

Available online at <http://www.mecs-press.net/ijeme>

## Impact of Climatic Change on Agricultural Product Yield Using K-Means and Multiple Linear Regressions

<sup>1</sup>Gbadamosi Babatunde, <sup>2</sup>Adeniyi Abidemi Emmanuel, <sup>3</sup>Ogundokun Roseline Oluwaseun, <sup>4</sup>Oladosu Bukola Bunmi, <sup>5</sup>Anyaiwe Ehiedu Precious

<sup>1,2,3,4,5</sup> *Department of Computer Science, College of Pure and Applied Sciences, Landmark University, Omu-Aran, Kwara State*

Received: 12 September 2018; Accepted: 17 December 2018; Published: 08 May 2019

---

### Abstract

Adequate information about climate change helps farmers to prepare and helps boost crop yield. Over the years, crops prediction was performed by manually considering farmer's experience on the particular crop in relation to the weather. This method was inadequate, depends on the farmer's unreliable memory and grossly inaccurate. There is a need to introduce computational means to study and predict optimal climatic factors for improved crop growth and yield. The aim of this research work is to study the impact of climatic changes on the yield production of roots and tubers crops. K-means classification algorithm, Multiple Linear Regression, Python programming language, Flask Framework, Python machine learning packages numpy, matplotlib, Scikit-learn are the methodology used. While the obtained results show that CO<sub>2</sub> Emission and Temperature does not really play a key role on how climate impact yield of root and tubers, rainfall plays more role; therefore, the study concludes that the three variables (temperature, rainfall, and CO<sub>2</sub> Emission) are not enough to predict agricultural yield. It is therefore recommended that further research should be carried out to determine how other climatic factors such as soil type; humidity, sunlight etc. affect the yield of crops. The objective of this research is to study climatic change using data mining techniques, to design a predictive model using multiple linear regression to find the most optimal temperature and rainfall for effective crop yield and to simulate the multiple linear regression model design that achieve a high accuracy and a high generality in terms of climate change to crop yield.

**Index Terms:** Data mining, K-mean, Climate, Rainfall, Temperature, Multiple Linear Regression, Agricultural product.

© 2019 Published by MECS Publisher. Selection and/or peer review under responsibility of the Research Association of Modern Education and Computer Science.

---

\* Corresponding author.

E-mail address: <sup>1</sup>gbadamosi.babatunde@lmu.edu.ng, <sup>2</sup>adeniyi.emmanuel@lmu.edu.ng, <sup>3</sup>ogundokun.roseline@lmu.edu.ng, <sup>4</sup>oladosu.bukola@lmu.edu.ng, <sup>5</sup>ehiedu.precious@lmu.edu.ng

## 1. Introduction

Agricultural product yield prediction is becoming a huge problem battling with the agricultural organizations. Each agriculturist is eagerly to know how much yield production they will be expecting. Attempts to solve this problem dated back to the time when first farmers started cultivating soil. Problem of predicting production yield can be solved with data mining techniques. It should be considered that the sensor data are available for some past tense, in which appropriate production yield was recorded. All this information creates a set of data which can be used for learning ways of classifying future production yields because new sensor data are available [1].

Over the years, yield prediction was conducted based on the farmer's experience of certain agricultural cultures and crops. However, this information can be acquired with the use of modern technologies, like GPS. Today acquiring a large amount of sensor data is relatively quick, so agriculturists not only reap crops but also ever larger amounts of data. These data are processed and often interconnected [2].

In agricultural research, data mining begins with a hypothesis and the results are adjusted to fit that hypothesis. This is different than the standard data mining practice, which simply begins with a set of data with no apparent hypothesis. While traditional data mining is directed to patterns and trends in datasets, agricultural data mining is more focused on the majority which stands out from patterns and trends [3]. The fact that standard data mining is more focused on description and not on the explanation of patterns and trends is what deepens the difference between standard and agricultural data mining.

Various analytic methods used in data mining are in most cases well known mathematical techniques and algorithms [4]. Data mining is said to be a new method for the process of analyzing data in the agriculture field. The fact that connected those techniques and large databases were the cheapening of storage space and processing power. Data mining is used to obtain certain information from the huge amount of data or dataset. In large datasets, information is usually hidden, but data mining techniques can be used to discover, and transform into useful knowledge.

