Check for updates

Comparative analysis of classification algorithms for crop yield prediction

*Okengwu U.A., Onyejegbu L.N., Oghenekaro L.U., Musa M.O., Ugbari A.O.
Department of Computer Science
University of Port Harcourt
*ugochi.okengwu@uniport.edu.ng

Abstract

A machine learning model is an essential tool for deciding which crops to produce and what to do during those crops' growing seasons. The employment of various machine learning algorithms in research to forecast higher crop output has considerably benefited the agriculture sector. In this study, an appropriate crop recommendation solution is constructed utilizing a Kaggle dataset that incorporates several factors such as (N-Nitrogen, K-Potassium, P-Phosphorus, Humidity, pH value of the soil, rainfall and temperature). The major goal of this model is to estimate which crop would grow best on a given farm based on the parameters that were used to create the model. The evaluated models revealed random forest to have the highest prediction accuracy with a score of 99 percent, K-Nearest Neighbor was next with a score of 97 and logistic regression recorded 96 percent. Hence random forest produced the highest accuracy score of 0.99 in recommending appropriate crops to farmers especially in times of drought and low soil fertility.

Keyword

Machine Learning, Classification, Random Forest, Crop Yield forecast, Modeling.

1.0 Introduction

Crop yield can be defined as the amount of crop produced with a given area of land. It is a significant metric to consider when dealing with the issue of food security, and may be reported in kilograms/hectare, metric tons/hectare or bushels/acre. As the world population increases exponentially, so is the demand for food. Crop produce serves not only human beings but also livestock and it is further used for bio-energy. With a growing human population of 7 billion people, the need for crop produce is enormous, but the land and resources for crop cultivation is limited (Muruganantham et. al, 2022). Hence the need to adopt technologies that will enhance production and ensure efficient and smart farming practices. Of all issues persistent in the agricultural domain, the issue of increasing crop yield remains the issue of great interest to farmers, as they seek ways to increase their crop yield per acre. Fortunately, with the advent of precision agriculture, models have been built to predict crop yields and give farmers recommendations that will increase their yield.

A crop yield recommendation or prediction system is an application that suggests or recommends to farmers which crop to grow on his or her field based on the parameters on which the application system was developed. By parameters, we mean the factors that affects the growth of plants and crops. Some of these factors are; seed quality, soil type, climate conditions, pest infestation, seed treatment, fertilizer application techniques, season, rainfall and temperature etc.

© The Author(s) 2023
E. O. Nwaichi and F. O. Nduka (eds.), *Proceedings of the 6th Biennial International Conference for Organization for Women in Science for the Developing World Nigeria (OWSD 2023)*, Advances in Biological Sciences Research 37,
https://doi.org/10.2991/978-94-6463-306-1_22

Crop yield prediction is a fast-growing area in agriculture because it increases farm efficiency and plays an important role in estimating food availability for our fast-growing world population (Kheir et. al, 2021). It also contributes to the actualization of the second sustainable development goal that seeks to find sustainable solutions to end hunger in the world.

Although there has been a high increase in genetically modified crops which has made agricultural yields bigger in number, farmers and decision-makers need advanced tools that will help them make quick and timely decisions that will improve agricultural yields in their farms. This is where machine learning algorithms come in to play. Algorithms perform differently when subjected to different kind of data. In this research, we compared Random Forest, K-Nearest Neighbor and Logistic Regression on Kaggle dataset details of which are shown in figure 1.

This essay is organized as follows: an introduction, a literature review, a discussion of the findings, and a conclusion.

2.0 Review of Literature

Forecasting crop yields is a crucial role for decision- and policy-makers at all levels.

Farmers can choose what kinds of crops to raise and when by using an accurate crop forecast. It is a useful tool for policymakers, investors, agronomists, production businesses, manufacturers, and commodity traders in addition to farmers (Basso and Liu 2019; Chipanshi et al., 2015). Crop-specific traits and environmental factors are just two of the many variables that affect crop output.

Any predictive model for crop yield forecasting must take into account all of these components and their amount of influences on crop output.

