**APPLICATION OF MACHINE LEARNING FOR PREDICTING CASSAVA YIELD BASED ON VEGETATIVE GROWTH INDICES**

**By**

OMEJE CHINECHEREM JUSTUS

**2016/239173**

**DEPARTMENT OF AGRICULTURE AND BIORESOURCES ENGINEERING**

**FACULTY OF ENGINEERING,**

**UNIVERSITY OF NIGERIA NSUKKA**

**ADVISER: PROF. O. A. ANI**

**NOVEMBER, 2022**

TITLE PAGE

Application of machine learning technique for predicting cassava yield based on vegetative growth indices

DEDICATION

This work is dedicated to God almighty and my family as a whole.

ACKNOWLEDGMENT

As the will of God is been done, gratitude to God the creator, I appreciate my dear parents who made my studying possible up to this point and also my friend who brought out his time to put me through this project, for the knowledge been passed to me during the course of this project. I also want to use this opportunity to appreciate my supervisor in a very honorable way for his kind gestures, patience, guidance, directions, dedication and most importantly for his knowledge impact in the new course of the programming field, may almighty Godbless him immensely.

DECLEARATION

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Bachelor of Science in B.Eng. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

**Signed:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**Signed:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

OMEJE CHINECHEREM JUSTUS H.O.D :PROF.ANYADIKE C.C.

**Registration No.:** 2016/239173

**Date:**30th November 2022.

**Signed:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**Signed:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ADVISER:** PROF.O.A.ANIEXTERNAL SUPERVISOR

# **ABSTRACT**

This paper investigates the use of machine learning in agriculture, developing a prediction model to predict cassava tuber yields. The objectives are of work is to identify, adapt and apply a suitable machine learning algorithm for determining the relationship between the vegetative growth indices and yield of cassava. In this work, we performed an analysis using K-nearest neighbor (KNN) linear regression and random forest as the machine learning algorithm the data set was generated, artificially which is made up of 130 total dataset, split in two, 80% (104 data) for training sets and 20% (26 data) for test set imported in to the online python programming space (colab.research.google.com) installed on a PC 4gb ram, personal computer with Windows 10,Intel (R) Core (TM) i3-6500 CPU (2.50 GHz), 16.00 GB memory and NVIDIA 1060 GTX GPU. The total leaf length and stem diameter dataset in centimetre and the cassava tuber yield in kilogram (kg) were used in the training and testing of the models. Linear regression had an accuracy of 0.7610 k-nearest neighbour had an accuracy of 0.5596 while random forest had an accuracy of 0.8585. from the results of the three algorithms, k-nearest neighbour performed poorly, while random forest turn out to be good with an accuracy that is high enough up to 80%, therefore random forest was our chosen algorithm The value of the predicted value and the actual value were almost close. This shows that the model was good and can be improved when larger number of dataset is used to train the model. And also machine learning technique can be applied in agriculture.

Table of Contents

[**ABSTRACT** vi](#_Toc121821608)

[**CHAPTER ONE** 1](#_Toc121821609)

[**INTRODUCTION** 1](#_Toc121821610)

[**CHAPTER TWO** 3](#_Toc121821611)

[LITERATURE REVIEW 3](#_Toc121821612)

[2.1 Definition of Terms 23](#_Toc121821614)

[**CHAPTER THREE** 25](#_Toc121821615)

[MATERIAL AND METHOD 25](#_Toc121821616)

[3.0 Materials/resources: 25](#_Toc121821617)

[3.1 Algorithms 26](#_Toc121821618)

[**3.2 IMPLEMENTATION** 27](#_Toc121821619)

[3.2.1 Experimental Setup 27](#_Toc121821620)

[**CHAPTER FOUR** 32](#_Toc121821621)

[RESULT AND EVALUATION 32](#_Toc121821622)

[4.0. Results 32](#_Toc121821623)

[4.1 Correlation Analysis 35](#_Toc121821624)

[**CHAPTER FIVE** 36](#_Toc121821625)

[CONCLUSION AND RECOMMENDATION 36](#_Toc121821626)

[REFERENCE 37](#_Toc121821627)

[Appendix 45](#_Toc121821628)

# **CHAPTER ONE**

# **INTRODUCTION**

**1.0 Background Of Study**

New technologies such as machine learning have emerged over the last years that can potentially aid farmers’ decision making (Hoogenboom et al 2004) which Prediction models can contribute to better experimental planning and management decision making. There have been a few applications of modeling methods to predict the cassava yields comparing experiences and past historical data farmers rely on such as the crop yields and weather to make important decisions to increase short-term profitability and long-term sustainability of their operation (Arbuckle and Rosman 2014) mathematical model such as linear regression model was used to predict the crop yield. However, a linear relationship assumption must be assumed prior to fit the model. Hence this method might not be suitable for more complex relationship between input variables and output response. Therefore the science of training machines to learn and produce models for future predictions is widely used.

Machine learning approaches are used in many fields, ranging from supermarkets to evaluate the behavior of customers (Ayodele, 2010) to the prediction of customers’ phone use (Witten et al., 2016). Machine learning is also being used in agriculture for several years (McQueen et al., 1995). Crop yield prediction is one of the challenging problems in precision agriculture, and many models have been proposed and validated so far. This problem requires the use of several datasets since crop yield depends on many different factors such as climate, weather, soil, use of fertilizer, and seed variety (Xu et al., 2019). This indicates that crop yield prediction is not a trivial task; instead, it consists of several complicated steps. Nowadays, crop yield prediction models can estimate the actual yield reasonably, but a better performance in yield prediction is still desirable (Filippi et al., 2019a). Machine learning, which is a branch of Artificial Intelligence (AI) focusing on learning, is a practical approach that can provide better yield prediction based on several features. Machine learning (ML) can determine patterns and correlations and discover knowledge from datasets. The models need to be trained using datasets, where the out-come are represented based on past experience. The predictive model is built using several features, and as such, parameters of the models are determined using historical data during the training phase. For the testing phase, part of the historical data that has not been used for training is used for the performance evaluation purpose. An ML model can be descriptive or predictive, depending on the research problem and research questions. While descriptive models are used to gain knowledge from the collected data and explain what has happened, predictive models are used to make predictions in the future (Alpaydin, 2010). ML studies consist of different challenges when aiming to build a high-performance predictive model. It is crucial to select the right algorithms to solve the problem at hand, and in addition, the algorithms and the underlying platforms need to be capable of handling the volume of data. For many years, we have used computers to work, play, communicate, watch movies, and do plenty of other things.

**1.1Aim and specific objectives**

The aim of this work is to identify, adapt and apply a suitable machine learning algorithm for determining the relationship between the vegetative growth indices and yield of cassava.

**1.2 Specific Objectives**

**The specific objectives are to:**

1. Evaluate three machine learning algorithms (Random Forest, K-Nearest Neighbour-KNN and Linear Regression) for effectiveness in predicting cassava yield based on total leaf length (TLL) and stem diameter (SD)
2. To generate a suitable artificial dataset of TLL, SD and Yield by trial and error
3. To use the dataset to train and validate the machine learning algorithm
4. To compare the predicted and actual yield

# **CHAPTER TWO**

## LITERATURE REVIEW

Machine learning has the ability to learn the machine without defined computer programming, so it improves machine performance by detecting and characterizing the consistency and pattern of drive data. Machine learning can be classified into three categories according to the learning method –Supervised learning, Un-supervised learning and Reinforcement learning. For this project I will be making use of supervised learning algorithms to predict crop yield. This type of algorithms helps to build most accurate and effective model because here, the learning data comes with labels or desired outputs and the objective is to find a general rule of mapping input to output. It involves building a machine learning model that is based on labeled samples. The proposed system analyzes the application of supervised machine learning approaches in predicting the tuber yield. The SVM, KNN and Decision trees have been used for yield prediction. Every year, we develop computers that are better and faster. However, this is not enough for us—we dream of supercomputers with computing power that go beyond our imagination. Such computers would allow us to manipulate much more data in a short time. Scientists get inspiration for these powerful computers from the way the brain is built: full of cells called neurons, connected to each other in a huge net in which electrical impulses transfer data between the neurons. Scientists are trying to create a digital version of the brain’s natural neural network. They create artificial neurons that are connected within a huge network in which, instead of the electrical impulses used in our brains, data is represented by digital numbers in electronic circuits. This is called an [artificial neural network](https://kids.frontiersin.org/articles/10.3389/frym.2021.560631#KC2) and these networks can perform some tasks, like image recognition, very efficiently. But we have still not reached the limit of what computers can do with artificial neural networks. Maybe 1 day, computers will be able to think like humans!

Did you know that, in the span of just 1 s, the brain can receive billions upon billions of signals? That is a huge amount of data flowing through the brain every single second! The human brain is our on-board computer, and it is the most complex machine that has ever existed. It is much smaller than a soccer ball but has more cells than there are stars in the Milky Way! If the body were a ship, the brain would be the captain. The brain can adapt very quickly to completely new, unfamiliar situations. It can also recognize objects much faster than the world’s best computers can. When you see the face of your best friend, your brain recognizes that face faster than a mosquito flaps its wings. Computers can recognize multiple complicated patterns, but even the fastest computer cannot yet compete with the human brain. Scientists from various fields are still trying to understand all the processes that take place in the brain when it carries out tasks like pattern recognition.

**2.0 Fundamentals of Machine Learning:**

A Brief Overview In general, the objective of ML algorithms is to optimize the performance of a task, via exploiting examples or past experience. In particular, ML can generate efficient relationships regarding data inputs and reconstruct a knowledge scheme. In this data-driven methodology, the more data are used, the better ML works. This is similar to how well a human being performs a particular task by gaining more experience. The central outcome of ML is a measure of generalize ability; the degree to which the ML algorithm has the ability to provide correct predictions, when new data are presented, on the basis of learned rules originated from preceding exposure to similar data. More specifically, data involve a set of examples, which are described by a group of characteristics, usually called features. Broadly speaking, ML systems operate at two processes, namely the learning (used for training) and testing. In order to facilitate the former process, these features commonly form a feature vector that can be binary, numeric, ordinal, or nominal. This vector is utilized as an input within the learning phase. In brief, by relying on training data, within the learning phase, the machine learns to perform the task from experience. Once the learning performance reaches a satisfactory point (expressed through mathematical and statistical relationships), it ends.

Since the invention of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are been concentrated on cultivating artificial products that is hybrid products where there leads to an unhealthy life. Nowadays, modern people don’t have awareness about the cultivation of the crops in a right time and at a right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air, which lead to insecurity of food. By analyzing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India there are several ways to increase the economical growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining also useful for predicting the crop yield production. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information.