## 2. Related Work

People have been seeking patterns in data since human life began. Hunters seek patterns in animal migration behavior, farmers seek patterns in crop growth, politicians seek patterns in voter opinion, and lovers seek patterns in their partners' responses. A scientist's job (like a baby's) is to make sense of data, to discover the patterns that govern how the physical world works and encapsulate them in theories that can be used for predicting what will happen in new situations. The entrepreneur's job is to identify opportunities, that is, patterns in behavior that can be turned into a profitable business, and exploit them [7, 13].

Data mining is about solving problems by analyzing data already present in databases. Suppose, to take a well-worn example, the problem is fickle customer loyalty in a highly competitive marketplace. A database of customer choices, along with customer profiles, holds the key to this problem. Patterns of behavior of former customers can be analyzed to identify distinguishing characteristics of those likely to switch products and those likely to remain loyal. Once such characteristics are found, they can be put to work to identify present customers who are likely to jump ship. This group can be targeted for special treatment, treatment too costly to apply to the customer base as a whole. More positively, the same techniques can be used to identify customers who might be attracted to another service the enterprise provides, one they are not presently enjoying, to target them for special offers that promote this service. In today's highly competitive, customer-centered, service-oriented economy, data is the raw material that fuels business growth—if only it can be mined. Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities. How are the patterns expressed? Useful patterns allow us to make nontrivial predictions on new data. There are two extremes for the expression of a pattern: as a black box whose innards are effectively incomprehensible and as a transparent box whose

construction reveals the structure of the pattern. Both, we are assuming, make good predictions. The difference is whether or not the patterns that are mined are represented in terms of a structure that can be examined, reasoned about, and used to inform future decisions. Such patterns we call structural because they capture the decision structure in an explicit way. In other words, they help to explain something about the data [7, 13].

Milovic and Radojevic [5] classification and clustering are the major groups of data mining techniques. Different techniques are used in practice to classify data in subsets, predict outcomes based on data, group records in subgroups, or score tendencies for certain measures in the records [6]. In agricultural yield prediction, various applications of data mining can be used to forecast the condition of the weather. The K-means techniques can be applied in the prediction of pollution in the atmosphere while daily precipitations simulation is done using the K Nearest Neighbor Algorithm (KNN) [1]. Some of the data mining techniques are an Artificial neural network, Genetic algorithm, Condition tree, and Nearest neighbor method.

Data mining is relatively new in agriculture. This consists of the application of data mining techniques to various aspect of agriculture [7]. Recent Wikipedia article on Data mining for agriculture highlighted some examples of the application of data mining in agriculture. These examples include relationship between sprays and fruit defect, open doubt wine fermentations forecasting, detection of diseases from sounds issued by animals, optimizing pesticide use by data mining (Data mining in agriculture. (2016, October 5)).

Agwu, Nwachukwu, and Anyanwu [8] this study discussed how farmers in Nigeria depends solemnly on the weather because these following factors are the major tools for the development of crop production (water, sunlight, relative humidity, and temperature). The nature of weather in Nigeria agriculture states the vulnerable to climate change and variability [9].

Auroop and Karsten [10] also stated that the temperature, rainfall, sunlight and relative humidity affect the growth and development of crops and yield. These various elements of climate are the major factor that determines the production of agricultural produces. That is the increase in rainfall and normal temperature brings about the high productivity of agricultural produce. The drop in rainfall or scarcity of normal rainfall on some crops may cause failure in the growth and development of such crops. Also, excessive rainfall causes erosion and leaching of the soil nutrient which leads to low yield output. So, therefore, moderate rainfall is required to achieve better agricultural yield [11].

Ankita et al., [9] this researcher predicted the maximum temperature, rainfall, evaporation and wind speed with the aid of data mining techniques, he stated that there are various challenges that arise during the process of weather prediction which includes limited observation of the observation and the history of past data because there is a lot of processing in predicting weather which is not possible enumerate and measure. The rapid increase in communication systems makes it possible for the climate forecaster to integrate and share resources which give rise to the hybrid system. With this improvement in the communication system, the weather forecaster still finds it difficult to forecast the weather.

Rajesh [2] applied spatial data mining in agriculture. The k-means algorithm was used to select at random the number of k objects, which initially denote the cluster mean. And for each of other objects, another object is assigned to the cluster of the most similar object. From this, new mean for each cluster is computed; this process is performed until the criterion function converges. This method is applied in agriculture by supplying temperature an rainfall as the initial spatial data and analyzing the agricultural meteorology for the improvement of agricultural produce as well as reducing the losses of yield.