Recently, a variety of techniques have been used to predict agricultural production, including field surveys, remote sensing, statistical techniques, and crop growth models. Each of these approaches addresses various agricultural yield forecasting issues or areas of concern. The purpose of field surveys is to gather real-world data based on reported farmer surveys and field measurement surveys. Field surveys have a number of drawbacks and issues, such as decreasing feedback, a lack of reliability owing to sampling and non-sampling errors, and resource limitations.

Crop growth models (Basso et al., 2013; Chipanshi et al., 2015) reproduce the growth and development of crops utilizing fundamental plant principles, environmental conditions, and management practices.

However, these models do not explain factors responsible for reduction in yield and also are plagued by data and calibration requirements.

To assess the crop's current state and forecast the final yield, remote sensing techniques rely on satellite photography (Lopez-Lozano et al., 2015). Remotely sensed data are globally available and are devoid of human errors but remotely sensed observations only offer indirect or second-hand evaluations or assessments of crop yield such as observed radiance (Jones and Vaughan, 2010)

In order to create a straightforward rectilinear relationship between variables regarded as predictors and crop yield, statistical models are concerned with using weather variables, integrated results from remote sensing, crop growth models, and field surveys as forecasters (Bussay et al., 2015). The results of statistical models cannot be generalized or extrapolated to different spatio-temporal contexts, notwithstanding their precision and wide range of interpretations.

Numerous research has used both a traditional (shallow) approach to machine learning and deep learning in applications to crop yield. While Jeong et al. (2016) utilized Random Forests to estimate the global yield of wheat and potatoes in US, Gonzalez Sanchez et al. (2014) examined the performance of various machine learning algorithms on 10 different crops in Mexico.

You et al. (2017) used representation learning concepts and ideas to estimate soyabean yield in the US, and Crane Droesch (2018) used semiparametric deep neural networks to forecast maize yield in the US. These are only two examples of deep learning applications. These results imply that both shallow and deep machine learning methods can be used for crop yield predictions. While some studies focused on publicly accessible data (such as You et al. (2017)), others focused on multiple crops and locations (such as Gonzalez Sanchez et al. (2014); Joeng et al. (2016)), with the overall objective of making performance comparisons against statistical methods as opposed to replicable methods.

2.1 Artificial Neural Network for Predicting Crop Yield

Being similar to the biological process of the brain, the ANN structure, whose fundamental premise is the imitation of mathematical models, can be utilized to resolve complex problems (Thyagarajan et al., 1998). The bare minimum number of layers needed to construct an ANN system are the input layer, the hidden layer, and the output layer. The total number of hidden modes, which can be easily increased to include more hidden layers depending on how challenging the study is. According to Kaul, Hill, and Walthall (2005), the input comprises nodes that correspond to input variables, and the output contains nodes that link to output variables. The relationship through connection weight between the output of the input layer, the output layer's unit outputs, and the inputs of succeeding layers (Marchant et al., 2002). Through the input layer, the inputs are divided among numerous hidden layers. Since the node computes a weighted total of all of its net inputs and absorbs data from the layer before it, weighted connections facilitate the movement of data between layers. The sigmoidal function is the most typical transfer or activation function for the hidden and output layers (Marchant et al., 2002; Kaul, Hill, and Walthall, 2005). Kaul, Hill, and Walthall (2005) claim that a linear transfer function is generally used to move information from the input layer to the hidden layers. The dataset can then be "learned" through training (Marchant et al., 2002). Alvarez (2009) describes learning as "a process which consists of modifying the weights connected to the transfer functions between neurons" when comparing ANN output with observed data. The most popular method (BP) is back-propagation training. To reduce error (Shearer et al., 2000), the feedforward neural network is trained using the BP technique, where the difference between the calculated output and the desired value is referred to as the error (Wieland and Mirschel, 2008). However, a big network with an excessive number of nodes would overtrain, memorizing the training data, producing subpar predictions (Lawrence, 199) and using up a lot of memory (Shearer et al., 2000). Until a predetermined error limit is reached or all training cycles (epochs) have been completed, the process is repeated.

