



**Prediction of Tea Production in Rwanda using Data Mining Techniques**

**By**

**Clarisse UMUTONI**

**Registration Number: 213001359**

**A dissertation submitted in partial fulfillment of the requirements for the degree of  
Master of Science in Data Science.**

**University of Rwanda, College of business and Economics  
The African Center of Excellence in Data Science (ACE-DS)**

**Supervisor: Innocent NGARUYE, Ph.D.**

**September, 2020**

## **Declaration**

I declare that this dissertation entitled “Prediction of tea production in Rwanda using data mining techniques “ is the result of my own work and has not been submitted for any other degree at the University of Rwanda or any other institution.

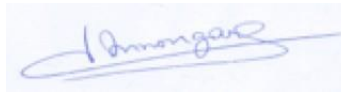
**Names : Clarisse UMUTONI**



**Signature:**

## Approval sheet

This dissertation entitled “Prediction of tea production in Rwanda using data mining techniques” written and submitted by Clarisse UMUTONI in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in Data Mining is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 16 % which is less than 20% accepted by ACE-DS.



---

**Dr. Innocent NGARUYE**

Supervisor



---

**Dr. Ignace KABANO**

Head of Training

## **Dedication**

This work is dedicated my Parents, my lovely Daughter Medica, Brothers and sisters for their Understanding, patience, encouragement and support while pursuing this course

## **Acknowledgments**

In the first place, I thank God the Almighty who has provided the breath of life to enable The successful completion of this research project. I am grateful to the Government of Rwanda in general and the National University of Rwanda in particular, for allowing me to undertake this master's degree course in Data science specialized in data mining.

Also, I highly appreciate the financial support provided by world bank through the African Center of Excellence in Data Science. Without this support, I would not have managed to complete my master's program. My heartfelt thanks also go to the entire teaching staff of the data science and particularly to my supervisor, Dr. NGARUYE Innocent who have shaped me to be what I am as far as data science are concerned.

My appreciation also goes to my parents who have provided all that is required to make me reach this academic level. Their parental guidance and moral support have been a cornerstone to my academic achievement. Finally, I would also wish to recognize the administrative staff as well as the entire team of my classmates who have been supportive in one way or another. May God bless them abundantly.

## **Abstract**

Agriculture is the main economic activity in Rwanda and tea is major cash crop in Rwanda. There has been extensive research on prediction of tea production but most of the methods applied were the traditional statistical analyzes with limited prediction capability. Data mining algorithm models, linear regression, K-Nearest Neighbor (KNN), Random Forest Regression, Extremely Randomised Trees are discussed in this study to identify critical features in different domains to facilitate accurate prediction of tea production in Rwanda. In this study also, I identified different factors which are strongly associated with tea production and developed data mining models for predicting tea production using training and test data from National Agricultural Export Development Board (NAEB) 2010-2019. The findings reveal that random forest is the best model among the others to predict tea production in Rwanda.

**Keywords:** Tea production, Data mining, model accuracy, Rwanda

## Contents

Declaration .....	i
Approval sheet.....	ii
Dedication .....	iii
Acknowledgments.....	iv
Abstract.....	v
CHAPTER ONE: GENERAL INTRODUCTION .....	1
1.1. Background of the study .....	1
1.2. Problem statement .....	2
1.3. Research objective.....	3
1.3.1. General Objective .....	3
1.3.2. Specific Objectives .....	3
1.4. Research Questions.....	3
1.5. Scope and limitation of the study .....	3
1.6. Report organization .....	4
CHAPTER TWO: LITERATURE REVIEW.....	5
2.2. Conceptual framework .....	6
2.3. Definition of concepts and variables specification .....	7
2.3.1. Dependent variable .....	7
2.3.2. Independent variable.....	7
2.3.2.1. Climate change .....	7
2.3.2.2. Fertilizer .....	7
2.3.2.3. Seedlings .....	8
2.3.2.4. Area under plantation .....	8
CHAPTER THREE: METHODOLOGY .....	9
3.2. Data.....	9
3.3. Data processing .....	9
3.4. Data Mining Models.....	9
3.4.1. Linear Regression Model.....	9
3.4.2. K-Nearest Neighbor regressor .....	10

3.4.3. Random Forest Regressor .....	11
3.4.4. Extremely Randomised Trees (Extra trees) Regressor .....	11
3.5. Evaluation Criterion .....	11
3.5.1. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) .....	11
3.5.2. Coefficient of Determination $R^2$ .....	12
CHAPTER FOUR: DATA ANALYSIS .....	13
4.1. Checking relationship using scatterplot .....	13
4.2. Ordinary Least Square Regression (OLS) .....	19
4.4.1. OLS regression for rainfall predictor .....	19
4.4.2. OLS regression for fertilizer predictor .....	20
4.1.3. OLS regression for seedling predictor .....	21
4.4.4. OLS regression for area under plantation predictor .....	23
CHAPTER FIVE: MAIN RESULTS AND DISCUSSION .....	25
5.1.1. Data Mining Models Comparison using Test Data before Hyperparameter Tuning .....	25
5.1.2. Feature Importance using Extra Tree Regressor .....	27
5.2. Discussion .....	27
CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS .....	29
6.2. Recommendations .....	30
REFERENCE .....	31
APPENDICES .....	33



## List of figures

Figure 1:Conceptual framework.....	6
Figure 2:Scatter plot of rainfall against production .....	13
Figure 3:Scatter plot of month against production .....	14
Figure 4:Scatter plot of fertilizer against production .....	15
Figure 5:Scatter plot of fertilizer against production .....	16
Figure 6:Scatter plot of area under plantation against production .....	17
Figure 7:Scatter plot of year against production .....	18
Figure 8:Feature importance.....	27
Figure 9:Correlation Matrix .....	33
Figure 10:Effect of Fertilizer on Tea production at area under plantation .....	34
Figure 11:Feature importance of tea production using random forest regressor .....	34
Figure 12:Using R2 to check relationship between actual values against predicted values .....	35

**List of Tables**

Table 1:OLS Regression outcome for rainfall ..... 19

Table 2:OLS Regression outcome for rainfall ..... 20

Table 3:OLS Regression outcome for seeding ..... 23

Table 4:DataMiningModelsComparisonusingTestDataBeforeHyperparameterTuning ..... 26

Table 5:Data Mining Models Comparison using Test Data after Hyper Parameter Tuning ..... 26

Table 6:Data Mining Models Comparison using Training Dataset..... 33

## CHAPTER ONE: GENERAL INTRODUCTION

### 1.1. Background of the study

Agriculture is main economic activity in Rwanda and the population engaged in this sector are at the rate of 70%, around 72% of the working population in Rwanda are employed in agriculture. Agricultural sector in Rwanda accounts for 33% of the national Gross Domestic Product (GDP). In general, Rwanda's GDP has been growing at the rate of 7% since 2014. Tea and coffee are the major export crops contributing to this economic growth and tea was introduced in Rwanda in 1961 and today it is currently grown on 26,897 hectares by 42,840 farmers located in 12 districts of mainly in the Northern, Western and Southern parts of the country (NAEB,2018).

The majority of tea seedlings are cultivated on large area under plantation, with a small contribution from tea cooperatives and private growers. Tea plants can be seen covering the whole rolling hills, their rich green are striking contrast to the blue skies, dirt roads and sunshine.

Tea leaves are processed in a dozen tea factories across the country. These factories are open to the public, enabling visitors to discover how tea is harvested and processed, with the opportunity to taste the results. Rwandan tea is planted on hillsides at a high altitude between 1,900m and 2,500m, and also on well drained marshes at an altitude between 1,550m and 1,800m. The production of tea has increased steadily, from 60 Metric Tons of tea in 1958 to about 30,000 metric tons annually nowadays. Rwanda tea is actually known to be better because of its high quality and it is among the best in the world. Some of the best tea qualities produced in Rwanda are the following: black tea, orthodox tea, white tea, green tea, organic tea and spiced tea.

Rwanda tea to day has become highly valued in the weekly East African Tea Trade Association auctions in Mombasa, fetching record prices over the past couple of years. Its major markets are actually in the Middle East, Pakistan, Kazakhstan and United Kingdom.

As tea is one of the major export in Rwanda, this study is aimed to predict tea production in Rwanda by using data mining techniques. Data mining is a process used to extract usable data from a large set of any raw data and Data mining techniques consist of the process for extracting important and useful information from large data sets and is a relatively new inter-disciplinary concept involving data analysis and knowledge discovery

from the databases(Jambekar et al., 2018). Also data mining is a process for analyzing the large datasets, it includes several steps such as analyzing, classification and clustering of the data(Veeresh and Saboji, 2019).

## **1.2. Problem statement**

Rwanda's tea production increased from 14,500 tons in 2000 to 25,128 tons in 2016/2017. From July 2016 to June 2017, Rwanda fetched \$74.5 million while in 2017-2018 the country exported 27,824 metric tons of made in Rwanda tea to 48 countries, generating \$88 million.

The statistics from the 2018/2019 report by the National Agricultural Export Development Board (NAEB) shows that there was a decline of 9.7 per cent in agricultural export earnings compared to the previous year. According to above findings it's clear that tea sector plays a big role in national GDP, for that increasing tea production at maximum is the main goal to this sector in general.

Several different traditional methods and models about tea prediction have been applied in research but they have limited prediction capability. Applying data mining techniques to analyze tea production in Rwanda will be useful for identifying the strongest factors associated with tea production and predicting future tea production in Rwanda. To my knowledge, there is no study conducted in Rwanda by using data mining techniques about prediction of tea production. Therefore, this study aims to analyze factors associated to tea production in Rwanda and predicting future tea production by using data mining techniques. The use of data mining techniques in tea sector will help to reveal hidden information that different partners in this sector can use in their daily decision making and for tea farmers in general.

