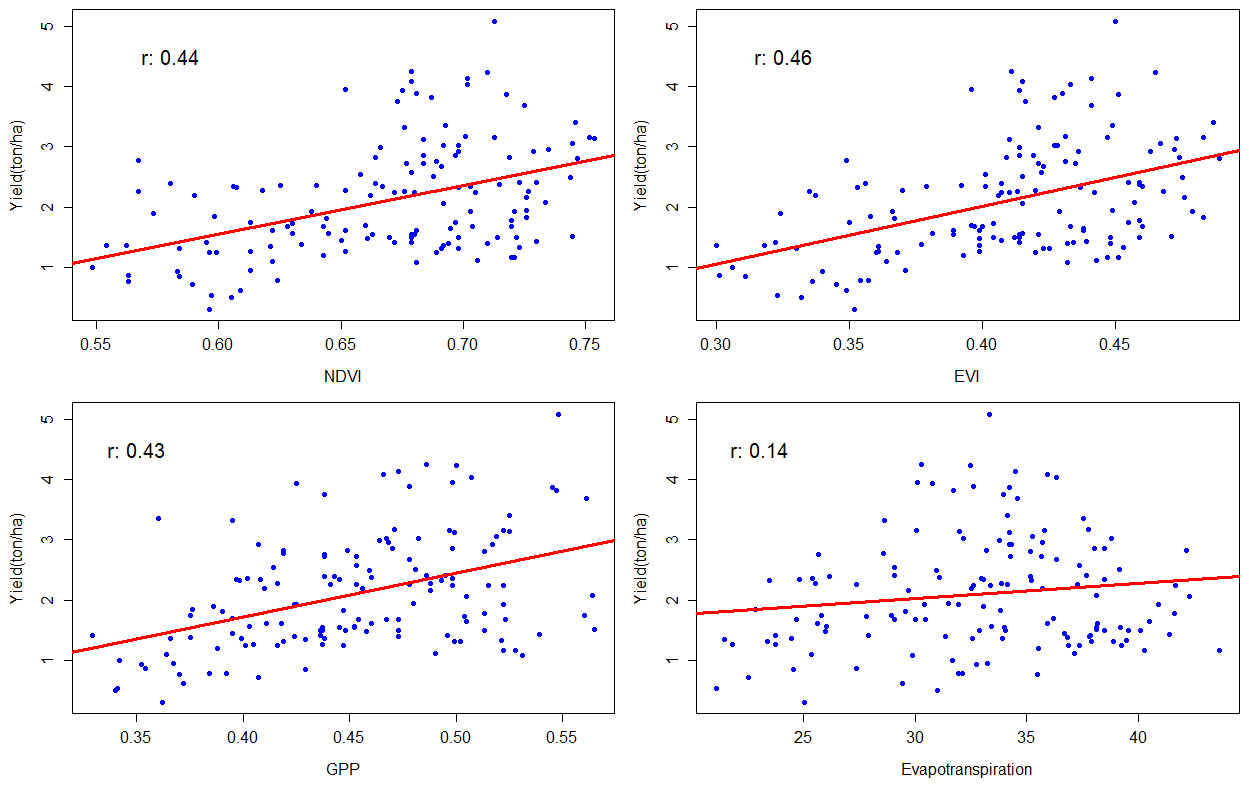
## Results

Results of research objectives implementation are presented sequentially in this chapter.

## Correlation results

Figure 3 shows the correlation results between the yields and the variables to be used for the prediction. There is a negative correlation which in other words can be termed as inverse correlation in which when one of the variables e.g. Wind speed which from the results has a correlation value of -0.29, this shows that if the speeds decrease the yields increase and vice versa is true in that when the speed increases the yields decrease. There is also a positive correlation of the yields with the vegetation indices especially EVI with a value of 0.46 showing that if it increases then yields also increase and if it decreases the yields will decrease too.

NDVI values from 0.55 – 0.75, have a positive relationship with a yield. For EVI, there exists a positive relationship between the values of 0.30 – 0.50 with the yields. For GPP there is a positive relationship for the values between 0.30 – 0.60 with the yields. For evapotranspiration, there is a stronger relationship for the value of 30 with the yields from 1 ton/ha to 5 ton/ha, but then the relation drops as the evapotranspiration value increases towards 40. For rainfall, there is a relationship between the rainfall amount between 0mm – 600mm and the yields, but a drop in the relation starts at 400mm toward 800mm. For temperature, there is a rise in the relation between 18℃ and 22℃ then followed by a drop in the relation from 22℃ towards 25℃. For the wind speed, there is generally a higher relation between speeds of 5km/h – 8km/h with the yields but it drops towards 12km/h. Finally, for air pressure, there is a relation for the pressure between 0millibars – 15millibars with yields between 1 ton/ha and 5 ton/ha. There is a decrease in the relationship as the pressure increases.



