A Rule Based Recommender System to Improve the Yield of Groundnut Crop Using Decision Tree with Backward Elimination, Principal Component Analysis

S. Muthukumaran[a], P.Geetha[b], E.Ramaraj[c]

[a]Research Scholar, Department of Computer Science, Alagappa University, Karaikudi, India.
Email:muthumphil11@gmail.com
[b]Associate Professor, Department of Computer Science, Dr.UmayalRamanathan College for Women, Karaikudi, India.
Email:geeth.ganesan@gmail.com
[c]Professor and Head, Department of Computer Science, AlagappaUniveristy, Karaikudi, India.
Email:eramaraj@rediffmail.com

Abstract: Precision farming is a new technological revolution that is presently happening in the agriculture field which aims at making an individual farmer to produce food for 155 people. Practicing precision farming push farmers to adopt new cropping technology that uses advanced software and IOT sensor equipped devices. In precision farming data is generated from various sensors in huge volume which requires special storage mechanism and machine learning algorithms to analyse the data. Decision Tree built using high dimension datasets takes more time to construct the tree, requires huge memory space and also produce complex rules. This paper proposed a machine learning model that has two decision tree algorithms namely **Decision Tree with Principal Component Analysis (DTPCA)** and **Decision Tree with Backward Elimination (DTBE)** that combines the Decision Tree Algorithm with feature selection techniques such as Principal Component Analysis (PCA), and Backward Elimination. This proposed framework aims to develop a rule based recommender system that will assist farmers in decision making and helps to improve the yield of Groundnut. The proposed algorithm was tested with a real time dataset that contains factors responsible for the growth of groundnut yield. The results showed that the Decision Tree Algorithm combined with Principal Component Analysis performs better and classify the dataset with higher accuracy and low error rate.

Keywords: Precision Farming, Machine Learning, Decision Tree, Feature Selection

1. Introduction

Agricultural sector is one of the major contributor to the Indian Economy by providing employment opportunity to two third of the overall population of India. The food products obtained through agriculture serves as the major raw materials for industries such as sugar industry, cotton industry and jute industry etc. Agricultural products produced in India are exported to foreign countries and 9% of the international trade of our country depends on Agricultural Sector[1]. Groundnut is one of the major oil crops which are cultivated all over the world since from the origin of humanity. In India groundnut is cultivated in 6.9 million hectares of land and produce a total capacity of 5.3 million tons of groundnut per year. The south Indian states like Tamilnadu, Gujarat and Andhra play a major role in the production of groundnut. The roots of the groundnut enrich the soil nutrition and also the stem and leaves are major food products for cattle. The cake produced from groundnut is a major food snack used by the rural people. Due to the nutty flavour and good taste in cooking, the oil produced from groundnut is in high demand. However the growth of groundnut is decreased due to lack in technical knowledge of crop management strategies among farmers, lack of disease tolerant varieties of seed, failure of pest control measures taken to control the pests, changes in the environment and global warming.

In the modern era, a new revolution in Agriculture sector is on-going by replacing the traditional crop production methods to smart farming with the help of Internet of things and Artificial Intelligence[2]. Use of sensors to monitor the environmental factors such as temperature, moisture content in air, relative humidity and water content in root zones of the cultivated plant helps to monitor the crop accurately in real time and helps to increase the yield. Analyzing the data collected through the IOT sensors with the help of machine learning techniques helps to forecast the factors responsible for crop yield. The use of sensors in collecting the real world data will result in producing a huge volume of data that needs a special mechanism to analyseit. Recording daily data with the help of IOT sensors will let to use of more attributes to construct the real time dataset. Applying machine learning algorithms to dataset having higher dimensions will reduce the performance of the algorithms and also it is very complex to recognize the patterns present in the dataset. To overcome the above mentioned difficulties, this research contributes the following:

- This paper proposed a Machine Learning Framework that has two decision tree algorithms namely **Decision Tree with Principal Component Analysis (DTPCA)** and **Decision Tree with Backward Elimination (DTBE)** that combines the Decision Tree Algorithm with feature selection techniques such as Principal Component Analysis (PCA), and Backward Elimination.
- The impact of Feature Selection Technique in analyzing the real time dataset was analyzed using a real time dataset that contains the factors responsible for groundnut growth and Meteorological factors such as Relative Humidity, Maximum and Minimum Temperature, Overall distribution of rainfall etc.
- The proposed machine learning framework aims to design a rule based recommender system that assist the farmers to take decisions recording the production of groundnut crop to improve the yield.

2. Literature Review

Anguraj.K et al[3] presented a paper and in it they collects the soil moisture, pH and humidity with the help of the IOT sensors and formed a database. Feature selection technique such as feature scaling method was applied to the input variables. The selected parameters were given as input to the Random Forest classifier and Naïve Bayes algorithms. The model analyzed the given dataset and then suggests the crop suitable for cultivation as a recommendation to the farmers. Vishal Nathgosavi et al[4] presented a paper and in it they presented a detailed survey of various researchers who use machine learning algorithms such as (SVM,ANN and Decision Tree) to analyze the agricultural data and suggest various recommendation to improve the yield of crops such as coffee, cherry, wheat and tomatoes. Samuel Asante Gyamerah et al[5] presented a paper and in it they proposed Quantile Random Forest model (QRF) that uses a Epanechnikov Kernel function to analyze the non-linear relationship between the meteorological parameters and the crop yield. The author uses groundnut and millet crops and identifies the features responsible for forecasting the crop yield. The performance of the QRF model was analyzed with the performance metrics such as RMSE, MAPE, R-Square and Bias. HaidiRao et al[6] presented a paper and in it they proposed a feature selection framework that uses bee colony and gradient boosting techniques to reduce the dimensionality of the datasets that was applied on real time. The author uses 8 datasets collected from public repository and applied the feature selection techniques. The authors proved that the accuracy of the classifiers is improved by using the feature selection techniques.

