Malaysian Journal of Computing, 6(1): 679-697, 2021

eISSN: 2600-8238 online

IMPLEMENTATION OF MACHINE LEARNING FOR PREDICTING MAIZE CROP YIELDS USING MULTIPLE LINEAR REGRESSION AND BACKWARD ELIMINATION

Stephen Gbenga Fashoto[1*], Elliot Mbunge[2], Gabriel Ogunleye[3] and Johan Van den Burg[4]

[1]*Department of Computer Science,*
University of Eswatini, Kwaluseni, Eswatini (formerly Swaziland)
[2]*Department of Computer Science,*
Federal University, Oye-Ekiti, Nigeria
[1*]gbengafash2@gmail.com, [2]mbungeelliot@gmail.com, [3]Ope992000@yahoo.com,
[4]johanvdburg93@gmail.com

ABSTRACT

Predicting maize crop yields especially in maize production is paramount in order to alleviate poverty and contribute towards food security. Many regions experience food shortage especially in Africa because of uncertain climatic changes, poor irrigation facilities, reduction in soil fertility and traditional farming techniques. Therefore, predicting maize crop yields helps policymakers to make timely import and export decisions to strengthen national food security. However, none of the published work has been done to predict maize crop yields using machine learning in Eswatini, Africa. This paper aimed at applying machine learning (ML) to predict maize yields for a single season in Eswatini. A ML model was trained and tested using open-source data and local data. This is done by using three different data splits with the open-source predictor data consisting of 48 data points each with 7 attributes and open-source response data consisting of 48 data points each with a single attribute, adjusted R^2 values were 0.784 (at 70:30), 0.849 (at 80:20), and 0.878 (at 90:10) before being normalized, 1.00 across the board after normalization, and 0.846 (at 70:30), 0.886 (at 80:20), and 0.885 (at 90:10) after backward elimination. At the second attempt, it is done by using the combined predictor data of 68 data points with 7 attributes each and combined response data of 68 data points with a single attribute each, with the same data splits and methods adjusted R^2 values were 0.966 (at 70:30), 0.972 (at 80:20), and 0.978 (at 90:10) before being normalized, 1.00 across the board after normalization, and 0.967 (at 70:30), 0.973 (at 80:20), and 0.978 (at 90:10) after backward elimination.

Keywords: *Agricultural technology, Backward elimination, Environmental factors, Linear Regression Machine Learning, Maize crop.*

Received for review: 14-04-2020; Accepted: 30-08-2020; Published: 29-01-2021

1. Introduction

Agricultural activities have evolved drastically, moving from traditional farming methods to smart farming, to ensure food security and alleviate poverty. Unfortunately, these advancements are generally limited to large multinational corporations operating in developed countries because of huge capital investment, the technological infrastructure required to introduce and practice smart farming, as well as policies that support the digitalization of farming activities. Since antiquity, agriculture has contributed to a significantly large and important part of the world's economy. It has allowed populations to flourish by providing means to sustain them, both through feeding the population directly as well as using excess food produced as a means to trade and provide an income for households.

Globally, agriculture plays important role in many economies including Africa, and it involves growing, harvesting and primary processing of crops, plants as well as breeding and raising animals, both aquatic and terrestrial (Oliver, 2014). Due to the various agricultural activities involved in farming, this paper looked at the floral aspect of agriculture, in particular, the maize (Zea mays) crop. Maize is now the most produced and most consumed grain in the world with South Africa being the world's biggest producers (Ranum *et al.*, 2014). The period of introduction of maize to Africa has not been definitively proven but the consensus is that it was brought to the continent by Europeans from the New World during the 16th century (Miracle, 1965; Walle & McCann, 2005), while the introduction of maize in Eswatini occurred during the mid-to-late 19th century (Mkhabela *et al.*, 2005). Currently, 77% of Eswatini's population relies on subsistence farming (*Eswatini | World Food Programme*, n.d.), most of which is maize farming is supported with maize imported from elsewhere while 63% of the country's population lives below the poverty line (Akinnuwesi *et al.*, 2020; Angelique & Nicholas, n.d.; Fielding-Miller *et al.*, 2014; Mwendera, 2006). In Eswatini, there are currently over 60 governmentally recognized maize varieties, including white and yellow maize varieties.

With such a large maize variety coupled and little published work, there is a need to apply machine learning to predict maize crop yields in Eswatini instead of relying on previous year yield experience. Changes in the climatic conditions and environmental factors make crop yields prediction difficult especially when relying on previous' harvest. To accurately predict crop yields, climatic parameters, maize crop specifies, environmental factors and phenotype information should be considered as well as machine learning models to process such information. Machine learning is a subset field in Artificial Intelligence which allows a system to conclude a result or find a solution to a task without being explicitly programmed to find that particular solution. Machine learning is usually categorized into three forms, namely unsupervised, supervised, and semi-supervised learning (Mbunge *et al.*, 2015). Unsupervised machine learning refers to a system identifying similarities between unclassified data in a set and then looking for further similarities, or lack thereof, in any new data provided to it. Supervised machine learning refers to training a system by feeding it test data for typical relationships between input and output data, and then predicting output for any new input data using the algorithm developed. Semi-supervised machine learning is seen as a combination of both the unsupervised and supervised methods. It involves using a set consisting of both classified and unclassified data to come up with an algorithm. The use of technology such as machine learning in the agricultural sector has allowed players in the industry to produce food more efficiently, thus enabling the rapid increase in the world's population, especially during the 20th century (Chlingaryan *et al.*, 2018; Wik *et al.*, 2008).

This study aimed at developing supervised machine learning model that predicts the yield of maize crop grown in Eswatini using multiple linear regression and backward elimination. The remainder of the paper is organized as follows: In section 2, we review the literature on crops prediction using computing tools, review of existing systems. Section 3 on methodology, describes the type of data used, methods of data collection, as well as explaining the proposed framework. The implementation and discussion of results are presented in section 4, while some conclusions are drawn in section 5.

2. Related Work

The earliest reference to modelling maize production comes from a simulation model which was used to monitor maize growth and development (Brown, 1987). The model monitored various factors which could affect growth such as fertilizer rates, weather, and soil. This CERES-Maize model was then used in a more wide-scale scenario to predict the total production of the entire United State (U.S.) Cornbelt, which at the time made up 85% of total maize production in the U.S (Hodges *et al.*, 1987). The CERES-Maize model has subsequently been updated a few times since, leading to the latest version 4.5.

The use of regression techniques has also been tried before where regression techniques were used to predict crop yields of several crops, including maize, in conjunction with data mining (Khaki & Wang, 2019). Also, more recently, together with the more mainstream use of Artificial Intelligence in all industry sectors, there have been increased suggestions on how to utilize machine learning techniques to predict crop yields all over the world. In Mexico, the use of multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression and k-nearest neighbour methods (Gonzalez-Sanchez *et al.,* 2014) was justified for that scenario since there were ten different crops, including maize. However, in this study, the focus was solely on multiple linear regression because, as shown by the results for the Mexican study, to be the best method for predicting maize yield.