2.2 Algorithms and Parameters affecting Crop Yield

The yield of any crop is predicted based on a variety of criteria. These are essentially the characteristics that aid in estimating a crop's annual yield. Following are some of the key factors: 1. Season, 2. Rainfall, 3. Temperature, 4. Area.

Temperature and rainfall are the factors that have the biggest impact on crop production forecasts. Time series machine learning methods are applied to temperature and rainfall data since they are consecutive. These are the algorithms:

1) RNN with nodes, first: In an artificial neural network (ANN), recurrent neural network (RNN) connections between nodes form a directed graph along a temporal sequence (Mandic and Chambers, 2001). As a result, it may display temporal dynamic behavior.

2) LSTM: Long short-term memory (LSTM) is a type of artificial recurrent neural network (RNN) architecture used in deep learning (Hochreiter and Schmidhuber, 1997). Unlike LSTM, conventional feed forward neural networks lack a feedback loop. LSTM is advantageous for "general purpose computers" as a result (Sak, Senior, and Beaufays, 2014).

The agricultural production dataset is provided to classification and regression algorithms to forecast the name and yield of the crop. Along with KNN Classifier (Cover and Hart, 1967), Logistic Regression (Kleinbaum et al., 2002), Linear Regression (Seber and Lee, 2012), and Artificial Neural Networks (Urada, 1992), ensemble learning techniques like Random Forest Classifier (Liaw and Wiener, 2002) and XGBoost (Chen and Guestrin, 2016) are also used.

2.3 Taxonomy for Algorithm-Based Crop Yield Analysis

For effective decision-making, it is essential for national and regional decision-makers to forecast crop yields. Farmers can choose what crops to grow and when to plant them with the help of an accurate crop production projection model. Numerous methods can be used to forecast crop yields.

Many researchers from around the globe have used machine learning algorithms to predict crop yields.

Tseng (2019) tracked agricultural yield projections using Internet of Things (IoT) technology for intelligent agriculture. The proposed models anticipated crop production in farms where weather damage to crops was a frequent occurrence by using big data in intelligent agriculture. The created model made use of an Internet of Things (IoT) sensor that sensed air pressure, humidity, moisture content, temperature, and soil salinity while monitoring the entire agricultural land. The study made use of big data analysis in IoT to assess and comprehend the farmers' crop-growing practices as well as environmental abnormalities. The 3D cluster evaluation of the relationship between environmental components and subsequent examination of the recommendations from the farmers was a benefit of the proposed model. However, when exposed to potential risks in air, temperature, humidity, and soil moisture content, the developed model had an unusual distribution.

Tiwari and Shukla (2018) combined CNN and the Geographical Index to create a model for predicting agricultural productivity. According to the dominant paradigm, agricultural drifts for crop production were always failing since they weren't ideal for environmental elements including temperature, weather, and soil quality. The created CNN model was trained using BPNN and employed geographical information as input for error prediction. The utilization of a real-time dataset gathered from reliable geographic sources was advantageous for the suggested model. The established method did, however, lessen the accuracy of crop yield predictions while reducing the relative error.

Robust Deep-Learning was used by Fuentes et al. (2022) to identify tomato plant diseases and insect infestations in crops. Due to pests and illnesses in crops that significantly increased economic loss, the current model had trouble predicting crop yields. The created model incorporates a complex meta-architecture to forecast plant pests. Three essential characteristics of indicators are taken into account by the created model:

Deep meta-architecture is also known as Region-Based Fully CNN, Single Shot Multibox Detector (SDD), and Faster Region-Based CNN. A method for a global and local period explanation was also made available by the use of deep meta-architecture and feature extractors. The addition of data raises accuracy while also reducing the percentage of false positives during training. The created model's advantage was its success in identifying various pests and diseases by handling challenging local circumstances. The robust deep learning method uses sophisticated pre-processing algorithms, which takes more time and costs more money to compute.