Xin Zhang et al[7] presented a paper and in it they used KNN to analyze canopy datasets that was used to identify the apple that are either mechanically harvested or mechanically unharvest. The author uses PCA to reduce the dimensions of the canopy dataset and apply KNN with 30 distance measures such as Spearman, Hamming, Jaccard and Minkowskietc and predict the harvesting type of the apple. K.Suganya Devi et al[8] presented a paper and in it they used KNN with the combination of image processing techniques such as binary mask, HSV segmentation and proposed a disease prediction system for groundnut crop. The pixel of the images were extracted and analyzed with KNN and the types of disease infected leaves were identified for groundnut crop. Zhenbo Li et al[9] presented a paper and in it they used three machine learning algorithms namely CNN, SVM, HoG and Hu invariant algorithms to classify the images of groundnut having one, two and three peanuts. The author proved that among the three algorithms used, SVM classifies the images better than the remaining algorithms. Vinita Shah et al[10] presented a paper and in it they used KNN, ANN and Linear Regression to predict the yield of groundnut crop. The authors collected the data that contains weather parameters and yield parameters having 20 attributes belonging to the time period 2006-2013. The role of the input attributes on influencing the yield of groundnut crop was analyzed with the four machine learning algorithms and the results were compared.

3. Theoretical Background

3.1 Principal Component Analysis (PCA)

While concentrating on designing an intelligent computer system that can learn things like humans from the real world applications. The major difficulties are the redundancy and insignificancy of data that will decrease the learning capacity of the classifier. The important factor that improves the learning capacity of the classifier is the input features provided to the algorithms. When the input variables given to a classifier are efficient and low then the machine learning algorithms performs better and predicts the output effectively. Principal Component Analysis (PCA) is one of the dimensionality reduction techniques that removes the inconsistencies present in the dataset and select the highly correlated features from the dataset. PCA first converts all input variables to a standard form and then compute the covariance matrix for all input variables. Eigen Vectors and Eigen values are calculated form the covariance matrix for all input variables and the principal components are selected. The principal components are the new set of variables that are derived from the original set of input variables. The variables that are most influencing in the dataset is ordered by numbers such as PCA1, PCA2,..etc. The covariance matrix for the input variables in the dataset is calculated using the formula

$$s_{ij} = \frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_j)(x_{jk}-\bar{x}_j)}{n-1} \qquad \text{Equation (1)}$$

The equation for Eigen vector for n*n matrix is given by

$$\Sigma Y = \beta^T \Sigma \beta \qquad \text{Equation (2)}$$

Here β is the Eigen value and β^T is the Eigen vector. The equation for constructing the principal components in a dataset is given by the following equation.

$$Y = W_1 * PC_1 + W_2 * PC_2 + \ldots + W_{10} * PC_{10} + C \qquad \text{Equation (3)}$$

3.2 Backward Elimination

Artificial Intelligence aims to design computer programs to think like humans. Backward Elimination is introduced to design machine learning algorithms in smarter way that it will identify the most important features in the dataset and ignore the least significant attributes automatically without any human intervention. Backward Elimination improves the precision and time complexity of the machine learning algorithms by setting a significance level. The independent attributes are correlated with the dependent attributes and the model computes the slope and intercept for the regression line. The dependent variable which is out of the regression is removed from the attribute list and process continued until all the independent attributes are within the significance level. Let us consider Y as the dependent variable and the line of intercept isβ_0, and then the linear regression line for plotting the input variable X_i is framed by the following equation

$$Y=\beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \qquad \text{Equation (4)}$$

3.3 Decision Tree Induction Algorithm

A Decision Tree Induction Algorithm is used to create a tree like flowchart that follow IF-Then Rules to make decisions[11]. The dataset is divided into smaller subsets based on some attribute selection criterion and arranged in a root node and leaf node having branches. The node is constructed with the input variable after applying a testing condition and the branches are the results of the test conducted on the root node. The association rules were formed from the induction trees by traversing from the root node to the leaf node and the information hidden in the dataset is extracted from the association rules of the decision tree. In order to avoid the anomalies present in the decision tree the rules are pruned to extract the efficient information present in the dataset. This paper uses C4.5 algorithm to construct the decision tree and the Entropy and Information gain were used for selecting the best attribute as a root node, and it is given by the following equation.