The study carried out in Bangladesh focused on small-scale farmers and used a reasonable set of machine learning techniques, utilizing supervised machine learning, decision trees, and K-nearest neighbour regression (Shakoor *et al.*, 2017). Also, the crops used by the authors were their local subsistence crops whereas the maize crops used in this study are the main subsistence crop grown in Eswatini. The use of machine learning has also been applied in the United States of America to create statistical models for forecasting crop yield productions (Cai *et al.,* 2017; Forkuor *et al.,* 2017; Palanivel & Surianarayanan, 2019); however, this was in the context of a developed country with the resources available to keep the soil well managed, as well as the regular use of remote sensing. In India, a system created by Liakos *et al.,* (2018) applied multi-linear regression and machine learning to predict the yields of several different crops ranging from barley, beans, carrots, as well as other local crops through the implementation of an Android and web application and sought to find the most profitable crop for the farmer.

The framework shown in Figure 1 depicts the architecture of the system used by Zingade *et al.* (2018) which uses multiple linear regression to predict crop production for several different crops through the implementation of an Android application. While there is a good use of external factors in the description of inputs, some aspects such as the presence of fertilizer (socio-economic) and the presence of plants in the immediate surrounding area (biotic) are not considered. Finally, the output given by the model is different to the outcome of this study as the study work towards different ends; the output for this study is solely predicted crop yield for a certain variety of maize.

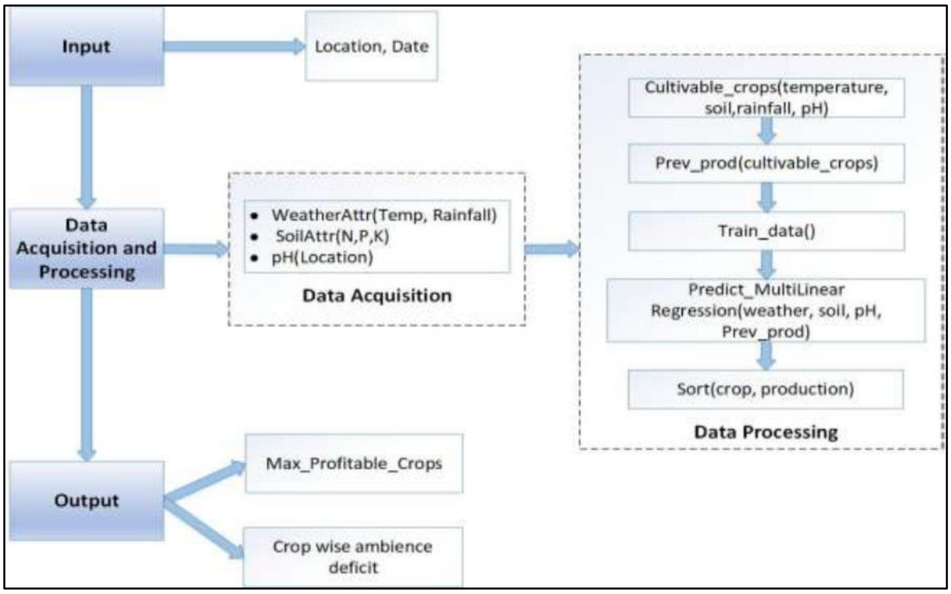

Figure 1: Framework applied by Cai *et al*., (2017)

In the study conducted by Gonzalez-Sanchez *et al.*, (2014), several factors were chosen to be used as predictor attributes for the study (Planting area, Irrigation water depth, Solar radiation, Rainfall, the time to maturity of the cultivar, and Maximum, Average, and Minimum temperatures), however, other factors which could prove to be significant in Eswatini were lacking. The authors further proceeded to use multiple linear regression, regression trees, artificial neural networks, support vector regression, and k-nearest neighbour methods in comparison to finding the method that most accurately predicts yield for all the crops considered.

3. Methodology

3.1 Data Collection

The data were obtained from the National Agricultural Library of the United States Department of Agriculture and Ministry of Agriculture in Eswatini (formerly Swaziland). The factors to be taken into consideration when defining the inputs which can be used to produce the output, i.e. crop yield, can be generally classified into internal and external or environmental factors (Singh *et al.*, 2016; Tamil Nadu Agricultural University, 2020). Internal factors are generally determined by the varieties of maize through their genetic makeup while the external factors are provided through the use of secondary data sets sourced, firstly, from open datasets sourced from the National Agricultural Library of the United States Department of Agriculture where experiments were carried out by the USDA-Agricultural Research Service to be used as the training set for the machine learning model (Comas *et al.*, 2018). Local data came from the Ministry of Agriculture in Eswatini for crop yield, land use, and fertilizer use variables, and the Centre for Environmental Data Analysis based in the UK for satellite data providing values for precipitation, evapotranspiration, temperature, and vapour pressure, as well as the free solar radiation data gathered by satellites and provided by SoDa's (www.soda-pro.com) HelioClim-1 web service.

Table 1: Sample of Local dataset

Year	**Month**	**Day**	**Irradiance**	**Uncertainty**	**Nb hours**	**Nb data**	**Top of Atmosphere**	**Clear-Sky**	**Irradiation**	**Irradiation J/cm2**	**Irradiance in MJ per m2 per day**
1985	1	1	353	34	31	9	501	391	8472	3049	30.4992
1985	1	2	343	34	31	9	501	391	8232	2963	29.6352
1985	1	3	378	35	31	9	501	391	9072	3265	32.6592
1985	1	4	306	32	31	9	500	390	7344	2643	26.4384
1985	1	5	241	28	31	9	500	390	5784	2082	20.8224
1985	1	6	195	25	31	9	500	390	4680	1684	16.848
1985	1	7	280	31	31	9	499	389	6720	2419	24.192
1985	1	8	134	19	31	9	499	389	3216	1157	11.5776
1985	1	9	264	30	31	9	498	388	6336	2280	22.8096
1985	1	10	370	34	23	9	498	388	8880	3196	31.968

Internal factors are the aspects that are defined by the genetic makeup or internal workings of the seed when it is put into the ground. These factors include:

a) Resistance to diseases, which refers to the seed’s ability to withstand attacks from bacteria and viruses
b) Resistance to pests, which refers to the seed’s ability to either repel, hinder or withstand attacks from insects, birds, or other creatures which may try to eat the seed
c) Resistance to salinity, which refers to the ability to withstand higher than average levels of alkalinity or acidity in the soil which could affect growth rate
d) Resistance to drought or floods, which refers to the plant’s ability to survive or even thrive during periods of low or excessively high rainfall
e) An innate time to maturity for the plant, which determines the earliest time for the plant to reach full maturity given perfect conditions

External or environmental factors, on the other hand, are the aspects that define what happens around the seed or plant which may affect its growth and yield potential. These include:

a) Climatic factors, which in turn can be broken down into variables and their persistence such as precipitation, temperature, atmospheric humidity, solar radiation, wind velocity, and atmospheric gases
b) Edaphic factors, which refers to the factors related to the soil such as moisture, temperature, mineral and organic content, the presence of other organisms, and the soil’s pH level
c) Biotic factors, which refers to the positive or negative effects posed by plants and animals in the immediate surrounding area
d) Physiographic factors, which refer to the elements such as topography (how to level the land is), altitude, and the exposure to light and wind
e) Socio-economic factors, which can be seen as the human factor i.e. the ability to provide inputs such as fertilizers, insecticides, the land to be used for planting, as well as how well the farming process is managed

3.2 Proposed System Architecture

This study was carried out by having various environmental or external factors from both the open-source and local datasets fed into the system as attributes and using crop yields from previous years to form relationships between the data during the training phase. The open-source and local data being used for training and testing was split 70:30, 80:20, and 90:10 to ensure the most accurate and reliable relationships are formed. The factors which were considered when defining the inputs for the machine learning model are as follows:

- Water input (rainfall and irrigation)
- Nitrogen added through fertilizer
- Crop evapotranspiration
- The area used to grow the plants in the study
- Average air temperature
- Vapour pressure
- Solar radiation

An important aspect to note is there was missing data in the local dataset with regards to the amount of fertilizer used. This was taken care of through imputation using the mean as the strategy to fill in the missing data. The response variable was given by the attribute Grain Yield (normalized to 15.5% moisture) in the open-source dataset, and Yield in the local dataset.

