INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 17-02-2023

Accepted: 26-04-2023

Published: 25-05-2023

Citation: Pakkala PGR, Rai BS, Rai P (2023) Statistical Driven Feature Selection for Prognostic Reasoning and Insight Exploration of Areca Nut Crop using Data Analytics Approach. Indian Journal of Science and Technology 16(20): 1495-1504. https://doi.org/10.17485/IJST/v16i20.2508

* **Corresponding author**.

rpakkala01@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Statistical Driven Feature Selection for Prognostic Reasoning and Insight Exploration of Areca Nut Crop using Data Analytics Approach

Permanki Guthu Rithesh Pakkala[1*], **Bellipady Shamantha Rai**[1], **Prakhyath Rai**[1]

1 Sahyadri College of Engineering & Management, Mangaluru, Karnataka and Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India

Abstract

Objectives: To develop a prognostic reasoning model for discovering insights about the areca nut crop and to recommend an optimal strategy for improving crop productivity and heightening the lifetime of the areca nut tree using a statistical feature selection technique namely Kolmogorov–Smirnov (KS) test and data analytics approach. **Methods:** Data for the study was gathered by distributing questionnaires to farmers cultivating the areca nut crop in the Mangaluru region. Farmers can plan ahead of time to improve crop yield and estimate the lifetime of the tree with this strategy. The Kolmogorov–Smirnov test is employed for the pre-processed data, and optimal features are selected. To forecast crop yield and tree lifetime, various classifiers namely decision tree, support vector machine (SVM), and artificial neural network (ANN) are applied to 300 test samples and their performance is evaluated. **Findings:** The findings of the experiment show that the decision tree works better than other classifiers for crop yield and tree lifetime with a prediction accuracy of 96 % and 94.66 % respectively. **Novelty:** The proposed study performs the extraction of significant features of the areca nut crop using the KS test that results in the prediction of agricultural production and forecasting tree lifetime.

Keywords: Arecanut Crop; Crop Yield; Tree Lifetime; Feature Selection; Kolmogorov-Smirnov Test; Data Analytics; Prognostic Reasoning Model

1 Introduction

Crop production estimation is an analytical technique for predicting agricultural yield prior to actual harvest[1]. To ensure farmers' year-round financial security and a stable level of living independent of market and climate uncertainties, it is essential to research the productivity of plantation crops[2]. For both attracting new farmers and keeping hold of present ones, models for managing agricultural risk must be developed[3,4]. Since the areca nut is the main source of income for farmers in southern India, it is crucial to research to obtain useful insights that will aid the farmers in making proactive decisions. In predicting areca nut crop yield, very few studies have been conducted[5].

Existing models can only predict crop yield through statistical analysis[6]. As a result, there is a huge need for systematic investigation and reporting.

In the proposed study, a feature selection approach and data analytics algorithms are used to present a prediction framework on the areca nut crop for scientific decision-making that supports agricultural forecasts. Farmers can gain insights of the cultivation strategy of areca nut crop using this framework well in advance. The proposed system is developed by considering various climatic and agronomic factors influencing areca nut crop to get insights of crop yield and tree lifetime. An actual data set gathered from farmers growing areca nut in the Mangaluru locale is used in this research to evaluate the correlation between variables and study the impact of these variables on crop yield.

Contributions: The following contributions to the field of agricultural crop yield prediction are made by this research work:

• First, in order to create a predictive model for crop yield and tree lifetime prediction using data analytics, the most recent datasets on the areca nut crop have been created.

• Second, the feature importance problem has been explored by comparing the different feature selection techniques and evaluating the performance that which the feature selection technique has given effective results with different classifiers.

• Finally, limited work was done on the areca nut crop and thus exploring more insights in terms of crop yield and tree lifetime of the Arecanut crop.

Data preparation, feature selection, classification, and predictive analysis are the key principles in data analytics that are used in this study. Only essential components are determined through filtering after preprocessing raw agricultural data. The best traits are found using the statistical test known as the Kolmogorov-Smirnov (KS) test to discover the knowledge about the crop[7]. In this study, predictive analysis of crop yield and tree lifetime is done with the use of classifiers namely decision tree, support vector machine (SVM), and artificial neural network (ANN). Based on classifier accuracy metrics, the performance of various methods is compared. Finally, the most comprehensive information about the plantation strategy for improving crop yield and achieving maximum tree lifetime is identified, analyzed, and reported.