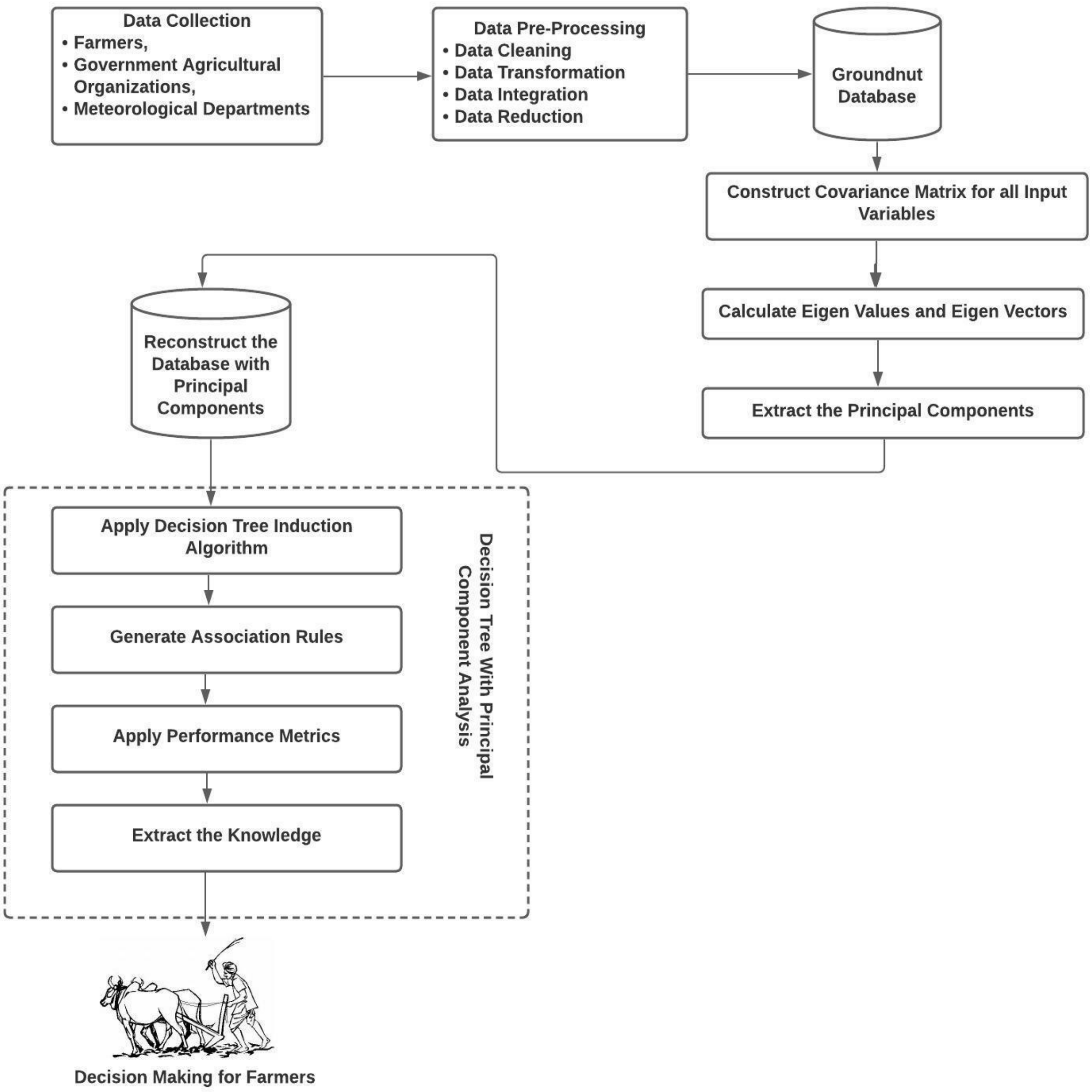
$$\text{Information Gain (InfoA(D)=-}\sum_{j=1}^{V} \frac{|D_j|}{|D|} \times \text{Info(Dj)} \qquad \text{Equation (5)}$$

$$\text{Gain (A) =Info (D) –InfoA(D)} \qquad \text{Equation (6)}$$

4. Proposed Framework

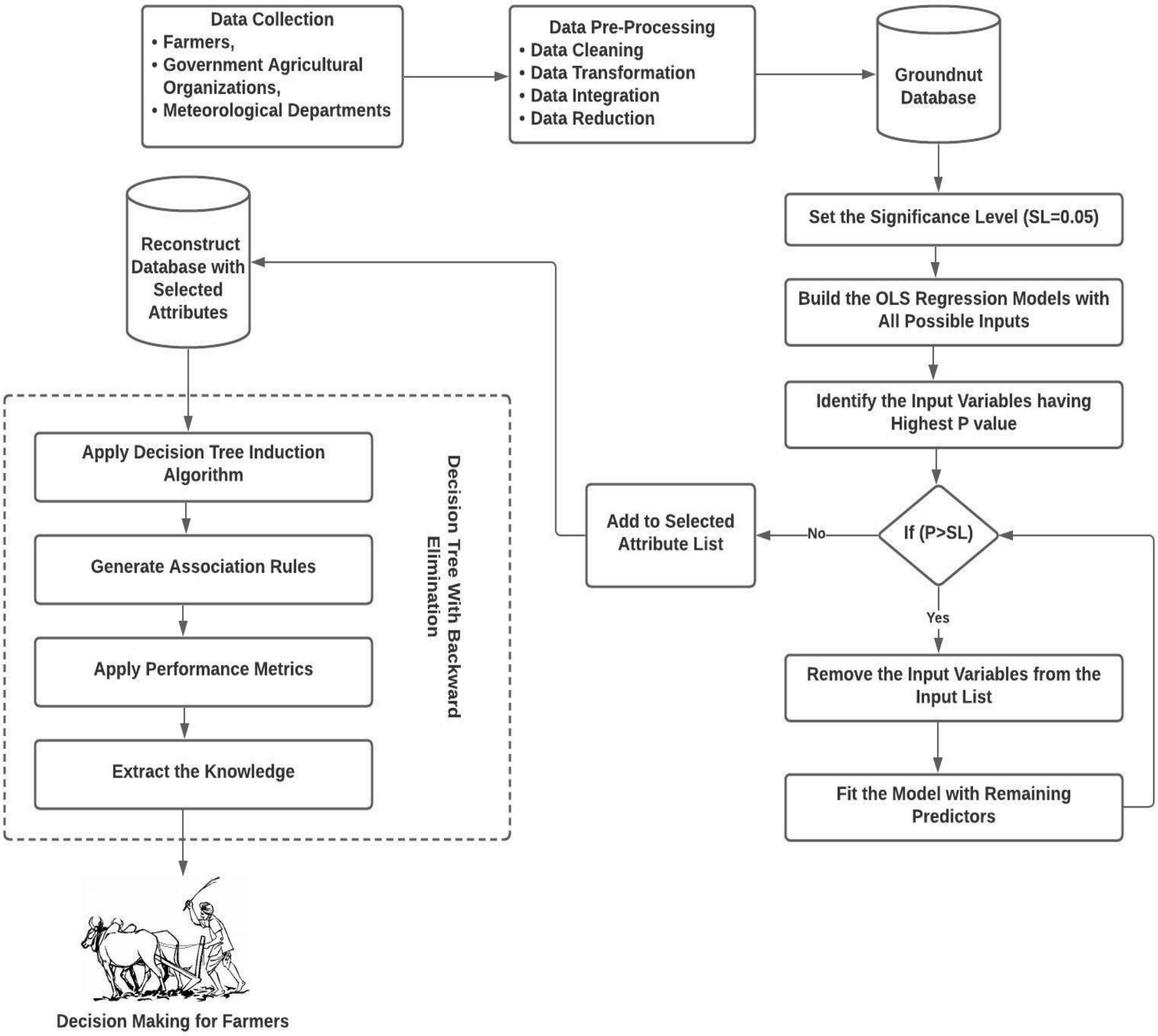
The data needed for this study is collected from four agricultural block namely Cuddalore, Virudhachalam, Panruit, Kurinjipadi situated in Cuddalore district of Tamilnadu. The total number of farmers participated in this study is 2789 and the time period of this study is from November 2019 to January 2020. In Cuddalore agricultural block 450 farmers were selected and from Kurinjipadi block 899 farmers were selected. From Panruti and Virudhachalam 1030 and 410 farmers were selected for this study. The variety of groundnut used for this study is Virudhachalm2, JRL and Gujarat7. The dataset used in this study is a secondary dataset that contains parameters responsible for groundnut growth as well as parameters related to environmental factors. The parameters such as manure used for land preparation, pesticides, seed rate, yield and fertilizers used were collected from directly interviewing the farmers. The meteorological parameters such as Instant Wind Speed (measured using Anemometer), Relative Humidity (measured using Hygrometer), Maximum and Minimum temperature were collected as a secondary data from Regional Meteorological Centre situated at Chennai, Tamilnadu. The overall distribution of rainfall is measured with the help of rain gauges during the period November-2019 to January-2020 (Pre-flowering stage(0-35days), Peak-flowering stage(35-45days),Post-flowering stage(45-65days), Pod-development stage(65-90days)) is collected as a secondary data from the Rain gauge station situated at the respective agricultural blocks. The collected data is preprocessed and a dataset is constructed that contains the input variable responsible for groundnut crop production and input variable related to environmental factors. To overcome the performance loss due to large dimensions in the real time, the significant features were selected first time using Principal Component Analysis (PCA) and for the second time Backward Elimination was used. The PCA first selects the number of components used to select from the dataset and then it construct the covariance matrix for all the input variables. Eigen vectors and Eigen values are computed from the covariance matrix and the Principal components were extracted from the dataset. The original groundnut dataset is reconstructed with the parameters selected as the principal components. The proposed framework for groundnut yield prediction using Decision Tree with Principal Component Analysis is shown in Figure 1.