Sun et al. (2019) used the Deep CNN-LSTM approach to project the soybean yield estimation. The yield projection has significant effects on crop market planning, crop insurance, harvest management, and remote sensing. The CNNLSTM method was also supported by the model, which increased its applicability and success in predicting the PM2.5 concentration. Utilizing historical data, including the cumulated wind speed, the duration of the rain, and the concentration of PM 2.5, the DNN structure—which incorporated LSTM and CNN—was created. The most recent research in this field shows that both LSTM and CNN can deliver phonological information and more spatial factors, both of which are crucial for predicting crop production. As a result of the method's usage of histogram-based tensor modification to combine different remote sensing data, feature extraction was still challenging.

ML methods were used by Bondre and Mahagonkar (2019) to forecast agricultural production and manure recommendations. Yield prediction was a big issue in agriculture that was resolved by developing a machine learning system. The effectiveness of the developed model was evaluated for the purpose of computing crop yield in agriculture. The use of past data for crop prediction and the prescription of the proper fertilizer for each crop using ML algorithms like random forest and SVM were two additional advantages of the developed model. However, the intelligent farm irrigation technology that would have enhanced yields was never implemented.

Data mining techniques were used by Devika and Ananthi (2018) to forecast the annual yield of important crops. Farmers were reluctant to harvest the crop because of limited water resources and unanticipated weather changes, but these issues were fixed by using a data mining system. The developed model assembled historical crop-growing data that was then assessed for accurate crop yield prediction. The training data can be gathered from the gathered documents and employed in the training phase that has to be exploited in various data mining procedures. The developed model had the advantage that only sugarcane, cotton, and turmeric had the highest level of crop production prediction. The range was small for other crops like wheat, rice, etc.

Using hybrid ML approaches, T. Senthil Kumar (2020) created a data mining-based marketing decision support system that addresses the issues with the corresponding financial and marketing applications. Decisions are made using a decision support system that analyses the actual situation, improving organizational performance. Globalization, privatization, and liberalization increased the organization's competitiveness in the preexisting paradigms. The competition is fair and strong enough to support implementing carefully planned and implemented marketing tactics. However, because the model presented challenges throughout the process and had poor assessment performance, an optimization model was needed. Through examination of the research, a number of feature categories related to soil information, including soil maps, soil types, and production areas, were looked at. The location of the soil as well as the sorts of nutrients that can be found there are both detailed in the maps of the soil. In terms of crop density, weight growth, and leaf area index, the features of crops, including mustard plants, wheat, rice, tomato plants, etc., were evaluated. Similar meteorological characteristics include humidity, precipitation, rainfall, and forecasted precipitation. The components of the nutrients play a significant impact in relation to various environmental conditions. Nitrogen, potassium, magnesium, zinc, boron, and other elements are among the nutrients. The properties, which are also linked to temperature and radiation (gamma), shortwave radiation, solar radiation, and degree days, are estimated using the solar data. Less is used of the features. The calculations take into account pressure, imagery, and wind speed. All these is geared to provide to an integrated and effective predictive model for various crop yield for various crops.

3.0 Materials and Methods

Data used for the study was obtained from Kaggle data repository. It holds soil properties and climate properties as its features, and as its label; recommended crop. The data, titled crop-recommendation.csv, is a comma separated value file, with 2200 instances, 7 independent features and 1 dependent label as seen in figure 1. Temperature, humidity, soil ph, and rainfall are all aspects of the climate. Included in the soil characteristics are phosphorus, nitrogen, and potassium. There are 22 different crops stated on the label, including rice, maize, chickpeas, kidney beans, pigeon peas, moth beans, mungbeans, blackgram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee.