Meng et al.[8] aimed to estimate the yield of maize crop by combining the data of climate, satellite, fertilizer, and soil from multiple sources and evaluate the efficiency of these input variables to yield forecast. Findings of the study demonstrated that employing random forests and adaptive boosting to combine all of the datasets can improve yield prediction results with an R^2 of above 0.87. It was discovered that knowing about the fertilizer does not significantly improve forecasts, and the accuracy varies depending on the system. Additionally, the research region's locality is limited.

Kuradusenge et al.[9] employed machine learning techniques and considered historical weather and yield data for predicting the yields of Irish potatoes and maize. Support vector regression, polynomial regression, and random forest were used to examine the data that had been obtained. Temperature and rainfall were utilized as forecasters. A root mean square error (RMSE) of 510.8 is achieved for the potato and 129.9 for the maze. Also, R^2 of 0.875 and 0.817 is obtained for the same agricultural datasets. The findings show that Random Forest is the best model but the study is limited by only two variables.

The goal of Krithika et al.[10] was to find the most effective machine-learning model to forecast the yield of the groundnut crop. Only four variables—rainfall, area, irrigation, and production—were taken into account in the study. The experiment's results demonstrated that the Least Absolute Shrinkage and Selection Operator (LASSO) and ElasticNet delivered the best outcome with the lowest Root mean square error (RMSE) and Relative Root Mean Squared Error (RRMSE) values. Through the process of feature selection, the ideal timing for sowing and watering of a region is also determined. The study solely relates to Tamil Nadu and is data-specific. As a result, when it comes to the various regions of the country, the dataset pattern and accuracy vary.

A model for forecasting the yields of six crops at the country level in West Africa was put forth by Cedric et al.[11]. K-nearest neighbor models, multivariate logistic regression, and decision trees were all employed in the analysis of the study. To improve the model, they used a hyperparameter tuning strategy during cross-validation. According to the findings, crop k-Nearest Neighbor (Ck-NN) outperforms Crop Decision Tree (CDT) and Crop Multivariate Logistic Regression (CMLR) with an R^2 score of 95.03% and a Mean Absolute Error (MAE) of 0.160 kg/ha. The proposed prediction models can be applied to the whole of West Africa, and other factors can be included to boost the model's quality.

Based on actual productivity and weather data, Liu et al.[12] created a unique crop harvest time prediction model that integrates feature selection and long short-term memory techniques to forecast harvest times accurately and minimize resource waste for improved sustainability. In comparison to long short-term memory (LSTM) and recurrent neural networks (RNN), the developed model long short-term memory with feature selection (LSTMFS) demonstrated superior accuracy for forecasting harvest time with 0.199 root mean square error (RMSE) and 4.84% mean absolute percentage error (MAPE). The study does not use any data-gathering sensors and is only concerned with actual production and climatic variables.

For the areca nut crop, Krishna et al.[13] generated a dataset of weather characteristics and a disease prediction algorithm. Several regression models, including support vector (SV), decision tree (DT), random forest (RF), and multilayer perception

(MLP), are used to validate the dataset. It is discovered that random forest regression (RFR) predicts the incidence of areca nut fruit rot disease with a very high degree of accuracy and a minimal error rate of 0.9. It has been noted that the error rate has been impacted by the elimination of features, but it can still be reduced.

The characteristic comparison of existing methods with the proposed method is shown in Table 1.