### **1.3. Research objective**

#### **1.3.1. General Objective**

The general objective of this project is to predict future tea production in Rwanda by using data mining techniques

#### **1.3.2. Specific Objectives**

The target specific objectives of this study are the following:

- To examine relationship between fertilizers, climatic change and tea production in Rwanda
- To compare various data mining techniques in predicting tea production
- To Find most robust data mining model in predicting tea production in Rwanda
- To Find important features contributing tea production in Rwanda

### **1.4. Research Questions**

This research attempts to provide answers to the following research questions:

1. Is there any relationship between climate change, fertilizer, tea seedlings, area under plantation and tea production?
2. What are important features contributing tea production in Rwanda?
3. Is there any predictive model among data mining techniques that performs better than others in predicting tea production?

### **1.5. Scope and limitation of the study**

This study focuses on the use of data mining techniques to predict tea production in Rwanda by using secondary data collected from National Agriculture Export development

Board (NAEB) 2010-2019. By building predictive model, I choose to use different data mining models like linear regression, K-Nearest Neighbor (KNN) Regressor, Random Forest Regressor and Extremely Randomized Trees (Extra trees) Regressor in order to check which one will be better to be used for predicting tea production in Rwanda.

## 1.6. Report organization

The structure of this dissertation is composed by 6 chapters as follow:

- **Chapter 1-** In chapter we will talk about back ground of the study what is a contribution of agriculture in economic in general and also we talk about how data mining have model which is best to predict future tea production.
- **Chapter 2 - :** In this chapter we will focus on the views of other authors and scholar about tea production prediction by using different methods and models.
- **Chapter 3 - :** This chapter deals with data will be used in this thesis, data pre processing, data mining models which is used to predict tea production in Rwanda and also evaluation criterion which will be used to evaluate our models.
- **Chapter4-:**This chapter will used to extract useful information that will be used by making decision that is why we will use scatter plot by checking relationship between different features, and it contain also OLSRegression results for all features.
- **Chapter 5 - :** This chapter will show the result of our analysis and the discussion of those results.
- **Chapter 6 - :**In this chapter will focus on the conclusion of our study and what to recommend to the future researchers.

## CHAPTER TWO: LITERATURE REVIEW

### 2.1.Introduction

This part is concerned with the views of other authors on the prediction of tea production using data mining technique in different country and on other related studies. In their study on Prediction of Tea Production in Kenya Using Clustering and Association Rule Mining Techniques, Nzuva and Lawrence (2017) explained how Kenya depends on agriculture sector more than the other economic sectors, where the production of food remains a top priority in the whole country. Therefore, the main focus of their study was to outset any relationship in tea production of different months of the year, from 2003 to 2015. As result, they found out that in order to enhance tea production, plan for the future production and increase profitability of the ventures, the tea farmers need more to understand the trends in the production, how it is consumed and the process of exportation.

According to their findings in the paper about data mining discussion in the agricultural discipline, Rudyy (2001) explain how data mining can greatly help in linking the knowledge gained from the mined data to agricultural yields estimation. This is affirmed by Vamanan and Ramar(2011) who assert that classification approach in data mining can be applied to soil and crop datasets to establish any meaningful association between variables in the dataset applied different mining techniques on the identified variables to outset the existence of any meaningful relationships.

In a study conducted by Sitienei et al. (2017) about using regression model to predict tea crop yield response to climate change in Nandi Kenya county, authors explain how they use statistical model. The statistical model was trained on historical tea yields, and how they related to the past data on maximum temperature, minimum temperature and precipitation over Nandi East Sub-County. Scatter diagrams for selected months, a multiple linear model was developed to predict tea yield using climatic variables and the objectives of their study were to examine the ability of regression models to predict tea yield responses to changes in maximum, minimum temperature and precipitation

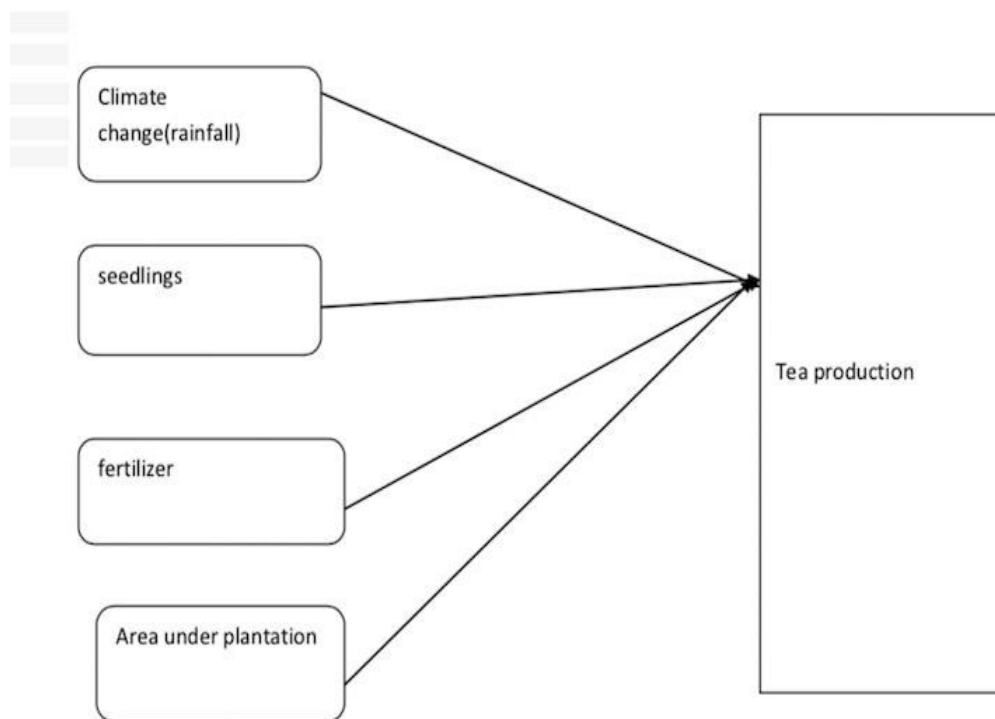
Data mining techniques are also used in India in the paper by Jambekar et al. (2018) about prediction of crop production in India by application of data mining techniques. Authors have explained clearly how agriculture is the most important sector in India so that it is

better to apply data mining techniques which is a process of extracting useful information from the data to help them to improve performance of agriculture sector in the country. That is why their study was focusing on application of the data mining techniques to predict future production of crops such as Rice, Wheat and Maize with respect to various climatic conditions. They have been trying to use different data mining algorithms in order to choose the best one. They conclude by saying that, in the future, one would apply these algorithms for prediction of crop production and find the accuracy of these algorithms to compare which one gives better result.

Following the methods and techniques used by fore mentioned authors, I want to conduct my study by using regression algorithms to predict tea production in Rwanda.

## 2.2. Conceptual framework

Conceptual framework is a structure which the researchers believe can be the best to explain the natural progression of the phenomenon to be studied.



**Figure 1: Conceptual framework**



## **2.3. Definition of concepts and variables specification**

### **2.3.1. Dependent variable**

**Tea production:** is a dependent variable which is defined by how many tones of tea produced per month Tea production is in the form of green tea leaves that are harvested and then processed to be ready for consumption. Tea production is cultivated in 3 provinces in Rwanda. Those are the following: Northern province, Southern province and Western province where climate conditions are favorable for growing tea. The tea green leaves have been harvested on both industrial block and out growers' blocks.

Tea is usually at auctions in the country of origin or at commodity markets in Europe or the United States and simple sold in deals made between producers and buyers. In the old days tea was sold in chests and buyers drilled holes in the chests to sample the qualityfactsanddetails.com (2020).

### **2.3.2. Independent variable**

Those are variables related to dependent variables in regression equation. Independent variables in this study are climate change, seedlings, area of plantation and fertilizer.

#### **2.3.2.1. Climate change**

Climate is the average of the weather conditions at particular point on the earth. Typically, climate is expressed in terms of expected temperature, rainfall and wind conditions based on historical observations. "Climate change" is a change in either the average climate or climate variability that persists over an extended period.(Chris Riedy,2016) In my study I was consider maximum rainfall and minimum rainfall for compare which one influencing tea production positively.

#### **2.3.2.2. Fertilizer**

Mineral fertilizer is the fertilizer made by chemical products processed to meet crop requirements to supply plant nutrients in exact, scientifically formulated quantities. It should be also used together with organic fertilizers which improve the structure of the soil and soil water holding capacity. So, the precision that manufactured mineral fertilizers offer helps to overcome the limitation of organic fertilizer.

Organic fertilizer: These are fertilizers that are made in products which are derived from

the remaining of living organisms or other products like trash of plants and animals.

#### **2.3.2.3. Seedlings**

These are seeds of tea they use to plant in plantation

#### **2.3.2.4. Area under plantation**

It is the surface where they have to plant tea seedling

## **CHAPTER THREE: METHODOLOGY**

### **3.1. Introduction**

The parts that will be handled in this section include data description and its source, data pre-processing, models and techniques used to achieve at the objectives and software to be utilized. This study used various data mining techniques to achieve its objectives.

### **3.2.Data**

The data for this study was obtained from National Agriculture Export Development Board (NAEB), the government organization in Rwanda in charge of managing exportation of agricultural products. The original identified dataset consisted of 114 observations collected between 2010 and 2019. The features in data set which were used to predict tea production were: year, month rainfall (mm), seedling(seed), fertilizer (kg) and area under plantation ( ha) The response variable was the production.

### **3.3. Data processing**

Data was obtained in excel format; however, missing values, noise and outliers were common in the data. Therefore, the data was cleaned removing noise and outliers. Outliers was removed by using z core method and missing values were handled by imputation. Moreover, cross validation was carried out in order to ease analysis. The 10 Kfolds were identified to be more optimal and give better accuracy than others. Cleaning the dataset reduced misclassification and ensure improved model performance.

### **3.4.Data Mining Models**

This study used supervised data mining techniques to predict tea production in Rwanda Some of data mining models that were used include; linear regression (multiple linear regression), K-Nearest Neighbor, random forest and Extra trees.