	N	P	K	temperature	humidity	ph	rainfall	label
2180	80	18	31	24.029525	58.848806	7.303033	134.680397	coffee
2181	101	31	26	26.708975	69.711841	6.861235	158.860889	coffee
2182	103	33	33	26.717174	50.501485	7.131436	126.807398	coffee
2183	93	26	27	24.592457	56.468296	7.288212	137.704405	coffee
2184	104	35	28	27.510061	50.666872	6.983732	143.995555	coffee
2185	116	36	25	27.578476	58.525343	6.172090	156.681037	coffee
2186	107	38	29	26.650693	57.566957	6.351182	145.105065	coffee
2187	101	33	33	26.972516	62.018363	6.908671	142.861079	coffee
2188	107	31	31	23.171246	52.978412	6.766184	153.120164	coffee
2189	99	16	30	23.526521	65.443409	6.392792	186.172820	coffee
2190	103	40	30	27.309018	55.196224	6.348316	141.483164	coffee
2191	118	31	34	27.548230	62.881792	6.123796	181.417081	coffee
2192	106	21	35	25.627355	57.041511	7.428524	188.550654	coffee
2193	116	38	34	23.292503	50.045570	6.020947	183.468585	coffee
2194	97	35	26	24.914610	53.741447	6.334610	166.254931	coffee
2195	107	34	32	26.774637	66.413269	6.780064	177.774507	coffee
2196	99	15	27	27.417112	56.636362	6.086922	127.924610	coffee
2197	118	33	30	24.131797	67.225123	6.362608	173.322839	coffee
2198	117	32	34	26.272418	52.127394	6.758793	127.175293	coffee
2199	104	18	30	23.603016	60.396475	6.779833	140.937041	coffee

Figure 1: Crop yield dataset

The data science methodology was adopted for the work and was executed in 5 main stages. The methodology ensured that development stages were ran iteratively to produce optimal results. The different aspect of the data science methodology has been highlighted in figure 2;

a. **Problem Understanding:** The problem of what crops can be planted in times of drought and low fertility, was identified as the problem. The requirements for the data science solution were also outlined to be data format and programming language to be adopted and selected models to be used from the scikit-learn library of the python programming language.

b. **Data Preprocessing:** This stage handled multiple sub-stages; such as putting the data in a dataframe to get a wholistic view of the features, describing the data to review the statistics behind the data such as the variance, standard deviation and frequency of instances of the labels. The description of the data demonstrated a class balance as seen in figure 3, hence the data was fit for training. A correlation matrix function was used to visualize the patterns present, by showing the correlation coefficient between the features of the dataset. The correlation was demonstrated with a heatmap from seaborn library of the python programming language. From the heatmap as shown in figure 4, it can be seen that the features; phosphorous and potassium are 73% correlated, as the correlation between every two feature is also recorded on the heatmap. All features were used for training given that no two features were highly correlated.

c. Model Development: The dataset is split into train and test data on an 80:20 ratio at this step, and the independent and dependent variables are defined.

d. Model Training: The K-Nearest Neighbor Classifier model is trained and fitted with the training data. The Random Forest model and the Logistic Regression model underwent the similar procedure.

e. Model Evaluation: The test data were used to gauge how well each model predicted crop yields. The categorization report function and confusion matrix function together produced the test's outcome.

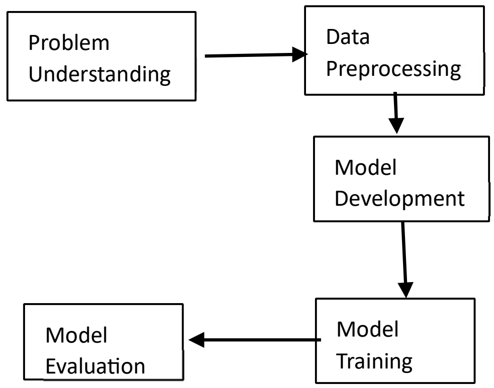

Figure 2: Work flow of the predictive model

```
cotton          100
coffee          100
pomegranate     100
blackgram       100
coconut         100
pigeonpeas      100
rice            100
mango           100
chickpea        100
mothbeans       100
papaya          100
maize           100
grapes          100
orange          100
jute            100
banana          100
mungbean        100
watermelon      100
lentil          100
apple           100
muskmelon       100
kidneybeans     100
Name: label, dtype: int64
```

Figure 3: The dependent variables (crops) for prediction

Figure 4: Correlation between the seven variables in the dataset

4.0 Results Discussion

Random Forest, K-Nearest Neighbour (KNN) and
Logistic Regressions classification algorithms were used to build the model using dataset from Kaggle