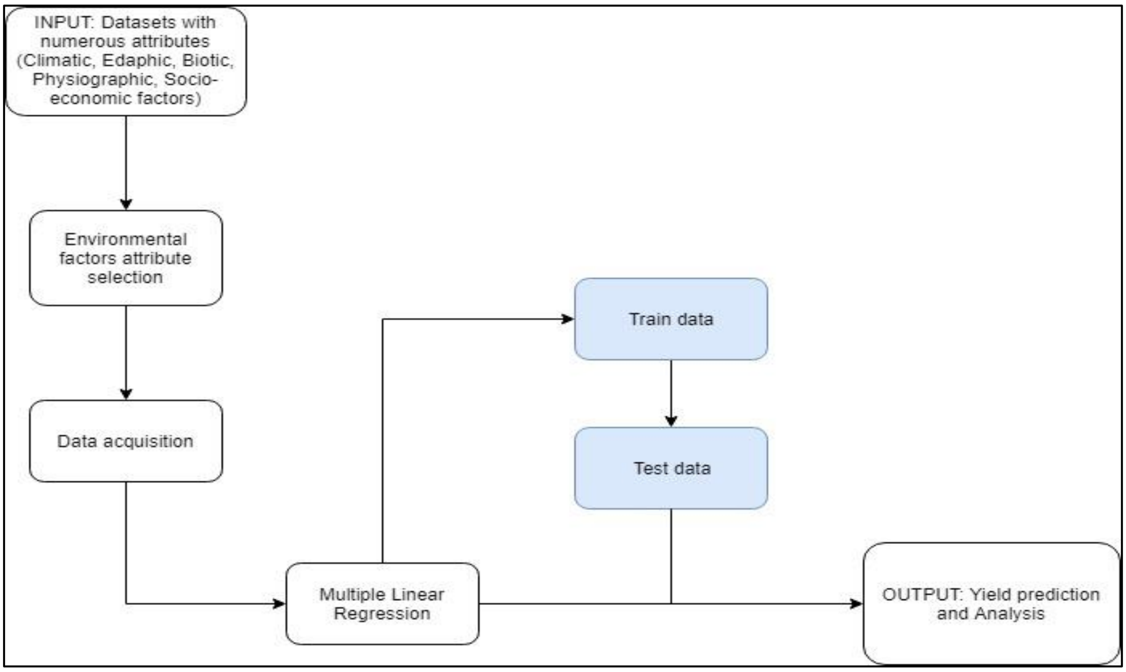

Figure 2: Proposed framework for this study

Our proposed framework consists of input dataset sources, attribute selection, data acquisition, multiple linear regression and output. The input dataset sources are the part of the framework which dealt with collecting all the relevant datasets from their various sources. The Attribute Selection part is responsible for the selection of the relevant environmental factors, or attributes had to be selected. This was done by looking through the open-source and local datasets to see which attributes were present in both and utilizing them as the independent variables for the model. The Data Acquisition part involved the reading of the data from their respective files and ordering them into an array consisting of independent variables and an array consisting of the response variable (Yield). The multiple linear regression was then carried out by splitting the data in the arrays into training and testing data and then predicting the response variable using the model to compare with the actual response variable given in the testing split. Following the prediction of the crop yield, the results were analyzed using adjusted R2 scores and Root Mean Squared Error. The dataset was then normalized to determine whether normalization can improve the accuracy of the yield, and therefore the results were compared to the unnormalized dataset using the same performance matrix. Finally, backward elimination was implemented on the dataset to determine the factors which affect the yield the most and the least.

3.3 Regression Models in Machine Learning

In the course of this study, the method used to predict or estimate the crop yield for a certain variety of maize was multiple linear regression. As a result, it is pertinent to describe what multiple linear regression is and how it is performed. Multiple linear regression is used to model the relationship between one or more independent (or predictor) variables and a single dependent (or response) variable by relating the observed input data to the resulting value through determining a linear equation which will relate the two, i.e. every independent variable xi is associated with the dependent variable y. The model for multiple linear regression is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon \tag{1}$$

where we have ϵ represents the residual terms of the model, k representing the number of observation or features, and each β_i representing the regression coefficient or feature weight.

In order to train the multiple linear regression model defined in equation (**1)** by fitting the input values x_i, a cost function was employed to determine and minimize the difference between the observed or true y values and the fitted or predicted y values. The cost function which was used by this study was the Root Mean Square Error function in equation (2),

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{2}$$

Where we have

- n determining the number of data points in the sample
- y_j representing the observed or true values, and
- $\hat{y}_j$ representing the predicted or fitted values

The R^2 value or coefficient of determination is used to determine how well the predictor values fit the model being assessed and is given by equation (3):

$$R^2 = 1 - \frac{SS_{regression}}{SS_{total}} \tag{3}$$

where, $SS_{regression}$ denotes the sum of squares as a result of the regression, and SS_{total} represents the total sum of squares. While using R^2 in linear regression is perfectly acceptable, when it comes to multiple linear regression the formula needs slight alteration because the greater the number of independent variables added, the more the R^2 value improves regardless of whether the newly added variables make a significant difference or not. As a result, for this study it is more useful to use the adjusted R^2 formula in equation (4):

$$Adjusted\ R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-(k+1)}\right) \tag{4}$$

In equation (4), n denotes the size of the sample used, and k is given by the number of predictors in the regression equation.

4. Implementation Results

The system for predicting crop yield was developed using Python on Anaconda 3 bundled with Spyder. Anaconda is an open-source program that contains numerous Python packages which can be used during programming to make implementation easier for the developer through the use of predetermined functions. Spyder is an Integrated Development Environment (IDE) that gives the developer an alternative to using the terminal to write programs and carry out instructions. The libraries used in this study are Numpy, Scikit-learn, Panda. Matplotlib and Statsmodels. Applying multiple linear regression on the US Department of Agriculture dataset over the period 2008-2013 resulted in output in Table 2 before normalization and backward elimination:

Table 2: Results of training and testing before normalization and backward elimination with open dataset

Test size	**Train size**	**R^2**	**Adjusted R^2**	**RMSE**
30%	70%	0.820	0.784	989.31
20%	80%	0.870	0.849	739.15
10%	90%	**0.900**	**0.878**	**693.28**

In an attempt to improve the accuracy of the model the data was normalized and produced the results in Table 3.