Table 1. Characteristic Comparison of the Existing Methods

Characteristics	**Meng et al.** [8]	**Kuradusenge et al.** [9]	**Krithika et al.** [10]	**Cedric et al.** [11]	**Liu et al.** [12]	**Krishna et al.** [13]	**Proposed Method**
Dataset Size	Moderate	Small	Small	Moderate	Moderate	Small	Moderate
Variable Selection	Intrinsic (automatic) feature selection	Intrinsic (automatic) feature selection	Intrinsic (automatic) feature selection	Intrinsic (automatic) feature selection	Intrinsic (automatic) feature selection	Intrinsic (automatic) feature selection	Variables are selected based on statistical score.
Study Region	Locally Limited	Limited by only two variables	The study is limited to Tamil Nadu	The study is limited to the whole of West Africa	The study is concerned with only production and climate variables	The study is focused on weather characteristics.	The study is concentrated in the Mangaluru region and considered various agronomic and meteorological variables.
Error Rate	Moderate	Moderate	Moderate	Moderate	Minimal	Minimal	Minimal
Computational Cost	Moderate	Moderate	Expensive	Moderate	Expensive	Moderate	Moderate
Performance	Relatively fast	Relatively fast	Moderate	Relatively fast	Moderate	Relatively fast	Fast and effective

2 Methodology

2.1 Proposed Model

The proposed predictive framework entails discovering optimal combinations of features to obtain high crop yield and tree lifetime of areca nut with the aid of the KS test and data analytics approach, which is shown in Figure 1.

The proposed insight exploration model employs the data analytics tasks namely:

- Data Collection
- Data Preprocessing
- Feature Selection
- Classification
- Knowledge Discovery

2.2 Data Collection

The research study is focused on the Mangaluru area and a real data set was created through interactions with farmers who were growing an areca nut crop. The key features that were the subject of the questionnaires were area and type of land, variety of crop, type of plantation, sunshine rate, Information on shade facility, availability of water, type of irrigation, usage, and type of fertilizer, usage of pesticide and drainage facility[14].

The agronomic features of the areca nut crop[15] considered for the study are listed in Table 2.

2.3 Data Pre-processing

The collected data is cleaned in a pre-processing stage after data collection since it is noisy, fragmented, and irrelevant, which leads to poor prediction accuracy. For category attributes, the null or missing values are filled up with the most common attribute values. Null or missing values of numerical attributes are filled using weighted average values[16].

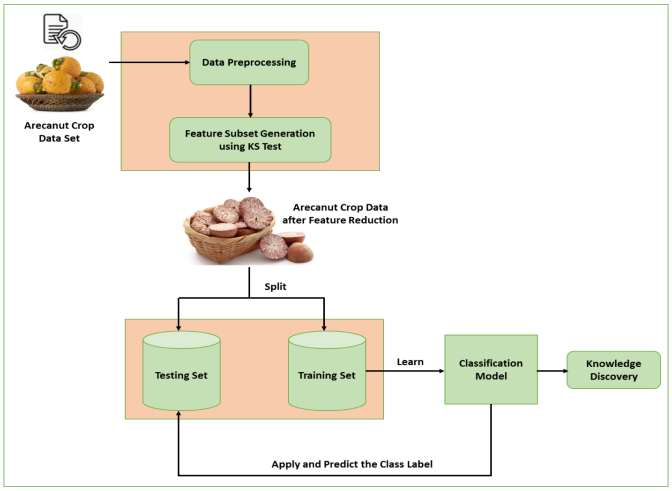

Arecanut Crop Data Set
Data Preprocessing
Feature Subset Generation using KS Test
Arecanut Crop Data after Feature Reduction
Split
Testing Set
Training Set
Learn
Classification Model
Knowledge Discovery
Apply and Predict the Class Label

