A Comparative Analysis of Machine Learning Algorithms for Crop Yield Prediction Utilizing Agricultural Dataset in Myanmar

1st Aye Thida Win
University of Computer Studies (Kyaing Tong)
Kyaing Tong, Myanmar
ayethidawin@ucskt.edu.mm

2nd Dr.Thin Thu Naing
University of Computer Studies (Kyaing Tong)
Kyaing Tong, Myanmar
thinnthunaing@ucskt.edu.mm

3rd Dr. Myint Myint Lwin
University of Computer Studies (Kyaing Tong)
Kyaing Tong, Myanmar
myintmyintlwin@ucskt.edu.mm

Abstract— **The agriculture sector is one of the most important things for human beings because they can't live without food. As a result, farmers need to know how much their crops will produce from their yields. Traditionally, the production rate of agricultural yield is dependent on its soil type, climate, and how it is farmed. Most of the recent researchers proposed Machine Learning (ML) methods to predict the production rate based on soil types, climate, etc. However, the crop production rate of some regions especially Eastern Shan State, Myanmar is only dependent on the regions, sown area, and harvest area because the climate conditions of these regions are stable. Therefore, this research aims to conduct a comparative analysis of the K-Nearest Neighbors (KNN) and linear regression methods to forecast paddy production in Eastern Shan State. According to the experimental results, both methods achieved satisfactory levels of accuracy and precision. This article offers important insights into the potential of KNN and linear regression algorithms for predicting of paddy yield.**

Keywords— **Paddy Crop, Machine Learning, Linear Regression, K-Nearest Neighbors, Comparative Study, Yield Prediction System.**

I. Introduction

To meet the needs of a rapidly growing global population, the contemporary of agricultural sector must enhance its food production capabilities. Based on the present matter, the proposed system attempts to construct a decision assistance system via the utilization of machine learning methods. The selection of an appropriate algorithm for a certain application area is a crucial determinant in the development of a decision support system. As a result, the agricultural sector is adopting cutting-edge technologies to increase net productivity through the collection and processing of data. Moreover, climate change hints at the unavoidable need to modernize the agricultural sector with cutting-edge equipment and techniques. Consequently, in the modern era, agriculture and farming sectors are adopting modern technologies such as machine learning to increase net productivity and maximize agricultural resource utilization. It also shaped the concept of agriculture, which opened up new avenues for innovative agricultural research. KNN and linear regression are discussed in this paper. KNN is a fundamental, supervised machine learning algorithm applicable to classification and regression problems. KNN identifies the k-nearest neighbors based on the shortest distance between the query instance and the training samples. Then, the straightforward majority of the k-closest neighbors are selected as the prediction query object. For distance calculations in the KNN algorithm, Euclidean distance is predominantly employed. Linear regression is one of the simplest and most prevalent Machine Learning algorithms. It is a supervised algorithm that predicts future outcomes based on labeled data. Linear regression aims to determine the line that best represents the relationship between two variables: an independent variable (x) and a dependent variable (y). This paper intends to compare the performance of both methods which are based on the same dataset, the paddy dataset of the Eastern Shan State region.

This paper presents five sections: introduction, literature review, problem statement, methodology, and Experimental results.

II. Literature Review

Myanmar's economy is dependent on the export and import of agricultural products. Agriculture is an essential component of Myanmar's economy. Due to crop yield uncertainty, there is a significant decline in economic status. Myanmar's main commodities are Rice, Wheat, Pulses, and Grains. Myanmar's population is increasing daily, necessitating an increase in crop yields to sustain the expanding population. One of the most effective ways to predict unknown data is to use machine learning algorithms. The goal of this study is to build a machine learning-based model for predicting rice yield. In this section, previous works on machine learning algorithm methods and implementations are presented for literature reviews.