#### **3.4.1.Linear Regression Model**

There are two types of linearity that are linear to variable and linear to parameter. Model is said to be linear model if it is linear to the parameter (Permai and Tanty, 2018). The model which may not be linear to the variable, but as long as it is linear to the parameter, then it is still a linear model (Permai and Tanty, 2018). Multiple linear regression can be utilized to evaluate the relationship or correlation between response

variables (tea production) with two or more input features (independent variables in our case are rainfall, weather). When modeling linear regression, it should be ensured that correlation or relationship between each predictor (independent variable) to the response variable is linear. In this study we have one response variable (tea production of the  $i$ -month) denoted by  $y_i$  and  $k$  predictors (independent variables) denoted by  $(x_1, x_2, \dots, x_k)$ . The multiple linear regression model can be expressed by

$$y_i = b_0 + b_1x_1 + \dots + b_kx_k + e_i, i=1, \dots, n \quad (3.1)$$

or in matrix form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \quad (3.2)$$

where

- $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$  is the vector of response variable for  $n$  sample units
- $\mathbf{X}: n \times (k+1)$  is the matrix of regressors (independent variables)
- $\mathbf{b} = (b_0, b_1, \dots, b_k)^T$  is the vector of regression model parameters
- $\mathbf{e} = (e_1, e_2, \dots, e_n)^T$  is the vector of random errors

### 3.4.2. K-Nearest Neighbor regressor

K-Nearest Neighbor (KNN) regression is an instance grounded lazy learning algorithm. It is non-parametric regression which does not make any supposition on the distribution of data, thus stimulating training phase (Goyal et al., 2014). KNN learns complex label function rapidly without losing information. For a given input features  $x$  of training set,  $K$  observations with  $x_j$  in the proximity are considered and the average of the rejoinder of those  $K$  predictors (independent variables) gives the predicted output  $\hat{y}$  (Goyal et al., 2014).

$$\hat{y} = \frac{1}{N} \sum_{X_j \in N_p} Y_j$$

(3.3)

Where  $N_p$  represents  $K$  closest points in the neighborhood of  $x$

### 3.4.3. Random Forest Regressor

A random forest is a tree-based ensemble which contains many weak decision tree learners. These weak learners are grown in parallel to minimize the bias and reduce the variance of the model as well (Breiman,2001).To train a random forest,  $n$  bootstrapped sample of datasets are drawn from the novel dataset. Each sample which has been bootstrapped is then utilized to grow an un-pruned regression tree(Ahmad et al., 2018). Instead of utilizing all predictors which are available in this step, a small and fixed number of  $K$  predictors which have randomly been sampled are chosen as split candidates. There should be recurrence of these two steps till  $C$  trees are grown, and new data is projected by combining the projection of the  $C$  trees (Ahmad et al., 2018). Random forest uses bagging to upsurge the trees diversity by growing them from diverse training datasets, and thus the overall variance of the model is reduced (Rodriguez-Galiano et al.,2015).

### 3.4.4. Extremely Randomised Trees (Extra trees) Regressor

The extra trees regressor is as an extension of random forest algorithm and has low chance of overfitting a dataset(Geurts et al., 2006). Extra trees regressor utilizes a random subset of input features for training each base estimator just like random forest. Nevertheless, it randomly chooses the feature which is the best along with the conforming value for splitting the node(John et al., 2015). Extra tree utilizes the entire dataset to train each regression tree (Ahmad et al., 2018).

## 3.5.Evaluation Criterion

In this work, a various evaluation metrics were used to evaluate the machine learning models. The criteria are Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Regression coefficient value ( $R^2$ ).

### 3.5.1.Mean Absolute Error (MAE) and Root Mean Square Error(RMSE)

The Mean absolute error is the absolute difference between the response variable and the value projected by the model. The MAE does not penalize the errors extremely like MSE thus deemed to be more robust to outliers. The MAE is linear score since it weighs all the individual differences equally. It is not appropriate for applications to pay more

consideration to the outliers.

The Root Mean Square Error is the most prevalent used evaluation criterion for regression errands and is the square root of the aggregated squared difference between the response variable and the value projected by the model. It is more ideal in some cases since the errors are first squared before aggregating which stances a high penalty on large errors. This indicates that RMSE is useful when large errors are undesired..

$$MAE = \frac{1}{N} \sum_{j=1}^N |\hat{Y}_j - y_j|,$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^N (Y_j - \hat{Y}_j)^2}$$

2

where  $\hat{y}_j$  is the predicted value and  $y_j$  is the actual value(Ahmadetal.,2018).

### 3.5.2. Coefficient of Determination $R^2$

The Coefficient of Determination  $R^2$  is a metric utilize to evaluate the performance of a regression model. The evaluation criterion aids in comparing the current model with a constant baseline and communicates the extent of model robustness. The constant baselines selected by taking the average of the data and sketching a line at the mean. Coefficient of determination is a scale-free score which infers that even if the values are too big or small, the coefficient of determination will always be a number between zero and one. If we consider the multiple regression model defined in (3.2), then the  $R^2$  is defined by

$$R^2 = \frac{(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}))}{(\sum_{i=1}^n (y_i - \bar{y})^2)(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2)}$$

## CHAPTER FOUR: DATA ANALYSIS

After cleaning tea production data set, the data was explored to comprehend deep insights. There was a need to check relationship between input features (rainfall, fertilizers, seedling, area under plantation, month and year) and tea production. This study utilizes scatter and Ordinary Least Squares regression to establish the relationship between the predictors and response variable.

### 4.1. Checking relationship using scatterplot

The plot on Figure (4.1) depicts a fairly strong positive relationship between rainfall against production. Production increases with increase in rainfall (mm) upto certain limit (2500 mm in our case).

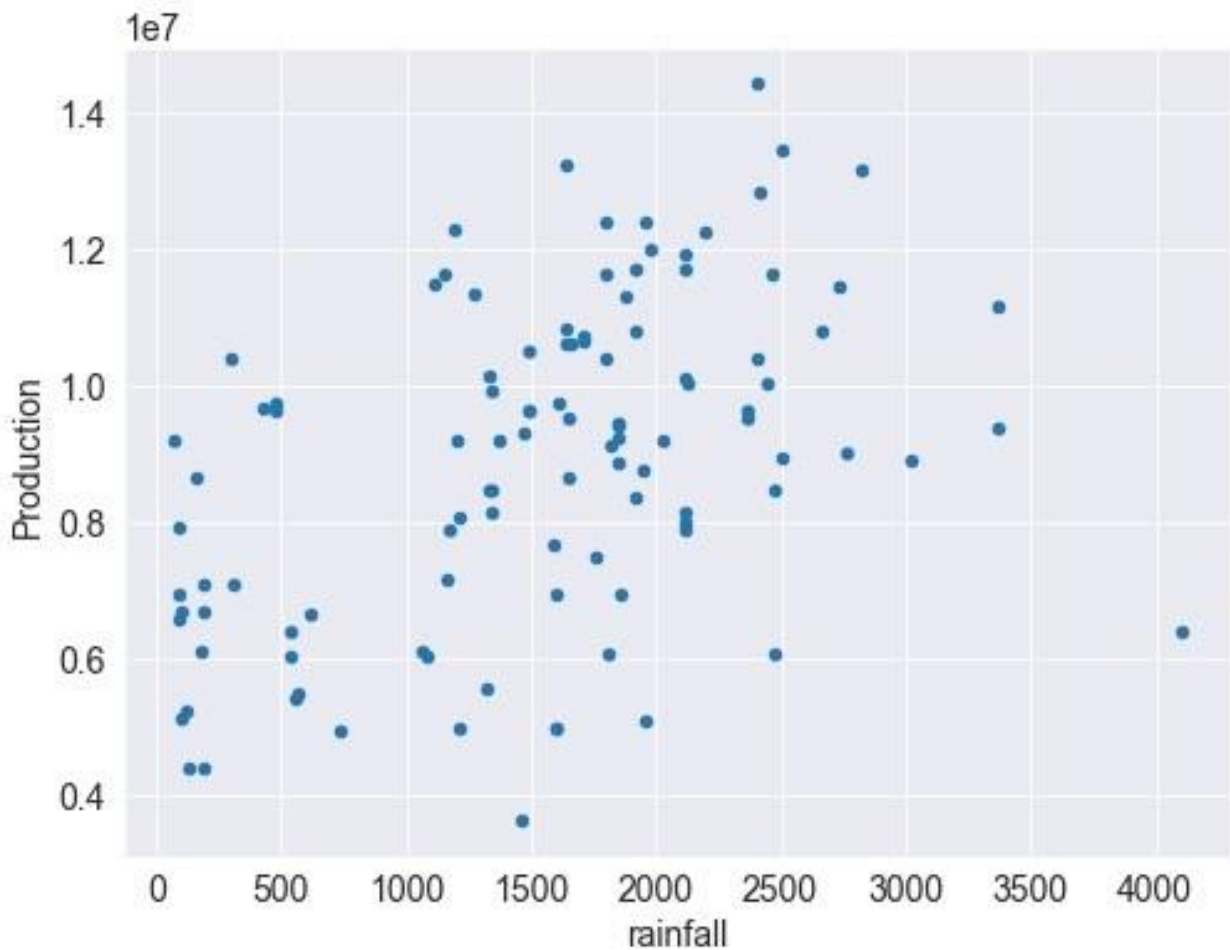


Figure 2: Scatter plot of rainfall against production

TheFigure(4.2)showsthattherearesomemonthswhenproductionofteaishighwhile other months are low. For example, the 5th and 11-th month the production of tea is high while in 8-th month the production islow.