Fig 1. Proposed Model for Insight Exploration of Areca nut Crop

Table 2. Influencing Features of Arecanut Crop

Sl. No.	Attribute/Feature	Values	Description
1	Land_Area	{5, 7, 10,20,30,40 etc...}	Specification of land area in cents
2	Type of Land	{Dry, Wet}	Nature of Land
3	Variety of Crop	{Sygon, Mohitnagar, Mangala, Inter Mangala, Local}	Type of Arecanut
4	Spacing between palms	{2.75,3}	The preferred gap between the palms in meters
5	No. of Palms Per Cents	{5,6}	Fitting palms as per spacing
6	Type of Plantation	{Deep, Shallow}	Nature of plantation
7	Rate of Sunlight	{High, Moderate, Low}	Amount of sunlight
8	Shade Facility	{Yes, No}	Provision of Shading
9	Proper usage of irrigation	{Yes, No}	Proper supply of water
10	Type of Irrigation	{Sprinkler, Drip, Flood}	Nature of irrigation
11	Proper Drainage Facility	{Yes, No}	Proper management of drainage for water flow
12	Proper usage of fertilizer	{Yes, No}	Usage of fertilizer
13	Type of Fertilizer	{Organic, Chemical, Mixed}	Nature of fertilizer used
14	Proper usage of pesticide	{Yes, No}	Usage of pesticides for pest control
15	Yield Per Acre	{Good, Average, Not Good}	If productivity > 2000 kg, then Good, if productivity is between 1200 and 2000, then average, otherwise not good
16	Tree Lifetime	{High, Low}	If the lifetime is between 40 and 55 years, then high otherwise low

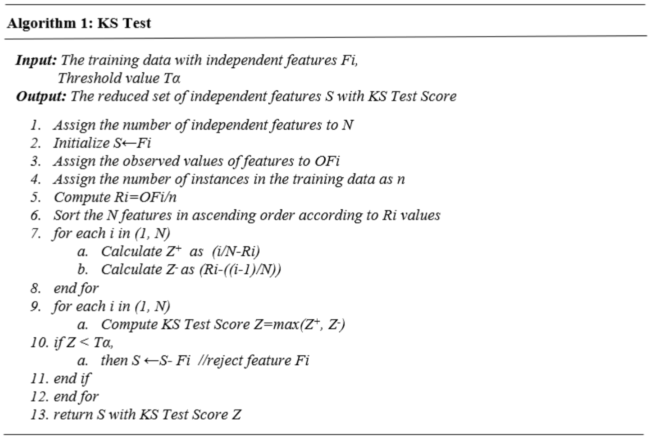
Each value x_i in a set can have a corresponding weight w_i for i=1, 2,..., N. The weights indicate the significance, relevance, or frequency of occurrence attached to each value. Here, the weighted average $\bar{x}$ is calculated as

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N} \tag{1}$$

Additionally, further analysis may be conducted during the filtering to obtain accurate information. For the feature selection process, highly enriched data is taken into account.

2.4 Feature Selection

The appropriate features are chosen from the list of features using the Kolmogorov-Smirnov Test once the preprocessing task yields clean data. This statistical test identifies attributes that contribute more to the mining process and thus reduces the irrelevant attributes[17]. The KS test algorithm is presented in as follows.



Algorithm 1: KS Test

```
Input: The training data with independent features Fi,
       Threshold value Tα
Output: The reduced set of independent features S with KS Test Score

Assign the number of independent features to N
Initialize S←Fi
Assign the observed values of features to OFi
Assign the number of instances in the training data as n
Compute Ri=OFi/n
Sort the N features in ascending order according to Ri values
for each i in (1, N)
    Calculate $Z^+$ as (i/N-Ri)
    Calculate $Z^-$ as (Ri-((i-1)/N))
end for
for each i in (1, N)
    Compute KS Test Score Z=max($Z^+$, $Z^-$)
if Z < Tα,
    then S ←S- Fi  //reject feature Fi
end if
end for
return S with KS Test Score Z
```

The KS test algorithm proceeds by taking the training data with independent features as input. Also, the threshold value Tα will consider as the requisite for the algorithm. Initially, the number of independent features is assigned to the variable N. Also, the reduced feature set S is initialized with all the independent features in the beginning. The observed values for each feature with all possible feature labels against the values will be determined. The obtained observed values will be divided by the number of instances in the training data and values will be ordered from smallest to largest. After ranking, for each feature, KS Test score Z will be calculated and compared with threshold value Tα. If the value of Z is less than the Tα, then that feature will be removed from S and the process continues for all the features. Finally, the reduced feature set will be returned with its score.

2.5 Classification

Following the feature selection, the predictive analysis entails classifying areca nut crop data and acquiring insights about plantations. Depending on the values of the attribute and the objectives of the study, a classification technique organizes the values of the data into classes with specific labels[18].