## Data mining software is an analytical tool that allows users to analyze data from many different dimensions or angles, categorize, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The patterns, associations, or relationships among all this data can provide information. Information can be converted into knowledge about historical patterns and future trends.

**Convolutional Neural Networks**

CNN (Convolutional Neural Networks) is playing an important role in many sectors like image processing. It has a powerful impact on many fields. Even, in Nano-Technologies like manufacturing semiconductors, CNN (Convolutional Neural Networks) is used for fault detection and classification. Handwritten digit and characters recognition has become an issue of interest among researchers. There are a large number of papers and articles are being published these days about this topic. In research, it is shown that Deep Learning algorithm like multilayer CNN (Convolutional Neural Networks) using Keras with Theano and TensorFlow gives the highest accuracy in comparison with the most widely used machine learning algorithms like SVM, KNN & RFC, Darmatasia and Mohamad Ivan Fanany (2017). Because of its highest accuracy, Convolutional Neural Network (CNN) is being used on a large scale in image classification, video analysis, etc. Many researchers are trying to make sentiment recognition in a sentence. CNN (Convolutional Neural Networks) is being used in natural language processing and sentiment recognition by varying different parameters. It is pretty challenging to get a good performance as more parameters are needed for the large-scale neural network. Many researchers are trying to increase the accuracy with less error in CNN (Convolutional Neural Networks). In another research, they have shown that deep nets perform better when they are trained by simple back-propagation. Their architecture results in the lowest error rate on MNIST compare to NORB and CIFAR10, Nair. Pranav P, Ajay James, and C Saravnan (2017). Researchers are working on this issue to reduce the error rate as much as possible in handwriting recognition. In one research, an error rate of 1.19% is achieved using 3-NN trained and tested on MNIST.

Machine learning in Agriculture is a Novel field and a great deal of work has been done in field of Agriculture utilizing Machine learning. There are diverse guaging philosophies created and assessed by the specialists everywhere throughout the world in the field of farming or related sciences. Agricultural scientists in Pakistan have demonstrated that endeavors of harvest yield amplification through expert pesticide state strategies have prompted a hazardously high pesticide use. These examinations have revealed negative relationship between's pesticide use and harvest yield. In their investigation they have explained that how data mining incorporated farming information including irritation exploring, pesticide utilization and meteorological information are helpful for streamlining of pesticide use. Topical data identified with agribusiness which has spatial properties was accounted for in one of the study. Their research went for perceiving patterns in farming creation with references to the accessibility of information assets. K-means method turned into applied to carry out gauges of the contamination in the air, the k- nearest neighbor become connected for mimicking day by day precipitations and other climate elements, and numerous ability changes of the weather situations are dissected utilizing SVM. Statistics mining techniques are often used to have a look at soil qualities. As example, the kmeans method is used for segmenting soils in mixture with GPS-based technology. A decision tree classifier for agriculture information turned into proposed. This new classifier uses new facts expression and can address each entire records and in entire records. Inside the test, 10-fold cross validation technique is used to check the dataset, horse-colic dataset and soybean dataset. Their results showed the proposed selection tree is capable of classifying all styles of agriculture records. A yield prediction version turned into proposed in one of the take a look at which makes use of data mining techniques for category and prediction. This model worked on enter parameters crop name, land location, soil type, soil ph, pest information, climate, water stage, seed type and this model anticipated the plant boom and plant diseases and therefore enabled to select the nice crop based on climate information and required parameters. There are few research works about sugarcane yield prediction which can be associated with our work. Sugarcane yield prediction technique with use of Random forest became proposed in one of the survey, the features used in this study consist of biomass index, climate statistics (e.g., rainfall) and yields from previous years. Two predictive tasks are provided in:

1. the category problem for predicting whether or not the yield can be above or underneath the found median yield, and
2. the regression hassle for predicting the yield estimates in two distinct time intervals. In addition, support vector system for rice crop yield prediction become proposed, the dataset used in this method are precipitation, minimum, maximum and common temperature, place, evapotranspiration and manufacturing. The sequential minimal optimization classifier is implemented on the dataset. The dataset is processed through WEKA tool to build the set of rules on the current dataset. The results were generated in python by using SVM algorithm. In based on the C4.5 algorithm, decision tree and decision rules have been developed in their study they have developed a website called Crop Advisor: This is an interactive website for discovering the affect of weather and crop production by using C4.5 algorithm. This gives the idea of how different climatic parameters impact the growth of the crop. The selections were made based on the area under the chosen crop. The information regarding the associated years climatic parameters like rainfall, high and low temperature, wet day frequency where collected. The id3 algorithm were developed to get good quality and improved cassava tuber yield which is implemented in PHP platform and uses csv as data sets.

Agriculture plays a critical role in the global economy. With the continuing expansion of the human population understanding worldwide crop yield is central to addressing food security challenges and reducing the impacts of climate change. Crop yield prediction is an important agricultural problem. The Agricultural yield primarily depends on weather conditions (rain, temperature, relative humidity, radiation, Co2 and VPD), pesticides. Accurate information about history of crop yield is important for making decisions related to agricultural risk management and future predictions. Yield development is influenced mainly by fruit temperature. This parameter is inversely related to rate and shows a linear relationship with air temperature. To understand production levels, yield prediction is carried out which involves predicting the yield of the crop based on the existing data. Previously, crop yield estimates were based on farmer's specific crops and cultivation experience. There are many ways to enhance and improve crop yield and quality. Data mining techniques are also helpful for predicting crop yields. In general, data mining analyzes knowledge from various approaches and summarizes it as profitable information. Data mining software is an analytical tool that allows users to classify and summarize identified relationships as well as analyze data at various angles or dimensions. Technically, data mining is finding correlations or patterns of fields in large relational databases. All of these data can provide information between models, connections, or relationships. Knowledge can be transformed into historical patterns and knowledge of future trends. For example, a survey of agricultural products helps farmers to suggest and prevent future crop losses. Researches have been conducted to develop an efficient method for yield prediction but focus has been always on statistical techniques and not much has been done in machine learning approach. The crop production depends on various factors which change with every square meter and depends on:

1. Geography of region,

2. Weather (Temperature, humidity, precipitation),

3. Soil type (saline, alkaline, sodic, non-alkaline),

4. Soil composition (ph, N, P, K,EC,OC, Zn, F).

Various subsets of these parameters are used in different prediction models for various crops. Prediction models are essentially two main types.

* 1. Statistical models, which use a single prediction function that includes all sample spaces.
  2. Machine learning technology, a new technology for knowledge search that connects input and output variable models.

Machine learning algorithms have been applied individually using the Cross Validation techniques with 10 folds and accuracy of prediction has been observed for each of them. In this paper the accuracy for SVM was calculated for two different kernels i.e, svm \_rbf and svm\_linear among these two RBF kernel was showing more error rate. The decision tree is giving more accuracy with very less MSE.

This study provides information on how to apply data analysis to the data set of sugar cane crops. There are three sets of data: soil data sets, rainfall data sets, and crop yield data sets. This data set consists of a variety of parameters that are useful for identifying status of crops and conducting supervisory training on data sets collected from agriculture domain to divide information into multiple classes. This paper shows the comparison of three different algorithms like, decision tree, KNN and SVM. These algorithms were used to train the 0.8 or 80 percentage of the input data and are tested with the remaining 0.2 or 20 percentage of test dataset and results of the algorithms were compared based on accuracy and mean square error. Here, the decision tree algorithm is giving more accuracy of 99% and also the mean square error for this algorithm is very less. This system will help to reduce the problems faced by farmers and will serve as an intermediary to provide farmers with the information they need to earn high profits and maximize profits.

KefayaQaddoum, E. L. Hines, and D. D. Iliescu 2013, proposed an automatic tomato yield predictor to assist the human operators in anticipating more effectively weekly fluctuations and avoid problems of both over demand and overproduction if the yield cannot be predicted accurately. The parameters used by the predictor consist of environmental variables inside the greenhouse, namely, temperature, CO2, vapour pressure deficit (VPD), and radiation, as well as past yield. Greenhouse environment data and crop records from a large scale commercial operation, Wight Salads Group (WSG) in the Isle of Wight, United Kingdom, collected during the period 2004 to 2008, were used to model tomato yield using an Intelligent System called “Evolving Fuzzy Neural Network” (EFuNN). Our results show that the EFuNN model predicted weekly fluctuations of the yield with an average accuracy of 90%

IS techniques were robust in dealing with imprecise data, and they have a learning capability when presented with new scenarios and need to be tested on different tomato cultivars. Experimentation results show that EFuNN performed better than other techniques like ANN in terms of low RMSE error and less computational loads (performance time). ANN training needs more epochs (longer training time) to achieve a better performance.EFuNN makes use of the knowledge of FIS and the learning done by NN. Hence, the neurofuzzy system is able to precisely model the uncertainty and imprecision within the data as well as to incorporate the learning ability of NN. Even though the performance of neurofuzzy systems is dependent on the problems domain, very often the results are better while compared to a pure neural network approach. Compared to NN, an important advantage of neurofuzzy systems is its reasoning ability (If-Then rules) within any particular state. A fully trained EFuNN could be replaced by a set of If-Then rules.

**2.1 Review of Related Literature**

Gil, Hassner, Haider and Hayder(2013) Neural networks are playing a major role in machine learning applications, since neural networks work like human brain, they try to mimic the working of neurons in human brain that is why they are nowadays used to build self-driven cars that can park themselves without any driver. Google deep minds are also working in this field to develop board playing games.

Dr. Kusum Gupta (2016) In machine learning applications, we need different features for the prediction. So, the feature recognition helps us in extracting features and their parameters. If we have more features then it will sometimes lead to over fitting. When there is large difference between train accuracy and test accuracy, then model will over fit that is why feature specification is very important. We have to select those features only which will help us in better prediction in the application.

Faisal and Kamran (2007) Before 1980’s and 1990’s the data sets are very less and the neural networks are fringe but in 21st century the data sets are present in large amount that is why neural networks are making a big come back and neural networks helps us to make large number of machine learning applications. Later in 1980’s developing of speech recognition was a very difficult task but now this is not a very difficult task because of the increase in the data set.

Sherif, Maha and HanyAhmed (2012) Neural networks can be used as powerful tool for signal and image processing. In this research paper the online Arabic handwritten digits recognition is done. The neural networks have lot more potential to do things by increasing the number of hidden layers in the network, all the processing is done by the hidden layer nodes.

Yusuf and Ashish, (2011) To recognize the alphabets we can represent the English alphabets in the binary format in which the input is given to the simple feature extraction system and the output is fed the neural network that we have created.