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	13
banana	1.00	1.00	1.00	17
blackgram	0.94	1.00	0.97	16
chickpea	1.00	1.00	1.00	21
coconut	1.00	1.00	1.00	21
coffee	1.00	1.00	1.00	22
cotton	0.95	1.00	0.98	20
grapes	1.00	1.00	1.00	18
jute	0.89	0.86	0.87	28
kidneybeans	0.93	1.00	0.97	14
lentil	0.96	1.00	0.98	23
maize	1.00	0.95	0.98	21
mango	1.00	1.00	1.00	26
mothbeans	1.00	0.89	0.94	19
mungbean	1.00	1.00	1.00	24
muskmelon	1.00	1.00	1.00	23
orange	1.00	1.00	1.00	29
papaya	1.00	1.00	1.00	19
pigeonpeas	1.00	0.94	0.97	18
pomegranate	1.00	1.00	1.00	17
rice	0.76	0.81	0.79	16
watermelon	1.00	1.00	1.00	15
accuracy			0.97	440
macro avg	0.97	0.98	0.97	440
weighted avg	0.98	0.97	0.98	440

Figure 5a: Model evaluation for KNN

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	13
banana	1.00	1.00	1.00	17
blackgram	0.87	0.81	0.84	16
chickpea	1.00	1.00	1.00	21
coconut	1.00	1.00	1.00	21
coffee	1.00	1.00	1.00	22
cotton	1.00	1.00	1.00	20
grapes	1.00	1.00	1.00	18
jute	0.84	0.93	0.88	28
kidneybeans	1.00	1.00	1.00	14
lentil	0.82	1.00	0.90	23
maize	1.00	1.00	1.00	21
mango	0.96	1.00	0.98	26
mothbeans	1.00	0.79	0.88	19
mungbean	0.96	1.00	0.98	24
muskmelon	1.00	1.00	1.00	23
orange	1.00	1.00	1.00	29
papaya	0.94	0.84	0.89	19
pigeonpeas	1.00	0.94	0.97	18
pomegranate	1.00	1.00	1.00	17
rice	0.79	0.69	0.73	16
watermelon	1.00	1.00	1.00	15
accuracy			0.96	440
macro avg	0.96	0.95	0.96	440
weighted avg	0.96	0.96	0.96	440

Figure 5b: Model evaluation for Logistic Regression

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	13
banana	1.00	1.00	1.00	17
blackgram	1.00	1.00	1.00	16
chickpea	1.00	1.00	1.00	21
coconut	1.00	1.00	1.00	21
coffee	1.00	1.00	1.00	22
cotton	1.00	1.00	1.00	20
grapes	1.00	1.00	1.00	18
jute	0.88	1.00	0.93	28
kidneybeans	1.00	1.00	1.00	14
lentil	1.00	1.00	1.00	23
maize	1.00	1.00	1.00	21
mango	1.00	1.00	1.00	26
mothbeans	1.00	1.00	1.00	19
mungbean	1.00	1.00	1.00	24
muskmelon	1.00	1.00	1.00	23
orange	1.00	1.00	1.00	29
papaya	1.00	1.00	1.00	19
pigeonpeas	1.00	1.00	1.00	18
pomegranate	1.00	1.00	1.00	17
rice	1.00	0.75	0.86	16
watermelon	1.00	1.00	1.00	15
accuracy			0.99	440
macro avg	0.99	0.99	0.99	440
weighted avg	0.99	0.99	0.99	440