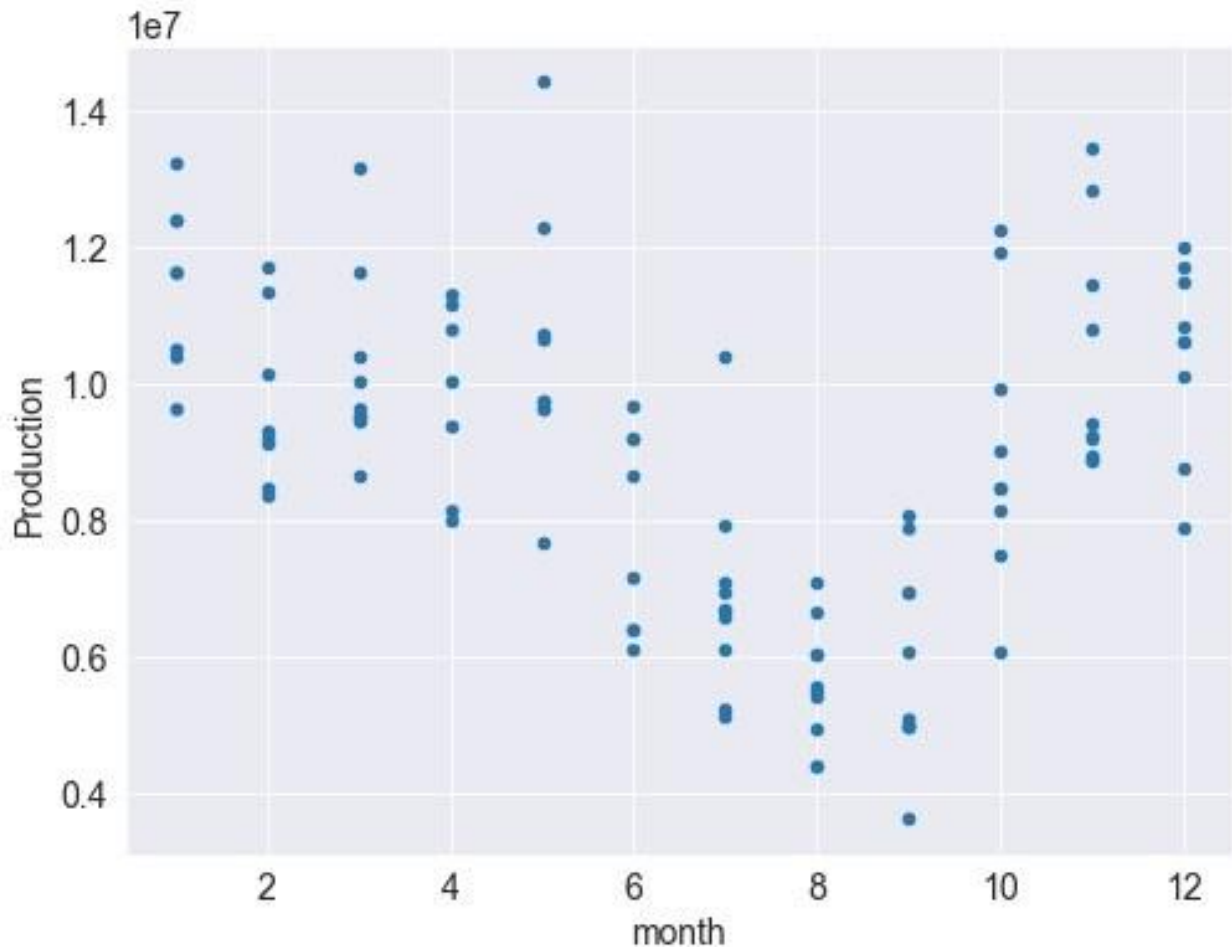
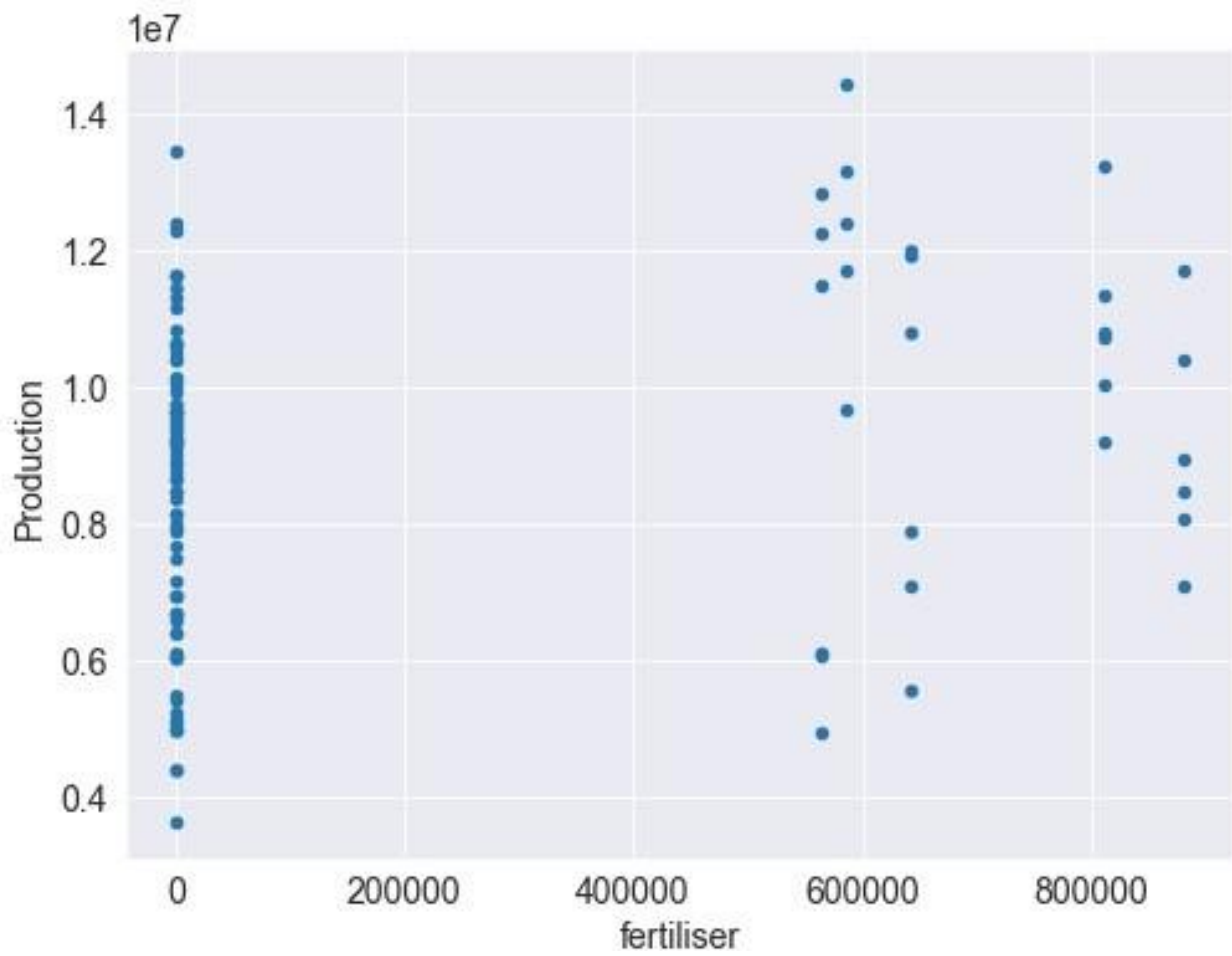


Figure 3:Scatter plot of month against production

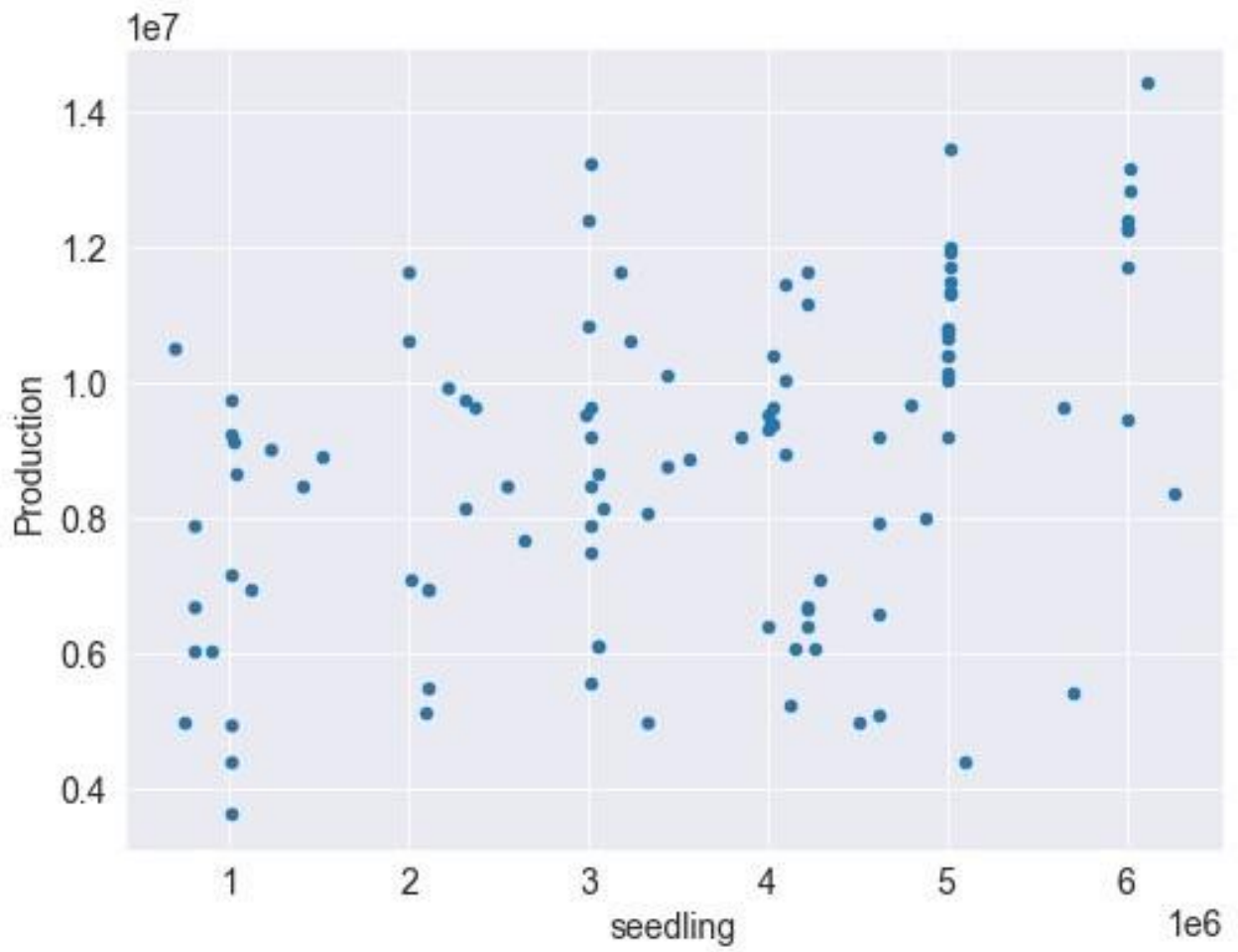
The plot on Figure (4) shows uneven distribution between fertilizer against production. At some point the production increases even with no fertilizer and at some point (600000 units) of fertilizer the production is very high.