Gregory, Saeed and Jonathan, (2017) In this research paper MNIST is introduced which is a dataset which is extended version of the NIST. This dataset contains about 60k images and the images are only the black and white images and because of its large data set this dataset help us to train our algorithm more accurately. MNIST also solves the problem of data set for the neural networks.

Diederik and Jimmy, (2017) Adam Optimizer is a variant of stochastic gradient descent algorithm. Adam is an abbreviation for adaptive moment estimation. It is the current best choice among gradient based convex optimization algorithm. One advantage is that hyper parameters require less tuning. This method is very efficient, requires very little memory requirements and is best for problems which are having large data or parameters.

Zhang and Woodland, (2015) The creation of the hidden layers is always an issue and the function that can be used in the hidden layer is also an issue, the most common function is the sigmoid function and ReLU function. In this paper we have studied about the various parameters of the ReLU function that has also helped in our project.

Glorot and Bengio, (2010) In this research paper we have studied that there are some problems in finding the local minima of the algorithm and there can be problem of over fitting and under fitting, the training of deep neural networks is also a major issue for this we have to do random initialization.

Hinton and Salakhutdinov, (2006) In this research paper we have studied that sometimes the dimensionality of the features may be a major issue we have to keep the features as small as possible for this there are many dimensionality reduction techniques to do a better prediction.

Hush and Horne (1993) In this research paper we have studied that the supervised learning can be of two types static and dynamic this paper tells us that the dynamic supervised learning algorithms have memory while the other one that is static supervised learning algorithm have no memory.

Cun and Boser, (1992) In this paper an application for the handwriting using back propagation is given and this method is shown to have 1% error and 9% reject rate. We have to perform back propagation for output layer to the input layer.

Kessab1 and Daoui (2013) In this research paper the hand writing detection is shown through the optical character reorganization (OCR) and the success rate through this will be 80%.

LeCun and Bottou (2001) In this research paper it is shown that the number of layers is varied in the neural network and the gradient descent back propagation algorithm is applied for the application.

K. G. Liakos et al., (2018) presented a comprehensive review of research dedicated to machine learning applications in agriculture domain. Various parameters on which work was analysed were: crop management, livestock management, water management and soil management. ML models have applied for crop yield prediction and disease detection. ML based detection can be extracted without the need of fusion of data from other resources. Author claims that farm management systems are evlvong into real artificial intelligent systems, with the ultimate scope of production improvement. Author motivates to use ML for the benefit of agriculture as it is the basic need amongst all other needs for survival.

Crane Droesch (2018), has used data on corn yield from the USMidwest, and shown that the approach of using semi-parametric variant of deep neural network, accounting for complex non-linear relationship in high dimensional dataset, the model will outperform both classical statistical methods and fully non-parametric neural networks in predicting yield of years withheld during model training. Authors have developed a novel approach for augmenting parametric statistical model with deep neural networks, they have termed it as semi parametric neural networks. It is used as a crop yield modeling framework, the SNN achieves better out of sample predictive performance than anything else yet published. it uses prior knowledge of functional phenomenon and functional form relating them to the outcome. So the SNN improves statistical efficiency over typical neural networks. They found that combining ML with domain area knowledge from empirical studies improves predictive skills, while altering conclusions about climate change impact to agriculture.

P. Priya et al., (2018) has proposed a random Forest Algorithm for predicting the crop yield of particular area considering various parameters such as rainfall, seasonal crop (Rabi and Kharif) district-wise, temperature (max.), crop production in terms of Kgs/tonnes. Area for doing research was Tamil Nadu. Dataset record were collected from Indian Government over 15years for rice production. They proved in experimental results that prediction analysis done using Random Forest Algorithm – a supervised machine learning algorithm will help farmer to predict the yield of the crop before cultivating onto the agricultural field. This algorithm runs efficiently on large databases with high classification accuracy.

A. L. Ismail et al.(2018) [1] created a framework to predict preparedness of a country to face the climate change using machine learning approach. The study is done for South East Asia. Steps for calculating the predictive index are data acquisition, data training, data testing, index predicting, index validation and index visualization. The study is a precautionary measure to alert the regions and verify its vulnerable index using deep learning.

Zhen Nan Liu, et al., (2018), In this paper, authors have compared different machine learning algorithms for calculating, Standardized Precipitation Index (SPI) and SPEI. After data collection, Extreme learning methods, online sequential extreme learning machine, Selfadaptive evolutionary extreme learning machine. Authors claimed that all three algorithms can be applied successfully on drought forecasting. However, OS-ELM and SADE- ELM performs better than ELM.

Chaoyun Zhang, et al., (2018) presented an ample survey of the crossovers between the two areas. A brief study of applications of networking using deep learning techniques is done. We then discuss several techniques and platforms that facilitate the efficient deployment of deep learning onto mobile systems. Authors focus on how can deep learning can be useful for mobile and wireless networking. This is a survey paper surfacing the issues and challenges in deep learning in wireless and mobile networks. N. Zhu, et al.(2018) This article summarizes DL algorithms, considering concepts, constraints, implementation, procedure of training, and sample codes, to aid researchers in agriculture to facilitate with DL techniques quickly. Research on DL applications in agriculture is summarized and analysed, and future opportunities are discussed in this paper, which is expected to help researchers in agriculture to better understand DL algorithms and learn major DL techniques quickly, and further to facilitate data analysis, enhance related research in agriculture, and thus promote DL applications effectively.

S. Rasp, et al. (2018), presents a different perspective to sub-grid parameterizations to a DDA that influences the benefits of high-resolution modeling. Challenges to overcome, but advances in computing capabilities and deep learning in recent years present novel opportunities that are just beginning to be investigated. Authors believe that machine-learning approaches have huge potential to be explored connection with development of traditional model.

Jinyoung Rhee et al., (2018) has targeted all officials whose main duties include water resources and agricultural management. The final beneficiaries of the output are residents of the area; water users and farmers for whom decision-making can be helped by drought prediction information with finer spatial resolution The models provide spatially distributed detailed drought prediction data of the 6-month Standardized Precipitation Index for the case study area, Fiji. They used Weather Research Forecasting (WRF) model as reference data for overcoming the limitations of non-dense monitoring network. Also they used Performance measures of the mean absolute error as well as classification accuracy. The WRF outputs reflect the topography of the area. Hybrid models showed better performance than simply bias corrected forecasts in most cases. The model based on Extra-Trees trained using the WRF model outputs performed the best in most cases.

S. D. Patil et al., (2017) suggests that according to their results, direct prediction of spectral band information is highly beneficial due to the ability it provides for deriving ecologically relevant products which can be used to analyse land cover change scenarios from multiple perspective. Aim of the authors, is to enhance the use of machine learning based land cover change models to predict the spectral band information of satellite based land cover images. Experimental areas covered by authors are in some portion of United States. They used data from two large sites in US to train model RF machine learning model to spectral values from bands. They used the trained model to explore the look of land cover for a climate change scenario. The demonstrative results show that the direct prediction ofspectral band information is helpful for deriving ecological products. They have considered this as a major strength of their proposed approach as it has enabled the analysis of land cover change from multiple viewpoints The authors have made a comment through their literature survey, that in 2081-2100 projected rise in temperature will be 1.5o – 4.8oC than 1986-2005 era. This will impact global landcover. timely and accurate prediction may provide useful solutions. Author has chosen RF model of ML in the study as it ensembles constitutes are comprised of DT models that offers variety of attractive features over other statistical learning techniques. Parameter selection was done based on their judgement of importance of factors. They urge research scholars to continue the work by analysing other predicted variables or modifying the data sources.

AnnelieHolzkamper (2017) gives a systematic literature review in modeling for adaptation in planning in agricultural production systems. The author has studied five types of models namely empirical crop model, regional suitability model, Biophysical model, meta model, decision model. According to the author, the key challenge of adaptation plantation is the risk of maladaptation – adaptation that implies negative consequences in long term or in wide context. The five approaches differ in terms of their applicability for decision support in short term and long term adaptation planning. The main value lies in the ability to predict the climate change impact on yield potential at all level. For short term, reactive adaptation responses, statistical and biophysical models are less useful. Authors say adaptive management cycles should be institutionalised, within which adaptation behaviour, consequences of adaption responses and changes in impacts are continuously monitored.

Dr. Pushpa Mohan et al, (2017) has given analysis of the techniques employed and parameters achieved with limitation that every technique and experiment faced. This paper helps to have a crisp view of Regression Analysis, Linear regression by Sellam (2016). Limitations say that it is more complex to predict the optimized number of input parameters.

Evan Racah et al., (2017), The authors have used deep learning for weather prediction and climate change. for this they have used. For calculating the extreme weather projection values, the labels for extreme weather events namely Tropical Depressions (TD) Tropical Cyclones (TC), Extra-Tropical Cyclones (ETC) and Atmospheric Rivers (AR) using TECA (Prabhat et al., 2012) are identified. 3D semi-supervised learning architecture is used. For experimentation, frame-wise reconstruction is done, Detection and localization and feature exploration is done.

Zaki Ahmad Khan, et al., (2017) suggests different machine learning strategies for Wireless Sensor Networks (WSN). It presents a brief idea about supervised and unsupervised learning and its respective types. The author has suggested machine learning solutions for some operational, functional issues such as - query processing and event recognition, Medium Access Control, routing in WSN, object targeting and localization, Clustering and Data collection. Some other challenges highlighted are non-operational and application-specific challenges to address the WSN challenges.

Amir Ghaderi et al., (2017) mainly contributed to obtain forecasts of all nodes of the graph at the same time based on one framework. They studied the results of a case study on recorded time series data from a collection of wind mills in the north-east of the U.S. and shown that the proposed DL-based forecasting algorithm significantly improves the short-term forecasts compared to a set of widely-used benchmarks models. They used LSTM and RNN for their work Aized Amin Soofi et al., (2017) Classification is a data mining (machine learning) technique used to predict group membership for data instances. There are several classification techniques that can be used for classification purpose. Researchers use the basic classification techniques. Later usage of some major types of classification method including Bayesian networks, decision tree induction, k-nearest neighbour classifier and Support Vector Machines (SVM) with their strengths, weaknesses, potential applications and issues with their available solution had been applied. Their ultimate goal was to provide a comprehensive review of different classification techniques in machine learning. This provided platform for both academia and new comers in the field of machine learning to further strengthen the basis of classification methods.