Figure 1: Proposed Framework for Decision Tree with Principal Component Analysis (DTPCA)



The proposed method secondly used Backward Elimination to select the important parameters by setting a significance level. The OLS Regression model is fit with all the input variables and the input variables with highest p value is removed from the attribute list. This process continues until all the input variables that satisfy the significance level is remained in the dataset. Decision Tree Induction algorithm was implemented on the new datasets and decision tree was constructed. Association rules were generated from the decision tree and factors that is responsible for the growth of groundnut crop is identified from the association rules. Performance metrics were used to measure the performance of the classifier and the results were compared. The proposed framework for Decision Tree with Backward Elimination (DTBE) is shown in Figure 2.

Figure 2: Proposed Framework for Decision Tree with Backward Elimination (DTBE)

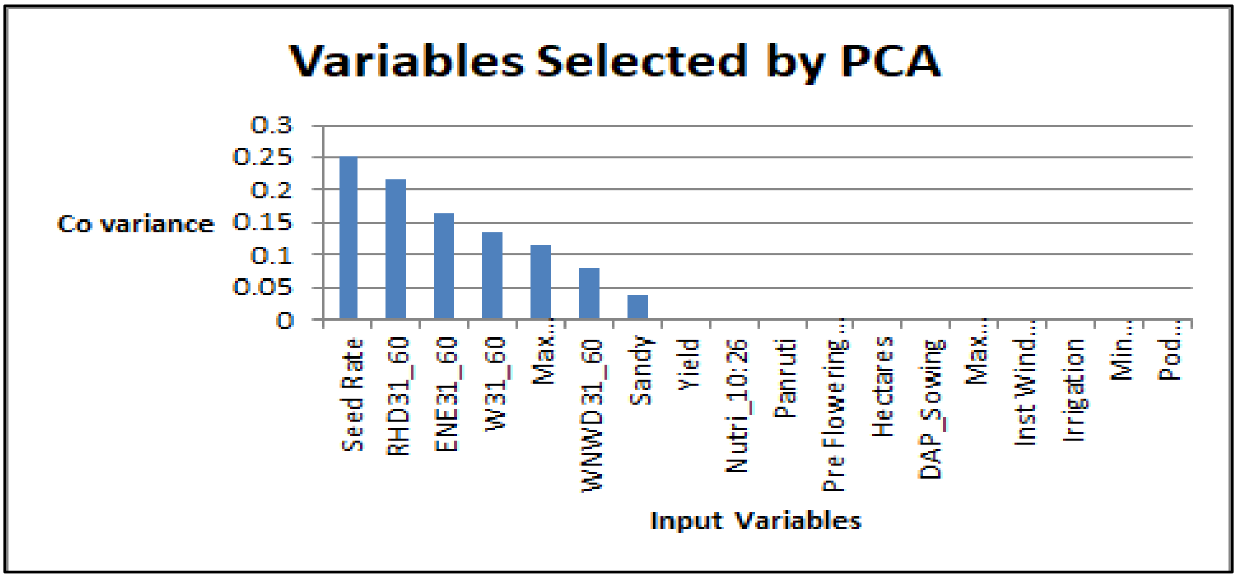


5. Experimental Results

5.1 Principal Component Analysis (PCA)

The proposed framework is implemented in Python 3.8 version with Spyder 4.2.5 as the code editor present in Anaconda Navigator. To implement the Principal Component Analysis (PCA), the dataset which contains categorical attributes such as Agricultural Block, Soil Types, Wind Direction_D1_D30, Wind Direction_D31_D60, Wind Direction_D61_D90 were transformed to numerical attributes using Label Encoding technique and assign as a new column using ColumTransformer. The dataset is split into training and testing and all the input variables are normalized using StandardScaler method present in sklearn.preprocessing package. The Principal Component Analysis is implemented using the sklearn.decomposition package and the number of components used to select were chosen as 18. The attributes selected by the PCA method is ranked by the covariance ratio and is shown in Figure 3.