In the initial phase of the classification, data used for the training are evaluated, and a model for the classification is built. In the later stage, the classification accuracy is measured by giving the test data. The classification algorithms SVM, decision tree, and ANN are used in this research study to analyze the data. The accuracy of each classifier's predictions is used to test its results.

2.5.1 Support Vector Machine (SVM)

The equation for kernel in SVM is given by:

$$f(X1,X2) = \exp(-\text{ gamma } * \|X1.X2\| \wedge 2) \tag{2}$$

Here, gamma describes the influence a single training point has on nearby data points. f(X1, X2) provides the polynomial decision boundary that will split the data. The dot product between features is ||X1. X2||.

2.5.2 Decision Tree

Entropy, which is represented by the symbol H(S) for a finite collection S of data, is a measure of how random or uncertain data are and given by:

$$H(S) = \sum_{x \in X} p(x)\, log_2 \frac{1}{p(x)} \tag{3}$$

For a set, information gain is denoted by IG(S, A). S represents the actual change in entropy after choosing a specific attribute A. In relation to the independent variables, it quantifies the relative change in entropy.

$$IG(S,\,A) = H(S) - \textstyle\sum_{i=0}^{n} p(x) * H(x) \tag{4}$$

where IG(S, A) denotes the information acquired from applying feature A. H(S) is the total set's entropy, and the second term determines the entropy after employing the attribute A, where p(x) is the probability of the event x.

2.5.3 Artificial Neural Network (ANN)

The sigmoid function, which is employed in ANN as the activation function, is given by:

$$F(x) = \frac{1}{1+e^{-sum}} \tag{5}$$

where

$$sum = \textstyle\sum_{i=1}^{n} x_i W_i \tag{6}$$

Here $x_1,\ldots,x_n$ are the inputs to the neuron and $W_{0,}\ldots,W_n$ are the weights.

2.6 Knowledge Discovery

The information for increasing crop yield and areca nut tree lifetime is finally found, processed, and sent to the farmer in an appropriate form based on classification and predictive analysis. With this information, the farmer can make a proactive decision regarding plantation, and several feature combinations are identified that will result in a high yield and a long tree lifetime.

The suggested model achieves its superiority in forecasting areca nut crop yield and tree lifetime by identifying the best features using KS Test, which helps in precise classification with the highest accuracy. When compared to the other existing approaches, the KS Test automatically separates features from different distributions. Also, the suggested method chooses the most important elements that are more valuable for classification than the other methods already in use.

3 Results and Discussion

3.1 Quantitative Analysis

The study is based on a data set created via questionnaires distributed to various farmers. The most impacting features for the classification are determined using the KS test.

Initially, the KS Test score will be calculated for all independent features. After calculation, it will be compared with the threshold value. Here the threshold value is set to 0.5 and thus 11 features are selected among 14 independent features which are shown in Table 3.

Table 3. Optimal Feature Selection using KS Test

Sl. No.	Independent Feature	KS Test Score
1	Land Type	0.6066
2	Crop_Variety	0.8875
3	Palm Spacing	0.5354
4	Plantation_Type	0.6456
5	Sunlight Rate	0.6614
6	Proper Irrigation Use	0.6403
7	Irrigation_Type	0.7819
8	Proper Drainage Facility	0.7683
9	Proper Fertilizer Use	0.7546
10	Fertilizer Type	0.7758
11	Proper Pesticide Use	0.7412

Following feature selection, the holdout approach is used to divide the dataset with the best features into training and test sets. The experiment is carried out on a total of 1500 instances, out of which training instances are 70 % and testing instances are 30 %. The effectiveness of the induced model on the test set can be used to assess the classifier's accuracy.

The suitable class for the crop yield (high, average, or low) is determined using the classification method. Similarly, the appropriate class for a lifetime (lifetime = high) or (lifetime = low) is identified.

The prediction accuracy of crop yield and the tree lifetime of areca nut is measured using SVM, decision tree, and ANN model. Table 4 shows the different statistical parameters used for evaluating different classifiers for the crop yield and tree lifetime. The decision tree classifier results in a 96 % prediction accuracy for the crop yield and 94.66 % for the tree lifetime.