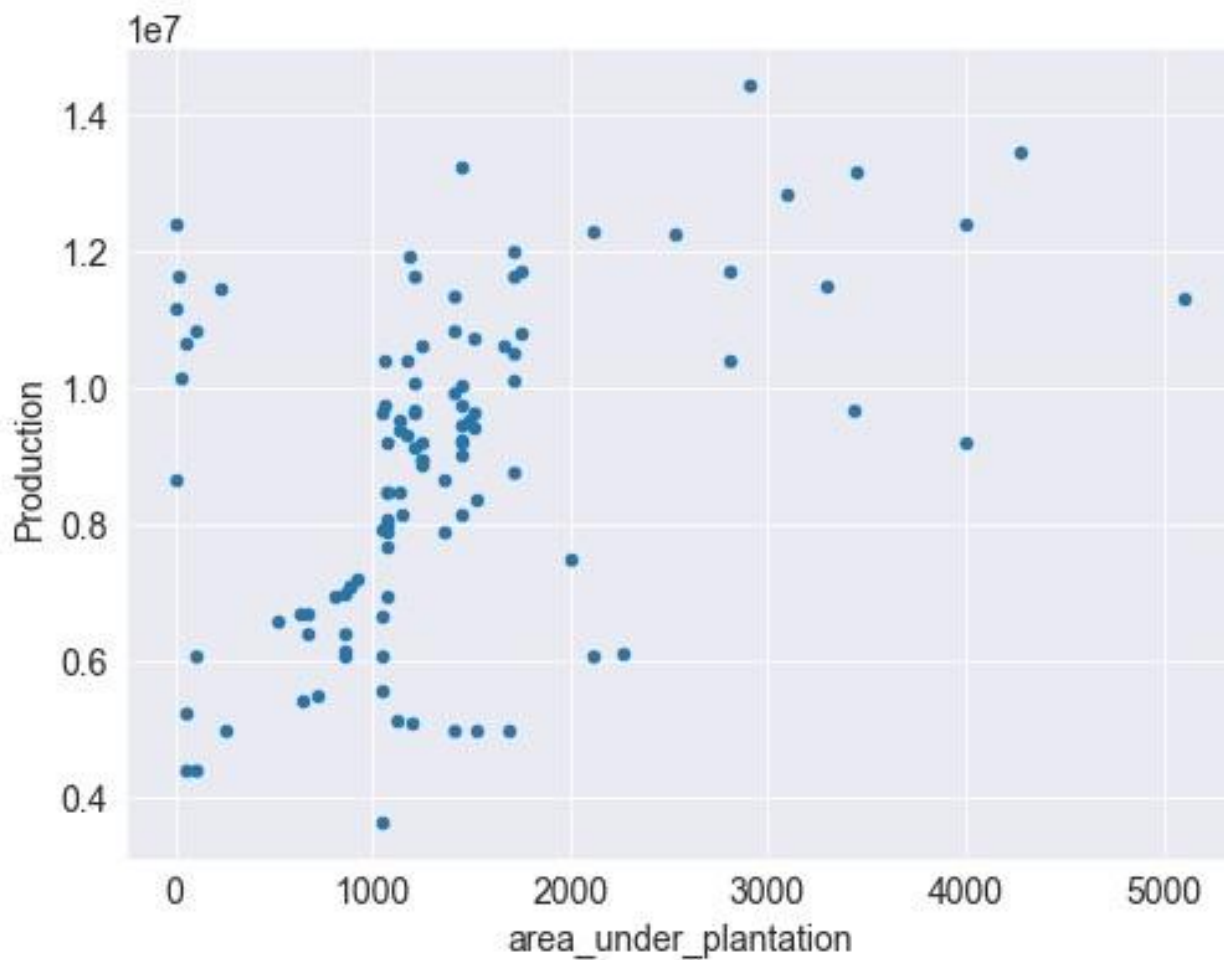
**Figure 4:Scatter plot of fertilizer against production**

The plot on Figure (5) depicts a fairly strong positive relationship between seedling against production. Production increases with increase in seedlings.



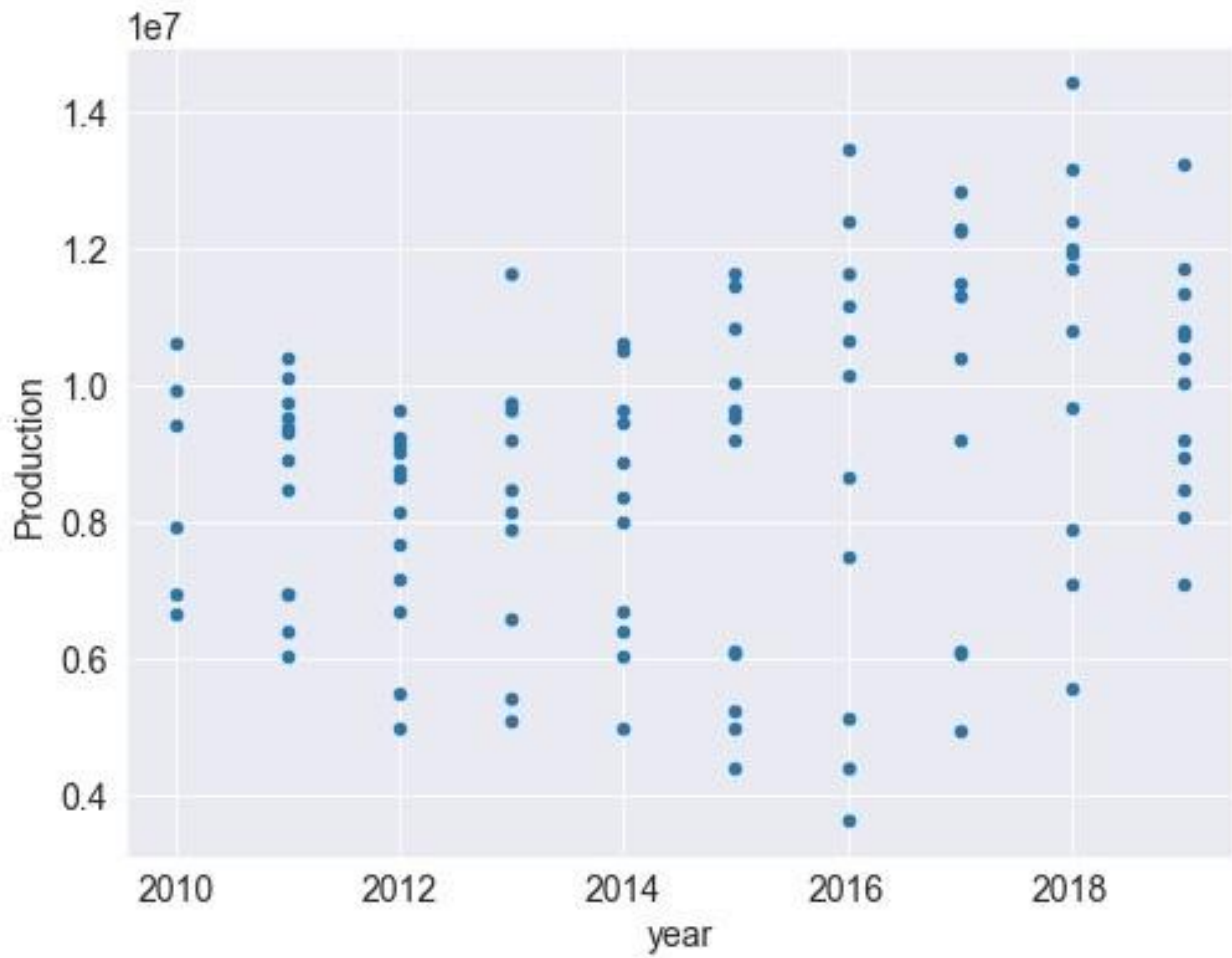
**Figure 5:Scatter plot of fertilizer against production**

The ploton Figure (6)depict safairly strong positive relationship between area under plantation against production. Production increases with increase in area under plantation (ha).



**Figure 6:Scatter plot of area under plantation against production**

From Figure (7) we can see that tea production has been increasing over the years. The production was lowest in 2012 but highest in the year 2018.



**Figure 7:Scatter plot of year against production**

## 4.2. Ordinary Least Square Regression (OLS)

Ordinary Least Squares regression (OLS) which is also linear regression was utilized to estimate the relationship between input features and tea production (response variable). OLS was utilized to estimate the relationship by abating the sum of the squares in the variance between the true and projected values of the response variable constituted in form of a straight line. The key indicators were the slope  $\beta_1$  and p-value which was utilized to check relation and the effect of input feature (rainfall, fertilizers, seedling, area under plantation, month and year) on tea production is statistically significant (using  $p < 0.05$  as a rejection rule)..

### 4.4.1. OLS regression for rainfall predictor

Hereafter is the output of regression analysis for rainfall predictor Note that the notation  $xe + n$  and  $xe - n$  stands for  $x * 10^n$  and  $x * 10^{-n}$  respectively.