Figure 3: Variables Selected by PCA



The feature importance of the selected variables is measured using the covariance ratio and from Figure 3, it is clear that among the 18 variables the first 7 variables plays a vital role. The percentage of importance of the variable is (Seed Rate (PCA1)=25%, Relative Humidity from Day 31 to Day 60(PCA2)=22%, Direction of Wind(East North East) form Day 31 to Day 60(PCA3)=16%, Direction of Wind(West) form Day 31 to Day 60(PCA4)=13%, Maximum Temperature Day 1 to Day 30(PCA5)=12%, Direction of Wind(West North West) form Day 31 to Day 60(PCA6)=8%, Soil Type (Sandy) (PCA7)=4%.

5.2 Decision Tree with Principal Component Analysis (DTPCA)

The dataset constructed with Principal Component Analysis was used to construct the decision tree. The variety of groundnut crop is used as the target variable and all the remaining 21 attributes are given as input to the decision tree algorithm. The dataset is divided into training set having 1859 records and testing set having 930 records. The decision tree was constructed with the DecisionTreeClassifier method present in the sklearn.tree library. The hyper parameters used to construct the decision tree was criterion is entropy, random state is set to zero, max_depth is set to 7, max_depth is set to 7, min_samples_leaf is set to 20, max_features is set to log$_2$, min_samples_split is set as 50. The decision tree constructed for the above parameters is given below in table 1.

Table 1: Decision Tree Constructed for Groundnut Dataset using DTPCA Algorithm

<table>
<tr>
<td>|--- MaxtempD1_30 <= 34.50
| |--- Inst Wind SpeedD1_30 <= 5.00
| | |--- DAP_Sowing<= 687.50
| | | |--- Seedrate<= 300.00
| | | | |--- class: JRL
| | | |--- Seedrate> 300.00
| | | | |--- Yield <= 9028.50
| | | | | |--- Sandy <= 0.50
| | | | | | |--- class: Virudhachalam2
| | | | | |--- Sandy > 0.50
| | | | | | |--- class: Virudhachalam2
| | | | |--- Yield > 9028.50
| | | | | |--- Yield <= 12028.00
| | | | | | |--- class: JRL
| | | | | |--- Yield > 12028.00
| | | | | | |--- Seedrate<= 900.00
| | | | | | | |--- class: Gujarat7
| | | | | | |--- Seedrate> 900.00
| | | | | | | |--- class: Gujarat7
| | |--- DAP_Sowing> 687.50
| | | |--- class: Virudhachalam2
|--- MaxtempD1_30 <= 34.50
| |--- Inst Wind SpeedD1_30 > 5.00
| | |--- Seedrate<= 1100.00
| | | |--- Relative Humidity_D31_60 <= 93.00
| | | | |--- ENE31_60 <= 0.50
| | | | | |--- Pod development Rain <= 60.60
| | | | | | |--- Yield <= 5960.00</td>
<td>|--- MaxtempD1_30 <= 34.50
| |--- Inst Wind SpeedD1_30 > 5.00
| | |--- Seedrate<= 1100.00
| | | |--- Relative Humidity_D31_60 > 93.00
| | | | |--- Hectares <= 1.50
| | | | | |--- class: Gujarat7
| | | | |--- Hectares > 1.50
| | | | | |--- DAP_Sowing<= 437.50
| | | | | | |--- Seedrate<= 500.00
| | | | | | | |--- class: Gujarat7
| | | | | | |--- Seedrate> 500.00
| | | | | | | |--- class: Gujarat7
| | | | | |--- DAP_Sowing> 437.50
| | | | | | |--- Nuti_10:26 <= 562.50
| | | | | | | |--- class: Gujarat7
| | | | | | |--- Nuti_10:26 > 562.50
| | | | | | | |--- class: Gujarat7
| | |--- Seedrate> 1100.00
| | | |--- Inst Wind SpeedD1_30 <= 9.00
| | | | |--- class: Virudhachalam2
| | | |--- Inst Wind SpeedD1_30 > 9.00
| | | | |--- Panruti<= 0.50
| | | | | |--- class: Gujarat7
| | | | |--- Panruti> 0.50
| | | | | |--- class: Gujarat7
|--- MaxtempD1_30 > 34.50
| |--- Yield <= 16597.00
| | |--- Yield <= 6068.50</td>
</tr>
</table>

```
|   |   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |   |   |--- Yield >  5960.00
|   |   |   |   |   |   |   |   |--- class: JRL
|   |   |   |   |   |--- Pod development Rain >  60.60
|   |   |   |   |   |   |--- Yield <= 3034.50
|   |   |   |   |   |   |   |   |--- class: JRL
|   |   |   |   |   |   |   |--- Yield >  3034.50
|   |   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |--- ENE31_60 >  0.50
|   |   |   |   |   |--- Yield <= 12109.50
|   |   |   |   |   |   |--- Nuti_10:26 <= 187.50
|   |   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |   |--- Nuti_10:26 >  187.50
|   |   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |--- Yield >  12109.50
|   |   |   |   |   |   |--- class: Gujarat7

|   |   |   |--- Hectares <= 1.50
|   |   |   |   |--- class: Virudhachalam2
|   |   |   |--- Hectares >  1.50
|   |   |   |   |--- class: JRL
|   |   |--- Yield >  6068.50
|   |   |   |--- Nuti_10:26 <= 562.50
|   |   |   |   |--- DAP_Sowing<= 437.50
|   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |--- DAP_Sowing>  437.50
|   |   |   |   |   |--- Sandy <= 0.50
|   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |--- Sandy >  0.50
|   |   |   |   |   |   |--- class: JRL
|   |   |   |--- Nuti_10:26 >  562.50
|   |   |   |   |--- Lateritic <= 0.50
|   |   |   |   |   |--- class: JRL
|   |   |   |   |--- Lateritic >  0.50
|   |   |   |   |   |--- class: Gujarat7
```

The association rules generated for the above decision tree is given below in the following Figure 4.