Ramesh and Vardhan [1] these researchers discussed the importance of data mining techniques to explore agricultural yield data. They laid emphasis on yield prediction and how it is a major problem in agriculture and how it remains to be solved based on the available data. They later concluded that agricultural yield prediction can be solved by engaging various techniques of Data Mining. In his research work, they aim at finding suitable data models that achieve a high generality in terms of agricultural yield capabilities and different types of data mining techniques were evaluated on different datasets.

Olaiya and Adesesan [3, 12] these researchers studied and try to examine the importance of data mining techniques predicting the best temperature, rainfall, evaporation and wind speed. The researchers make use of Artificial Neural Network and Decision Tree algorithms to explore the meteorological data collected from

Ibadan city, Nigeria within the year 2000 to 2009. The researchers make use of classifier algorithms on the meteorological data model collected and compared the performance of the algorithm using standard performance metrics. This algorithm gave the best results and was used to generate classification rules for the mean weather variables. This shows that data mining techniques can be used for weather forecasting.

Multiple linear regression and K-means algorithm will be adopted for this research as there is more than one independent variable (Temperature, Rainfall, CO<sub>2</sub> emission) determining the outcome of our dependent variable (Yield).

Data mining is a relatively new concept in agriculture. This research aims at using data mining technique to study how climatic factors like temperature and rainfall affect crop yield. This research will help form a model for predicting the number of variables that produces the highest quantity of crop. This will also serve as a form of motivator for upcoming researchers to focus more on the application of data mining in the field of agriculture.

### **3. Material and Methods**

The material used in this research work was based on the four modes which are: Data capture, Data normalization, K-means, and multiple linear regression. Historic data collected include average monthly rainfall, temperature, CO<sub>2</sub> emission, and yield.

**K-means classification algorithm** was used to group the dataset and select the most optimum group or combination of rainfall and temperature that produce yields more.

**Multiple Linear Regression** was used to model the linear relationship between a dependent variable (yield) and independent variables or predictors (rainfall and temperature).

**Python programming language** was used to implement the algorithms.

**Flask Framework** was used to implement a web interface to show visualizations of results.

Python machine learning packages numpy (numerical python), matplotlib (plotting for data visualization), Scikit-learn (An efficient tool for machine learning) was also used to implement the algorithms

#### *3.1. Model System Design and Motivation*

Data mining is a relatively new concept in agriculture. This research work aims at using data mining technique to study how climatic factors like temperature and rainfall affects crop yield. This will help farmer form a model for predicting the amount of variables that produces the highest quantity of crop. This will also serve as a form of motivator for upcoming researchers to focus more on the application of data mining in the field of agriculture.

#### **Mode 1:**

1. Data capture: Historic data from FAO (Food and Agriculture Organization) and World Bank Group was for this project. Data from FAO includes yield of roots and tubers and CO<sub>2</sub> emission from 1961 to 2012. Data from World Bank Group include Temperature (Celsius) and Rainfall (mm) from 1961 to 2012.

#### **Mode 2:**

2. Data collected from mode 1 above is normalized in an excel spreadsheet saved using CSV (comma separated values) format. Data normalization includes arranging in rows and columns in which a row describes all columns.

**Mode 3:**

3. K-means: Initial cluster centroids are selected. Each row in the normalized data is then analyzed to determine which cluster they belong too.

**Mode 4:**

4. Multiple linear regressions: To get the predictive model, temperature, CO2 emissions, rainfall column serve as independent variables used in finding intercept and coefficients of the predictive model.

Multiple linear regression equation used for this research work is shown below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (1)$$

Where:

$X_1 = \text{Temperature}$

$X_2 = \text{Rainfall}$

$X_3 = \text{CO}_2 \text{ Emission}$

$Y = \text{Yield of root and tubers.}$

$\beta_0$  = Mean value of Y when all dependent variables (Temperature, rainfall, Co2 emissions) are zero

$\beta_1$  = Regression Co-efficient of X1 (Temperature)

$\beta_2$  = Regression Co-efficient of X2 (Rainfall)

$\beta_3$  = Regression Co-efficient of X3 (Co2 Emission)

$\varepsilon$  = Residual which is the difference between  $Y - \hat{Y}$

The term ‘linear’ is used because in multiple linear regression we assume that y is directly related to a linear combination of the explanatory variables

### 3.2. Method and Experiment Model

The following steps are taken to get the raw data gathered from the FAO (Food and Agriculture Organization) and World Bank group into a normalized data in CSV format.