Table 3: Results after normalizing open dataset

Test size	**Train size**	**R^2**	**Adjusted R^2**	**RMSE**
30%	70%	1.00	1.00	0.00
20%	80%	1.00	1.00	0.00
10%	90%	1.00	1.00	0.00

An alternative method was then executed to try to improve the accuracy by implementing backward elimination instead of normalization, removing the explanatory variable with the highest p-value greater than 0.5 as shown in Table 4.

Table 4: Results after backward elimination on open dataset only

Test size	**Train size**	**R^2**	**Adjusted R^2**	**RMSE**
30%	70%	0.870	0.846	846.31
20%	80%	**0.900**	**0.886**	**650.95**
10%	90%	0.900	0.885	681.64

Graphically, the predicted results in Table 2 to 4 are displayed in Figure 3 to 8:

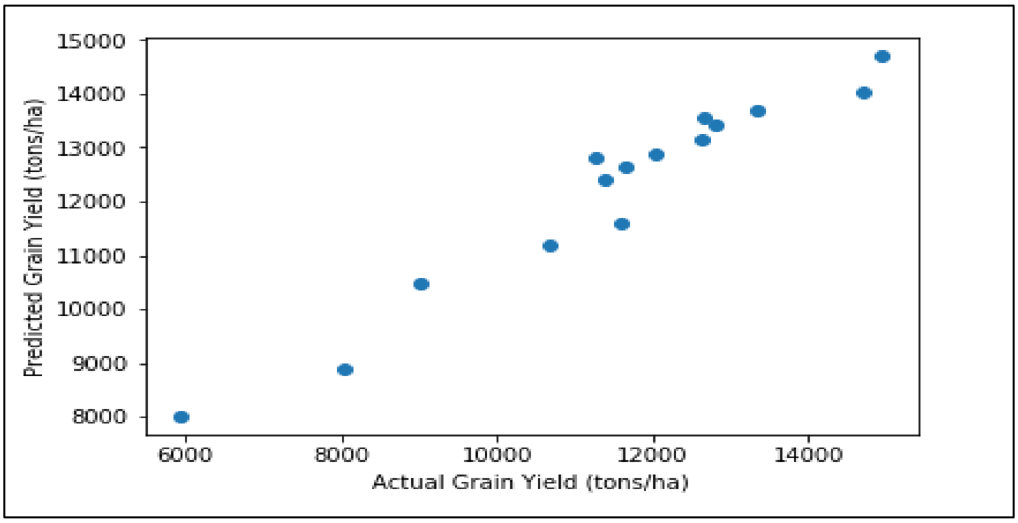

Figure 3: Before backward elimination on the open dataset at 70:30 split

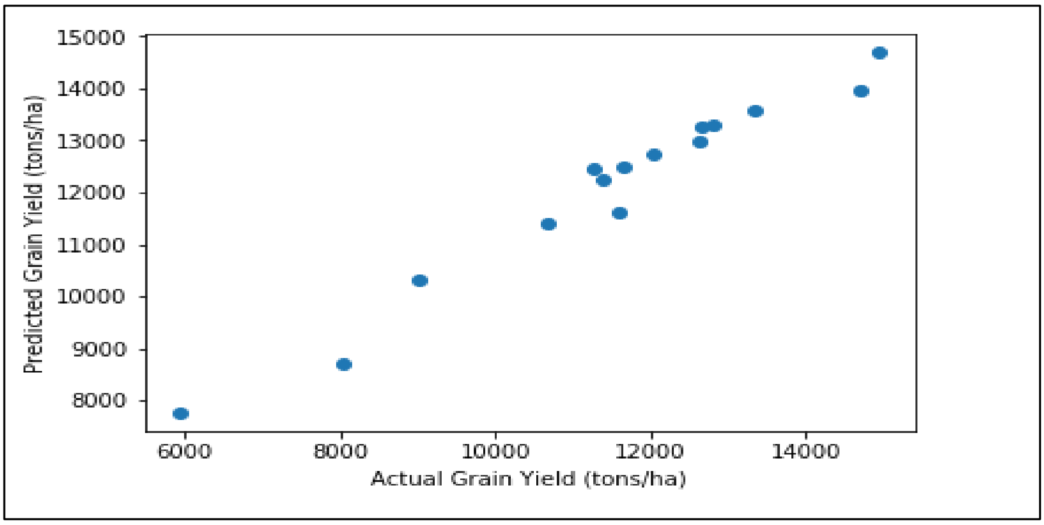

Figure 4: After backward elimination on the open dataset at 70:30 split

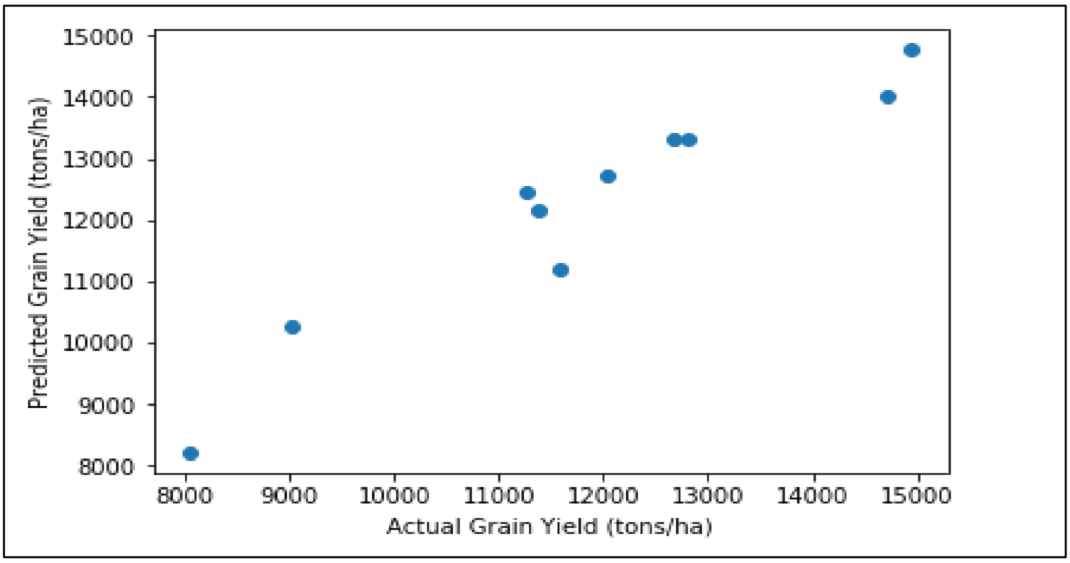

Figure 5: Before backward elimination on the open dataset at 80:20 split

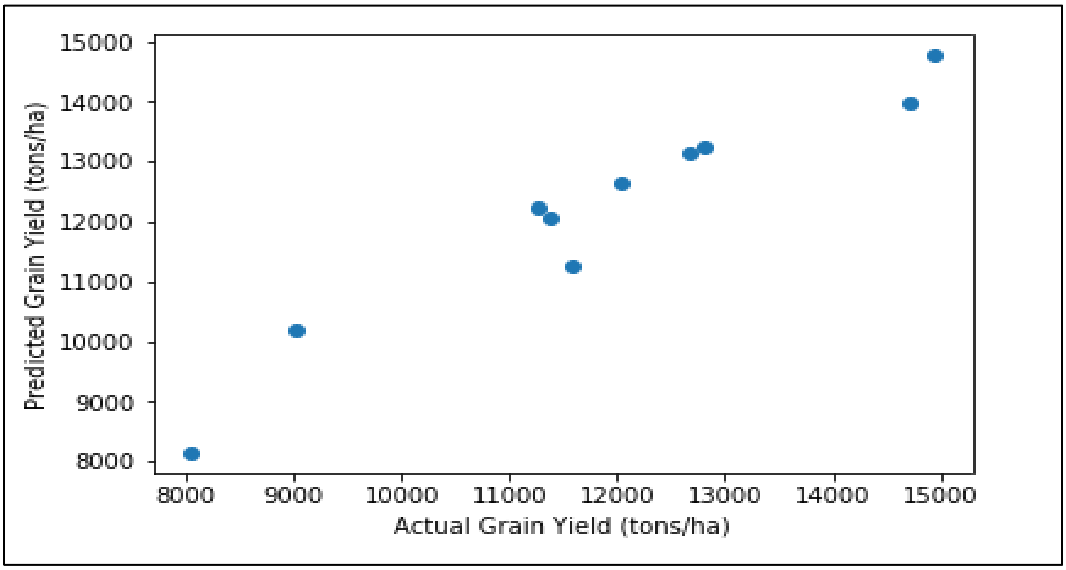