Farmers can make more informed decisions about crop production before harvesting the field by using a system based on the Random Forest algorithm in the agricultural domain. In the data collection phase, a variety of sources were used to compile datasets that illustrate the relationship between chaotic time series behavior and several factors, including meteorological conditions, temperature, humidity, precipitation, and moisture levels. In their study, the Random Forest Algorithm was used to predict the crop yield for Indian farmers via a web-based application [1].

The authors of the research demonstrated the use of many machine-learning methodologies, such as long short-term memory (LSTM), simple RNN, random forest, and xgboost, to forecast agricultural production in India. During the data collecting phase, agricultural data were gathered from the paddy fields located in Maharashtra State, India, with the assistance of both farmers and agricultural specialists [2].

The comprehensive use of image recognition in intelligent agriculture is essential for the identification of agricultural maladies. Agricultural diagnostics has seen the adoption of several machine learning techniques, including

convolutional neural networks (CNN) and transfer learning. The data collection process was conducted specifically in palm fields Their dataset consisted of various variables such as leaf and fruit information, irrigation details, soil properties, climatic data, crop management practices, historical yield records, fertilization information, cropland characteristics, and satellite data [3].

The CNN framework was used to utilize the recurrent Q-Network model to provide predictions within the agricultural sector. Their data collection included the acquisition of agricultural data, with a special focus on the Vellore district, situated in the southern portion of India. The dataset was obtained from agriculture professionals and the meteorological service in India. [4].

This study evaluates the most essential classification characteristics for generating accurate results. The machine learning (ML) algorithms Artificial Neural Network, Support Vector Regression, K-Nearest Neighbor, and Random Forest (RF) were proposed for increased precision. 70% of the data were selected at random and used to train the model, while the remaining 30% were used to evaluate the model's predictive ability. Using the same agricultural training data, their findings revealed that the RF algorithm obtains the highest accuracy based on error analysis values for all distinct feature subsets [5].

III. Dataset of the System

Food crops are subsistence produce that are consumed by humans [6]. The information that was utilized for this investigation was produced by the Eastern Shan State Myanmar Agriculture Department in conjunction with a variety of other places situated within the Eastern Shan State region.

The dataset is depicted in TABLE I. This is an example of the dataset utilized by the proposed method. It is comprised of four factors: Township p-Code, Sown Area of Paddy (acres), Harvested Area of Paddy (acres), and Paddy as a proportion of All Harvested that can be produced. The target will be the production value of the crop that will be harvested from the Sown Paddy Area (acres).

TABLE I. SAMPLE DATASET OF CROP AND YIELD DATA

	Township P-Code	Sown Area of Paddy (acres)	Harvested Area of Paddy (acres)	Paddy as percentage of All Harvested	Production of Paddy (tin)
2	1001	27042	27042	0.679225379	1744822
3	1001	48215	48215	0.625202609	3138747
4	1001	4418	4418	0.995493466	177446
5	1001	16630	16583	0.939599977	1078712
6	1001	1972	1972	0.455111932	87347
7	1001	1288	1288	0.798017348	60213
8	1001	109554	105972	0.878859503	8454832
9	1001	72626	72454	0.911852803	5243869
10	1001	28006	27718	0.97440765	1890287

IV. Methodology and Implementation

This research involves dataset development, classification, and classifier performance assessment. All of these procedures are essential for crop yield prediction. Paddy crop statistics from Eastern Shan State, Myanmar, are used to create the dataset. In this study, KNN and LR are utilized for classification. During system analysis and assessment, Mean Square Error, Mean Absolute Error, Median Absolute Error, Explained Variance Score, and R2 Score are generated to assess classification algorithm performance. The scheme of the proposed system is described in Fig. 1. There are three main steps in this system. They are data preprocessing, training, classification, and performance assessments. The first step is the data preprocessing. The cleaned dataset is used to train and test for the evaluation of classifier performance. The proposed system estimates the crop yields for new places.