1. Gather the raw data
2. Raw data for temperature and rainfall is normalized by getting the average from January to December
3. Rows are converted to columns in the CO2 emission data
4. .csv extension file is created
5. Data are imported into the file

## 4. Result and Discussion

In the interface above, the user will select the no of cluster needed to perform the operation and a normalized CSV file containing the raw data is uploaded.

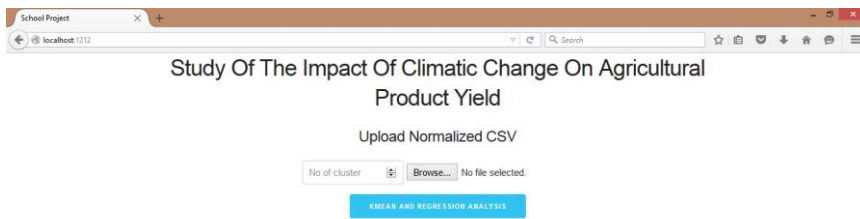


Fig.1. Interface of the Test System

Normalized data consist of data in rows and column. There are four columns respectively. The four columns are Yield, Temperature, Rainfall and CO2 Emission. The first column yield serves the Y dependent variable for multiple linear regression analysis. Initial K-means clusters were selected randomly depending on the choice of the user. K-means clustering was done on 52 rows of data to group them into classes.

	A	B	C	D
1	Roots and Tubers Yield(hg/ha)	Temperature(Celsius)	Rainfall (mm)	CO2 Emission
2	85540	26.31616833	92.83625192	0
3	78775	26.73195617	106.973266	0
4	86340	27.08257375	110.2657637	0
5	83548	26.56858958	97.93592238	0
6	87438	26.57504725	96.7386054	0
7	88814	26.69579725	100.4155989	0
8	89661	26.49689008	95.94890759	0.244682891
9	83274	26.70000017	103.4321691	0.123667699
10	92248	27.2801515	102.5596641	0.220785224
11	97525	26.86894592	93.24743641	0.383738649
12	84533	26.41210858	90.8678757	0.561853839
13	80482	26.91759217	87.69594151	0.704174376
14	80867	27.2230765	81.20377497	0.822384797
15	95906	26.2714155	94.48079393	1.007021127
16	100739	26.21939675	96.42497714	0.745623049
17	95925	26.387963	91.46436333	0.844407389
18	100388	26.49335217	86.9931845	0.749983
19	110490	26.52872725	101.4314568	0.694761021
20	103815	26.93345192	96.88257692	0.981427331
21	96018	26.87613975	100.0645533	0.92478457
22	74543	26.77628158	91.5725998	0.870971897
23	75783	26.86008992	85.6780157	0.843982981

Fig.2. Uploaded Normalized CSV file

### KMeans Result

Row From Excel sheet	Class Label	Class (Centroid)
['85540', '26.31616833', '92.83625192', '0']	2	[87132.05555555556, 26.92611057722222, 96.01842755444444, 0.4053122658888889]
['78775', '26.73195617', '106.973266', '0']	0	[76865.66666666667, 27.03790848222222, 88.58594881666667, 0.6692872495555555]
['86340', '27.08257375', '110.2657637', '0']	2	[87132.05555555556, 26.92611057722222, 96.01842755444444, 0.4053122658888889]
['83548', '26.56858958', '97.93592238', '0']	2	[87132.05555555556, 26.92611057722222, 96.01842755444444, 0.4053122658888889]
['87438', '26.57504725', '96.7386054', '0']	2	[87132.05555555556, 26.92611057722222, 96.01842755444444, 0.4053122658888889]
['88814', '26.69579725', '100.4155989', '0']	2	[87132.05555555556, 26.92611057722222, 96.01842755444444, 0.4053122658888889]

Fig.3. Result of K-means Analysis

Fig. 3 shows the result of K-means analysis. K-mean is an unsupervised learning algorithm.