Figure 6: After backward elimination on the open dataset at 80:20 split

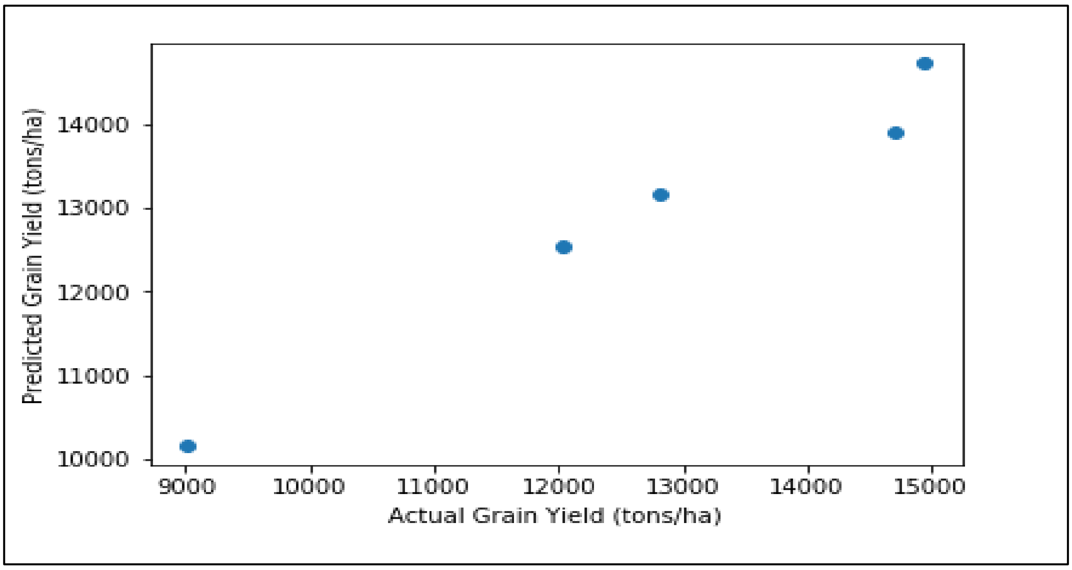

Figure 7: Before backward elimination on the open dataset at 90:10 split

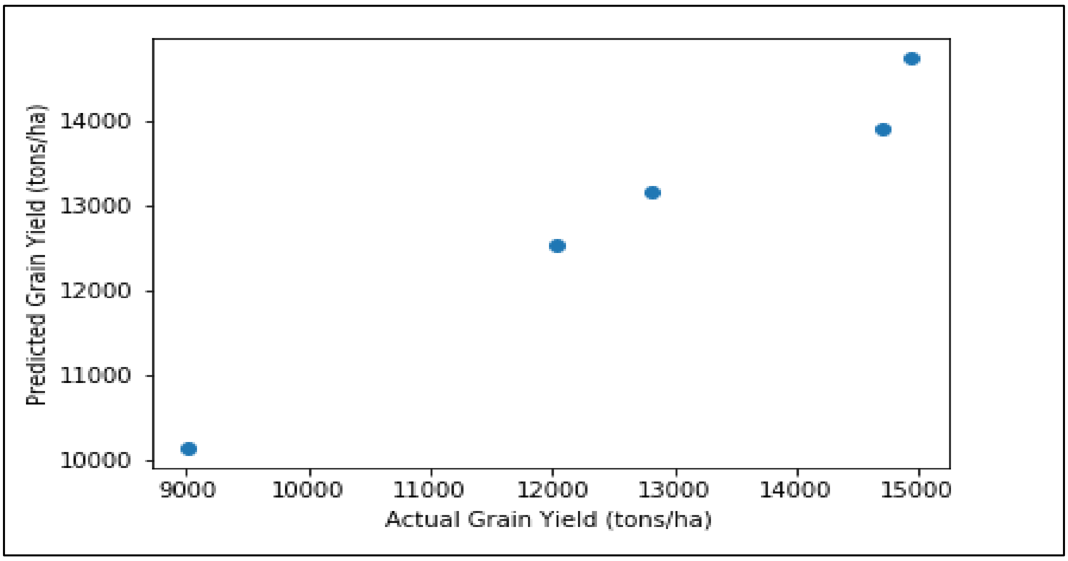

Figure 8: After backward elimination on the open dataset at 90:10 split

Eswatini data obtained for the years 1985 to 2005 were then added and evaluated similarly as shown in Table 5.

Table 5: Results of training and testing before normalization and backward elimination using the local data

Test size	Train size	R^2	Adjusted R^2	RMSE
30%	70%	0.970	0.966	934.21
20%	80%	0.980	0.972	828.53
10%	90%	**0.980**	**0.978**	**569.95**

Table 6: Results after normalizing the combined dataset

Test size	Train size	R^2	Adjusted R^2	RMSE
30%	70%	1.00	1.00	0.00
20%	80%	1.00	1.00	0.00
10%	90%	1.00	1.00	0.00

Table 7: Results after backward elimination on the combined dataset

Test size	Train size	R^2	Adjusted R^2	RMSE
30%	70%	0.970	0.967	925.88
20%	80%	0.980	0.973	826.71
10%	90%	**0.980**	**0.978**	**568.45**

Graphically the results of the predictions were shown in Figure 9-14.