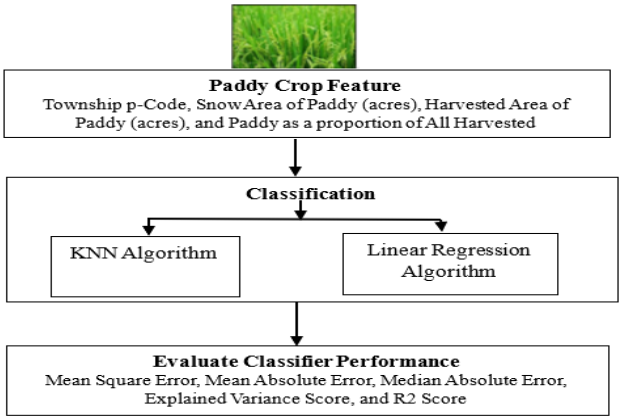

Fig. 1. Scheme of Comparative Study for Crop Yield Prediction System

The experimental results are accomplished by inputting the necessary data, which enables the learning algorithm to make a prediction and classification. To find the best classifier, the proposed system evaluates the performance of the classifiers with error methods such as Mean Square Error, Mean Absolute Error, Median Absolute Error, Explained Variance Score, and R2 Score. According to the evaluation results, the Linear Regression Algorithm is chosen to predict the crop yield for new coming data to applied real-world agriculture domain. The implementation of the proposed system is shown in Fig. 2.

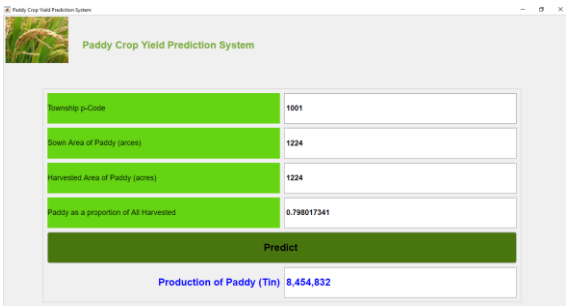

Fig. 2. Implementation of Paddy Crop Yield Prediction System

A. Classification Method

Machine learning is an Artificial Intelligence (AI) technique and branch of computer science that employs data and algorithms to mimic how humans learn and gradually enhance their accuracy. The machine learning algorithm is normally defined as supervised, unsupervised, and reinforcement and is an evaluation of the regular algorithm. There are numerous types of machine learning algorithms, including linear regression, K-Nearest Neighbor (K-NN), Random Forest, Naive Bayes, and Artificial Neural Network [7].

Classification is a supervised machine learning technique in which the algorithm learns from the provided data input and then applies this learning to classify new

observations. K-nearest neighbor (K-NN) is the algorithm used in the aforementioned classification method.

B. K-NN Algorithm

The K-Nearest Neighbor (KNN) technique is regarded as one of the top five data mining techniques. In this paper, we consider each of the characteristics in our training set to be a distinct dimension in some space, and we use the value an observation has for each dimension as its coordinate in that dimension, thereby obtaining a set of points in space. The distance between two points in this space, as measured by an appropriate metric, can then be regarded as their similarity. The algorithm determines which points from the training set are sufficiently similar to be considered when selecting the class to predict for a new observation by selecting the k closest data points to the new observation and selecting the class with the highest frequency among these. This is why the algorithm is named k Nearest Neighbors.

The definition of Euclidean distance is the distance between two points. Because the Euclidean distance can be calculated using coordinate points and the Pythagorean Theorem, it is sometimes referred to as the Pythagorean distance.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

Where,
"d" is the Euclidean distance
(x_1, y_1) is the coordinate of the first point
(x_2, y_2) is the coordinate of the second point.