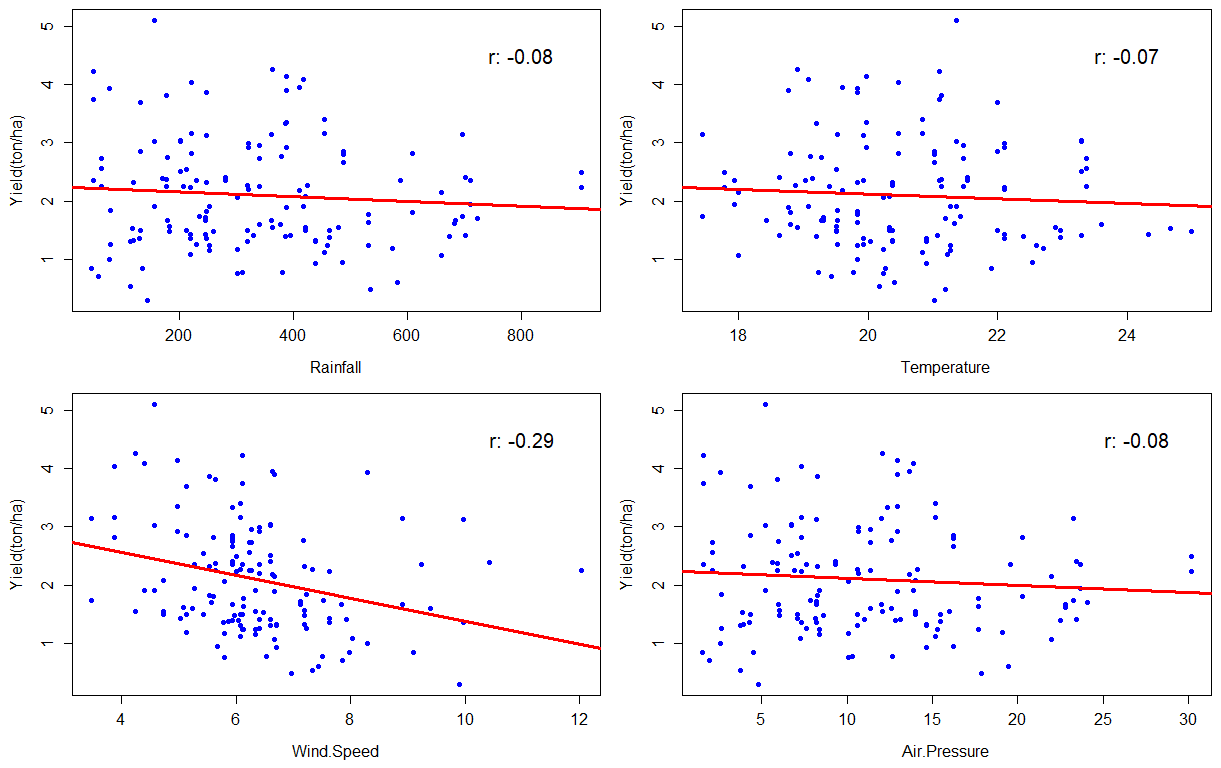


Figure 3: Correlation of the maize yield and the variables used for prediction

## Yield prediction results

Table 2 shows the predicted yields for the year 2021 by the different prediction models with both the vegetation indices and the weather variables in ton/ha. Kiambu has the lowest yields for the prediction done by LR of 0.40 ton/ha, for RF, Kiambu also has the lowest yields of 1.03 ton/ha while for the SVM, Kiambu still has the lowest yield of 0.42 ton/ha, and Trans-Nzoia has the highest yields for the models used. The yields are 3.43 ton/ha, 4.17 ton/ha, and 3.46 ton/ha respectively.

Table 3 shows the predicted yields for the year 2021 by the different prediction models with the vegetation indices in ton/ha. Nyeri has the lowest yields for the prediction done by LR of 0.99 ton/ha, for RF, Kiambu has the lowest yields of 1.08 ton/ha while for the SVM, Kiambu also has the lowest yield of 0.97 ton/ha, and Trans-Nzoia again has the highest yields for the models used. The yields are 3.24 ton/ha, 4.31 ton/ha, and 3.34 ton/ha respectively.

Table 4 shows the predicted yields for the year 2021 by the different prediction models with the Weather Variables in ton/ha. Kiambu has the lowest yields for the prediction done by LR of 0.38 ton/ha, for RF, Kiambu has the lowest yields of 1.22 ton/ha while for the SVM, Kiambu still has the lowest yield of 0.54 ton/ha, and Trans-Nzoia again has the highest yields for the models used. The yields are 3.32 ton/ha, 4.16 ton/ha, and 3.21 ton/ha respectively.

Table 2: Predicted yields for the year 2021 by the different prediction models with both the vegetation indices and the weather variables in ton/ha.

|  |  |  |  |
| --- | --- | --- | --- |
| **COUNTY** | **LR** | **RF** | **SVM** |
| Baringo | 0.44 | 1.45 | 0.54 |
| Bomet | 1.90 | 2.10 | 2.13 |
| Bungoma | 2.43 | 2.88 | 2.45 |
| Busia | 1.07 | 1.60 | 1.08 |
| Elgeyo-Marakwet | 1.91 | 2.24 | 1.94 |
| Kakamega | 2.11 | 1.92 | 2.14 |
| Kericho | 2.26 | 2.46 | 2.25 |
| Kiambu | 0.40 | 1.03 | 0.42 |
| Kisii | 2.75 | 1.91 | 2.83 |
| Migori | 2.28 | 1.85 | 2.34 |
| Nakuru | 2.49 | 2.08 | 2.42 |
| Nandi | 2.31 | 2.81 | 2.35 |
| Narok | 1.93 | 2.57 | 2.13 |
| Nyamira | 2.64 | 1.92 | 2.69 |
| Nyandarua | 1.34 | 1.51 | 1.32 |
| Nyeri | 1.13 | 1.42 | 1.09 |
| Trans Nzoia | 3.43 | 4.17 | 3.46 |
| Uasin Gishu | 1.99 | 2.95 | 2.05 |
| Vihiga | 1.09 | 1.57 | 1.15 |
| West Pokot | 2.15 | 2.11 | 2.08 |

Table 3: Predicted yields for the year 2021 by the different prediction models with the vegetation indices in ton/ha.

|  |  |  |  |
| --- | --- | --- | --- |
| **COUNTY** | **LR** | **RF** | **SVM** |
| Baringo | 1.20 | 1.44 | 1.22 |
| Bomet | 1.85 | 1.66 | 2.11 |
| Bungoma | 2.91 | 2.91 | 2.88 |
| Busia | 1.60 | 1.71 | 1.46 |
| Elgeyo-Marakwet | 1.78 | 2.34 | 1.84 |
| Kakamega | 2.62 | 1.96 | 2.58 |
| Kericho | 2.24 | 2.55 | 2.26 |
| Kiambu | 1.04 | 1.08 | 0.97 |
| Kisii | 2.72 | 1.81 | 2.86 |
| Migori | 2.14 | 1.83 | 2.21 |
| Nakuru | 2.40 | 2.15 | 2.34 |
| Nandi | 2.87 | 2.97 | 2.84 |
| Narok | 2.45 | 2.65 | 2.59 |
| Nyamira | 2.64 | 2.00 | 2.74 |
| Nyandarua | 1.26 | 1.51 | 1.27 |
| Nyeri | 0.99 | 1.41 | 0.98 |
| Trans Nzoia | 3.24 | 4.31 | 3.34 |
| Uasin Gishu | 2.75 | 3.29 | 2.75 |
| Vihiga | 1.81 | 2.00 | 1.75 |
| West Pokot | 2.07 | 2.18 | 1.96 |

Table 4: Predicted yields for the year 2021 by the different prediction models with the Weather Variables in ton/ha.

|  |  |  |  |
| --- | --- | --- | --- |
| **COUNTY** | **LR** | **RF** | **SVM** |
| Baringo | 0.60 | 1.47 | 0.71 |
| Bomet | 1.91 | 2.27 | 1.84 |
| Bungoma | 2.13 | 2.79 | 2.20 |
| Busia | 1.02 | 1.53 | 1.13 |
| Elgeyo-Marakwet | 1.98 | 2.16 | 1.88 |
| Kakamega | 1.86 | 1.73 | 1.92 |
| Kericho | 1.86 | 2.16 | 1.80 |
| Kiambu | 0.38 | 1.22 | 0.54 |
| Kisii | 2.71 | 1.96 | 2.60 |
| Migori | 2.78 | 1.98 | 2.69 |
| Nakuru | 2.49 | 2.22 | 2.45 |
| Nandi | 1.67 | 1.83 | 1.74 |
| Narok | 1.87 | 2.76 | 1.94 |
| Nyamira | 2.46 | 1.68 | 2.36 |
| Nyandarua | 1.29 | 1.54 | 1.27 |
| Nyeri | 1.45 | 1.55 | 1.45 |
| Trans Nzoia | 3.32 | 4.16 | 3.21 |
| Uasin Gishu | 1.57 | 2.83 | 1.69 |
| Vihiga | 0.72 | 1.46 | 0.85 |
| West Pokot | 2.01 | 2.17 | 1.91 |