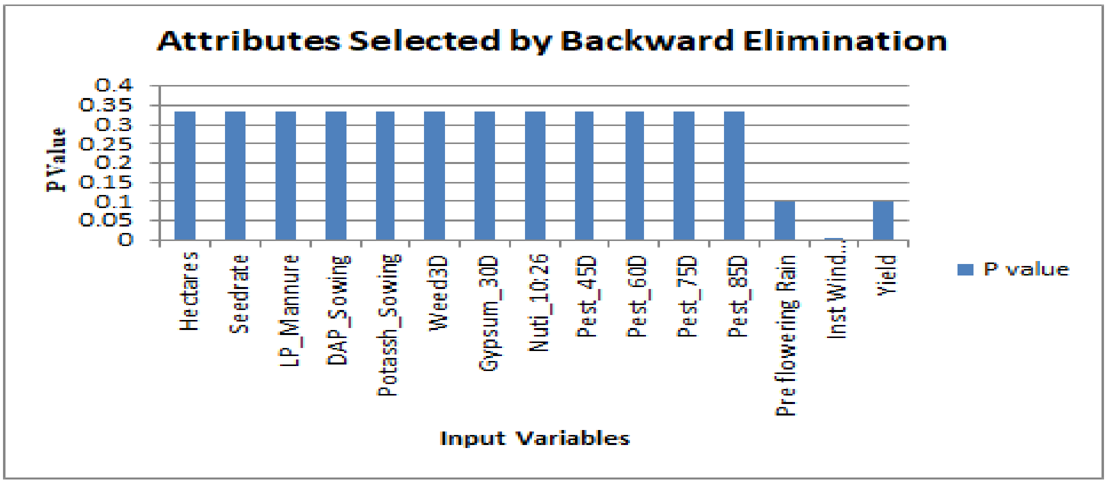
Figure 4: Association Rules Generated from the Decision Tree for DT_PCA Algorithm

R1:if(MaxtempD1_30 <= 34.5°C)^(Instant Wind SpeedD1_30 <= 5knots)^(DAP_Sowing<= 687.50kgs)^(Seed rate <= 300kgs)-->(Variety=JRL)
R2:if(MaxtempD1_30 <= 34.50°C)^(Instant Wind SpeedD1_30 <= 5knots)^(DAP_Sowing<= 687.50kgs)^(Seed rate > 300kgs)^(Yield <= 9028.50kgs)^(Soil type=Sandy)-->(Variety=Virudhachalam2)
R3:if(MaxtempD1_30 <= 34.50°C)^(Instant Wind SpeedD1_30 <= 5.00knots)^(DAP_Sowing<= 687.50kgs)^(Seed rate > 300kgs)^(Yield > 9028.50kgs)^(Yield <= 12028kgs)-->(Variety=JRL)
R4:if(MaxtempD1_30 <= 34.50°C)^(Instant Wind SpeedD1_30 <= 5knots)^(DAP_Sowing<= 687.50kgs)^(Seed rate > 300kgs)^(Yield > 9028.50kgs)^(Yield > 12028kgs)^(Seedrate <= 900kgs)-->(Variety=Gujarat7)
Likewise 30 rules were generated from the decision tree

5.3. BACKWARD ELIMINATION

The Backward Elimination is implemented using the statsmodels.api package. The dataset containing categorical variables are converted to numerical using Label Encoding and OneHotEncoding. The input variables are fit with the LinearRegression model present in the sklearn.linear_model package and the regression line was formed. All the 52 input variables are then fit the OLS Regression model present in the statsmodels.api. During the first iteration the variable (soil type=sandy) and (soil type=lateritic) has the highest p value and it is removed from the attribute list. During the second iteration (Wind Direction_D61_90=East North East) and (Agricultural block=Kurinjipadi) has the highest p value so it is removed from the list. Totally the dataset has undergone 26 iteration and the attributes with highest p value is removed from the attribute list for each iterations.The final attributes selected by the Backward Elimination algorithm is given below in Figure 5.