Figure 5c: Model evaluation for Random Forest

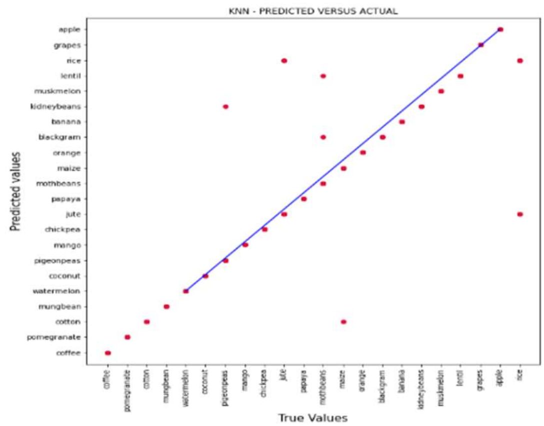

Figure 6a: Scatter plot of predicted and actual values of KNN model

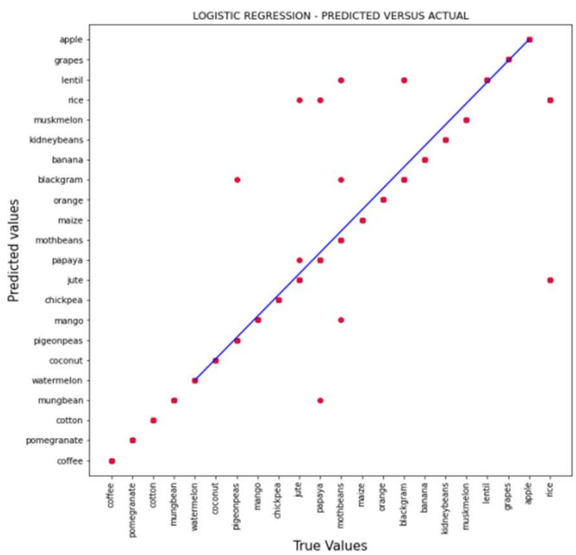

Figure 6b: Scatter plot of predicted and actual values of
Logistics Regression model

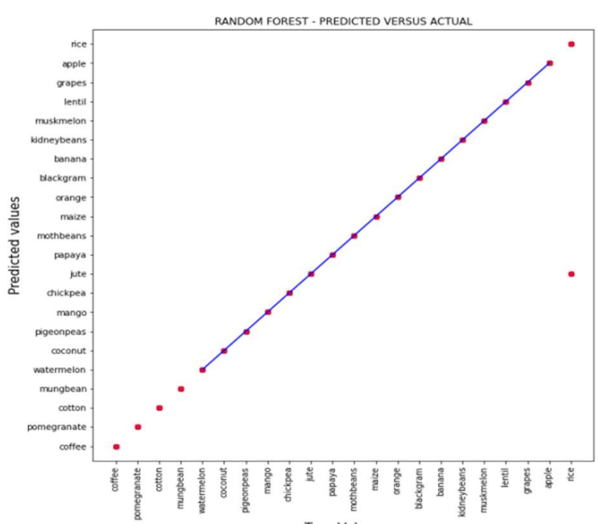

Figure 6c: Scatter plot of predicted vs actual values of Random Forest model

Figure 5(a), (b)and (c) shows the classification report that was used to measure the quality of the predictions of the three classification algorithms, based on the precision, recall and f1-score metrics. The precision score gave the information that 97% of KNN prediction were correct, and it recorded 96% and 99% for logistics regression and random forest respectively. The recall gave the information that KNN predicted the true positives up to a tune of 98%, while it was 95% and 99% for logistics regression and random forest respectively. Random forest had the best f1-score of 0.99, and was closely followed by KNN (0.97) and logistic regression (0.96).

The graphical representation of the quality of prediction by the classification models have been presented as a scatter plot as seen in figure 6(a), (b) and (c). The linear graph of the random forest model, figure 6c, was observed to have most captured most of its data points, hence making it the most accurate amongst the three models.

5.0 Conclusion

Three selected classification algorithms were trained with labeled data. The evaluated models demonstrated different levels of prediction accuracy. The Random Forest classification model statistically outperformed other classification algorithms used with an accuracy of 99%, whereas when K- Nearest Neighbor (KNN) and Logistic Regression recorded accuracy of 97% and 96% respectively. This work therefore recommends the use of random forest algorithm as the classification algorithm for crop recommendation systems.

References

Fuentes, A., Yoon, S., Kim, C. & Park, D. S. (2017). "A robust deep learning-based detector for real-time tomato plant diseases and pests' recognition," Sensors, 17(9) 2022, 1 – 21.