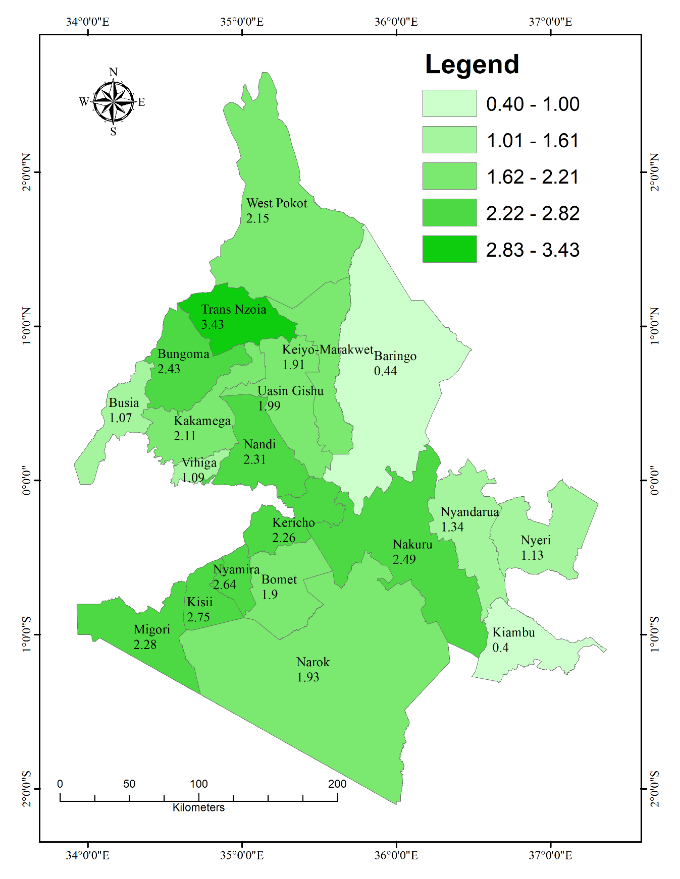
## Prediction maps

Figure 4 shows the prediction map for the results obtained from the LR model. The distribution of the expected yields at the end of the year can be visualized in figures (a), (b), and (c). Figure 4 (a) shows the yields from the prediction done by the combined vegetation indices and the weather variables. The counties along the center of the study area are seen to have a high yield with Trans-Nzoia leading with 3.43 ton/ha. Counties to the west i.e. Baringo, Nyeri, Nyandarua, and Kiambu have the lowest yield. Busia on the extreme west also has low yields. From Figure 4 (b), which were the predictions done with the vegetation indices, counties with high yields are clustered together in the northwest and south regions. Again the counties in the east have low yields apart from Nakuru. In Figure 4 (c), which are the predictions done with the weather variables, most of the counties have more than 1.5 ton/ha predicted yields still counties in the east apart from Nakuru have low yields. In the western region, Busia and Vihiga counties have the lowest yields of 1.02 ton/ha and 0.72 ton/ha respectively.

Figure 5 shows the prediction maps for the results of the prediction done using RF. From Figure 5 (a) which is from the prediction with the combined vegetation indices and the weather variables, the western region has the lowest yields in the counties of Baringo, Nakuru, Nyandarua, Nyeri, and Kiambu. Most of the counties have an average yield of between 1.6 tons/ha to about 2.9 ton/ha. Only Trans-Nzoia and Uasin Gishu have yields of more than 2.9 ton/ha with values of 4.17 ton/ha and 2.95 ton/ha respectively. Figure 5 (b), predictions by the vegetation indices has almost the same distribution as in Figure 5 (a) where the average yields are between 1.7 ton/ha and 3.0 ton/ha with still Trans-Nzoia and Uasin Gishu having the high yields of 4.32 ton/ha and 3.29 ton/ha respectively. The same distribution is repeated in Figure 5 (c) with the counties in the east of the study area together with the counties of Busia and Vihiga having the lowest yield while the others range between 1.8 ton/ha and 2.9 ton/ha. Still, Trans-Nzoia and Uasin Gishu had high yields of 4.16 ton/ha and 2.83 ton/ha respectively.

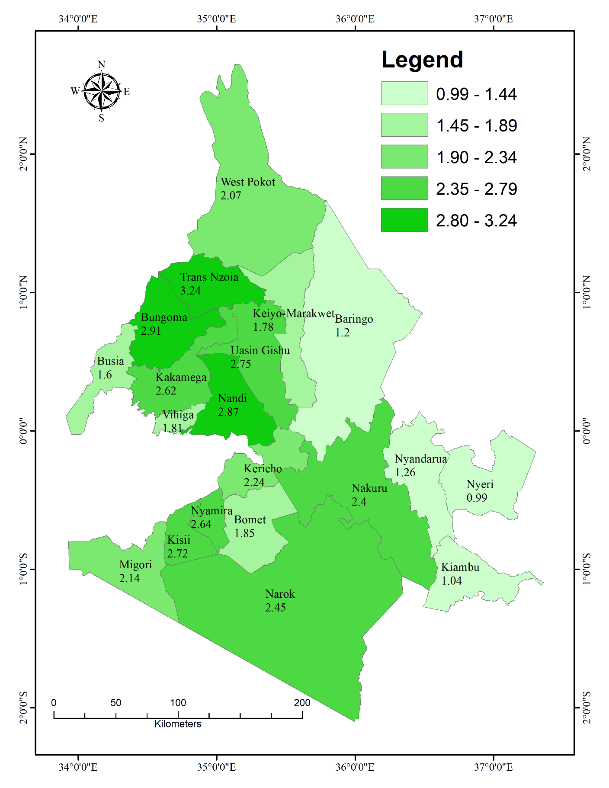
Figure 6 shows the prediction maps for the prediction done by the SVM model. Similar to figures 4 and 5. The counties in the east i.e. Baringo, Nyandarua, Nyeri, and Kiambu, have low yields together with Busia and Vihiga. Figure 6 (a), showing the prediction by the combined vegetation indices and the weather variables counties in the center, north-west, and the south-west regions have yields above 2.2 ton/ha. In Figure 6 (b), predictions using the vegetation indices, counties with high yields are clustered together in the northwest and south regions of the map. The highest being Trans-Nzoia with 3.34 ton/ha followed by Bungoma with 2.88 ton/ha. Figure 6 (c), prediction with the weather variables, counties along the center have an average yield of between 1.6 ton/ha and 2.2 ton/ha. Trans-Nzoia, Nakuru, Migori, and Kisii have yields above 2.4 ton/ha (3.21 ton/ha, 2.45 ton/ha, 2.69 ton/ha, and 2.6 ton/ha respectively).