This figure displays the result of Multiple Linear Regression Model with the Y (yield) while the scatter plots are displayed below.

### Multiple Linear Regression Model

Intercept: 25187.0645302

Co-efficient for X1 - Temperature(Celsius): 653.002861067

Co-efficient for X2 - Rainfall (mm): 482.380210578

Co-efficient for X3 - CO2 Emission: 7028.45816707

$$Y \text{ (Yield)} = 25187.0645302 + 653.002861067X_1 + 482.380210578X_2 + 7028.45816707X_3$$

Fig.4. Result of Multiple Linear Regressions.

### 5. Scatter Plots

This figure displays the scatter plots for the temperature of the regression analysis with roots and tuber yields on the y-axis while the temperature on the x-axis.

#### Scatter Plots

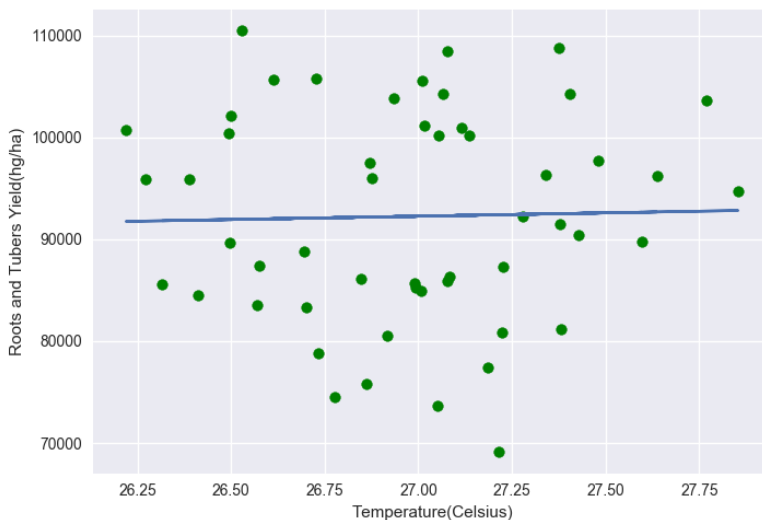


Fig.5. Scatter plots for Temperature

Fig. 6 displays roots and tuber yields on the y-axis while rainfall on the x-axis for a scatter plots of rainfall regression analysis.

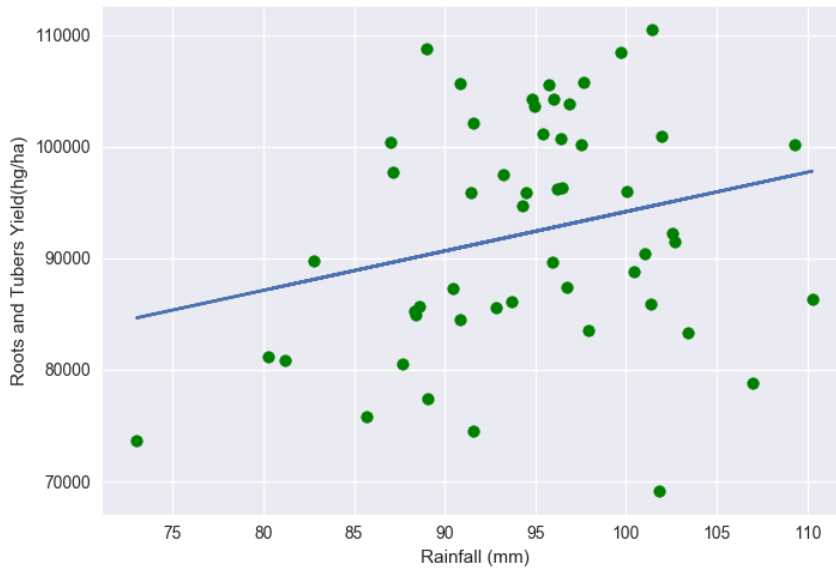


Fig.6. Scatter plot for Rainfall

Fig. 7 displays roots and tuber yields on the y-axis while CO2 on the x-axis for a scatter plots of CO2 Emission regression analysis.