Devika, A. & Ananthi, B. (2018). "Analysis of crop yield prediction using data mining technique to predict annual yield of major crops," International Research
Journal of Engineering and Technology, 5(12) 14601465,

Basso, B., & Liu, L. (2019). Seasonal crop yield forecast: Methods, applications, and accuracies. advances in agronomy, 154, 201-255.

Bussay, A., van der Velde, M., Fumagalli, D., & Seguini, L. (2015). Improving operational maize yield
forecasting in Hungary. Agricultural Systems, 141, 94106.

Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 785-794. ACM.

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., ... & Reichert, G. (2015). Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agricultural and Forest Meteorology, 206, 137-150.

Cover, T. M., Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

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

Bondre, D. A. & Mahagaonkar, S. (2019) “Prediction of Crop Yield and Fertilizer Recommendation Using Machine Learning Algorithms,” International Journal of Engineering Applied Sciences and Technology,4(5)371376.

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda Bustamante, W. (2014). Attribute selection impact on linear and nonlinear regression models for crop yield prediction. The Scientific World Journal. 1 – 11.

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Lawrence J. (1994), Introduction to Neural Networks, California Scientific Software Press, Nevada City, CA,.

Sun, J., L, Di, Sun, Z., Shen, Y. & Lai, Z. (2019). “County-level soybean yield prediction using deep CNN-LSTM model,” Sensors, 19(20) 4363,

Marchant, J. A. & Onyango, C. M. (2002) Comparison of Bayesian classifier with multilayer feed-forward neural network using example of plant/weed/soil discrimination, Computers and Electronics in
Agriculture 39 3-22.

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), e0156571.

Jones, H. G., & Vaughan, R. A. (2010). Remote sensing of vegetation: principles, techniques, and applications. Oxford university press.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., Klein, M. (2002). Logistic regression. New York: SpringerVerlag.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

López-Lozano, R., Duveiller, G., Seguini, L., Meroni, M., García-Condado, S., Hooker, J., ... & Baruth, B. (2015). Towards regional grain yield forecasting with 1 km-resolution EO biophysical products: Strengths and limitations at pan-European level. Agricultural and Forest Meteorology, 206, 12-32.

Kaul, M., Hill, R. L., & Walthall, C. (2005) Artificial neural network for corn and soybean prediction, Agricultural System 85 1-18.

Mandic, D. P., &

Chambers, J. (2001). Recurrent neural networks for prediction: learning algorithms, architectures and stability. John Wiley Sons, Inc..

Tiwari, P., & Shukla, P., (2018). “Crop yield prediction by modified convolutional neural network and geographical indexes,” International Journal of Computer Sciences and Engineering, 6(8) 503-513, Alvarez, R., (2009) Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach, European Journal of Agronomy 30 70-77.

R. Wieland, W. Mirschel, (2008) Adaptive fuzzy modelling versus artificial neural network, Environmental Modelling & Software 23 215-224.

S.A. Shearer, T.F. Burks, J.P. Fulton, S.F. Higgins, J.A. Thomasson, T.G. Mueller, et al., (2000). Yield prediction using a neural network classifier trained using soil landscape features and soil fertility data, in: Proceeding of Annual International Meeting, Midwest Express Center, Milwaukee, Wisconsin,

Sak, H., Senior, A., Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.

Seber, G. A., Lee, A. J. (2012). Linear regression analysis (Vol. 329). A John Wiley & Sons Publication. 227 -232

Senthil Kumar, T., (2020)."Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm," Journal of Artificial Intelligence, 2(03) 185-193,

Thyagarajan, T., Shanmugam, J., Panda, R. C., Ponnavaiko, M., & Rao, R. G., (1998) Artificial neural networks: Principle and application to model based control of drying systems—A review, Drying Technology 16 (6) 931-966.

urada, J. M. (1992). Introduction to artificial neural systems (Vol. 8). St. Paul: West publishing company

You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017, February). Deep gaussian process for crop yield prediction based on remote sensing data. In Thirty-First AAAI conference on artificial intelligence

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.