### Linear regression prediction maps



**(tons/ha)**

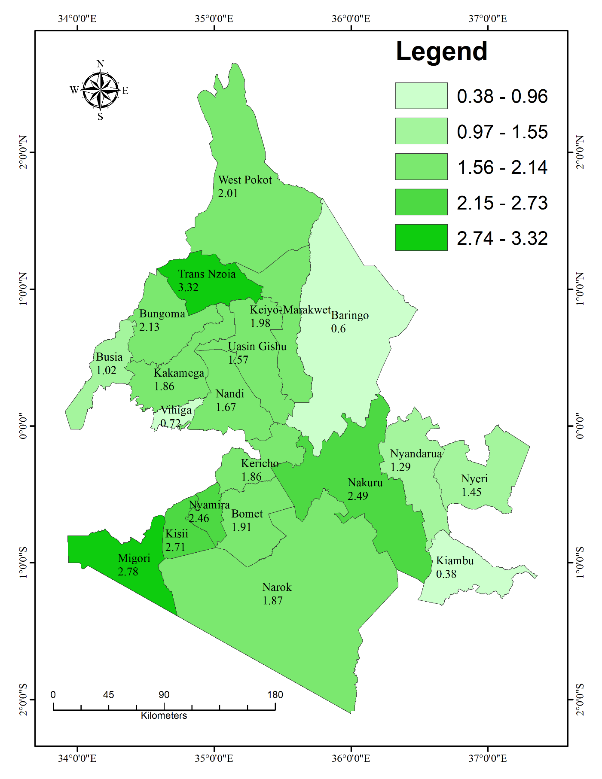
(a)



**(tons/ha)**

Figure 4: Shows the prediction results done by using Linear regression. (a) Shows the prediction map for the prediction done by the combined vegetation indices and the weather variables in ton/ha. (b) Shows the prediction map for the prediction done by the vegetation indices in ton/ha. (c) Shows the prediction map for the prediction done by the weather variables in ton/ha.

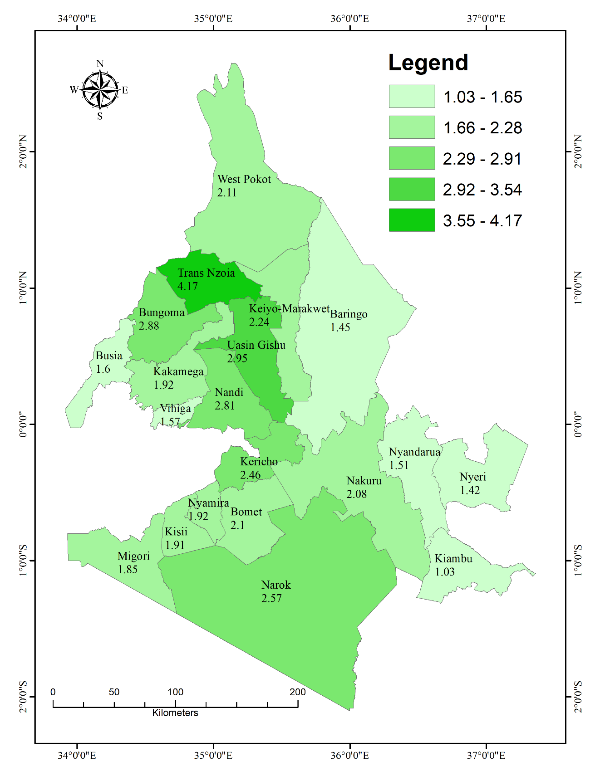
(b)



(c)

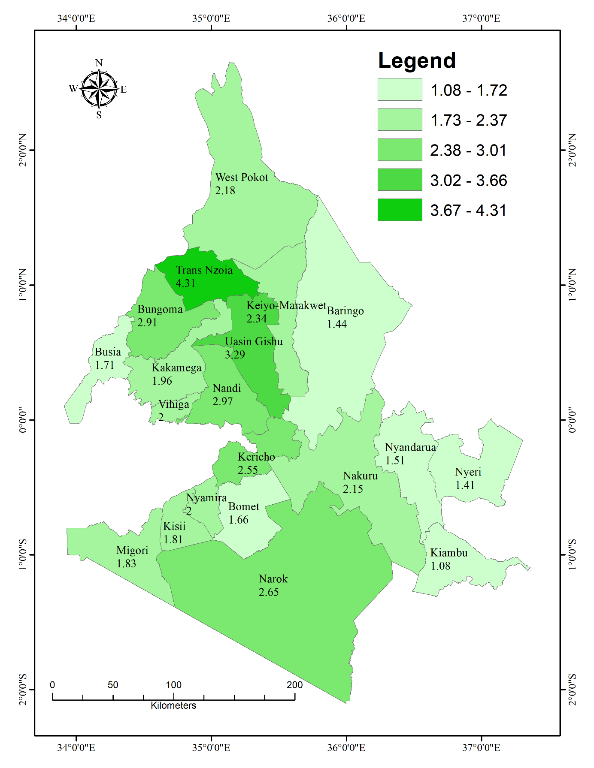
**(tons/ha)**

### Random Forest Prediction maps



**(tons/ha)**

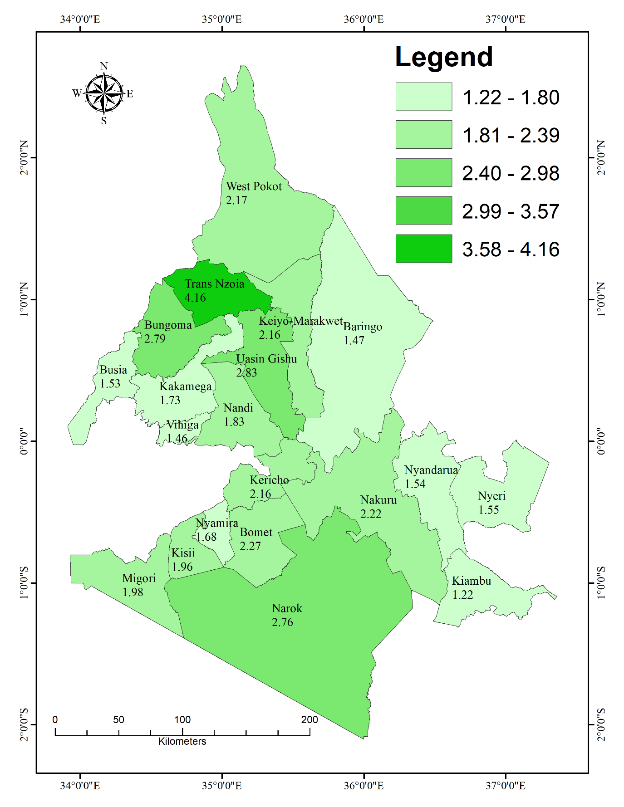
(a)