Donghyun Lee et al., (2017) describes Artificial Intelligence and deep learning as a promising futuristic concept of technological advancements. Authors used deep learning’s recurrent neural network (RNN) model algorithms to predict pro-environmental consumption index based on Google search query data. Advanced research on ANN and RNN development processes is done. 84 different datasets were used by the author for verification of reliability of data by doing repeated experiments. Authors have used the data for experimentation on different human parameters, and a comparative analysis of ANN and RNN is done.

M. Shah et al., (2016) The proposed model provides forecast of the monsoon at a long lead time which supports the government to implement appropriate policies for the economic growth of the country. The monsoon of the central, north-east, north-west, and south-peninsular India regions are predicted with errors of 4.1%, 5.1%, 5.5%, and 6.4%, respectively. The identified predictors show high skill in predicting the regional monsoon having high variability. The proposed model is observed to perform better than most of the prediction models.

KarandeepKaur (2016), The author has tried to provide a brief overview of various machine learning applications in Indian agriculture, to help farmers advance their work manifolds. Author has described what Machine learning is and its technique such as reinforcement learning, supervised and unsupervised learning. While studying the applications in agriculture author has taken into consideration various parameters such as crop selection, crop yield prediction, weather forecasting, smart irrigation system and crop disease prediction and hence deciding the minimum support price. Considering all these parameters, the best suited algorithms are suggested respectively with the help of literature survey. Author has concluded his research saying the high accuracy of AI machines is the result of machine learning algorithms. one of the example is sensor based farming system for increased precision. Prescriptive solution for more complex problem in case of large data and field size is yet to be done.

P. Mondal, et al., (2015) recommends use of Enhanced Vegetation Index (EVI) over other remotely sensed vegetation indices as it better adjusts for background soil and canopy reflectance. Authors used around 25 climate variables in their study and finalised the data sets in to 4 sets, Monsoon and winter season for both central and Western India. They concluded both central and western sites showed strong sensitivity day time and night time temperature for both seasons, especially to winter daytime warming. Western site was less sensitive to monsoon precipitation variability, likely due to increased access to groundwater level irrigation. This groundwater irrigation is sensitive to climate variability. Authors suggest that heat tolerant high yield varieties to be added for better crop cover.

Aditya Grover et al. (2015) explores a new way against the traditional predictive model. They call it as the hybrid approach - a combination of partially trained predictive model with Deep Neural Network (DNN) for joining weather-related variables. They have evaluated the methods by experimenting on the real- world meteorological data assuring the results found to be promising. They have collected weather forecast data from the National Oceanic and Atmospheric Administration (NOAA). For preparing logs, Integrated Global Radiosonde Archive (IGRA) is used. The authors have thrown light on the challenges faced and overcome by them e.g first challenge was relationship of the weather parameters as tightly coupled and author suggests to clear the concepts from physics-based tight statistical couplings. some excerpts from the same are pressure and temperature follows natural gas law and relative humidity and temperature follow tight relationship, second challenge was when space and time is considered, the variable dependencies may have long range impacts. Authors have generated hybrid model for weather related spatiotemporal inferences, data-driven kernel function for prediction according to physical law, efficient inference procedure and experimental results. The scope for work still lies in increasing the spatio-temporal dimension, and testing the results of the same in Maharashtra, India.

R. Kumar, et al., (2014), focuses on the impact of climate change on crop productivity. The particulars considered are geographical area, average rainfall, variation in rainfall, annual precipitation, available water resources, utilizable surface and ground water, present water utilization and per capita availability.

R.C. Deo, M. Şahin (2014) presents a drought prediction algorithm for region of eastern Australia. They have devised a novel algorithm called extreme learning machine (ELM). Prediction done by them is on Effective Drought Index (EDI). Variables considered were mean, min and max. air temperatures and rainfall. Performance of ELM was evaluated over ANN models and ELM excelled in terms of mean absolute error, root mean square errors, Coefficients of determination and Willmott’s Indices of Agreement.

M. Senapati et al., (2013) This paper considers the needed adaptation measures including changes needed for mitigation to improve agriculture sector in India. It considers the likely changes that climate change will bring in temperature, precipitation and extreme rainfall, drought, flooding, storms, sea-level rise and environmental health risks and the overall impact on agriculture. The agricultural sector is the major source of employment in India. Climate change has adverse impacts on agriculture, hydropower, forest management and biodiversity. Anticipated impacts on agriculture from climate change and its various aspects have been studied.

A. Belayneh et al., (2013) worked for the Awash river basin of Ethiopia. They chose Standard Precipitation Index (SPI) as the drought index. Artificial neural networks (ANNs), support vector regression (SVR), and coupled wavelet-ANNs, which pre-process input data using wavelet analysis (WA) was used. Using RMSE and R2, forecasted results were compared. coupled wavelet neural network (WA-ANN) models were the most accurate models for forecasting SPI 3 (3-month SPI) and SPI 6 (6-month SPI) values over lead times of 1 and 3 months in the Awash River Basin in Ethiopia.

Dr. ParagKulkarni, (2012) highlights on the concepts of supervised and unsupervised learning. Defines what Machine learning is and also explains the learning paradigms. It says that empirical learning method has three different approaches to modeling problems based on observation, data and partial knowledge about problem domains. They are more specific to problem domain. They are generative modeling, discriminative modeling and imitative modeling. Author has also quoted some similarities and differences between reinforcement learning and systematic learning.

K.S. Kavi Kumar (2009) The paper emphasises on agricultural impacts of spatial features that influences the agricultural climate sensitivity. The authors quote that dependent variable, net revenue at farm level and error term has significant positive spatial autocorrelation working on it can improve the accuracy of climate impact studies. The book also refers to the climate change projections for India.

Seneviratne et al., (2008) provides relative material on what are extreme events, compound events, the relevance of feedbacks for extremes. The author discusses on requirements and methods for analysing changes in climate extremes, assessments regarding changes in the climate variables, phenomena, and impacts. Analysis of regional to globally scaled data is done. The extremes are calculated based on the changes observed since 1950 till the 20th century, which is really an effort to be appreciated. This chapter has really helped to understand the reasons of extreme climatic changes.

ErnstKussul and TatianaBaidyk (2004) In this research paper it is shown that the how we can improve the handwriting algorithm using the MNIST database. In the best cases the error rates can be 0.7%,0.42% and 0.63%.

Handwriting recognition has already achieved impressive results using shallow networks. Many papers have been published with research detailing new techniques for the classification of hand written numerals, characters and words. The deep belief networks (DBN) with three layers along with a greedy algorithm were investigated for the MNIST dataset and reported an accuracy of98.75% Pham, Bluche, Kermorvant and Louradour, (2014). Pham et al. applied a regularization method of dropout to improve the performanceof recurrent neural networks (RNNs) in recognizing unconstrained hand writing. Hinton,Osindero and Teh, (2004). The author reported improvement in RNN performance with significant reduction in the character error rate (CER) and word error rate (WER).

The convolutional neural network brings a revolution in the handwriting recognition field anddelivered the state-of-the-art performance in this domain.Tabik, Alvear-Sandoval, Ruiz, Sancho-Gómez, Figueiras-Vidal and Herrera, (2020). In 2003, Simardetal.introduceda general convolutional neural network architecture for visual document analysis and weeded outthe complex method of neural network training.Simard, Steinkraus and Platt, (2003). Wang et al. proposed a novel approach forSensors 2020, 20, 3344 4 of 18end-to-end text recognition using multilayer CNNs and achieved excellent performance on benchmarkdatabases, namely, ICDAR 2003 and Street View Text.Wang,Wu, Coates and Ng, (2012). Recently, Shi et al. integrated the advantagesof both the deep CNN (DCNN) and recurrent neural network (RNN) and named it conventional recurrent neural network (CRNN). They applied CRNN for scene text recognition and found itto be superior to traditional methods of recognition. Shi, Baiand Yao, (2017). Badrinarayanan et al. proposed a deep convolution network architecture for semantic segmentation. The segmentation architecture is knownas SegNet and consists of an encoder network, a decoder network and a pixel-wise classification layer. The proposed method used max-pooling indices of a feature map while decoding and observed good performance. The method is also analyzed and compared with existing techniques for road scene and indoor understanding. Long,Shelhamer and Darrell, (2015). CNN has shown remarkable abilities in online handwritten character recognition of Arabic language; handwritten Tamil character recognition; Telugu character recognition, handwritten Urdu text recognition, handwritten character recognition in Indicscripts and Chinese handwritten text recognition. Recently,Gupta, Sarkhel,Das and Kundu, (2019).proposed a novel multi-objective optimization framework for identifying the most informative local regions from a character image. The work was also evaluated on isolated handwritten English numerals, namely, MNIST images, along with three other popular Indic scripts, namely, handwritten Bangala numerals and handwritten Devanagari characters. The authors used features extracted from a convolutional neural network in their model and achieved 95.96% recognition accuracy. The work of Nguyen et al. used a multi-scale CNN for extracting spatial classification features for handwritten mathematical expression (HME). The local feature sand spatial information of HME images were used for clustering HME images. The work observed high performance for the CROHME dataset. They (authors) also concluded that classification can be improved by training the CNN with a combination of global max pooling and global attentive pooling.Ziranetal.developed a faster R-CNN-based framework for text/word location and recognitionin historical books. The authors evaluated these deep learning methods on Gutenberg’s Bible pages. The handwritten character recognition problem is intelligently addressed in the work of Ptucha etal.by the introduction of an intelligent character recognition (ICR) system using a conventional neural network. The work was evaluated on French-based RIMES lexicon datasets and English-basedIAM datasets, showing substantial improvement.

The performance of CNNs depends mainly on the choice of hyper-parameters, which are usually decided on a trial-and-error basis. Some of the hyper-parameters are, namely, activation function, number of epochs, kernel size, learning rate, hidden units, hidden layers, etc. These parameters are very important as they control the way an algorithm learns from data. Hyper-parameters from model parameters and must be decided before the training begins.

ResNet-52, Google net, VGG-16 and Alex Net are some popular CNN models that have a total of 150, 78, 57 and 27 hyper-parameters, respectively. A bad choice fo hyper-parameter scan incur a high computation cost and lead to poor CNN performance. The researcher’s expertise plays an important role in deciding on the configuration of hyper-parameters and requires an intelligent strategic plan. This creates several questions about CNN design for handwriting recognition tasks. How is CNN better in extracting distinct features from handwritten characters? What effect do different hyper-parameters have on CNN performance? What is the role of design parameters in improving CNN performance? In order to guide future research in the handwriting recognition field, it is important to address these questions.

Convolutional neural networks (CNN) are one of the most popular models used today. This neural network computational model uses a variation of multilayer perception and contains one or more convolutional layers that can be either entirely connected or pooled. These convolutional layers create feature maps that record a region of image which is ultimately broken into rectangles and sent out for nonlinear processing.