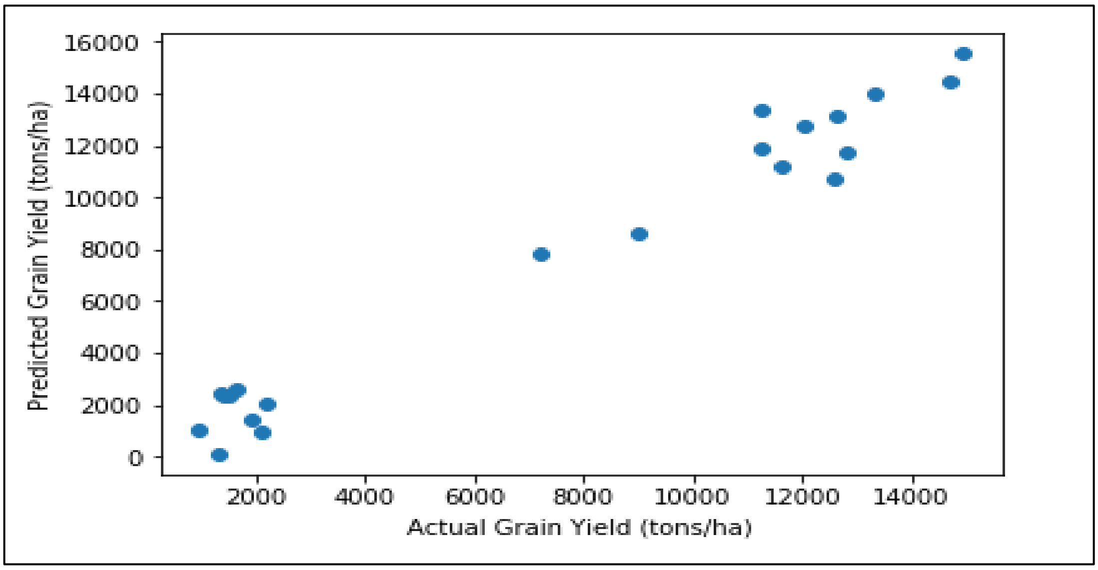

Figure 9: Before backward elimination on the combined dataset at 70:30 split

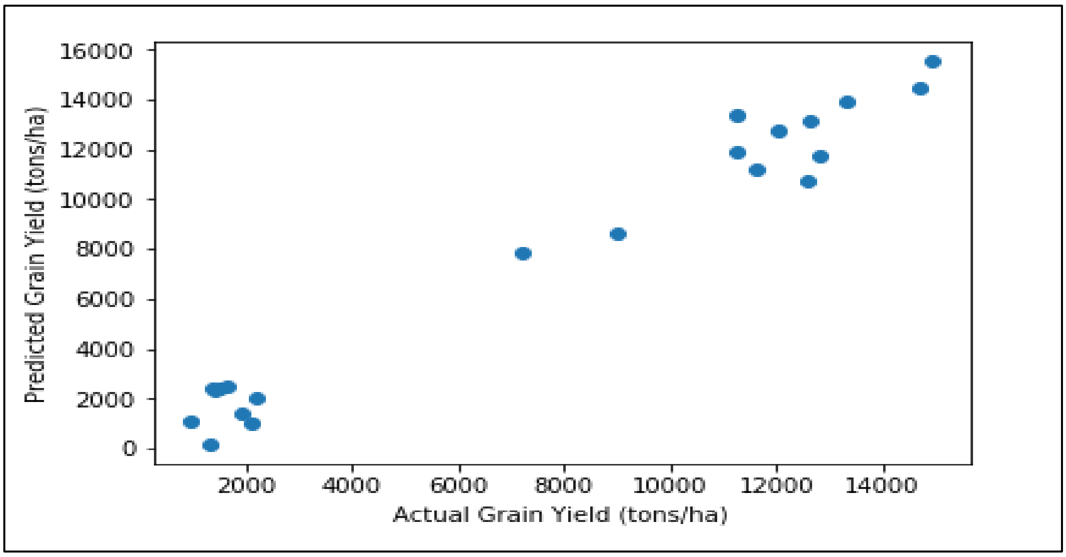

Figure 10: After backward elimination on the combined dataset at 70:30 split

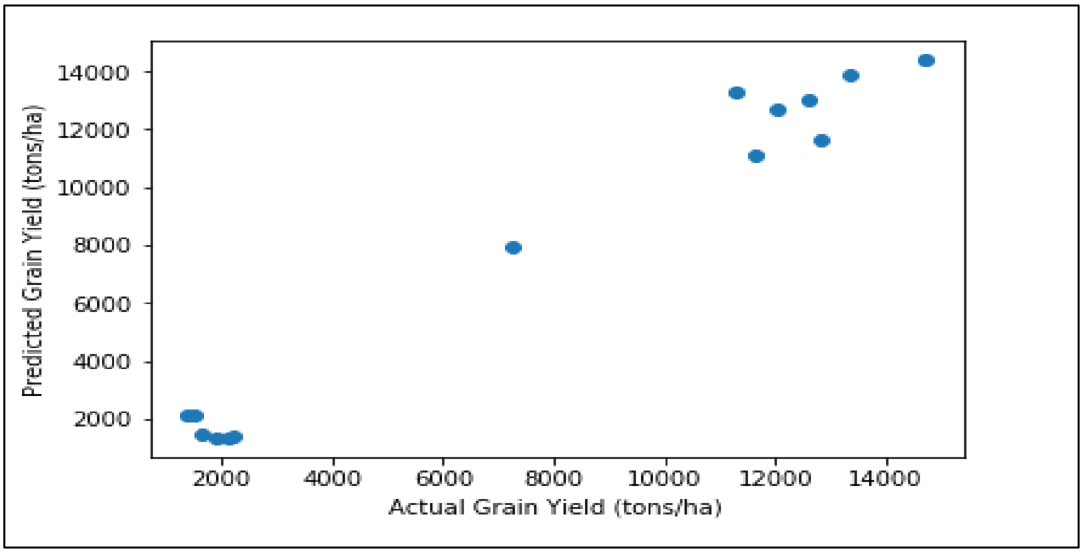

Figure 11: Before backward elimination on the combined dataset at 80:20 split

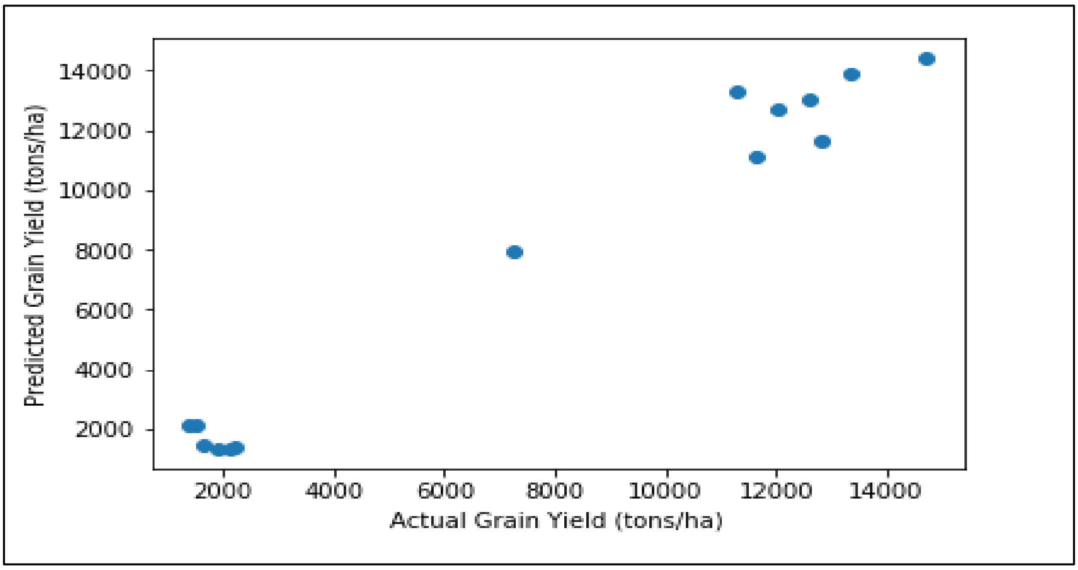

Figure 12: After backward elimination on the combined dataset at 80:20 split

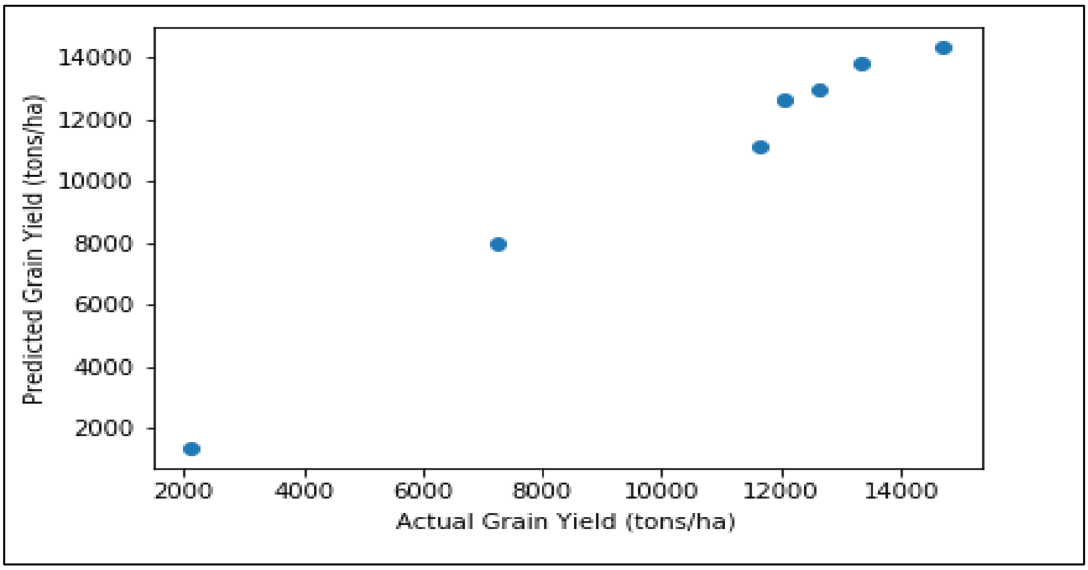

Figure 13: Before backward elimination on the combined dataset at 90:10 split

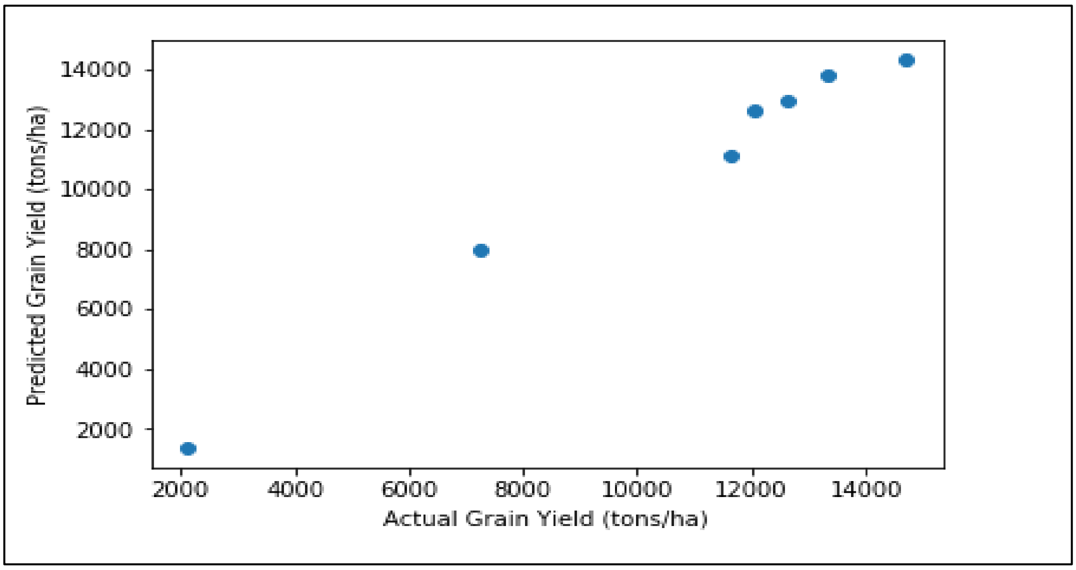

Figure 14: After backward elimination on the combined dataset at 90:10 split

5. Discussion of Results

The results on Table 2 shows that the data split of 90:10 before normalization and backward elimination outperform the data-split of 80:20 and 70:30 in terms of adjusted R-square and RMSE while Table 4 results reveal that the data-split of 80:20 after backward elimination outperform the data-split of 90:10 and 70:30 in terms of adjusted R-square and RMSE. The results on Table 5 shows that the data split of 90:10 before normalization and backward elimination outperform the data-split of 80:20 and 70:30 in term of RMSE while Table 7 results reveal that the data-split of 80:20 after backward elimination outperform the data-split of 90:10 and 70:30 in term of RMSE. Also, Tables 3 and 6 shows that the result for all the data-split is the same after the normalization of the dataset. The Figures 3 to 8 shows the results of the open dataset while figures 9 to 14 shows the results of the combined dataset before and after the introduction of backward elimination method on the three-percentage split carried out.

6. Conclusion

In this study, the R^2 and adjusted R^2 values found both before and after backward elimination show a strong correlation between the predictors used in the model and the response variables, thus showing that the predictor variables fit the model well. The use of normalization helped to improve the Root Mean Squared Error score; however, the results obtained did not give a clear picture of how well normalization helped since they returned perfect values. From the use of backward elimination, it was revealed that the predictor variable that affected the yield prediction the least was the amount of fertilizer used, while the factors that affected the prediction the most were water input, relative humidity, and solar radiation. Limitations observed from implementation included the use of relatively small datasets, which in turn also limited the number of predictor attributes assessed and need for greater processing power to complete the running of the python script promptly.