**(tons/ha)**

(b)

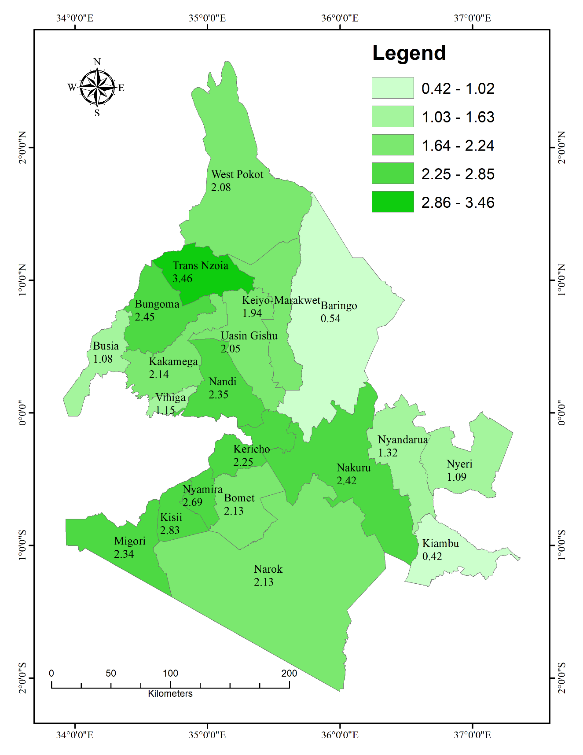
Figure 5: Shows the prediction results done by using Random Forest. (a) Shows the prediction map for the prediction done by the combined vegetation indices and the weather variables in ton/ha. (b) Shows the prediction map for the prediction done by the vegetation indices in ton/ha. (c) Shows the prediction map for the prediction done by the weather variables in ton/ha.



**(tons/ha)**

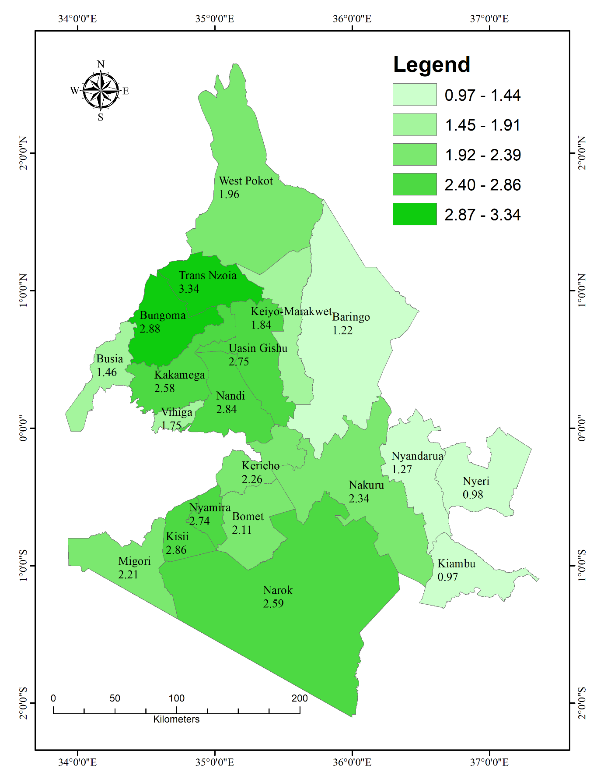
(c)

### Support Vector Machines Prediction maps



**(tons/ha)**

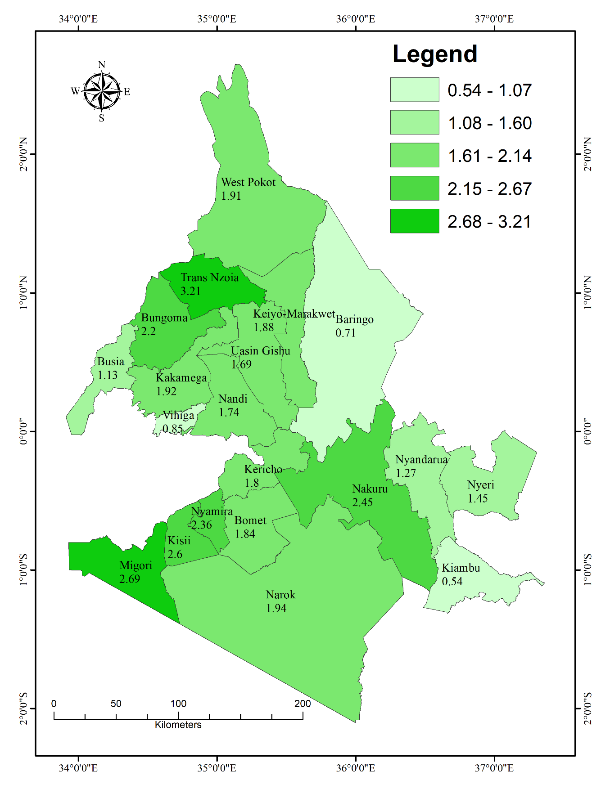
(a)



**(tons/ha)**

Figure 6: Shows the prediction results done by using Support Vector Machines. (a) Shows the prediction map for the prediction done by the combined vegetation indices and the weather variables in ton/ha. (b) Shows the prediction map for the prediction done by the vegetation indices in ton/ha. (c) Shows the prediction map for the prediction done by the weather variables in ton/ha.

(b)



**(tons/ha)**

(c)

## Validation and comparison results for the respective models used

Table 5 shows the MAPE for the models used for prediction. LR has a value of 32.46% for the prediction done with vegetation, when using the weather variables the value increases to 34.95% and then drops when both the indices and the weather variables are used to a value of 31.79%. RF has a value of 29.00% for the prediction done with vegetation, when using the weather variables the value increases to 33.05% and then drops when both the indices and the weather variables are used to a value of 29.52%. SVM has a value of 31.17% for the prediction done with vegetation, when using the weather variables the value increases to 33.97% and then drops when both the indices and the weather variables are used to a value of 32.35%.

Table 6 shows the RMSE for the models used for prediction. LR has a value of 0.70 ton/ha for the prediction done with vegetation, when using the weather variables the value increases to 0.73 ton/ha and then drops when both the indices and the weather variables are used to a value of 0.72 ton/ha. RF has a value of 0.62 ton/ha for the prediction done with vegetation, when using the weather variables the value increases to 0.65 ton/ha and then drops when both the indices and the weather variables are used to a value of 0.60 ton/ha. SVM has a value of 0.69 ton/ha for the prediction done with vegetation, when using the weather variables the value increases to 0.71 ton/ha and then drops when both the indices and the weather variables are used to a value of 0.68 ton/ha.