#### Table 1: OLS Regression outcome for rainfall

OLS Regression Results

Dep. Variable: Production R-squared: 0.203

Model: OLS Adj. R-squared: 0.196

Method: Least Squares F-statistic: 26.81

Prob (F-statistic): 1.09e-06

Log-Likelihood: -1710.4

No. Observations: 107 AIC: 3425.

Df Residuals: 105 BIC: 3430.

Df Model: 1

Covariance Type: nonrobust

coef std err t P>|t| [0.025  
0.975]

Intercept 6.933e+06 4.32e+05 0.000 6.08e+06 7.79e+0  
16.055 6

rainfall 1277.7115 246.757 5.178 0.000 788.438 1766.985

Omnibus: 1.514 Durbin-Watson: 1.126

Prob(Omnibus): 0.469 Jarque-Bera (JB): 1.582

Skew: -0.254 Prob(JB): 0.453

Kurtosis: 2.688 Cond. No. 3.65e+03

## Table 2: OLS Regression outcome for rainfall

Standard Errors take an assumption that there is correct specification of the covariance matrix errors .

From Table , it is depicted that:

The intercept  $\beta_0=6.933e+06$ .

The slope  $\beta_1= 1277.7115$ .

The estimate of rainfall (1277.7115) parameter is positive, this infers that rainfall has a positive effect on tea production, as depicted in Table 4.1.

The p-value for rainfall is 0.000 infers that the effect of rainfall on tea production is very significant (where  $p \leq 0.05$  was utilized as rejection rule).

The R-squared value of 0.203 depicts that variation of about 20.3% in log rainfall explained by protection against expropriation. There is large condition number 3.65e+03, depicting that there might be strong multicollinearity or numerical glitches

### 4.4.2.OLS regression for fertilizer predictor

Hereafter is the output of regression analysis for fertilizer predictor

Standard Errors take an assumption that there is correct specification of the covariance matrix errors.

The intercept  $\beta_0=8.504e+06$ .

The slope  $\beta_1 =2.0581$ .

The estimate of fertilizer (2.0581) parameter is positive, this infers that fertilizer has a positive effect on tea production, as depicted in Table 4.2 above.

The p-value for fertilizer is 0.004 infers that the effect of fertilizer on tea production

is very significant (where  $p < 0.05$  was utilized as rejection rule).

The R-squared value of 0.076 depicts that variation of about 7.6% in log fertilizer explained by protection against expropriation.

There is large condition number  $4.32e+05$ , depicting that there might be strong multicollinearity or numerical glitches.

#### 4.1.3. OLS regression for seedling predictor

Hereafter is the output of regression analysis for seedling predictor

Standard Errors take an assumption that there is correct specification of the covariance matrix errors..

The intercept  $\beta_0 = 6.609e+06$ .

The slope  $\beta_1 = 0.6460$

The estimate of seedling (0.6460) parameter is positive, this infers that seedling has a positive effect on tea production, as depicted in Table 4.3 above.

OLS Regression Results

Dep. Variable: Production R-squared: 0.076

Model: OLS Adj. R-squared: 0.067

Method: Least Squares F-statistic:

8.640 Prob (F-statistic): 0.00404

Log-Likelihood: -1718.4

No. Observations: 107 AIC: 3441.

Df Residuals: 105 BIC: 3446.

Df Model: 1

Covariance Type: nonrobust

coef std err t P>|t| [0.0250.975]

Intercept 8.504e+06 2.59e+05 32.778 0.000 7.99e+06  
9.02e+06 fertilizer 2.0581 0.700 2.939 0.004 0.670 3.446

Omnibus: 6.455 Durbin-Watson: 0.800

Prob(Omnibus): 0.040 Jarque-Bera (JB): 3.368

Skew: -0.196 Prob(JB): 0.186

Kurtosis: 2.224 Cond. No. 4.32e+05

OLS Regression Results

Dep. Variable: Production R-squared: 0.181

Model: OLS Adj. R-squared: 0.174

Method: Least Squares F-23.28  
statistic:

Prob (F-statistic): 4.77e-06

Log-Likelihood: -1711.9

No. Observations: 107 AIC: 3428.

Df Residuals: 105 BIC: 3433.

Df Model: 1

Covariance Type: nonrobust

coef std err t P>|t| [0.0250.975]

Intercept 6.609e+06 5.18e+05 12.756 0.000 5.58e+06 7.64e+0



seedling 0.64600.134 4.8250.000 0.3800.911

Omnibus: 3.168 Durbin-Watson: 1.000

Prob(Omnibus): 0.205 Jarque-Bera (JB): 3.109

Skew: -0.368 Prob(JB): 0.211

Kurtosis: 2.605 Cond. No. 9.57e+06

### **Table 3: OLS Regression outcome for seeding**

The p-value for seedling is 0.000 infers that the effect of seedling on tea production is very significant (where  $p < 0.05$  was utilized as rejection rule).

The R-squared value of 0.181 depicts that variation of about 18.1% in log seedling explained by protection against expropriation. There is large condition number 9.57e+06, depicting that there might be strong multicollinearity or numerical glitches.

#### **4.4.4. OLS regression for area under plantation predictor**

Hereafter is the output of regression analysis for area under plantation predictor

Standard Errors take an assumption that there is correct specification of the covariance matrix errors.

The intercept  $\beta_0 = 7.348e+06$ .

The slope  $\beta_1 = 1131.9577$ .

The estimate of area under plantation (1131.9577) parameter is positive, this infers that area under plantation has a positive effect on tea production, as depicted in Table 4.4 above.

The p-value for area under plantation is 0.000 infers that the effect of area under plantation on tea production is very significant (where  $p < 0.05$  was utilized as rejection rule).

depicts that variation of about 19.5% in log area under plantation explained by protection against expropriation.

There is large condition number  $2.94e+03$ , depicting that there might be strong multi-collinearity or numerical glitches.

#### OLS Regression Results

Dep. Variable: Production R-squared: 0.195

Model: OLS Adj. R-squared: 0.188

Method: Least Squares F-statistic:

25.51 Prob (F-statistic):  $1.86e-06$

Log-Likelihood: -1710.9

No. Observations: 107 AIC: 3426.

Df Residuals: 105 BIC: 3431.

Df Model: 1

Covariance Type: nonrobust

coef std err t P>|t| [0.025 0.975]

Intercept  $7.348e+06$   $3.7e+05$  19.850 0.000  $6.61e+06$   $8.08e+06$

area\_under\_plantation 1131.9577 224.100 5.051 0.000 687.608 1576.307

Omnibus: Durbin-Watson:

1.081 0.852

Prob(Omnibus): 0.583 Jarque-Bera (JB): 1.005

Skew: -0.034 Prob(JB): 0.605

Kurtosis: 2.530 Cond. No.  $2.94e+03$

## CHAPTER FIVE: MAIN RESULTS AND DISCUSSION

### 5.1. Main results

After carrying out data preprocessing and feature engineering, the relevant predictors were identified. Further, cross validation was carried out by splitting the dataset into training set and test set in order to train data mining models by utilizing training set and validate their performance utilizing test set. For training set 80% of dataset were used to fine-tune the algorithms. For the test set 10% of dataset were hold back from training of the model in order to be utilized to evaluate performance of model on unseen data. After training the selected data mining models, their performance was compared by utilizing test data. The initial data mining comparison was done before carrying out hyper parameter tuning (by utilizing default parameters of data mining models) on test set. Later, it was compared after carrying out hyper parameter tuning (by utilizing default parameters of data mining models) on test set. The error metric such as RMSE, MAE and  $R^2$  was utilized to depict model performance and find the most robust data mining model.

#### 5.1.1. Data Mining Models Comparison using Test Data before Hyperparameter Tuning

Table 4 shows comparison of various models on test data before hyper parameter tuning. The error metric such as RMSE, MAE and  $R^2$  was used to find most robust model. It is noted that  $R^2$  was most preferred metric for this study. The  $R^2$  for each of the models were: random forest (0.7894), extra tree (0.7709), lastly linear regression (0.7029) and KNN (0.6734). The RMSE for each of the models were: random forest (1.16E+06), extra tree (1.21E+06), linear regression (1.37E+06) and lastly KNN (1.44E+06). The most robust model before hyper parameter tuning was random forest regressor.

	<b>Model</b>	<b>MAE</b>	<b>MSE</b>	<b>RMSE</b>	<b>Rsquared</b>
1	Linear	1.18E+0	1.89E+1	1.37E+0	0.702852
		6	2	6	
2	KNN	1.13E+0	2.07E+1	1.44E+0	0.73428
		6	2	6	
3	Random Forest	1.02E+0	1.34E+1	1.16E+0	0.789402
		6	2	6	
4	Extratree	1.04E+0	1.45E+1	1.21E+0	0.770851
		6	2	6	

**Table 4: Data Mining Models Comparison using Test Data Before Hyperparameter Tuning**

Table 5. shows comparison of various models on test data after hyper parameter tuning. The error metric such as RMSE, MAE and R was used to find most robust model. However,  $R^2$  was most preferred metric for this study. The  $R^2$  for each one of the models were: extra tree (0.8953), random forest (0.8340), KNN (0.7302), and lastly linear regression (0.7029). The RMSE for each one of the models were: extra tree (8.15E+05), random forest (1.03E+06), KNN (1.31E+06), and lastly linear regression (1.37E+06). The most robust model before hyper parameter tuning was extra tree regressor.

	<b>Model</b>	<b>MAE</b>	<b>MSE</b>	<b>RMSE</b>	<b>Rsquared</b>
1	Linear	1.18E+0	1.89E+1	1.37E+0	0.7029
		6	2	6	
2	KNN	1.09E+0	1.71E+1	1.31E+0	0.7302
		6	2	6	
3	RandomForest	9.15E+0	1.05E+1	1.03E+0	0.8340
		5	2	6	
4	Extratree	6.06E+0	6.64E+1	8.15E+0	0.8953
		5	1	5	

**Table 5: Data Mining Models Comparison using Test Data after Hyper Parameter Tuning**

### 5.1.2.Feature Importance using Extra Tree Regressor

Extra tree regressor was used to identify important features that contributes to tea production. Season in Month(s) was most important feature, followed by rainfall, area under plantation, seedling, fertilizer and lastly year.