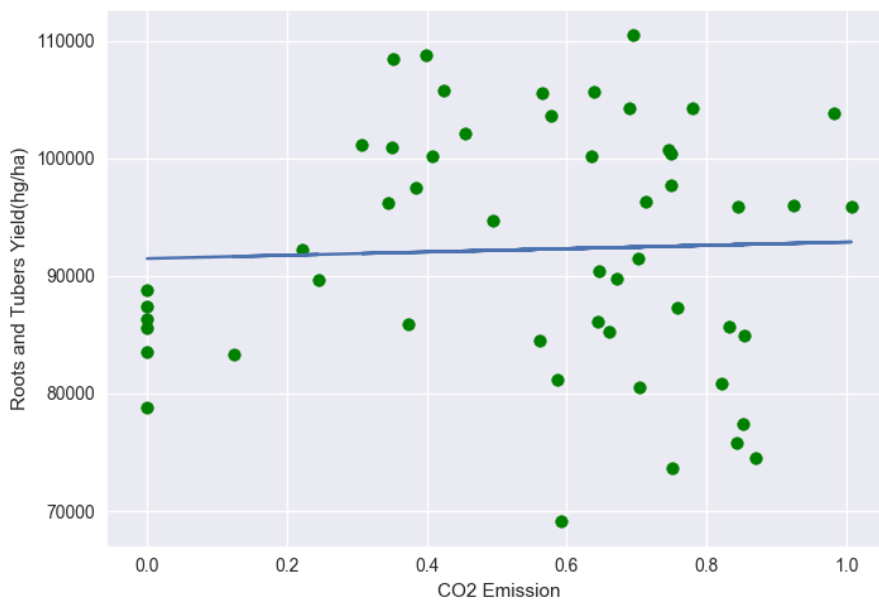


Fig.7. Scatter plot for CO2 Emission



## 6. Experiment and Innovation Interpretation

### 6.1. K-means Analysis Results

From the normalized CSV file, it shows that 18 rows, belonged to cluster 2, 25 rows belonged to cluster 1 and 9 rows belonged to cluster 0 after K-means analysis assuming the number of initial centroids selected is 3. The final updated centroids are as follows:

Cluster 2 = [87132.05555555556, 26.92611057722222, 96.01842755444444, 0.4053122658888889]

Cluster 1 = [101483.28, 26.9902397176, 95.5801255584, 0.60865230008]

Cluster 0 = [76865.66666666667, 27.03790848222222, 88.58594881666667, 0.6692872495555555]

We can say from the K-means analysis, after clustering, that member in cluster 1 has the highest yield/optimum yield because, from the centroids, cluster 1 has the highest yield so members or rows in these clusters have high optimal yield.

### 6.2. Multiple Linear Regression Analysis Result

Multiple Linear Result includes the following:

Intercept: 25187.0645302

Co-efficient for X1 - Temperature (Celsius): 653.002861067

Co-efficient for X2 - Rainfall (mm): 482.380210578

Co-efficient for X3 - CO2 Emission: 7028.45816707

**Y (Yield) = 25187.0645302 + 653.002861067X1 + 482.380210578X2 + 7028.45816707X3**

**R2 (R-squared) = 0.092**

From these results, we can predict the value of Y if values for X1, X2, and X3 are known

The graph is a scatter plot the two-dimensional graph. Other x-axis is the x independent variable while other y-axis is the dependent variable which is the yield. Three graphs were created separately so we can individually study the relationship between x variable and y variable. In other words, study how temperature, CO2 Emission, and rainfall affect the yield of root and tubers. The plot shows a straight linear line to signify relationship. This relationship is stronger in rainfall as compared to others.

## 7. Conclusion and Recommendation

In conclusion, the implementation of the adopted predictive model shows that the three variables (temperature, rainfall, and CO2 Emission) are not enough to predict agricultural yield. These can be confirmed by regression analysis and the scatter plot. Only rainfall shows a strong linear relationship as compared to CO2 emission and temperature changes. Also from the  $R^2$  (RSquared), this low value further confirms that the final model cannot be used to predict agricultural yield. This is due to the fact that other factors are involved in the yield of root and tuber crops. These factors include soil type, humidity, sunlight, human activities etc.

Variables like human activities play an important role in determining the yield of root and tuber crops in Nigeria. Further research should be carried out to determine how other climatic factors such as soil type, humidity, sunlight etc. affect the yield of crops.