Table 7 shows the MAPE for the models used for prediction. LR has a value of 0.47 for the prediction done with vegetation, when using the weather variables the value drops to 0.43 and then increases when both the indices and the weather variables are used to a value of 0.49. RF has a value of 0.57 for the prediction done with vegetation, when using the weather variables the value drops to 0.53 and then increases when both the indices and the weather variables are used to a value of 0.61. SVM has a value of 0.46 for the prediction done with vegetation, when using the weather variables the value drops to 0.44 and then increases when both the indices and the weather variables are used to a value of 0.46.

Table 5: Table showing the MAPE (Mean Absolute Percentage Error) of the models used

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Vegetation indices** | **Weather variables** | **Combined** |
| Linear Regression | 32.46% | 34.95% | 31.79% |
| Random Forest | 29.00% | 33.05% | 29.52% |
| Support Vector Machine | 31.17% | 33.97% | 32.35% |

Table 6: Table showing the RMSE (Root Mean Square Error) of the models used

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Vegetation Indices** | **Weather Variables** | **Combined** |
| Linear Regression | 0.70 | 0.73 | 0.72 |
| Random Forest | 0.62 | 0.65 | 0.60 |
| Support Vector Machine | 0.69 | 0.71 | 0.68 |

Table 7: Table showing the R2 of the models used

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Vegetation Indices** | **Weather Variables** | **Combined** |
| Linear Regression | 0.47 | 0.43 | 0.49 |
| Random Forest | 0.57 | 0.53 | 0.61 |
| Support Vector Machine | 0.46 | 0.44 | 0.46 |

# Chapter 5



## Discussion

The correlation results depict that there is a negative correlation between all the variables being used and the yields ranging from -0.29 to 0.46. From this, we can see that the physical weather variables have minimal impact on the yield. The respective correlation coefficients were; NDVI (0.44), EVI (0.46), GPP (0.43), Evapotranspiration (0.14), Rainfall (-0.08), Temperature (-0.07), wind speed (-0.29) and Air pressure (-0.08). There is a positive correlation of the yields with the vegetation indices. As visualized from the results there is an increase in the yields as the vegetation indices increase too. This is similar to the correlation results obtained by (Kenduiywo et al., 2020) who found a positive correlation between the yields and vegetation indices.

Meteorological data had a negative correlation. This occurs since, for example, the rainfall amount received after the planting season has little input to the germination of the crop because for germination the crop depends on the moisture that is already available in the soil. For temperature, since the crops are still young at this stage, higher temperatures would increase the rate of transpiration which when prolonged will lead to water stress in the crop and it may die hence the yields would drop, again low temperatures may lead to frost damage on the crops hence the study has revealed the optimum temperature ranges as being 22℃ to 25℃.

There was a high negative correlation between the yields and the wind speed since high/strong wind speeds blow over the crops and this leads to the death of the crop or even rotting of the crops before the harvesting season. The study in which the wind speed is negatively correlated to the yields implies that more yields are obtained for the lower wind speeds after the planting season during the germination stage of the crops. From the study wind speeds, more than 8km/h had a significant influence on the decrease of the yields.

Evapotranspiration which is the combination of transpiration and evaporation affects the yield in a manner that the more the evapotranspiration the lower the yields obtained since when the transpiration increases it leads to water stress for the small crops, the applies for the evaporation of the water from the soil which will, in turn, leave the crops with less water to absorb by the roots.

From the model validation results, the LR model was used as a baseline to check the performance of the other models being used and of the three models used. RF performed better in comparison with LR which was the baseline of the study with a MAPE of 29.00% with the vegetation indices. LR performed poorly with a MAPE of 34.95% for the predictions done with the weather variables. This was similar to what (Bouras et al., 2021) found where RF performed better than LR.

From the study, we can see that the addition of variables affects the performance of the different models. More variables can improve or reduce the accuracy of the model. As seen the addition of the weather variables reduces the performance of the model as used with only the vegetation indices.

In summary, the study revealed that rainfall does not have a huge impact on crop growth and that evapotranspiration plays an important role in crop growth which can also be tied to temperature. The wind has a large factor in the growth of the crops and air pressure has the least significance to the growth of the maize crop.

# Chapter 6



## Conclusion and outlook

Crop yield forecasting gives crucial and timely information to farmers, allowing them to make timely decisions to boost yields by optimizing agricultural practices during the growing season. Furthermore, it enables the modelling of global and local market pricing. The primary goal of our research was to create a method for forecasting cereal yield in Kenya using multi-source data and machine learning techniques. To that end, this paper proposes a methodology based on multiple machine learning approaches (LR, SVM, RF) for predicting maize production in Kenya before harvest in 2021 utilizing openly available information such as satellite-based vegetation indices and meteorological data. Our findings reveal that integrating satellite-based vegetation indices and meteorological data as maize yield predictors gave superior predicting accuracy than using any single data source. Several research e.g. (Bouras et al., 2021) and (Mupangwa et al., 2020) have found that using multi-source data improves the accuracy of the machine learning model for yield prediction, which is consistent with our findings. RF outperformed the other models.

The study's findings demonstrated that combining satellite-based vegetation indicators and meteorological data incorporated into machine learning algorithms is a potential technique for forecasting maize yields in Kenya. Furthermore, the proposed approach provides a source of timely information required for decision-making during agricultural budget allocation. Other variables could be added in future studies to improve the models' performance. Soil type and farm management strategies (water harvesting techniques, complementing irrigation, fertilizer inputs, planting dates, and so on) are important variables in crop growth.

# References

Amudavi, D., Hunziker, M., & Shiraku, H. (2017). *Maize | Infonet Biovision Home.* https://infonet-biovision.org/PlantHealth/Crops/Maize

Bouras, E. H., Jarlan, L., Er-Raki, S., Balaghi, R., Amazirh, A., Richard, B., & Khabba, S. (2021). Cereal yield forecasting with satellite drought-based indices, weather data, and regional climate indices using machine learning in morocco. *Remote Sensing*, *13*(16). https://doi.org/10.3390/rs13163101

Cleugh, H. A., Miller, J. M., & Böhm, M. (1998). Direct mechanical effects of wind on crops. *Agroforestry Systems*, *41*(1), 85–112. https://doi.org/10.1023/A:1006067721039

Climate, W. and. (2021). *World weather and climate*. https://tcktcktck.org/page/term

FAO. (2021). *WAPOR Actual Evapotranspiration and Interception*. https://www.fao.org/land-water/databases-and-software/wapor/en/

Folnovic, T. (2021). *Importance of Weather Monitoring in Farm Production*. https://blog.agrivi.com/post/importance-of-weather-monitoring-in-farm-production