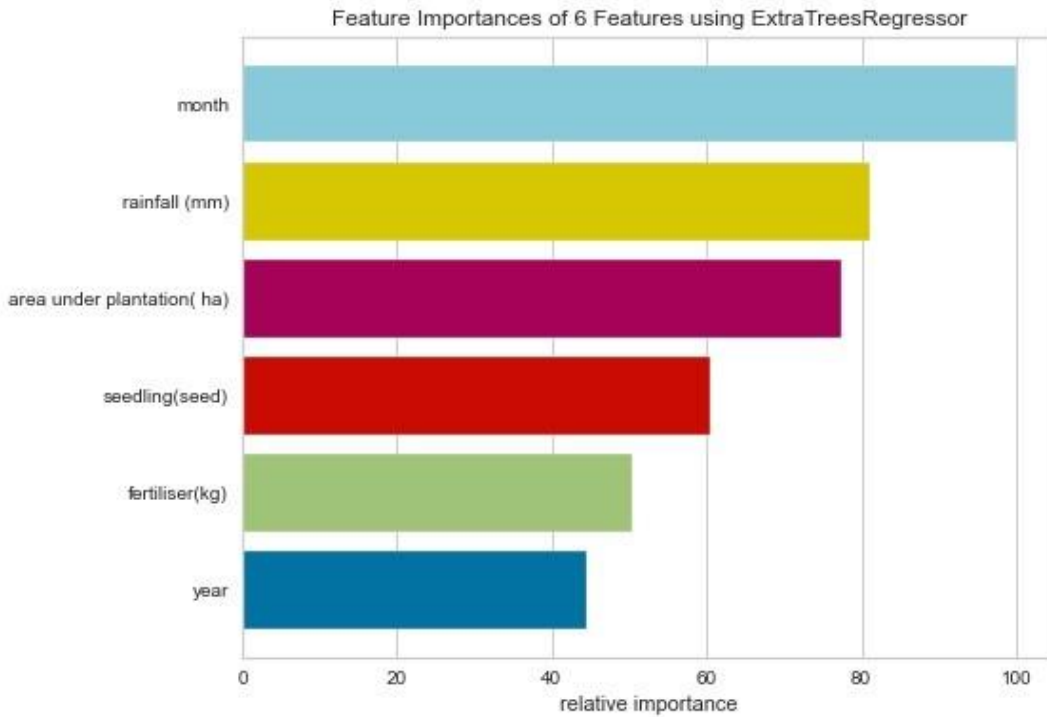


Figure 8:Feature importance

### 5.2.Discussion

In this study the main metrics utilized to evaluate the performance of data mining models were RMSE and  $R^2$ . Table 4 and Table 5 depicts the performance of prediction data mining models while varying  $R^2$  and error metrics. It was found that varying error criterion did not yield a significant improvement for tea production using our dataset. From Table 5 results, random forest have an outstanding performance as compared to other models. Since Table 4 results were obtained before the models were tuned, it shows that default parameters for random forest gives a better performance than default

parameters for other models in the data set used in this study. From Table 5 tree-based models resulted in slightly better performance than the non-tree-based models. The minimum number of samples required to split an internal node could be an important hyper-parameter depending on the problem.

There is variation in the results in Table 4 and Table 5 though same models were used. The model performance has improved in table after parameter tuning. Using optimal parameters in Table 5 leads to increase in  $R^2$  and reduced error metrics such as RMSE and MAE. The  $R^2$  results on Table 5 depicts all data mining models that were used in this study have a score above 0.70 which is viewed as good enough to make decision. However, the  $R^2$  for the extra tree regressor was outstanding with a score of 0.90 while least performing model was linear regression (0.70) since the larger the  $R^2$ , the better the regression model

fits in the observations. For extra tree regressor ninety percent of the variance in the true class (tea production) can be explained by the explanatory variables. The  $R$  was preferred in this study because when  $R$  is high it gives precise predictions. Based on  $R^2$  results, extra tree regressor could be used to predict tea production in Rwanda.

Table 5 results which was preferred to Table 4 because of RMSE and MAE results for data mining models on test data. From Table 5 extra tree regressor had the lowest RMSE and MAE values compared to other data mining models that were used in this study. The least performing model in this study was linear regression since it has the highest values of RMSE, and MAE based on Table 5 results. RMSE and MAE was used to measure the expanse of error in the dataset by comparing the predicted value with observed value. Based on RMSE and MAE results I find out that extra tree regressor is the most robust model and thus can be recommended to predict tea production based on this dataset.

Based on Table 5 results, extra tree regressor which has highest  $R^2$  and the least RMSE and MAE compared to other data mining models was considered as an optimal model and thus was used to find most important features as shown in Figure 8. From Figure 8 the most important feature in predicting tea production is associated with the months. It was found out that tea production is very high in some months, for instance the 5th and 11th month the production of tea is high while in 8-th month the production is low. The amount of rainfall is the second most important feature in predicting tea production. The rainfall which is part of weather conditions is viewed to contribute more to tea production. The tea production increases with increase in amount of rainfall (mm) up to certain limit (2500 mm in our case) not excess rainfall. The area under plantation was third most important feature. It was found that tea production increases with increase in area under plantation (ha). There is need to create or reclaim more land for tea plantation in order to increase production. Seedlings was fourth most important feature and it was found that tea production increases with increase in seedlings.

## CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

### 6.1. Conclusion

In this study, the feasibility of using data mining models (extra trees and random forest KNN and linear regression) to predict the tea production in Rwanda was evaluated. The capability of extra tree and random forest regressors for predicting the tea production has been verified with a better prediction accuracy of the models. To appraise the data mining models' prediction performance, different metrics of MAE, RMSE and R were used. It has been found that extra tree and random forest performed marginally better than the widely used data mining models KNN and linear regression. The study also proposed using extra tree regressor to give an insight into the analysis of the importance of each input feature. The presented analysis will enable researchers and practitioners in the industry to gain better understanding of the tea production. The developed data mining models can be applied to predict tea production based on different month, rainfall (climatic conditions) area under plantations, seedlings and fertilizer. The advantages of the extra tree and random forest regression are that they have only a few tuning parameters and, in most situations, default hyper-parameter can result in satisfactory prediction performance. Random forest performs internal cross validation that is utilizing out-of-bag samples and can be used to any nature of datasets. The proposed extra trees algorithm is computationally efficient and is more suitable for online or control applications. In future, the designed extra tree model can be used to predict tea production. There is also a need to investigate the performance of other data mining models such as extreme gradient boosted regression, cat boost regression and logistic regression. Future studies will also focus on assessing the performance of data mining models in other timescales and for different climate conditions. Development of separate models based on weather classification (i.e., classifying weather on different weather conditions such as temperature, clear sky, cloudy day, foggy day,) will also be investigated in future. There is also a necessity to explore Big Data technologies for training and deploying prediction models.

## **6.2.Recommendations**

This study recommends separate models based on weather classification (i.e., classifying weather on different weather conditions such as temperature, clear sky, cloudy day, foggy day) should be investigated in future. We also recommend that government should investigate other underlying factors during different months (seasons) which might have influenced tea production that were not captured in the provided dataset. For instance, it was found that tea production was high in some months and low in other months. Apart from suggestions provided in this study there is need to find more factors which can be enhanced by use of sensors and other data mining tools. The authors recommend the utilization of data mining models in predicting tea production especially the extra tree and random forest models which were found to be more effective in this study. Further authors recommend the government of Rwanda to sight see Big Data technologies to tune and deploy prediction models especially on forecasting tea production.



## REFERENCE

- Ahmad, M. W., Reynolds, J., and Rezgui, Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of cleaner production*, 203:810–821.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- factsanddetails.com (April 24, 2020). Tea Cultivation and Production  
<http://factsanddetails.com/asian/cat62/sub408/item2610.html>.
- Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.
- Goyal, R., Chandra, P., and Singh, Y. (2014). Suitability of knn regression in the development of interaction based software fault prediction models. *Ieri Procedia*, 6(1):15–21.
- Jambekar, S., Nema, S., and Saquib, Z. (2018). Prediction of crop production in india using data mining techniques. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–5.IEEE.
- John, V., Liu, Z., Guo, C., Mita, S., and Kidono, K. (2015). Real-time lane estimation using deep features and extra trees regression. In *Image and Video Technology*, pages 721–733. Springer.
- NAEB (2018). National Agricultural Export Development Board 2017-2018 AnnualReport.  
Report, Rwanda.
- Nzuva, M. and Lawrence, N. (2017). Prediction of tea production in kenya using clustering and association rule mining techniques. *Am J Compt Sci Inform Technol*,5(2).
- Permai, S. D. and Tanty, H. (2018). Linear regression model using bayesian approach for energypformanceofresidentialbuilding.*ProcediaComputerScience*,135:671–677.
- Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of

neural net-

works, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71:804–818.

Rudyy, R. (2001). On the classification of agricultural lands. *Archiwum Fotogrametrii, Kartografii i Teledetekcji*, 11.

Sitienei, B. J., Juma, S. G., and Opere, E. (2017). On the use of regression models to predict tea crop yield responses to climate change: A case of nandi east, sub-county of nandi county, kenya. *Climate*, 5(3):54.

Vamanan, R. and Ramar, K. (2011). Classification of agricultural land soils a data mining approach. *International Journal on Computer Science and Engineering*, ISSN, pages 0975–3397.

Veeresh, K. and Saboji, S. V. (2019). Agriculture price prediction using data mining. *International Journal of Engineering and Advanced Technology*, 8(6S):2249–8958.