C. Linear Regression

Regression is a widely applicable empirical statistical technique of data extraction. Predicting one variable from another is the simplest definition of regression, also known as simple linear regression. The following is the basic linear regression model.

$$Y=\beta\_0+\beta\_1\ x\_1+\varepsilon\_i,\ i=1,2,\ldots,n. \quad (2)$$

In this equation, Y represents the value of a predicted variable, _0 is a constant value, _1 is the predictor coefficient, the slope of the regression line indicating how much Y varies for each unit change in X, x_1 are predictor variables that explain the value of Y, and is an error term. As the number of dependent variables increases, so does the precision of the predicted value. The estimated regression line or regression line of least squares is:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} x_1, i = 1,2, \ldots, n \quad (3)$$

where

$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$, is regression coefficient and
$\widehat{\beta_0} = \hat{y} - \widehat{\beta_1}\bar{x}$
$\bar{x} = \frac{1}{n}\sum x_i$
$S_{xx} = \sum x_i^2 - n\bar{x}^2$
$S_{yy} = \sum y_i^2 - n\bar{y}^2$
$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$, is simple convariance

V. EXPERIMENT

A comparison of the experimental outcomes of the employed classification algorithms is presented in the tables and figures that follow. Different classifiers produce distinct outcomes for identical data. A comparison of Error Values for Applied Models is shown in Table 1. Table 2 compares the linear regression algorithm r2_scores and the k-nearest neighbor algorithm for crop prediction. Using Mean Absolute Error (MAE), Mean Square Error (MSE), Median Absolute Error, Explained Variance Score, and R2 Score, the quality of agricultural yield prediction techniques utilizing KNN and Linear Regression was evaluated and compared.

Mean Absolute Error (MAE): The average of absolute deviations between the target and predicted values are calculated as the Mean Absolute Error (MAE). The computation of MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|predicted\ value - actual\ value| \quad (4)$$

Mean Squared Error (MSE): MSE (mean squared error) or MSD (mean squared deviation) are measurements of the averages of the square error values. Its mean squared difference between both the actual and forecast values. The computation of MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|predicted\ value - actual\ value| \quad (5)$$

Median Absolute Error: This is the median of all the errors in the given dataset. The main advantage of this metric is that it's robust to outliers. A single bad point in the test dataset wouldn't skew the entire error metric, as opposed to a mean error metric.

Explained Variance Score: This score measures how well our model can account for the variation in our dataset. A score of 1.0 indicates that our model is perfect.

R2 score: This is pronounced as R-squared, and this score refers to the coefficient of determination. This tells us how well the unknown samples will be predicted by our model. The best possible score is 1.0, but the score can be negative as well.

A. Experimental Results

In our experiment, the dataset is divided into two sets called training and testing. The training is 80% and the testing is 20%. The training step involves the use of K-nearest neighbor and Linear Regression algorithms on datasets in order to produce classifiers. After getting the classifier, the testing dataset is tested with these trained classifiers to compare the classifiers' performance. The experimental results of these two classifiers are shown in Table II.

TABLE II. DEPICTS THE VARIOUS ERROR VALUES THAT ARE OBTAINED WHEN DIFFERENT ALGORITHMS ARE APPLIED TO THE CROP PREDICTION

Model	Mean Square Error	Mean Absolute Error	Median Absolute Error	Explained Variance Score	R2 Score
LR	0.0	0.0	0.0	1.0	1.0
KNN	473318543.08	12040	5864.5	0.93	0.91

In Table II, the mean square error of Linear Regression is 0.0 while KNN is 473318543.08 which means that LR is better than KNN according to the value of mean square error.

Also according to the results of Mean Absolute Error and Median Absolute Error, LR is better than KNN for paddy crop yield prediction system. In Explain Variance Score and R2 Score, the performance of both classifiers is acceptable because KNN has over 0.9 and LR has 1. The graphical representation of our comparative study is also represented in Fig. 3.