Larger training and testing datasets would be more beneficial to create better relationships between the independent variables and response variables, thus improving regression coefficients in the model. Using more regionally specific data to predict yields for certain areas in the country could help towards this end and also provide better insight into how well yield prediction performs in the different regions of the country. According to the Pareto principle, the data-split of 80/20 ratio should be used often (Harvey & Sotardi, 2018) but the findings in this study do not fully support that 80/20 ratio should be used in all scenarios which are corroborated in a study by Folorunso et al., (2020). Looking into the effect of using different maize varieties in the various regions of the country in yield prediction would also be recommended as this would allow for the testing of the impact of internal factors on a plant's yield.

Moreover, machine learning models outperform traditional methods for predicting maize crop yields (Palanivel & Surianarayanan, 2019).Some scholars including (Crane-Droesch, 2018; Gandhi *et al.,* 2016; Johnson *et al.,* 2016; Palanivel & Surianarayanan, 2019; van Klompenburg *et al.*, 2020; Veenadhari *et al.*, 2014) predicted crop yields using climatic parameters. However, the future work will focus on customized and automated decision support systems on the prediction of crop yield to aid decision making to complement farmers' experience is still nascent. This will assist farmers to make informed decisions as they progress with farming activities instead of only predicting crop yields. Also, the customized and automated decision support system will help farmers to take appropriate measures to improve crop yields and optimization of farming resources hence increase the chances of improving crop harvest. To optimize farming resources such as robot weeding, planting machines are imperative especially in smart farming; dynamic routing is also paramount to effectively utilize smart farming equipment (Ajit Kumar, 2020; Mbunge *et al.*, 2020).