**Advantages:**

1. Very High accuracy in image recognition problems.
2. Automatically detects the important features without any human supervision.
3. Weight sharing.

**Advantages of using ML in crop yield prediction:** (Arpit Jain, 2021)

1. Complex dataset \_ crop yield prediction involves enormousdataset composing of satellite data and/or historicdata. Faster and accurate predictions can be made byutilizing the AI techniques such as regression algorithms(SVR, RF) Neural networks (CNN).

2. Parameter variation \_ the crop yield depends on a lotof parameters, like climatic factors, soil quality, NDVI,altitude, air parameters. The AI based prediction systemshandle the parameters dependency efficiently.

3. Accurate prediction \_ prediction of parameters using MLexhibit low error indices such as RMSE, and R2 which arestandard measures of accuracy for statistical analysis.

**Challenges and limitations in prediction of crop yield:**

1. Varying parameters and complex datasets pose a challenge for universal design of the prediction algorithms.

2. Dataset selection is critical due to the complexity; as an

improper selection of data may result in underfit /overfit prediction pattern.

## 2.1 Definition of Terms

1. **Pattern recognition** is a data analysis method that uses machine learning algorithms to automatically recognize patterns and regularities in data. This data can be anything from text and images to sounds or other definable qualities. Pattern recognition systems can recognize familiar patterns quickly and accurately.
2. **Neural networks:** are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network Function is determined largely by the connections between elements. We can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output.
3. **Convolutional Neural Networks CNN:** are made up of a large number of interconnected neurons that have learnable weights and biases. In the architecture of CNN, the neurons are organized as layers. It consists of an input layer, many hidden layers and an output layer. If the network has a large number of hidden layers the same are generally referred as deep neural networks.
4. **Deep learning:** is an artificial intelligence (AI) function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabelled. Also known as deep neural learning or deep neural network.
5. **Machine learning:** is the concept that a computer program can learn and adapt to new data without human intervention. Machine learning is a field of artificial intelligence (AI) that keeps a computer’s built-in algorithms current regardless of changes in the worldwide economy.
6. **Data science:** provides meaningful information based on large amounts of complex data or big data. Data science, or data-driven science, combines different fields of work in statistics and computation to interpret data for decision-making purposes.
7. **Artificial intelligence (AI):** refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.
8. **Data-set**: is a collection of data that contains individual data units organized (formatted) in a specific way and accessed by one or more specific access methods based on the data set organization and data structure.
9. **kNN:** kNN is a simple supervised classification algorithm. In this algorithm first, the labeled dataset is divided in to different classes based on their outputs. Therefore, a new sample object is assigned a particular class based on its k-nearest neighbours. Also kNN is a simple non-parametric supervised machine-learning algorithm used to solve classification problems as well as regression problems. Implementation is easy which will simply create an imaginary boundary to identify the data and attempt to predict the nearest boundary line for new data points to come in.
10. **Random Forest**: random forest is the ensemble classification model which combines a number of decision tree classifier. The final class of a new object is found out based on the majority class predicted from different decision trees classifiers.

# **CHAPTER THREE**

## MATERIALS AND METHODS

## 3.0 Materials/Resources:

Carrying out this experiment, data gathering is one of the challenges faced. Materials that were supposed to be used in gathering data were a meter rule, micrometer screw gauge and a weighing balance. Looking out for alternative to get the dataset, I gathered some average measurement from google combined it with few measured leaf length and compute the data by adding plus and minus random 5.0 value for TLL and SD, and plus and minus 0.5 value for yield to the average of the measurement gotten from online source (google), in total I generated 130 datasets.After the data was generated it was converted in to a softcopy using a software called Microsoft excel, which enables the data input in to the python software work space for further processing and cleaning, and thereafter for the proper prediction.

The diagram below describes the general process of the prediction model development. It gives an over-view and the steps of the entire process. At each particular stage there are what is expected to take place in other to move to the next stage to arrive at the success of a working and accurate model for our prediction. There after the entire process were detailed accordingly.

Here we will be using three machine learning algorithms which are kNN and random forest and linear progression.

ML algorithm model

Yield prediction and accuracy

KNN, Random forest and linear regression

Data set

Data

processing

Figure 1, Process flow diagram

## 3.1 Algorithms

KNN is a simple supervised classification algorithm. In this algorithm first, the labeled dataset is divided in to different classes based on their outputs. Therefore, a new sample object is assigned a particular class based on its k-nearest neighbours. Also kNN is a simple non-parametric supervised machine-learning algorithm used to solve classification problems as well as regression problems. Implementation is easy which will simply create an imaginary boundary to identify the data and attempt to predict the nearest boundary line for new data points to come in.

Random Forest: random forest is the ensemble classification model which combines a number of decision tree classifier. The final class of a new object is found out based on the majority class predicted from different decision trees classifiers.

# **3.2 IMPLEMENTATION**

## 3.2.1 Experimental Setup

To accomplish the task of the study, a model of kNN and random forest is used and analysed for suitable different learning parameters to optimize prediction accuracy and processing time.

All the experiments were done using an online programming tool called colab.research.google.com, on a personal computer with Windows 10,Intel (R) Core (TM) i3-6500 CPU (2.50 GHz), 16.00 GB memory and NVIDIA 1060 GTX GPU.

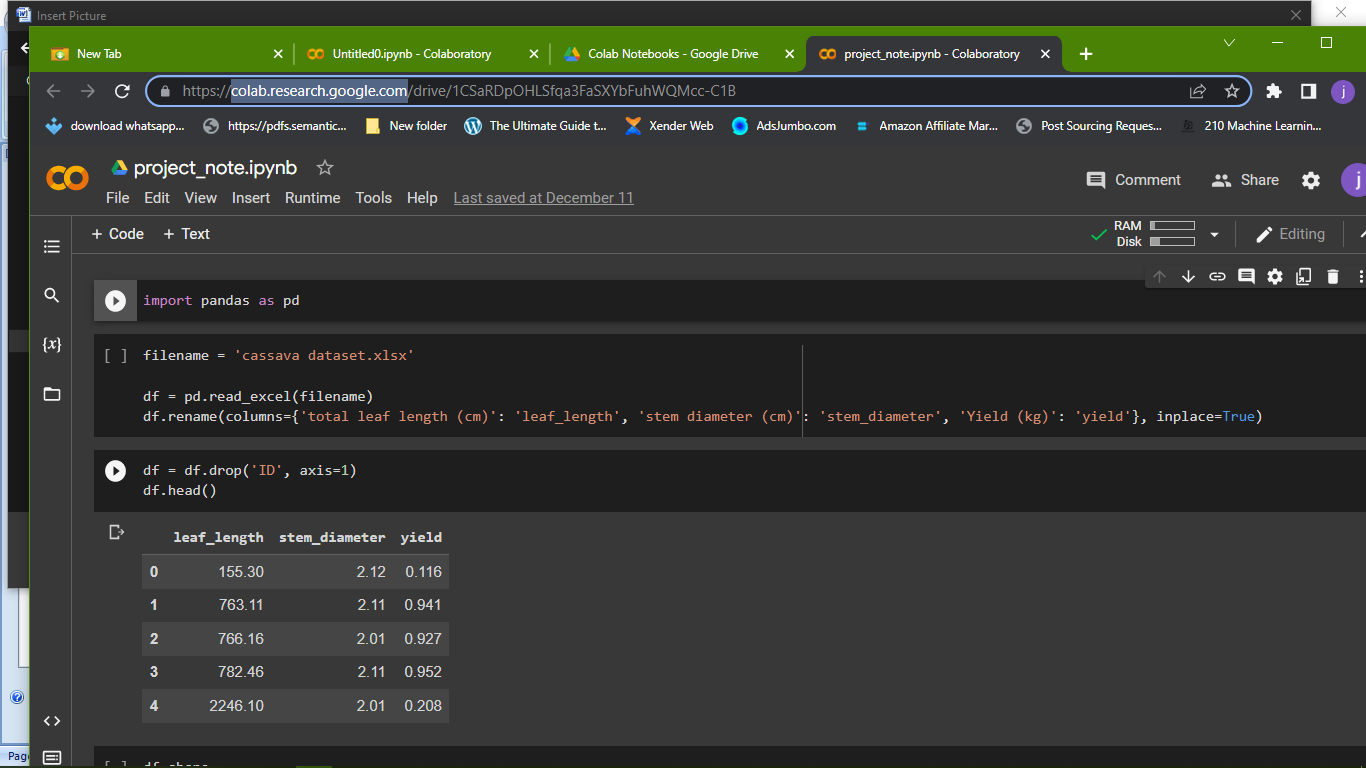
The online tool for programming is prefered due to the pre-installed python models libraries such as sklearn, numpy and pandas. The Total leaf length and stem diameter dataset measured in centimetres was involved in the training and testing models.

Figure 2, The working environment of the online tool used

**3.2.2 Importing Dataset**

The dataset was imported into the python workspace and split them into two parts, the training, target and testing sub dataset.

The dataset filename is called cassava dataset.xlsx

**3.2.3 Training the model**

In training the model, random forest regressor was imported from sklearn, with a line of code

The sklearn comes with algorithms, pre-trained models, to create, train, visualize, the result Through the use of the tools offered, we can perform classification, regression, clustering, dimensionality reduction, time series forecasting, and dynamic system modeling and control.