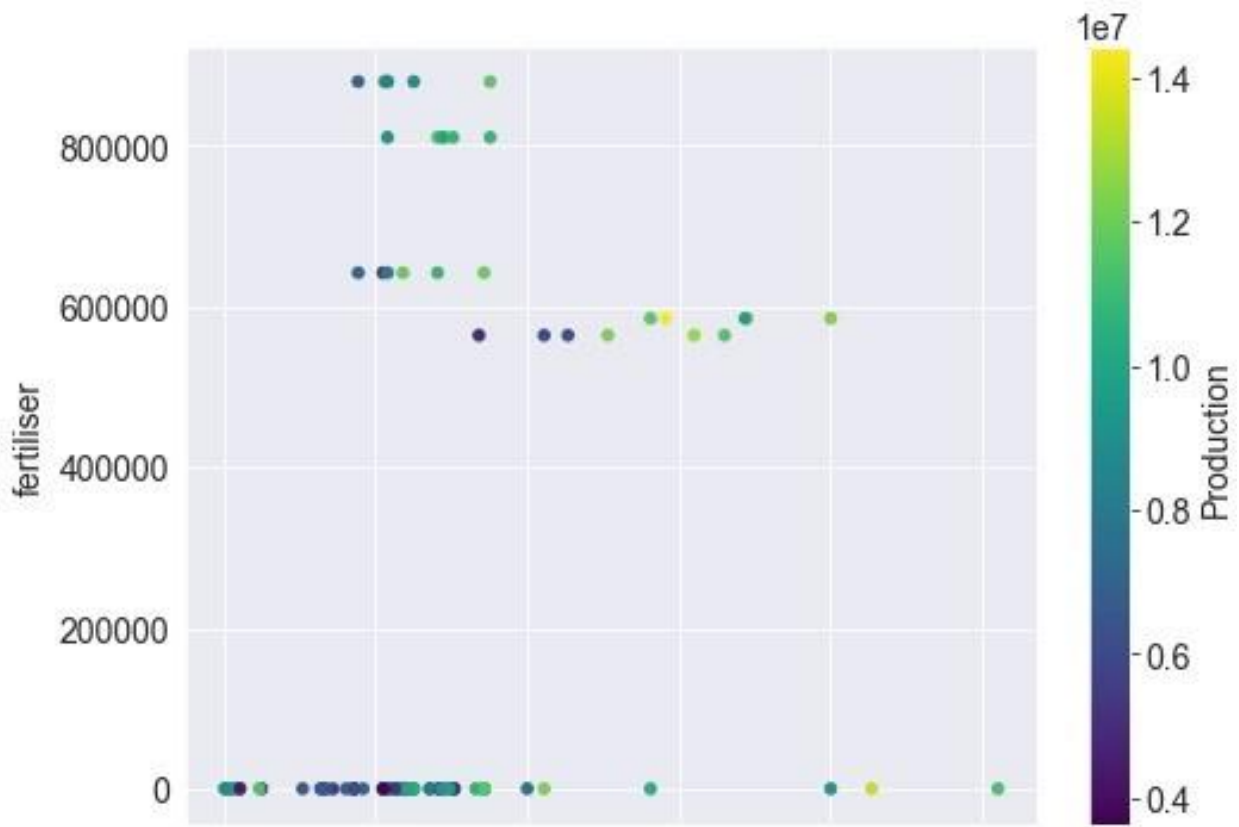
## APPENDICES

	Model	MAE	MSE	RMSE	Rsquared
1	Linear	1.42E+0	3.23E+1	1.80E+0	0.418116
		6	2	6	
2	KNN	1.13E+0	2.03E+1	1.42E+0	0.6345342
		6	2	6	
3	RandomForest	5.50E+0	5.97E+1	7.73E+0	0.892478
		5	2	5	
4	Extratree	6.79E-11	1.17E-19	3.43E-10	1

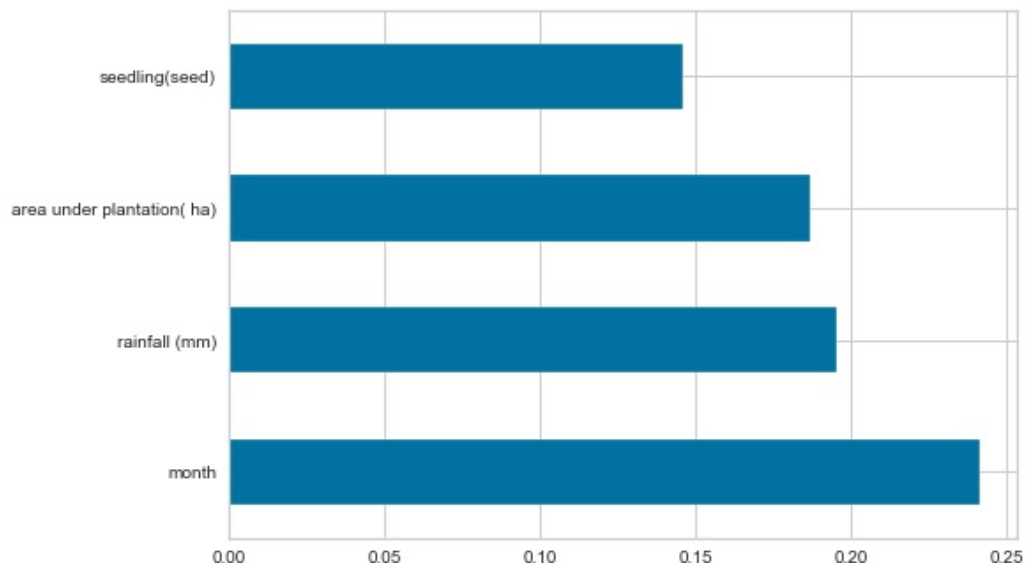
**Table 6:Data Mining Models Comparison using Training Dataset**

	year	month	rainfall (mm)	seedling(seed)	fertiliser(kg)	area under plantation( ha)	Production
year	1	-0.078869	0.00356271	0.372666	0.762804	0.270199	0.318367
month	-0.078869	1	-0.0340297	-0.107933	0.0740718	-0.12208	-0.239022
rainfall (mm)	0.00356271	-0.0340297	1	0.162093	0.0233637	0.138531	0.361027
seedling(seed)	0.372666	-0.107933	0.162093	1	0.287111	0.242639	0.406909
fertiliser(kg)	0.762804	0.0740718	0.0233637	0.287111	1	0.103645	0.282015
area under plantation( ha)	0.270199	-0.12208	0.138531	0.242639	0.103645	1	0.335957
Production	0.318367	-0.239022	0.361027	0.406909	0.282015	0.335957	1

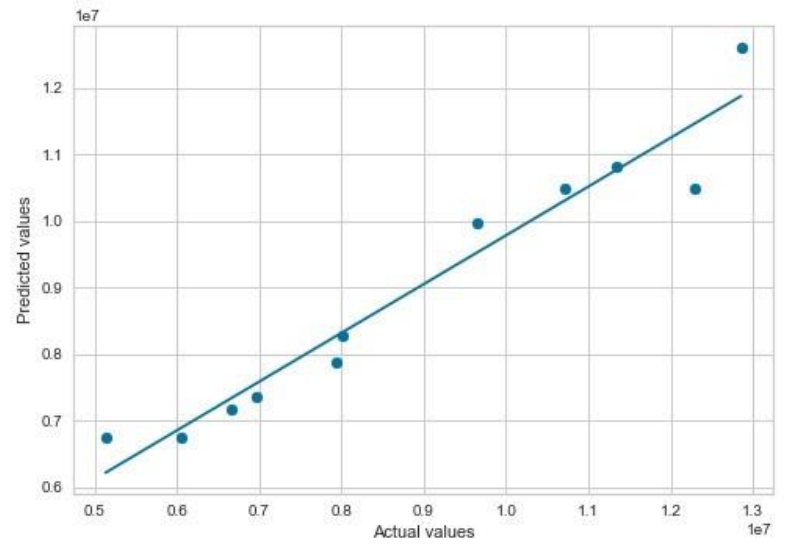
**Figure 9:Correlation Matrix**



**Figure 10:Effect of Fertilizer on Tea production at area under plantation**



**Figure 11:Feature importance of tea production using random forest regressor**



R-squared = 0.90

**Figure 12: Using R2 to check relationship between actual values against predicted values**

# Prediction of Tea Production in Rwanda using Data Mining Techniques

## ORIGINALITY REPORT

16%

SIMILARITY INDEX

14%

INTERNET SOURCES

9%

PUBLICATIONS

7%

STUDENT PAPERS

## PRIMARY SOURCES

1	Muhammad Waseem Ahmad, Monjur Mourshed, Yacine Rezgui. "Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression", Energy, 2018 Publication	2%
2	orca.cf.ac.uk Internet Source	2%
3	www.imedpub.com Internet Source	2%
4	postguild.org Internet Source	1%
5	rwandainspirer.com Internet Source	1%
6	Submitted to Harrisburg University of Science and Technology Student Paper	1%
7	Submitted to Intercollege Student Paper	

		1%
8	<a href="http://brettmontague.com">brettmontague.com</a> Internet Source	1%
9	Syarifah Diana Permai, Heruna Tanty. "Linear regression model using bayesian approach for energy performance of residential building", <i>Procedia Computer Science</i> , 2018 Publication	1%
10	Rinkaj Goyal, Pravin Chandra, Yogesh Singh. "Suitability of KNN Regression in the Development of Interaction based Software Fault Prediction Models", <i>IERI Procedia</i> , 2014 Publication	1%
11	<a href="http://factsanddetails.com">factsanddetails.com</a> Internet Source	1%
12	<a href="http://ieeexplore.ieee.org">ieeexplore.ieee.org</a> Internet Source	1%
13	<a href="http://www.researchgate.net">www.researchgate.net</a> Internet Source	1%
14	Submitted to University of Applied Sciences Berlin Student Paper	1%
15	Muhammad Waseem Ahmad, Jonathan Reynolds, Yacine Rezgui. "Predictive modelling	1%