Table 3 displays the statistical measures of different classifiers for crop yield and tree lifetime. The prediction accuracy of the decision tree classifier is 96% for crop yield and 94.66% for tree lifetime.

Table 4. The Statistical Measures of Different Classification Algorithms for the Crop Yield and Tree Lifetime

Particulars	**Decision Tree**		**SVM**		**ANN**	
	Crop Yield	Tree Lifetime	Crop Yield	Tree Lifetime	Crop Yield	Tree Lifetime
Correctly Classified Instances	288	284	268	264	282	276
Incorrectly Classified Instances	12	16	32	36	18	24
Kappa Statistic	0.9324	0.8892	0.8191	0.7527	0.8977	0.8352
MAE	0.06	0.0533	0.1067	0.1200	0.0833	0.0800
RMSE	0.3162	0.2309	0.4082	0.3464	0.3606	0.2828
RAE	0.0952	0.1111	0.2011	0.2500	0.1323	0.1667
RRSE	0.4503	0.4714	0.5814	0.7071	0.5135	0.5774
Test Instances	300	300	300	300	300	300

Here MAE, RMSE, RAE, and RRSE denote the Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, and Root Relative Squared Error respectively

3.2 Performance Analysis

In the proposed research, micro-averaged measures are considered for evaluating the performance of different classifiers. The recall score is determined by dividing the number of occurrences that were correctly predicted by the total number of examples in the class. Specificity is defined as the proportion of correctly detected negatives to the overall number. The proportion of accurately predicted occurrences to total expected instances is used to determine the precision score. By averaging the recall and precision scores, the F-Measure is calculated.

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

$$Specificity = \frac{TN}{TN+FP} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$F-Measure\ (F_1) = \frac{2*Recall*Precision}{Recall+Precision} \tag{10}$$

Here TP, TN, FP, and FN denote the True Positive, True Negative, False Positive, and False Negative respectively. A true positive result is one for which the model accurately identified the positive class. A true negative is an outcome that the model accurately predicted belongs to the negative class. An outcome when the model predicts the positive class inaccurately is known as a false positive. A false negative is a result that the model mispredicts as belonging to the negative class.

3.3 Comparative Analysis

The primary goal of feature selection algorithms is to improve classification precision. If the KS Test method eliminates the undesirable features that demonstrated difficulty during the classification, then the improvement was carried out during the classification. The proposed KS Test feature selection method has shown improvement in the classification that trains the model more quickly. Without any prior knowledge of the features, the feature selection algorithm identified the optimal subset. Table 5 and Table 6 show the results obtained for different classifiers using specific performance metrics with and without feature selection for the crop yield and tree lifetime respectively.

The training data samples lead to under performances in the SVM model when the number of features is exceeded. When the training process was finished, ANN decreased the sample's specific error value. The decision tree achieved good accuracy since it excludes all non-essential features during the training process. Because of the complexity that is formed among the features, results are only moderately achieved without the feature selection process. Thus decision tree achieved a 93.12% accuracy, 0.9012 of specificity, 0.9138 of precision, 0.8924 of recall, and 0.9029 of F1-score for the crop yield without feature selection. Similarly, 89.23% accuracy, 0.8712 of specificity, 0.8803 of precision, 0.8678 of recall, and 0.8740 of F1-score for the tree lifetime without feature selection.

Table 5. Comparative Analysis of Different Classifiers With and Without Feature Selection for Crop Yield

Classifier	Without Feature Selection					With Feature Selection				
	Accuracy (%)	Specificity	Precision	Recall	F1-Score	Accuracy (%)	Specificity	Precision	Recall	F1-Score
Decision Tree	93.12	0.9012	0.9138	0.8924	0.9029	96.00	0.9478	0.9549	0.9283	0.9401
SVM	66.58	0.6138	0.6323	0.6097	0.6207	89.33	0.8794	0.8935	0.8523	0.8690
ANN	90.08	0.8823	0.8918	0.8526	0.8717	94.00	0.9012	0.9394	0.8928	0.9123

Table 6. Comparative Analysis of Different Classifiers With and Without Feature Selection for Tree Lifetime