To make use of the tool a code “from sklearn.ensemble import randomforestregressor” is used to import the model.

fromsklearn.model\_selection import train \_test \_split

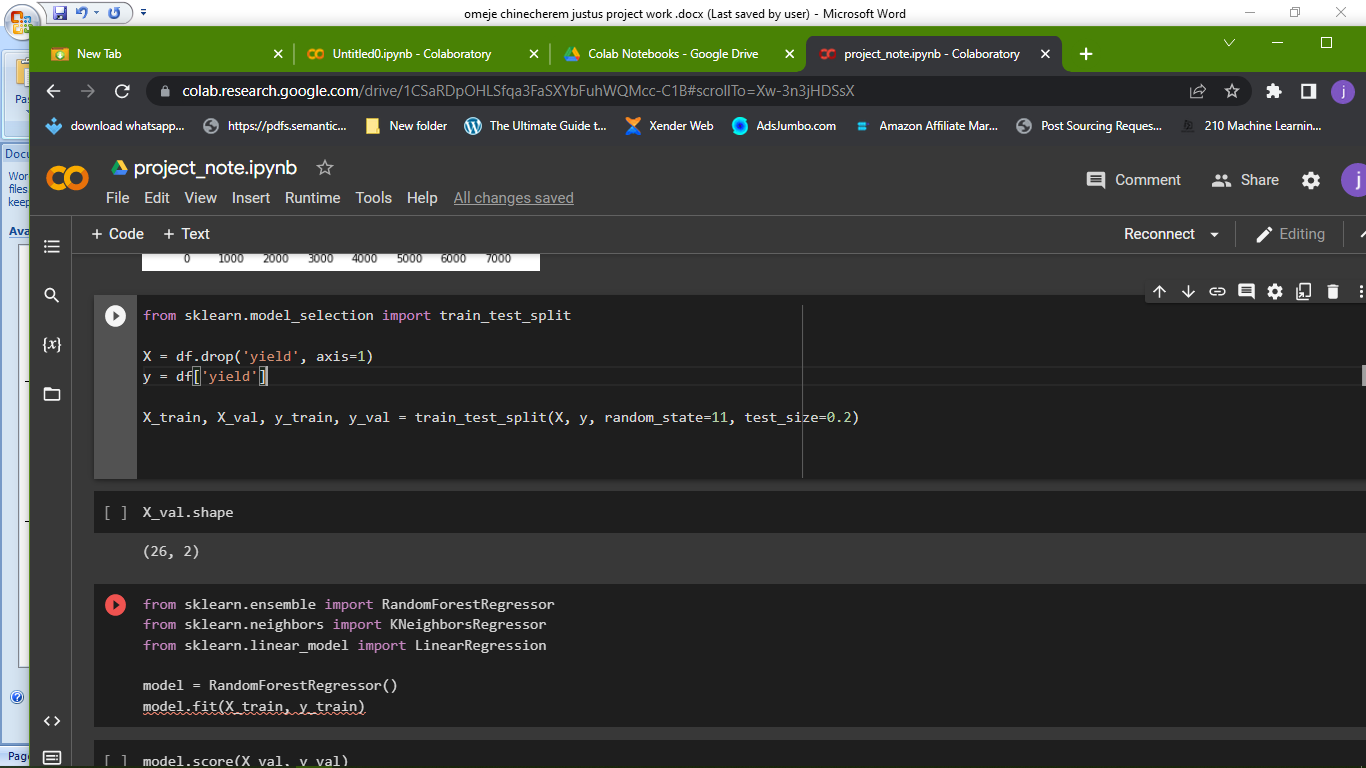
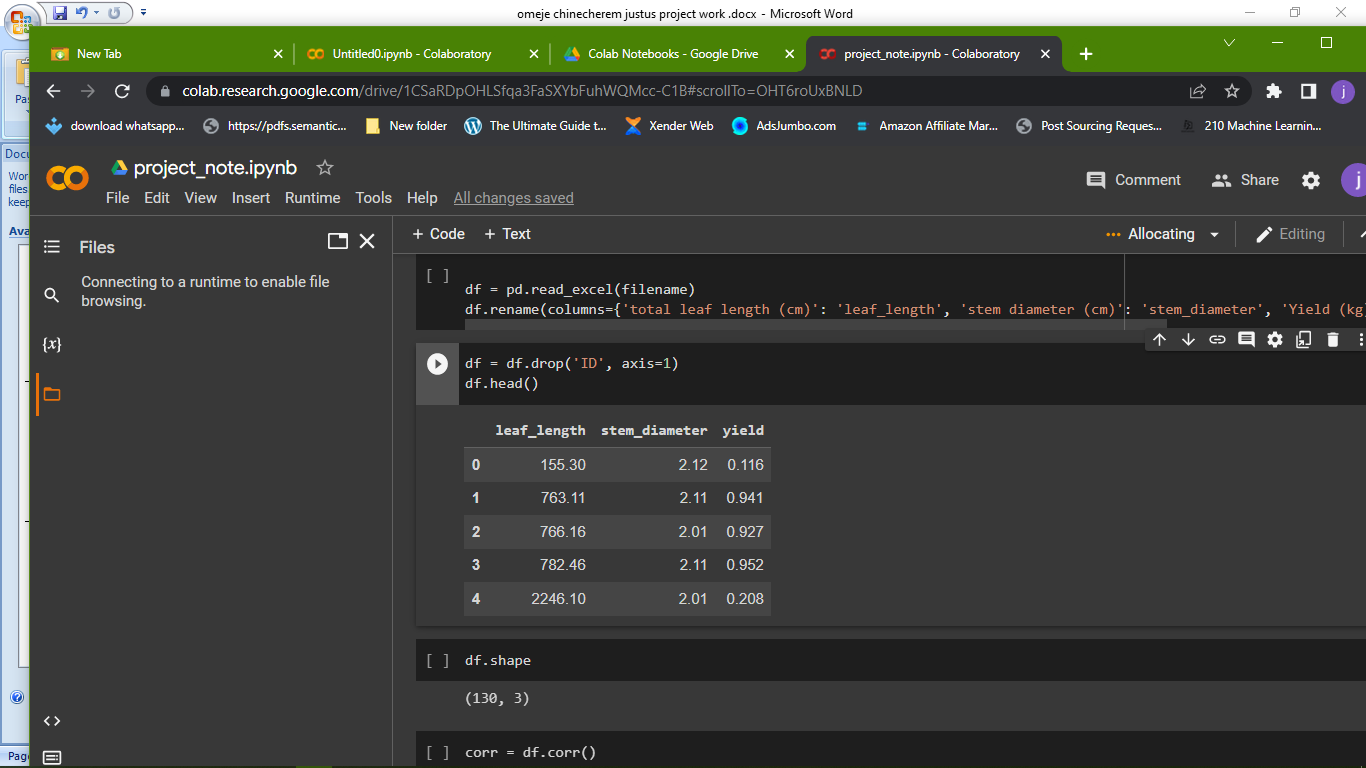


figure 3, The above command imports the model

**3.2.4 Connecting the Dataset to the model**

In this phase we connect our dataset to the model by using a click and drop window which is show below.

figure 4, Data importing

At this point our dataset is connected to the model the algorithm has been called upon, what happens next is the training proper

from sklearn.model\_selection import train\_test\_split

X = df.drop('yield', axis=1)

y = df['yield']

X\_train, X\_val, y\_train, y\_val = train\_test\_split(X, y, random\_state=11, test\_size=0.2)

Figure 5: splitting data in to training and testing

The above code block in figure 5 splits the data in to training and testing set at random state =11 and test size 0.2 which is 20% of the dataset

model = RandomForestRegressor()

model.fit(X\_train, y\_train)

Figure 6: Model selection

The above code block in figure 6 selects the random forest algorithm and trains it.

The model was trained, using this code “model.score(X\_val, y\_val)“ and we got an accuracy of 0.84149256 initial, run the code again to get an accuracy of 0.8585 which is better.

# **CHAPTER FOUR**

## RESULTSAND DISCUSSIONS

In machine learning more than one algorithm can be used to check for the one with higher accuracy. Therefore three algorithms where used in this case which are random forest, K­-nearest neighbour and linear regression. These three algorithms were been used to train and deployed for testing, linear regression had an accuracy of 0.7610, knearestneighbour had an accuracy of 0.5596 while random forest had an accuracy of 0.8585. from the results of the three algorithms, knearest neighbour performed poorly, while random forest turn out to be good with an accuracy that is high enough up to 80%, thereforerandom forest was our chosen algorithm. We had to make prediction to get the result and compare the prediction made by the model and the actual data collected in the field

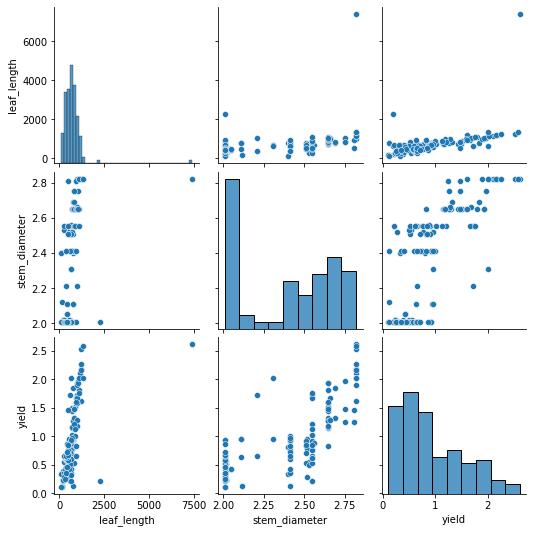
## 4.0. Results

The blow table shows the actual data used in training and the data predicted by random forest algorithm model.

Looking at the accuracy we can conclude the model worked fine because some of the figures were close enough.

Table 1: Generated Artificial Dataset of TLL, SD, Yield and predicted Yield

|  |
| --- |
| ID leaf length stem diameter actual yield predicted yield |
| 0 155.30 2.12 0.116 0.39676 |
| 16 277.68 2.01 0.331 0.27640 |
| 49 974.50 2.65 1.925 1.46182 |
| 2 766.16 2.01 0.927 0.77262 |
| 9 550.00 2.02 0.563 0.44477 |
| 65 836.94 2.65 1.311 1.38314 |
| 115 1188.20 2.82 2.167 2.03857 |
| 129 491.00 2.51 0.841 0.79018 |
| 30 620.81 2.01 0.204 0.77188 |
| 107 1131.13 2.82 2.105 2.01558 |
| 77 617.04 2.01 0.505 0.77188 |
| 56 679.52 2.55 0.755 0.68554 |
| 95 667.70 2.01 0.317 0.68371 |
| 106 499.00 2.81 1.466 1.12072 |
| 42 781.26 2.69 1.837 1.34097 |
| 111 1033.00 2.65 1.938 1.71124 |
| 6 237.50 2.02 0.232 0.26057 |
| 69 783.70 2.65 1.261 1.35059 |
| 101 907.81 2.01 0.642 0.83095 |
| 29 270.00 2.01 0.353 0.24672 |
| 119 446.90 2.01 0.450 0.64876 |
| 86 388.00 2.01 0.328 0.39146 |
| 87 610.65 2.41 0.714 1.12357 |
| 35 630.87 2.41 0.756 1.08779 |
| 114 412.41 2.01 0.856 0.54685 |
| 105 1104.00 2.82 2.011 2.01046 |

The graph of the result plotted using matlabplot

## Figure 7: Graph of the result

## 4.1 Correlation Analysis

It is important to learn about the data using the statistical tools for the successful training of ML

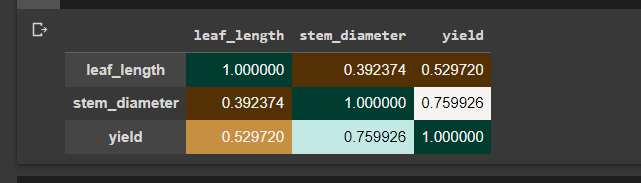
Algorithms. The Pearson correlation analysis results for all datasets are presented in figure 1 and two. The results depict that the stem diameter has the highest correlation for most of the dataset with cassava tuber yield.