Hassan, R. M. (1996). Planting strategies of maize farmers in Kenya: A simultaneous equations analysis in the presence of discrete dependent variables. *Agricultural Economics*, *15*(2), 137–149. https://doi.org/10.1016/S0169-5150(96)01194-2

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K. M., Gerber, J. S., Reddy, V. R., & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE*, *11*(6), 1–15. https://doi.org/10.1371/journal.pone.0156571

Kenduiywo, B. K., Ghosh, A., Hijmans, R., & Ndungu, L. (2020). Maize yield estimation in Kenya using modis. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*(3), 477–482. https://doi.org/10.5194/isprs-Annals-V-3-2020-477-2020

Laudien, R., Schauberger, B., Gleixner, S., & Gornott, C. (2020). Assessment of weather-yield relations of starchy maize at different scales in Peru to support the NDC implementation. *Agricultural and Forest Meteorology*, *295*(December 2019), 108154. https://doi.org/10.1016/j.agrformet.2020.108154

Ma, J., & Maystadt, J. F. (2017). The impact of weather variations on maize yields and household income: Income diversification as adaptation in rural China. *Global Environmental Change*, *42*, 93–106. https://doi.org/10.1016/j.gloenvcha.2016.12.006

MOALFI. (2020). *Kenya Maize Production By Counties*. 10 September 2019, 23:06 (UTC+03:00). http://kilimodata.developlocal.org/dataset/kenya-maize-production-by-counties

MODIS. (2021). *Moderate Resolution Imaging Spectroradiometer data*. https://modis.gsfc.nasa.gov/data/dataprod/mod13.php

Mupangwa, W., Chipindu, L., Nyagumbo, I., Mkuhlani, S., & Sisito, G. (2020). Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa. *SN Applied Sciences*, *2*(5), 1–14. https://doi.org/10.1007/s42452-020-2711-6

Onainor, E. R. (2019). *Why Evapotranspiration Affects Crop Yield*. https://aquaspy.com/why-evapotranspiration-affects-crop-yield/

Piao, S., Research; C, S. A. Z., Peng, S., Wallach, D., & Wu, D. (2017). *performed research and analyzed data*. *114*(35). https://doi.org/10.1073/pnas.1701762114

Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting Corn Yield With Machine Learning Ensembles. In *Frontiers in Plant Science* (Vol. 11). https://doi.org/10.3389/fpls.2020.01120

Sun, B. J., & Van Kooten, G. C. (2014). Weather effects on maize yields in northern China. *Journal of Agricultural Science*, *152*(4), 523–533. https://doi.org/10.1017/S0021859613000142

Sun, H., Zhang, X., Wang, E., Chen, S., Shao, L., & Qin, W. (2016). Assessing the contribution of weather and management to the annual yield variation of summer maize using APSIM in the North China Plain. *Field Crops Research*, *194*, 94–102. https://doi.org/10.1016/j.fcr.2016.05.007

Suresh, A., Manjunathan, N., Rajesh, P., & ... (2020). Crop Yield Prediction Using Linear Support Vector Machine. *… Journal of Molecular & …*, *7*(06), 2189–2195. https://ejmcm.com/article\_4024.html

Yin, X. G., Jabloun, M., Olesen, J. E., Öztürk, I., Wang, M., & Chen, F. (2021). *Effects of climatic factors, drought risk, and irrigation requirement on maize yield in the Northeast Farming Region of China*. https://doi.org/10.1017/S0021859616000150

# Appendix



## Combined vegetation indices and weather variables.

setwd("C:/Users/USER/Desktop/Project/Data/Maize data")  
  
library(caret)

library(tidyverse)

Yield = read.csv("Maize Yield April fin.csv")

#random seed  
set.seed(1988)  
  
#validation method  
fitControll = trainControl(method = "repeatedcv", number = 5,  
 repeats = 10, savePredictions = TRUE)  
  
#set random seed  
set.seed(1992)  
  
#model used -> linear regression  
LR <- train(YIELD ~ ., method = "lm",   
 trControl = fitControll, data = Yield, verbose = FALSE)

#MAPE  
mape\_LR = MAPE(LR$pred$pred, LR$pred$obs)  
  
#Random Forest  
set.seed(1254)  
  
RF <- train(YIELD ~ ., data = Yield, method = "rf",   
 trControl = fitControll, verbose = FALSE)  
print(RF)

#MAPE  
mape\_RF = MAPE(RF$pred$pred, RF$pred$obs)  
  
#Support Vector Machine  
set.seed(546)  
  
SVM <- train(YIELD ~ ., data = Yield, verbose = FALSE,  
 method = "svmLinear2", trControl = fitControll)

print(SVM)

#mape  
mape\_SVM = MAPE(SVM$pred$pred, SVM$pred$obs)  
  
#prediction  
Data <- read.csv("2021 weather.csv")  
Data = Data[,-1]  
  
preds\_LR <- predict(LR, newdata = Data)  
Data$LR <- preds\_LR  
  
preds\_RF <- predict(RF, newdata = Data)  
Data$RF <- preds\_RF  
  
preds\_SVM <- predict(SVM, newdata = Data)  
Data$SVM <- preds\_SVM  
  
#saving  
write.csv(Data, file = 'yields 2021 final.csv')

## Vegetation indices

setwd("C:/Users/USER/Desktop/Project/Data/Maize data/modis")  
  
indices = read.csv("data.csv")  
  
library(MLmetrics)

library(caret)

library(tidyverse)

colnames(indices)

#random seed  
set.seed(1787)  
  
#validation method  
fitControll = trainControl(method = "repeatedcv", number = 5,  
 repeats = 10, savePredictions = TRUE)  
  
#set random seed  
set.seed(2485)  
  
#model used -> linear regression  
LR\_I <- train(YIELD ~ ., method = "lm",   
 trControl = fitControll, data = indices, verbose = FALSE)

print(LR\_I)

#mape  
mape\_LR\_I = MAPE(LR\_I$pred$pred, LR\_I$pred$obs)  
  
  
#Random Forest  
set.seed(8954)  
  
RF\_I <- train(YIELD ~ ., data = indices, method = "rf",   
 trControl = fitControll, verbose = FALSE)  
print(RF\_I)

#MAPE  
mape\_RF\_I = MAPE(RF\_I$pred$pred, RF\_I$pred$obs)  
  
  
#Support Vector Machine  
set.seed(7546)  
  
SVM\_I <- train(YIELD ~ ., data = indices, verbose = FALSE,  
 method = "svmLinear2", trControl = fitControll)  
  
print(SVM\_I)