Acknowledgement

We would like to acknowledge the National Agricultural Library of the United States of America, the Department of Agriculture and Ministry of Agriculture in Eswatini (formerly Swaziland) for giving us modelling data.

References

Ajit Kumar, S. (2020). Applications of IoT in Agricultural System. *International Journal of Agricultural Science and Food Technology*. https://doi.org/10.17352/2455-815x.000053

Akinnuwesi, B. A., Fashoto, S. G., Metfula, A. S., & Akinnuwesi, A. N. (2020). Experimental Application of Machine Learning on Financial Inclusion Data for Governance in Eswatini. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-45002-1_36

Angelique, G. N., & Nicholas, M. O. (n.d.). *Poverty and Economic Growth in Swaziland: An Empirical Investigation.* Retrieved January 22, 2021, from www.mgt.fm-kp.si.

Benz, B. F. (2001). Archaeological evidence of teosinte domestication from Guilá Naquitz, Oaxaca. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.98.4.2104

Brown, D. M. (1987). CERES-Maize: A simulation model of maize growth and development. *Agricultural and Forest Meteorology*. https://doi.org/10.1016/0168-1923(87)90089-x

Cai, Y., Moore, K., Pellegrini, A., Elhaddad, A., Lessel, J., Townsend, C., Solak, H. & Semret, N. (2017). Crop Yield Predictions - High Resolution Statistical Model for Intra-season

Forecasts Applied to Corn in the US. *AGUFM, 2017*, GC31G-07. https://ui.adsabs.harvard.edu/abs/2017AGUFMGC31G..07C/abstract

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. In *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2018.05.012

Comas, L. H., Trout, T. J., Banks, G. T., Zhang, H., DeJonge, K. C., & Gleason, S. M. (2018). USDA-ARS Colorado maize growth and development, yield and water-use under strategic timing of irrigation, 2012–2013. *Data in Brief*. https://doi.org/10.1016/j.dib.2018.10.140

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*. https://doi.org/10.1088/1748-9326/aae159

Eswatini | World Food Programme. (n.d.). Retrieved January 22, 2021, from https://www.wfp.org/countries/eswatini

Fielding-Miller, R., Mnisi, Z., Adams, D., Baral, S., & Kennedy, C. (2014). "there is hunger in my community": A qualitative study of food security as a cyclical force in sex work in Swaziland. *BMC Public Health*, *14*(1), 1–10. https://doi.org/10.1186/1471-2458-14-79

Folorunso, S. O., Fashoto, S. G., Olaomi, J., & Fashoto, O. Y. (2020). A multi-label learning model for psychotic diseases in Nigeria. *Informatics in Medicine Unlocked*. https://doi.org/10.1016/j.imu.2020.100326

Forkuor, G., Hounkpatin, O. K. L., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using Remote Sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0170478

Gandhi, N., Petkar, O., Armstrong, L. J., & Tripathy, A. K. (2016). Rice crop yield prediction in India using support vector machines. *2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016*. https://doi.org/10.1109/JCSSE.2016.7748856

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*. https://doi.org/10.5424/sjar/2014122-4439

Harvey, H. B., & Sotardi, S. T. (2018). The Pareto Principle. *Journal of the American College of Radiology*. https://doi.org/10.1016/j.jacr.2018.02.026

Hodges, T., Botner, D., Sakamoto, C., & Hays Haug, J. (1987). Using the CERES-Maize model to estimate production for the U.S. Cornbelt. *Agricultural and Forest Meteorology*. https://doi.org/10.1016/0168-1923(87)90043-8

Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology*.

https://doi.org/10.1016/j.agrformet.2015.11.003

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*. https://doi.org/10.3389/fpls.2019.00621

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. In *Sensors (Switzerland)*. https://doi.org/10.3390/s18082674

Mbunge, E., Fashoto, S. G., & Dlamini, S. (2020). Modelling a Gossip-based Protocol for Enhanced Dynamic Routing . *Asian Journal of Information Technology* . https://www.researchgate.net/publication/343369008

Mbunge, E., Makuyana, R., Chirara, N., & Chingosho, A. (2015). Fraud Detection in E-Transactions using Deep Neural Networks-A Case of Financial Institutions in Zimbabwe. *International Journal of Science and Research*. https://doi.org/10.21275/ART20176804

Miracle, M. P. (1965). The introduction and spread of maize in Africa. *The Journal of African History*. https://doi.org/10.1017/S0021853700005326

Mkhabela, M. S., Mkhabela, M. S., & Mashinini, N. N. (2005). Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. *Agricultural and Forest Meteorology*. https://doi.org/10.1016/j.agrformet.2004.12.006

Mwendera, E. J. (2006). Rural water supply and sanitation (RWSS) coverage in Swaziland: Toward achieving millennium development goals. *Physics and Chemistry of the Earth*. https://doi.org/10.1016/j.pce.2006.08.040

Oliver, M. J. (2014). Why we need GMO crops in agriculture. In *Missouri medicine*.

Palanivel, K., & Surianarayanan, C. (2019). An Approach For Prediction Of Crop Yield Using Machine Learning And Big Data Techniques. *International Journal Of Computer Engineering And Technology*. Https://Doi.Org/10.34218/Ijcet.10.3.2019.013

Ranum, P., Peña-Rosas, J. P., & Garcia-Casal, M. N. (2014). Global maize production, utilization, and consumption. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.12396

Shakoor, M. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017). Agricultural production output prediction using Supervised Machine Learning techniques. *2017 1st International Conference on Next Generation Computing Applications, NextComp 2017*. https://doi.org/10.1109/NEXTCOMP.2017.8016196

Singh, H., Dahiphale, A., & Singh, R. K. (2016). Environmental factors affecting growth and productivity of crops. *ICCCICPFS*.

Tamil Nadu Agricultural University. (2020). Factors affecting crop production: climatic, edaphic, biotic, physiographic and socio-economic factors. In *Principles of Agronomy and Agricultural Meteorology*. http://eagri.org/eagri50/AGRO101/index.html

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. In *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2020.105709

Veenadhari, S., Misra, B., & Singh, C. D. (2014). Machine learning approach for forecasting crop yield based on climatic parameters. *2014 International Conference on Computer Communication and Informatics: Ushering in Technologies of Tomorrow, Today, ICCCI 2014*. https://doi.org/10.1109/ICCCI.2014.6921718

Walle, N. Van De, & McCann, J. C. (2005). Maize and Grace: Africa's Encounter with a New World Crop, 1500-2000. *Foreign Affairs*. https://doi.org/10.2307/20031766

Wik, M., Pingali, P., & Broca, S. (2008). Global Agricultural Performance: Past Trends and Future Prospects. *Background Paper for the World Development Report*.

Zingade, D.S., Buchade, O.,.Mehta, N., Ghodekar, S. & Mehta, C. (2018). Machine Learning-based Crop Prediction System. *International Journal of Emerging Technology and Computer Science, 3*(2).