Classifier	Without Feature Selection					With Feature Selection				
	Accuracy (%)	Specificity	Precision	Recall	F1-Score	Accuracy (%)	Specificity	Precision	Recall	F1-Score
Decision Tree	89.23	0.8712	0.8803	0.8678	0.8740	94.66	0.9496	0.9435	0.9458	0.9446
SVM	76.48	0.7745	0.7621	0.7814	0.7716	88.00	0.8703	0.8734	0.8523	0.8763
ANN	85.78	0.8216	0.8345	0.8198	0.8270	92.00	0.9178	0.9142	0.9222	0.9175

Table 7. Comparative Analysis of Proposed Method with Referred Existing Methods for the Crop Yield

Methods/Model	Accuracy (%)	R2	RMSE (kg/ha)	MAE (kg/ha)	RRMSE (%)	RMAE (%)
Random forests and Adaptive Boosting [7]	91.50	0.85~0.98	<1000	-	-	-
Random Forest Regressor [8]	-	0.875 – Potato 0.817 - Maize	510.8 – Potato 129.9 - Maize	418.699 - Potato 96.196 - Maize	-	-
LASSO [9]	81.08	0.549	491.603	333.154	20.68	14.02
ElasticNet [9]	81.34	0.550	490.931	331.827	20.66	14.00
Ck-NN + Hyper-Parameter Tuning [10]	94.86	0.9503	-	0.160	-	-
Proposed Method	96.00	0.9602	0.3162	0.06	-	-

The proposed method and the referred existing methods are compared in Table 7 and evaluated in terms of accuracy, R^2, RMSE, MAE, RRMSE, and RMAE.

Higher prediction performance is achieved by the model as R^2 approaches 1. The difference between the measured yield and the predicted yield is less when the RMSE is small. The R^2 value of the proposed method is very closer to 1 when compares to the R^2 value of other existing methods used in Table 5. Thus proposed method achieved the highest prediction accuracy of 96 % with an RMSE value of 0.3162 and MAE value of 0.06 for the crop yield than other referred existing methods. The R^2 value of the Ck-NN and Hyper-Parameter Tuning method is relatively closer to 1 and thus achieved a 94.86 % of prediction accuracy.

4 Conclusion

The precise extraction of unique knowledge involving a strategy for planting the areca nut crop is performed in this research. This knowledge involves applying the KS Test to identify the most important crop-related variables that affect areca nut yield and tree lifetime. The performance of different classification algorithms, namely SVM, decision tree, and ANN, are evaluated through the accuracy of prediction and different statistical metrics. The experimental results demonstrate that the decision tree achieved a superior over other classification algorithms on the areca nut crop information in terms of crop yield and tree lifetime. The decision tree performs a prediction of crop yield with 96 % accuracy, 0.9549 of precision, and 0.9283 of recall. Also, the prediction of tree lifetime is done with 94.66 % accuracy, 0.9435 of precision, and 0.9458 of recall using the decision tree.

The suggested model includes an exact test for identifying influencing features, so the feature selection method used in the proposed model does not depend on a sufficient sample size for approximations to be accurate. In the proposed study only 300 test samples were considered. Although this proposed model has strong generalizability, the test data were constrained by the study region. Thus, the model's ability for generalization requires further improvement.

In future work, some hypothesis testing can be used to prove the developed models using Data Analytics Approach are significantly the same or different. Also, even more, data samples can be employed to validate the model that has been provided, and other data collection sensors can be used to gather more distinct features that affect the growth of the areca nut crop, which will help to increase the model's prediction accuracy. Other significant insights, such as areca nut grading, market pricing, and disease classification, can also be explored in addition to crop yield and tree lifetime.