#mape  
mape\_SVM\_I = MAPE(SVM\_I$pred$pred, SVM\_I$pred$obs)  
  
setwd("C:/Users/USER/Desktop/Project/Data/Maize data/modis")  
  
mod\_ind = read.csv("indices.csv")  
mod\_ind = mod\_ind[,-1]  
  
preds\_ind = predict(LR\_I, newdata = mod\_ind)  
mod\_ind$LR = preds\_ind  
  
preds\_ind2 = predict(RF\_I, newdata = mod\_ind)  
mod\_ind$RF = preds\_ind2  
  
preds\_ind3 = predict(SVM\_I, newdata = mod\_ind)  
mod\_ind$SVM = preds\_ind3  
  
write.csv(mod\_ind, file = "PREDICTED.csv")

## Weather Variables

setwd("C:/Users/USER/Desktop/Project/Data/Maize data")  
  
Weather = read.csv("Weather.csv")  
  
library(MLmetrics)

library(caret)

library(tidyverse)

#random seed  
set.seed(1787)  
  
#validation method  
fitControll = trainControl(method = "repeatedcv", number = 5,  
 repeats = 10, savePredictions = TRUE)  
  
#set random seed  
set.seed(1945)  
  
#model used -> linear regression  
LR\_V <- train(YIELD ~ ., method = "lm",   
 trControl = fitControll, data = Weather, verbose = FALSE)

print(LR\_V)

#mape  
mape\_LR\_V = MAPE(LR\_V$pred$pred, LR\_V$pred$obs)  
  
  
#Random Forest  
set.seed(1254)  
  
RF\_V <- train(YIELD ~ ., data = Weather, method = "rf",   
 trControl = fitControll, verbose = FALSE)  
print(RF\_V)

#MAPE  
mape\_RF\_V = MAPE(RF\_V$pred$pred, RF\_V$pred$obs)  
  
  
#Support Vector Machine  
set.seed(546)  
  
SVM\_V <- train(YIELD ~ ., data = Weather, verbose = FALSE,  
 method = "svmLinear2", trControl = fitControll)  
  
print(SVM\_V)

#mape  
mape\_SVM\_V = MAPE(SVM\_V$pred$pred, SVM\_V$pred$obs)  
  
setwd("C:/Users/USER/Desktop/Project/Data/Maize data/weather")  
  
Weth = read.csv("weather 21.csv")  
Weth = Weth[,-1]  
  
preds\_we = predict(LR\_V, newdata = Weth)  
Weth$LR = preds\_we  
  
preds\_we2 = predict(RF\_V, newdata = Weth)  
Weth$RF = preds\_we2  
  
preds\_we3 = predict(SVM\_V, newdata = Weth)  
Weth$SVM = preds\_we3  
  
write.csv(Weth, file = "PREDICTIONS Final.csv")

## Plots

setwd("C:/Users/USER/Desktop/Project/Data/Maize data")  
#library  
library(ggplot2)  
  
#data  
Yield = read.csv("Maize Yield April fin.csv")  
  
par(mfrow=c(2,2), mai=c(0.75,0.75,0.1,0.1))  
  
#NDVI  
plot(Yield$NDVI, Yield$YIELD, pch = 16, col = "blue",   
 ylab= "Yield(ton/ha)", xlab="NDVI", cex = 0.9, cex.axis=1.2, cex.lab=1.2)  
abline(lm(Yield$YIELD ~ Yield$NDVI), col = "red", lwd = 3)  
text(paste("r:", round(cor(Yield$NDVI, Yield$YIELD), 2)),   
 x = 0.58, y = 4.5, cex=1.5)  
  
#EVI  
plot(Yield$EVI, Yield$YIELD, pch = 16, col = "blue",   
 ylab= "Yield(ton/ha)", xlab="EVI", cex = 0.9, cex.axis=1.2, cex.lab=1.2)  
abline(lm(Yield$YIELD ~ Yield$EVI), col = "red", lwd = 3)  
text(paste("r:", round(cor(Yield$EVI, Yield$YIELD), 2)),   
 x = 0.325, y = 4.5, cex=1.5)  
  
#GPP  
plot(Yield$GPP, Yield$YIELD, pch = 16, col = "blue",   
 ylab= "Yield(ton/ha)", xlab="GPP", cex = 0.9, cex.axis=1.2, cex.lab=1.2)  
abline(lm(Yield$YIELD ~ Yield$GPP), col = "red", lwd = 3)  
text(paste("r:", round(cor(Yield$GPP, Yield$YIELD), 2)),   
 x = 0.35, y = 4.5, cex=1.5)  
  
#Evapotranspiration  
plot(Yield$Evapotranspiration, Yield$YIELD, pch = 16, col = "blue",   
 ylab= "Yield(ton/ha)", xlab="Evapotranspiration", cex = 0.9, cex.axis=1.2, cex.lab=1.2)  
abline(lm(Yield$YIELD ~ Yield$Evapotranspiration), col = "red", lwd = 3)  
text(paste("r:", round(cor(Yield$Evapotranspiration, Yield$YIELD), 2)),   
 x = 23, y = 4.5, cex=1.5)

#Rainfall  
plot(Yield$Rainfall, Yield$YIELD, pch = 16, col = "blue",   
 ylab= "Yield(ton/ha)", xlab="Rainfall", cex = 0.9, cex.axis=1.2, cex.lab=1.2)  
abline(lm(Yield$YIELD ~ Yield$Rainfall), col = "red", lwd = 3)  
text(paste("r:", round(cor(Yield$Rainfall, Yield$YIELD), 2)),   
 x = 800, y = 4.5, cex=1.5)  
  
#Temperature  
plot(Yield$Temperature, Yield$YIELD, pch = 16, col = "blue",   
 ylab= "Yield(ton/ha)", xlab="Temperature", cex = 0.9, cex.axis=1.2, cex.lab=1.2)  
abline(lm(Yield$YIELD ~ Yield$Temperature), col = "red", lwd = 3)  
text(paste("r:", round(cor(Yield$Temperature, Yield$YIELD), 2)),   
 x = 24, y = 4.5, cex=1.5)  
  
#Wind speed  
plot(Yield$Wind.Speed, Yield$YIELD, pch = 16, col = "blue",   
 ylab= "Yield(ton/ha)", xlab="Wind.Speed", cex = 0.9, cex.axis=1.2, cex.lab=1.2)  
abline(lm(Yield$YIELD ~ Yield$Wind.Speed), col = "red", lwd = 3)  
text(paste("r:", round(cor(Yield$Wind.Speed, Yield$YIELD), 2)),   
 x = 11, y = 4.5, cex=1.5)  
  
#Air pressure  
plot(Yield$Air.Pressure, Yield$YIELD, pch = 16, col = "blue",   
 ylab= "Yield(ton/ha)", xlab="Air.Pressure", cex = 0.9, cex.axis=1.2, cex.lab=1.2)  
abline(lm(Yield$YIELD ~ Yield$Air.Pressure), col = "red", lwd = 3)  
text(paste("r:", round(cor(Yield$Air.Pressure, Yield$YIELD), 2)),   
 x = 27, y = 4.5, cex=1.5)