Figure 5: Attributes Selected by Backward Elimination Method



5.4. Decision Tree with Backward Elimination (DTBE)

Table 2: Decision Tree Constructed for Groundnut Dataset using DTBE Algorithm

```
|--- Pest_45D <= -0.14
|   |--- Preflowering Rain <= 0.51
|   |   |--- Yield <= -1.88
|   |   |   |--- class: Gujarat7
|   |   |--- Yield >  -1.88
|   |   |   |--- Seedrate<= -0.84
|   |   |   |   |--- Pest_75D <= -1.54
|   |   |   |   |   |--- Inst Wind SpeedD31_60 <= 0.17
|   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |--- Inst Wind SpeedD31_60 >  0.17
|   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |--- Pest_75D >  -1.54
|   |   |   |   |   |--- Preflowering Rain <= -0.66
|   |   |   |   |   |   |--- Yield <= -1.19
|   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |   |--- Yield >  -1.19
|   |   |   |   |   |   |   |--- class: JRL
|   |   |   |   |   |--- Preflowering Rain >  -0.66
|   |   |   |   |   |   |--- Yield <= -1.19
|   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |   |--- Yield >  -1.19
|   |   |   |   |   |   |   |--- class: JRL
|--- Pest_45D <= -0.14
|   |--- Preflowering Rain <= 0.51
|   |   |--- Yield <= -1.88
|   |   |   |--- Seedrate>  -0.84
|   |   |   |   |--- Yield <= -0.48
|   |   |   |   |   |--- Yield <= -0.49
|   |   |   |   |   |   |--- Preflowering Rain <= -0.66
|   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |   |--- Preflowering Rain >  -0.66
|   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |--- Yield >  -0.49
|   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |--- Yield >  -0.48
|   |   |   |   |   |--- class: JRL
```

```
|--- Pest_45D >  -0.14
|   |--- Yield <= 1.58
|   |   |--- Weed3D <= 1.25
|   |   |   |--- Nuti_10:26 <= 0.55
|   |   |   |   |--- class: Gujarat7
|   |   |   |--- Nuti_10:26 >  0.55
|   |   |   |   |--- Yield <= 0.90
|   |   |   |   |   |--- Yield <= 0.90
|   |   |   |   |   |   |--- Inst Wind SpeedD31_60 <= 0.79
|   |   |   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |   |   |--- Inst Wind SpeedD31_60 >  0.79
|   |   |   |   |   |   |   |--- class: JRL
|   |   |   |   |   |--- Yield >  0.90
|   |   |   |   |   |   |--- Inst Wind SpeedD31_60 <= -1.08
|   |   |   |   |   |   |   |--- class: Virudhachalam2
|   |   |   |   |   |   |--- Inst Wind SpeedD31_60 >  -1.08
|   |   |   |   |   |   |   |--- class: JRL
|   |   |   |   |--- Yield >  0.90
|   |   |   |   |   |--- Inst Wind SpeedD31_60 <= -0.15
|   |   |   |   |   |   |--- Inst Wind SpeedD31_60 <= -1.08
|   |   |   |   |   |   |   |--- class: JRL
|   |   |   |   |   |   |--- Inst Wind SpeedD31_60 >  -1.08
|   |   |   |   |   |   |   |--- class: JRL
|   |   |   |   |   |--- Inst Wind SpeedD31_60 >  -0.15
|   |   |   |   |   |   |--- class: Gujarat7
|   |   |--- Weed3D >  1.25
|   |   |   |--- Preflowering Rain <= -0.66
|   |   |   |   |--- class: JRL
|   |   |   |--- Preflowering Rain >  -0.66
|   |   |   |   |--- Preflowering Rain <= 0.51
|   |   |   |   |   |--- class: Gujarat7
|   |   |   |   |--- Preflowering Rain >  0.51
|   |   |   |   |   |--- class: Virudhachalam2
|   |--- Yield >  1.58
|   |   |--- class: Gujarat7
```

The same procedure and parameters used in DTPCA is used to build the decision tree for the groundnut dataset built by the Backward Elimination process, and the decision tree built for DTBE is given in Table 2. The Association rules generated from the constructed decision tree is given below in Figure 6.

Figure 6: Association Rules Generated from the Decision Tree for DT_BE Algorithm

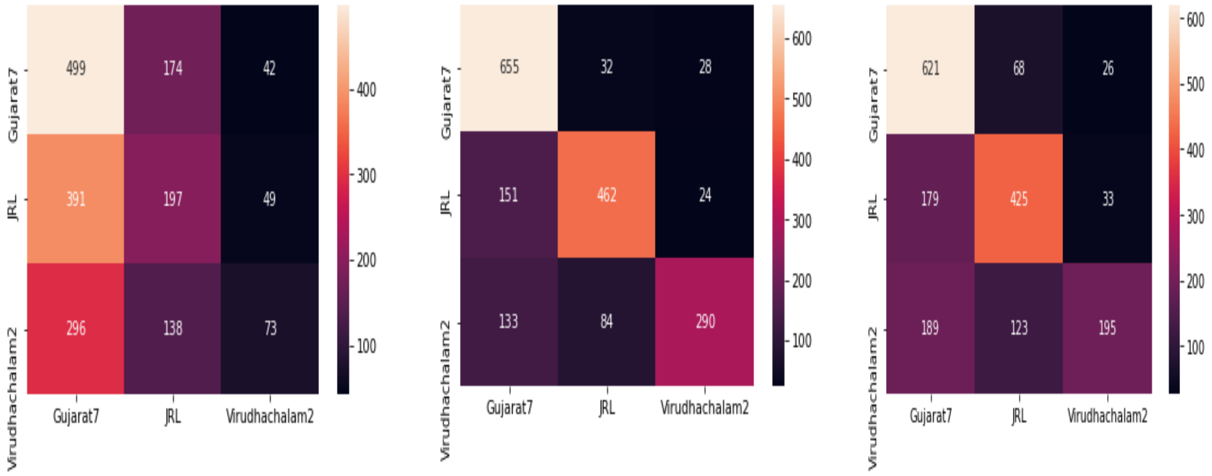
R1:if(Pest_45D <= 8.75litres)^(Preflowering Rain <= 19.05mm)^(Yield <= 2972kgs)**-->**(Variety=Gujarat7)
R2:if(Pest_45D <= 8.75litres)^(Preflowering Rain <= 19.05mm)^(Yield > 2972kgs)^(Seedrate <= 500kgs)^(Pest_75D <= 3.75litres)^(Instant Wind SpeedD31_60 <= 9knots)**-->**(Variety=Gujarat7)
R3:if(Pest_45D <= 8.75litres)^(Preflowering Rain <= 19.05mm)^(Yield > 2972kgs)^(Seedrate <= 500kgs)^(Pest_75D <= 3.75litres)^(Instant Wind SpeedD31_60 > 9knots)**-->**(Variety=Gujarat7)
R4:if(Pest_45D <= 8.75litres)^(Preflowering Rain <= 19.05mm)^(Yield > 2972kgs)^(Seedrate <= 500kgs)^(Pest_75D > 3.75litres)^(Preflowering Rain <= 18.30mm)^(Yield <= 5991kgs)**-->**(Variety=Gujarat7)
R5:if(Pest_45D <= 8.75litres)^(Preflowering Rain <= 19.05mm)^(Yield > 2972kgs)^(Seedrate <= 500kgs)^(Pest_75D > 3.75litres)^(Preflowering Rain <= 18.30mm)^(Yield > 5991kgs)**-->**(Variety=JRL)
Likewise 27 rules were generated from the decision tree constructed using Backward Elimination Dataset.