In Figure 3, the column represents the value in the range of 0 to 1 for the results and the horizontal axis represents the name of the methods for labeling results for testing data. The mean square error values have huge difference between LR and KNN because LR is the regression-based methods and it can generate least error for testing errors. Although the LR has the best results of mean square, mean absolute and median absolutes error, it has not good result in explained variance score and R2 score. Finally, LR has the minimum errors value over all results of classification performance measure methods. According to figure 3, LR has the better classification performance than KNN.

TABLE III DEPICTS THE VARIOUS R2_SCORES OBTAINED WHEN DIFFERENT ALGORITHMS ARE APPLIED TO THE CROP YIELD PREDICTION

Models	R2 Score	Accuracy	Comments
Linear Regression	1.0	97%	Best fit
K-Nearest Neighbour	0.91	91%	Good fit

Fig. 3. Comparisons of Linear Regression and K-Nearest Neighbors according to mean square error, mean absolute error, medium absolute error, explained variance scores, and R2 score

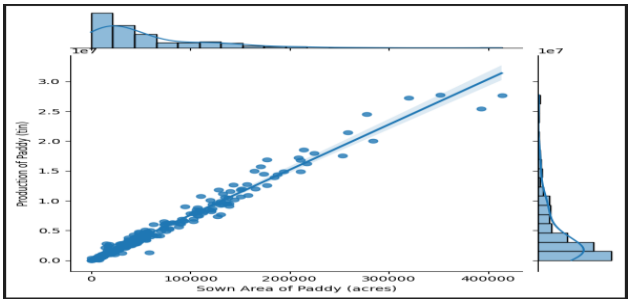

Fig. 4. Production of paddy (tin) value based on Sown Aera of paddy (acres)

In Table III, the accuracy and R2 score of Linear Regression is better than the K-Nearest Neighbour Algorithm. Our comparative study focuses not only on accuracy but also on error values to represent more accurate classification performance. According to all of the experimental results including accuracy, Linear Regression is chosen for the paddy crop yield prediction. The importance of crop yield prediction is highlighted in Fig. 4. In Fig. 4, the sown area of paddy (acres) is directly proportional to the target label called production of Paddy (tin).

VI. CONCLUSION

This paper demonstrates the effective use of KNN and LR algorithms for crop and yield prediction. Based on input values Township Code, Sown Area of Paddy (acres), Harvested Area of Paddy (acres), and Paddy as a percentage of All Harvested, crops and crop yield are predicted. In addition, the paper presents a comparative analysis of algorithms based on their accuracy and R2 scores and presents a suitable algorithm for predicting the crop and yield, respectively. Our comparative analysis is conducted on a real-world dataset (dataset from Eastern Shan State, Myanmar) and takes into account all of the essential performance evaluation factors. According to the experimental results, the accuracies of KNN and LR are 91% and 99%. Therefore, Linear Regression is more preferable for prediction because of its accuracy.

ACKNOWLEDGMENT

I would like to express my grateful thanks to Prof. Thinn Thu Naing, Rector, University of Computer Studies (Kyaing Tong) for her grade advices towards the successful, completion of this paper. I also thanks to Dr. Myint Myint Lwin, Associate Professor, University of Computer Studies (Kyaing Tong) who helps me with invaluable supporting.

REFERENCES

[1] Champaneri, Mayank, et al. "Crop yield prediction using machine learning." *Technology* 9.38 (2016).

[2] Elavarasan, Dhivya, and PM Durairaj Vincent. "Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications." *IEEE access* 8 (2020): 86886-86901.

[3] Hovey, Grosvenor Gilbert. "National Geographic." (2018).

[4] Nigam, Aruvansh, et al. "Crop yield prediction using machine learning algorithms." *2019 Fifth International Conference on Image Information Processing (ICIIP)*. IEEE, 2019.

[5] PS, Maya Gopal. "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms." Applied Artificial Intelligence 33.7 (2019): 621-642.

[6] Rashid, Mamunur, et al. "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction." *IEEE Access* 9 (2021): 63406-63439.

[7] https://www.ibm.com/topics/machine- learning

[8] https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/