## References

- [1] Ramesh D, and Vardhan V. Data Mining Techniques and Applications to Agricultural Yield Data. Int. JARCECE. 2013 Sept; 2(9).
- [2] Rajesh D. Application of Spatial Data Mining for Agriculture. Int. J. Com. App. 2011 Feb; 15(2):7-9.
- [3] Olaiya F, and Adesesan A. Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. Int. J. IEEB. 2012 Feb; 1(7): 51-59.
- [4] Mohammed JZ, Wagner MJ. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University 2014.
- [5] Milovic B, Radojevic V. Application of Data Mining in Agriculture. Bul. J. Agric. Sci. 2015; 21(1): 26-34.
- [6] Jiawei H, Micheline K, and Jian P. Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA. 2012.
- [7] Ian-HW, Eibe F. Data Mining Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems. 2005 2nd Edition; 560pp.
- [8] Agwu N, Nwachukwu, Anyanwu CI. Climate Variability: Relative Effect on Nigeria's Cassava Productive Capacity. Sci. Jor. Pub. 2012; 4(12): 11-14.
- [9] Ankita J, Bhagyashri K, Vaibhavi J, et al. Weather Forecasting and Climate Changing Using Data Mining Application. Int. JARCECE. 2015 March; 4(3).
- [10] Auroop G, Karsten S. Data Mining for Climate Change and Impacts. IEEE Int. Conf. Data Mining Workshops. 2008. 385-394
- [11] Shikonun N, El-Bolok H, Ismail MA. Climate Change Prediction Using Data Mining. IJICIS. 2005 July; 5(1).
- [12] Olaiya F. "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", (2012). <https://www.researchgate.net/publication/265750889>
- [13] Witten H. and Frank E. "Data Mining Practical Machine Learning Tools and Techniques". The Morgan Kaufmann Series in Data Management Systems (2005).

## Authors' Profiles



**Babatunde GBADAMOSI** is a Faculty member in the Department of Computer Science, Landmark University, Nigeria. He received his Ph.D. degree from Ladoke Akintola University of Technology, Omu-Aran, Nigeria. His current research mainly covers Mathematical Modeling of Infectious Disease and Mathematical Computations of nonlinear problems arising in Applied Mathematics.



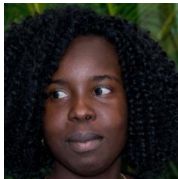
**Adeniyi, Abidemi Emmanuel** is a Staff in the Department of Computer Science, College of Pure and Applied Sciences, Landmark University, Omu-Aran, Nigeria. He received Postgraduate Diploma in Computer Science from University of Ilorin, Ilorin, Nigeria. His area of research includes Computer Security, Data Mining.



**Ogundokun Roseline Oluwaseun** is a Lecturer/Researcher in the Department of Computer Science, Landmark University, Omu-Aran. Her are of research includes Information Security, Data Mining and Information Science. Sh holds Bachelor of Science in Management Information Security from Covenant University, Ota; Master of Science in Computer Science from the University of Ilorin, Ilorin; Post Graduate Diploma in Education (PGDE) from the National Teachers' Institute (NTI), Kaduna and currently a PhD student in the Department of Computer Science, University of Ilorin, Ilorin. She is a member of Nigeria Computer Society (NCS); Computer Professionals (Registration Council of Nigeria) (CPN) and Graduate Member of Nigeria Institute of Management (NIM).



**Ehiedu Precious Anyaiwe** is a Staff in the Department of Computer Science, Landmark University, Omu-Aran, Nigeria. He obtained M.Sc. in Information Science in the year 2015. He is Career oriented and have the potentials of managing changes and crises, and can work under pressure.



**Oladosu Bukola** is a Graduate of Computer Science from Landmark University, Omu-Aran with specialization in data science and Business Intelligence, This statement should be deleted. She is a Strategic thinker and effective communicator. Graduate member of Computer Professional (Registration Council of Nigeria) (CPN) 2017 and a certified Project Manager. She is Currently a content/e-learning developer and resides in Abuja.

**How to cite this paper:** Gbadamosi Babatunde, Adeniyi Abidemi Emmanuel, Ogundokun Roseline Oluwaseun, Oladosu Bukola Bunmi, Anyaiwe Ehiedu Precious,"Impact of Climatic Change on Agricultural Product Yield Using K-Means and Multiple Linear Regressions", International Journal of Education and Management Engineering(IJEME), Vol.9, No.3, pp.16-26, 2019.DOI: 10.5815/ijeme.2019.03.02