References

1) Van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: A systematic literature review. 2020. Available from: https://doi.org/10.1016/j.compag.2020.105709.
2) Abbas F, Afzaal H, Farooque AA, Tang S. Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. *Agronomy*. 2020;10(7):1046. Available from: https://doi.org/10.3390/agronomy10071046.
3) Paudel D, Boogaard H, De Wit A, Janssen S, Osinga S, Pylianidis C, et al. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*. 2021;187:103016. Available from: https://doi.org/10.1016/j.agsy.2020.103016.
4) Nevavuori P, Narra N, Lipping T. Crop yield prediction with deep convolutional neural networks. 2019. Available from: https://doi.org/10.1016/j.compag.2019.104859.
5) Jin Y, Guo J, Ye H, Zhao J, Huang W, Cui B. Extraction of Arecanut Planting Distribution Based on the Feature Space Optimization of PlanetScope Imagery. *Agriculture*. 2021;11(4):371–371. Available from: https://doi.org/10.3390/agriculture11040371.
6) Pant J, Pant RP, Singh MK, Singh DP, Pant H. Analysis of agricultural crop yield prediction using statistical techniques of machine learning. *Materials Today: Proceedings*. 2021;46:10922–10926. Available from: https://doi.org/10.1016/j.matpr.2021.01.948.
7) Cardoso DO, Galeno TD. Online evaluation of the Kolmogorov–Smirnov test on arbitrarily large samples. *Journal of Computational Science*. 2023;67:101959. Available from: https://doi.org/10.1016/j.jocs.2023.101959.
8) Meng L, Liu H, Ustin SL, Zhang X. Predicting Maize Yield at the Plot Scale of Different Fertilizer Systems by Multi-Source Data and Machine Learning Methods. *Remote Sensing*. 2021;13(18):3760. Available from: https://doi.org/10.3390/rs13183760.
9) Kuradusenge M, Hitimana E, Hanyurwimfura D, Rukundo P, Mtonga K, Mukasine A, et al. Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture*. 2023;13(1):225. Available from: https://doi.org/10.3390/agriculture13010225.
10) Krithika KM, Maheswari N, Sivagami M. Models for feature selection and efficient crop yield prediction in the groundnut production. *Research in Agricultural Engineering*. 2022;68(3):131–141. Available from: https://doi.org/10.17221/15/2021-RAE.
11) Cedric LS, Adoni WYH, Aworka R, Zoueu JT, Mutombo FK, Krichen M, et al. Crops yield prediction based on machine learning models: Case of West African countries. *Smart Agricultural Technology*. 2022;2:100049. Available from: https://doi.org/10.1016/j.atech.2022.100049.
12) Liu SC, Jian QY, Wen HY, Chung CH. A Crop Harvest Time Prediction Model for Better Sustainability, Integrating Feature Selection and Artificial Intelligence Methods. *Sustainability*. 2022;14(21):14101. Available from: https://doi.org/10.3390/su142114101.
13) Krishna R, V PK, Gaonkar R. Areca nut disease dataset creation and validation using machine learning techniques based on weather parameters. *Engineered Science*. 2022;19:205–219. Available from: https://doi.org/10.30919/es8d712.
14) Pakkala PR, Rai BS. A Prognostic Reasoning Model for Improving Areacanut Crop Productivity using Data Analytics Approach. *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. 2022;p. 1–6. Available from: https://doi.org/10.1109/DISCOVER55800.2022.9974945.
15) Guthu RP, Bellipady SR. A Formal Statistical Data Modeling for Knowledge Discovery and Prognostic Reasoning of Arecanut Crop using Data Analytics. *International Journal of Software Science and Computational Intelligence (IJSSCI)*. 2022;14(1):1–27. Available from: https://doi.org/10.4018/IJSSCI.311447.
16) Pakkala R, Rai P, Bellipady SR. Impact of Syntactical and Statistical Pattern Recognition on Prognostic Reasoning. *Handbook of Research on Machine Learning Techniques for Pattern Recognition and Information Security 2021*;p. 38–55. Available from: https://doi.org/10.4018/978-1-7998-3299-7.ch003.
17) Raja SP, Sawicka B, Stamenkovic Z, Mariammal G. Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers. *IEEE Access*. 2022;10:23625–23641. Available from: https://doi.org/10.1109/ACCESS.2022.3154350.

18) Batool D, Shahbaz M, Asif HS, Shaukat K, Alam TM, Hameed IA, et al. A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning. *Plants*. 1925;11(15):1925–1925. Available from: https://doi.org/10.3390/plants11151925.