Figure 8: correlation table

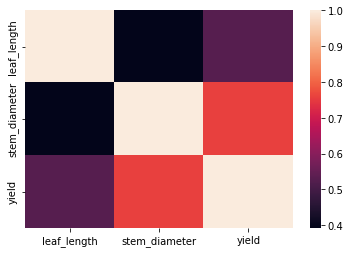


Figure 9: correlation visual graph

The correlation between stem diameter and tuber yield was found to be >70%

# **CHAPTER FIVE**

## CONCLUSION AND RECOMMENDATION

**5.0 Conclusion**

Using machine learning in agriculture is possible and recommendable as the world is evolving technologically the need to apply technological knowledge in to various fields is required. As we may have it, Building and Training three algorithms linear regression had an accuracy of 0.7610 knearest neighbour had an accuracy of 0.5596 while random forest had an accuracy of 0.8585. from the results of the three algorithms, knearest neighbour performed poorly, while random forest turn out to be good with an accuracy that is high enough up to 80%, therefore random forest was our chosen algorithmThe dataset consists of 130 data, split in two, 80% (104) for training data and 20% for testing data. This model can be improved when an accurate and large number of dataset is used to train the model.

**5.1Recommendation**

Machine learning model (Linear regression and random forest) could help in checking for yield prediction and making decision, and showed that the greater and healthier the stem diameter, the more yield of cassava tuber. However, all ML algorithms worked well by having over 80% accuracy. Furthermore, larger datasets may generateprecise and accurate results using either model. The information generated from this study may be

needed for creating site-specific management zones for cassava tuber yield.

## REFERENCES

CAMPBELL, G. M. 1976. Effect of ethephon and SADH on quality of clipped and non-clipped tomato transplants.J. Amer. Soc. Hortic.Sci 101: 648-651.

JAWORSKI, C. A., BRODIE, B. 8., GLAZE, N: C., McCARTER, S. M., GOOD, J. M. and WEBB, R. E. 1973. Research studies on field production of tomato transplants in southern Georgia. U.S. Dep. Agric. Production Res. Rep.

TAHA, A. A., KRETCHMAN, D. W. and JAWORSKI, C. A. 1980.Effect of daminozide and ethephon on transplant quality, plant growth and development and yield of processing tomatoes.J. Amer. Soc. Hortic. Sci. 105: 705-709.

Meirelles, G.; Manzi, D.; Brentan, B.; Goulart, T.; Luvizotto, E.J. **2017**Calibration Model for Water Distribution Network Using  Pressures Estimated by Artificial Neural Networks. *Water Resour. Manag.*,*31*, 4339–4351.

Shabani, A.; Ghaffary, K.A.; Sepaskhah, A.R.; Kamgar-Haghighi, A.A. **2017,**Using the artificial neural network to estimate leaf area. *Sci.*  *Hortic.*,*216*, 103–110.

López-Aguilar, K.; Benavides-Mendoza, A.; González-Morales, S.; Juárez-Maldonado, A.; Chiñas-Sánchez, P.; Morelos-Moreno, A.**2020** Artificial Neural Network Modeling of Greenhouse Tomato Yield and Aerial Dry Matter. *Agriculture*, *10*, 97.

Salazar, R.; Dannehl, D.; Schmidt, U.; López, I.; Rojano, A.**2017,** A dynamic artificial neural network for tomato yield prediction. *Acta* *Hortic.*,*1154*, 83–90.

Gholipoor, M.; Nadali, F.**2019,** Fruit yield prediction of pepper using artificial neural network. *Sci. Hortic.*, *250*, 249–253. [CrossRef]

Hu, Z.F.; Zhang, L.D.; Wang, Y.X.; Shamaila, Z.; Zeng, A.J.; Song, J.L.; Liu, Y.J.; Wolfram, S.; Joachim, M.; He, X.K. **2013,** Application  of BP Neural Network in Predicting Winter Wheat Yield Based on Thermography Technology. *Spectrosc.Spectr.Anal.*, *33*,  1587–1592.

Yin, G.H.; Gu, J.; Liu, Z.X.; Hao, L.; Tong, N. 2012,Analysis of Grain Yield Prediction Model in Liaoning Province. In *Advances in Future*  *Computer and Control Systems*; Jin, D., Lin, S., Eds.; Springer: Berlin, Germany,; Volume 159, pp. 355–360.

Zhang, J.Q.; Zhang, W.B.; He, Y.T.; Yan, Y.**2016,**Predicting the amount of coke deposition on catalyst pellets through image analysis  and soft computing. *Meas. Sci. Technol.*, *27*, 114006.

Lu, Z.J.; Zhu, L.; Pei, H.P. **2008,**The model of chlorophyll-a concentration forecast in the West Lake based on wavelet analysis and BP  neural networks. *Acta Ecol. Sin.*, *28*, 4965–4973.

Fang, J.; Zhang, Z. **2017,** Prediction of human blood pressure based on wavelet analysis and BP neural network.*Comput. Syst. Appl.*  , *26*, 157–161.

Li, S.X.; Yao, C.A.; Wen, J.; Huang, X.; Shao, X.H. March 2010,Forecasting of Basin Sediment Yield Based on Wavelet-BP Neural Network. In  Proceedings of the International Asia Conference on Informatics in Control, Automation and Robotics (CAR), Wuhan, China, 6–7  ; pp. 96–99.

Yd, A.; Zfa, B.; Yp, A.; Yz, A.; Hy, A.; Xla, B. **2019,** Precision fertilization method of field crops based on the wavelet-bp neural network  in China. *J. Clean. Prod.*, *246*, 118735.

Yang, X.H.; Liu, X.P.; Liu, H.S.; Guo, Y.; Xu, **2012,** S.P. Research based on the neural network of simulated annealing and genetic  algorithm in the precise fertilization. *Guangdong Agric. Sci.*, *39*, 60–69.

Ahamed, A.T.M.S., Mahmood, N.T., Hossain, N., Kabir, M.T., Das, K., Rahman, F., Rahman, R.M., 2015.2015, Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. In: 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD - Proceedings,

Ali, I., Cawkwell, F., Dwyer, E., Green, S., 2017. Modeling managed grassland biomass estimation by using multitemporal remote sensing data—a machine learning approach. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (7), 3254–3264.

Ananthara, M.G., Arunkumar, T., Hemavathy, R., 2013. CRY-An improved crop yield prediction model using bee hive clustering approach for agricultural data sets. In: Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, PRIME 2013,

Ayodele, T.O., 2010. Introduction to Machine Learning.Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning,

Baral, S., Kumar Tripathy, A., Bijayasingh, P., 2011. Yield Prediction Using Artificial Neural Networks, pp. 315–317.

Bargoti, S., Underwood, J.P., 2017.Image segmentation for fruit detection and yield estimation in apple orchards. J. Field Rob. 34 (6), 1039–1060.

Beulah, R., 2019. A survey on different data mining techniques for crop yield prediction. Int. J. Comput. Sci. Eng. 7 (1), 738–744.

Bhojani, S.H., Bhatt, N., 2020. Wheat crop yield prediction using new activation functions in neural network. Neural Comput. Appl. 1–11. Bose, P., Kasabov, N., Bruzzone, L., n.d. Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series. Ieeexplore.Ieee.Org.

Brownlee, J., 2016. Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras. Machine Learning Mastery.

Brownlee, J., 2017. Long Short-term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning. Machine Learning Mastery.

Brownlee, J., 2019. Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python. Machine Learning Mastery.

Cakir, Y., Kirci, M., Gunes, E.O., 2014. Yield prediction of wheat in south-east region of Turkey by using artificial neural networks. In: 2014 The 3rd International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2014.

Charoen-Ung, P., Mittrapiyanuruk, P., 2019. Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning, pp. 33–42.

Chen, Y., Lee, W.S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., He, Y., 2019.Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages.Remote Sens. 11 (13), 1584.Cheng, H., Damerow, L., Sun, Y., Blanke, M., 2017.

Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks. J. Imag. 3, 6.<https://doi.org/10.3390/jimaging3010006>.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput.Electron. Agric. 151, 61–69.

Chu, Z., Yu, J., 2020. An end-to-end model for rice yield prediction using deep learning fusion. Comput.Electron.Agric. 174.

Crane-Droesch, A., 2018.Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ. Res. Lett. 13 (11), 114003.

Črtomir, R., Urška, C., Stanislav, T., Denis, S., Karmen, P., Pavlovič, M., Marjan, V., 2012.Application of neural networks and image visualization for early forecast of apple yield.Erwerbs-Obstbau 54 (2), 69–76.

De Alwis, S., Zhang, Y., Na, M., Li, G., 2019. Duo attention with deep learning on tomato yield prediction and factor interpretation. In: Pacific Rim International Conference on Artificial Intelligence. Springer, Cham, pp. 704–715.

Elavarasan, D., Vincent, P.D., 2020. Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access 8, 86886–86901.

Elavarasan, D., Vincent, D.R., Sharma, V., Zomaya, A.Y., Srinivasan, K., 2018. Forecasting yield by integrating agrarian factors and machine learning models: a survey. Comput.Electron.Agric. 155, 257–28.

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D.S.N., Pozza, L.E., Ugbaje, S.U., Bishop, T.F.A., 2019a. An approach to forecast grain crop yield using multilayered, multi-farm data sets and machine learning. Precis. Agric. 1–15.

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D.S.N., Pozza, L.E., Ugbaje, S.U., Bishop, T.F.A., 2019b. An approach to forecast grain crop yield using multilayered, multi-farm data sets and machine learning. Precis.Agric.

Gandhi, N., Armstrong, L., 2016. Applying data mining techniques to predict yield of rice in humid subtropical climatic zone of India. In: Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016, 1901–1906.

Gandhi, N., Armstrong, L.J., 2016b. A review of the application of data mining techniques for decision making in agriculture. In: Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics.

Gandhi, N., Petkar, O., Armstrong, L.J., Tripathy, A.K., 2016. Rice crop yield prediction in India using support vector machines. In: 2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016.

Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W., 2014.Predictive ability of machine learning methods for massive crop yield prediction. Spanish J. Agric. Res. 12 (2), 313–328.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Kim, S.H., 2016. Random forests for global and regional crop yield predictions. PLoS ONE 11 (6)

Ji, S., Xu, W., Yang, M., Yu, K., 2012.3D convolutional neural networks for human action recognition.IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 221–231.

Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., Lin, T., 2020. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: a case study of the US Corn Belt at the county level. Glob. Change Biol. 26 (3), 1754–1766.

Johnson, M.D., 2013. Crop Yield Forecasting on the Canadian Prairies by Satellite Data and Machine Learning Methods.Master’s Thesis, University of British Columbia, Atmospheric Science.Retrieved from https://www.sciencedirect.com/science/ article/pii/S0168192315007546.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C.R., Anderson, M.C., 2020.Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest.

Environ. Res. Lett. Khaki, S., Wang, L., 2019.Crop yield prediction using deep neural networks. Front. Plant Sci. 10, 621. Khaki, S., Wang, L., Archontoulis, S.V., 2020.A cnn-rnn framework for crop yield prediction. Front. Plant Sci. 10, 1750.

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., Shearer, S., 2018. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. Comput.Electron. Agric. 153, 213–225.

Kitchenham, B., Charters, S., Budgen, D., Brereton, P., Turner, M., Linkman, S., Visaggio, G., 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering.Retrieved from https://userpages.uni-koblenz.de/~laemmel/ esecourse/slides/slr.pdf.

Kouadio, L., Deo, R.C., Byrareddy, V., Adamowski, J.F., Mushtaq, S., Phuong Nguyen, V., 2018. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. Comput.Electron. Agric. 155, 324–338.

Kunapuli, S.S., Rueda-Ayala, V., Benavidez-Gutierrez, G., Cordova-Cruzatty, A., Cabrera, A., Fernandez, C., Maiguashca, J., 2015. Yield prediction for precision territorial management in maize using spectral data. In: Precision Agriculture 2015 - Papers Presented at the 10th European Conference on Precision Agriculture, ECPA 2015 (pp. 199–206).

Lee, S., Jeong, Y., Son, S., Lee, B., 2019. A self-predictable crop yield platform (SCYP) based on crop diseases using deep learning. Sustainability 11 (13), 3637.

Li, B., Lecourt, J., Bishop, G., 2018. Advances in non-destructive early assessment of fruit ripeness towards defining optimal time of harvest and yield prediction—a review. Plants 7 (1).

Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: a review. Sensors (Switzerland) 18 (8).

Matsumura, K., Gaitan, C.F., Sugimoto, K., Cannon, A.J., Hsieh, W.W., 2015. Maize yield forecasting by linear regression and artificial neural networks in Jilin, China. J. Agric. Sci. 153 (3), 399–410.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Petersen, S., 2015.Human-level control through deep reinforcement learning. Nature 518 (7540), 529–533.

Mola-Yudego, B., Rahlf, J., Astrup, R., Dimitriou, I., 2016. Spatial yield estimates of fastgrowing willow plantations for energy based on climatic variables in northern Europe. GCB Bioenergy 8 (6), 1093–1105.Monga, T., 2018.Estimating vineyard grape yield from images, pp. 339–343.

Nevavuori, P., Narra, N., Lipping, T., 2019. Crop yield prediction with deep convolutional neural networks.Comput.Electron.Agric. 163.

Nguyen, L.H., Zhu, J., Lin, Z., Du, H., Yang, Z., Guo, W., Jin, F., 2019. Spatial-temporal multi-task learning for within-field cotton yield prediction. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, pp. 343–354.

Osman, T., Psyche, S.S., Kamal, M.R., Tamanna, F., Haque, F., Rahman, R.M., 2017. Predicting early crop production by analysing prior environment factors, pp. 470–479.

Pantazi, X.E., Moshou, D., Mouazen, A.M., Kuang, B., Alexandridis, T., 2014. Application of supervised self organising models for wheat yield prediction, pp. 556–565.

Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L., Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Comput.Electron. Agric. 121, 57–65.

Paul, M., Vishwakarma, S.K., Verma, A., 2015.Analysis of soil behaviour and prediction of crop yield using data mining approach. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, pp. 766–771.

Rao, T., Manasa, S., n.d. Artificial Neural networks for soil quality and crop yield prediction using machine learning.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99.

Romero, J.R., Roncallo, P.F., Akkiraju, P.C., Ponzoni, I., Echenique, V.C., Carballido, J.A., 2013. Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. Comput.Electron. Agric. 96, 173–179.

Ruder, S., 2017.An overview of multi-task learning in deep neural networks.arXiv preprint arXiv:1706.05098. Ruß, G., Kruse, R., 2010. Regression models for spatial data: an example from precision agriculture, pp. 450–463.

Ruß, G., Kruse, R., Schneider, M., Wagner, P., 2008. Data mining with neural networks for wheat yield prediction. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 5077 LNAI. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 47–56.

Saravi, B., Nejadhashemi, A.P., Tang, B., 2019. Quantitative model of irrigation effect on maize yield by deep neural network.Neural Comput. Appl. 1–14.

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M., Ebrahimie, E., 2014.Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture.PLoS ONE 9 (5), e97288.

Shidnal, S., Latte, M.V., Kapoor, A., 2019. Crop yield prediction: two-tiered machine learning model approach. Int. J. Inf. Technol. 1–9. Šmite, D., Wohlin, C., Gorschek, T., Feldt, R., 2010. Empirical evidence in global software engineering: a systematic review. Empirical Softw. Eng. 15 (1), 91–118.

Somvanshi, P., Mishra, B.N., 2015. Machine learning techniques in plant biology. In: PlantOmics: The Omics of Plant Science. Springer India, New Delhi, pp. 731–754.

Su, Y.X., Xu, H., Yan, L.J., 2017. Support vector machine-based open crop model (SBOCM): case of rice production in China. Saudi J. Biol. Sci. 24 (3), 537–547.

Sujatha, R., Isakki, P., 2016. A study on crop yield forecasting using classification techniques. In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering, ICCTIDE 2016.

Taherei-Ghazvinei, P., Hassanpour-Darvishi, H., Mosavi, A., Yusof, K.W., Alizamir, M., Shamshirband, S., Chau, K., 2018.Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network.Eng. Appl. Comput. Fluid Mech. 12 (1), 738–749.

Wang, A., Tran, C., Desai, N., Lobell, D., n.d. Deep transfer learning for crop yield prediction with remote sensing data. Dl.Acm.Org. Retrieved from https://dl.acm.org/ citation.cfm?id=3212707.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Data Mining: Practical Machine Learning Tools and Techniques.

Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Guanter, L., 2020. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. Environ. Res. Lett. 15 (2).

Yalcin, H., 2019. An approximation for a relative crop yield estimate from field images using deep learning. In: 2019 8th International Conference on Agro-Geoinformatics (Agro-

Ying-xue, S., Huan, X., Li-jiao, Y., 2017. Support vector machine-based open crop model (SBOCM): Case of rice production in China. Saudi J. Biol. Sci. 24 (3), 537–547.

Arun Kumar, Naveen Kumar, Vishal Vats.”Efficient crop yield prediction using machine learning algorithms”, IJRET Volume: 05 Issue: 06, June-2018, pp 3151-3159 2. P.Priya, U.Muthaiah ,M.Balamurugan. “ Predicting yield of the crop using machine learning Algorithm”, IJESRT et al., 7(4): April-2018, pp 2277-2284,

Vaneesbeer Singh, AbidSarwar, Vinod Sharma. “Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach”, IJARCS 8 (5), May-June 2017, pp 1254-1259. 4. A.S. Ponraj, Vigneswaran T. “Machine Learning Approach for Agricultural IoT”,IJRTE, Volume-7, Issue-6, March 2019,pp 383-391.

Abdullah, A., Brobst, S, Pervaiz.I.,UmerM.,andA.Nisar, “Learning dynamics of pesticide abuse through data mining”, Proceedings of Australian Workshop on Data mining and Web Intelligence, New Zealand, January .2004,pp 63-68.

KiranMai,C., Murali Krishna, I.V, an A.VenugopalReddy, “ Data Mining o f Geospatial Database for Agriculture Related Application”, Proceedings of Map India,New Delhi, 2006,pp 83-96.

Ravichandran G, Koteeshwari R S 2016 Agricultural Crop Predictor and Advisor using ANN for Smartphones IEEE

# Appendix

**Codes used**

import pandas as pd

filename = 'cassava dataset.xlsx'

df = pd.read\_excel(filename)

df.rename(columns={'total leaf length (cm)': 'leaf\_length', 'stem diameter (cm)': 'stem\_diameter', 'Yield (kg)': 'yield'}, inplace=True)

df.shape

corr = df.corr()

corr.style.background\_gradient(cmap='BrBG')

import seaborn as sns

sns.heatmap(df.corr())

sns.pairplot(df)

import matplotlib.pyplot as plt

plt.hist(df['yield'], bins=40)

from sklearn.model\_selection import train\_test\_split

X = df.drop('yield', axis=1)

y = df['yield']

X\_train, X\_val, y\_train, y\_val = train\_test\_split(X, y, random\_state=11, test\_size=0.2)

X\_val.shape

from sklearn.ensemble import RandomForestRegressor

from sklearn.neighbors import KNeighborsRegressor

from sklearn.linear\_model import LinearRegression

model = RandomForestRegressor()

model.fit(X\_train, y\_train)

model.score(X\_val, y\_val)

model.predict(X\_val)

y\_val.values

dummy = X\_val.copy()

dummy['correct\_label'] = y\_val

dummy['predicted\_label'] = model.predict(X\_val)

dummy

**Result of the work done**

ID leaf length, stem diameter, yield, predicted yield

0 155.30 2.12 0.116 0.39676

16 277.68 2.01 0.331 0.27640

49 974.50 2.65 1.92 1.46182

2 766.16 2.01 0.927 0.77262

9 550.00 2.02 0.563 0.44477

65 836.94 2.65 1.311 1.38314

115 1188.20 2.82 2.167 2.03857

129 491.00 2.51 0.841 0.79018

30 620.812.01 0.204 0.77188

107 1131.13 2.82 2.105 2.01558

77 617.04 2.01 0.505 0.77188

56 679.52 2.55 0.755 0.68554

95 667.70 2.01 0.317 0.68371

106 499.00 2.81 1.466 1.12072

42 781.26 2.69 1.837 1.34097

111 1033.00 2.65 1.938 1.71124

6 237.50 2.02 0.232 0.26057

69 783.70 2.65 1.261 1.35059

101 907.81 2.01 0.642 0.83095

29 270.00 2.01 0.353 0.24672

119 446.90 2.01 0.450 0.64876

86 388.00 2.01 0.328 0.39146

87 610.65 2.41 0.714 1.12357

35 630.872.41 0.756 1.08779

114 412.41 2.01 0.856 0.54685

105 1104.00 2.82 2.011 2.01046