6. Evaluation of Results

Figure 7: Confusion Matrix for Decision Tree Algorithm

Figure 8: Confusion Matrix for Decision Tree with PCA

Figure 9: Confusion Matrix for Decision Tree with Backward Elimination



The confusion matrix for the constructed decision tree for the three different algorithms is given above and it is used to measure the performance of the classifier. The performance such as Accuracy, Error Rate, Precision, Recall etc was calculated from the confusion matrix and the values are given below in the Table 3.

Table 3: Performance Metrics for three Decision Tree Algorithms

	Decision Tree	**Decision Tree with PCA**	**Decision Tree with Backward Elimination**
Accuracy	0.608	**0.837**	0.778
Error Rate	0.391	**0.163**	0.222
Recall	0.383	**0.737**	0.639
Specificity	0.691	**0.872**	0.825
Precision	0.417	**0.781**	0.694
Negative Predicted Value	0.698	**0.884**	0.841
False Positive Rate	0.307	**0.126**	0.173
False Negative Rate	0.616	**0.261**	0.359
False Discovery Rate	0.581	**0.218**	0.304

From the table above, it is clear that the performance of the Decision Tree Classifier is improved by applying the feature selection techniques. Among the two feature selection techniques used, Principal Component Analysis performs best in classifying the dataset with higher accuracy and low error rate. Next to PCA, the Backward Elimination performs better with accuracy higher than the existing decision tree algorithm. The decision boundaries for the two proposed algorithms were given below.

Figure 10: Decision Boundary for DTBE

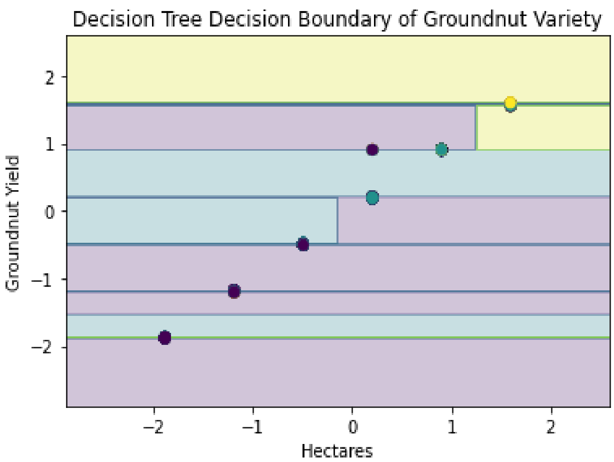
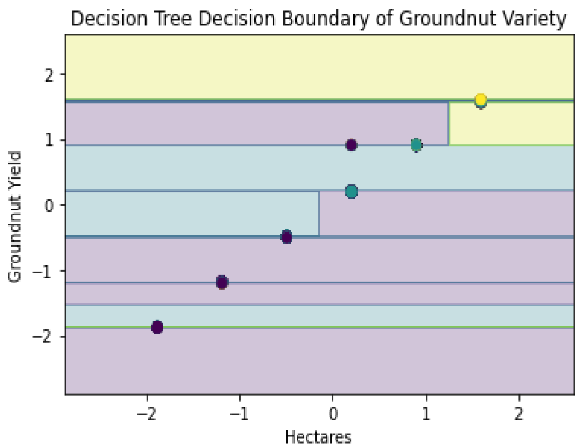
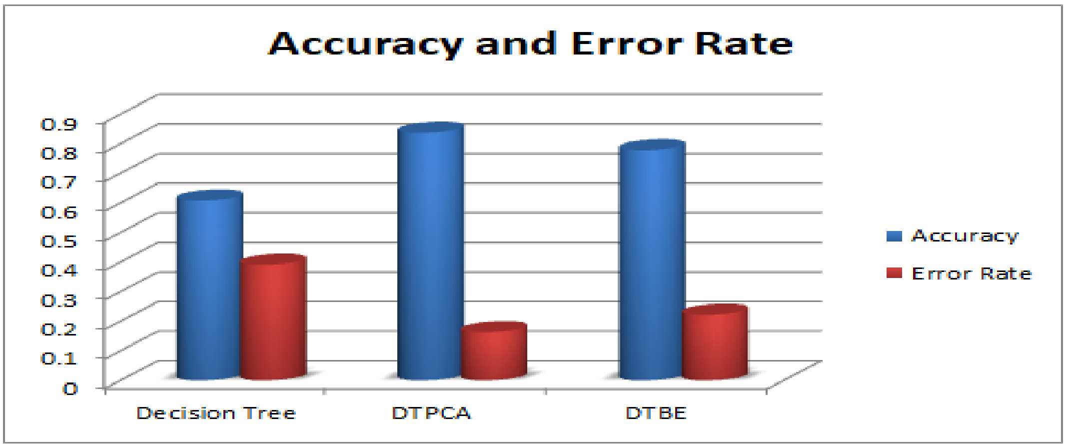


Figure 11: Decision Boundary for DTPCA



The Decision boundaries shown in Figure 10 and Figure 11 clearly tell that the dataset is non-linearly separable and it is complex in nature. For implementing machine learning algorithms to use in real time datasets, it is recommended to use feature selection using Principal Component Analysis to improve the performance of the classifier for obtaining best results. The accuracy and error rate for the two proposed algorithms were given below in Figure 12.

Figure 12: Accuracy and Error Rate for the Three Algorithms



6. Recommendations for Farmers

From this study it is proved that the yield of groundnut depends on the environmental factors such as Maximum Temperature, Intensity of Wind Speed during the pre-flowering phase. Application of fertilizer, DAP with potash and Gypsum with NPK10:26 during the sowing of groundnut plays a vital role in improving the yield of groundnut crop. The direction of wind during the 31st day to 60th day helps in increasing the carbon dioxide needed for groundnut crop for photosynthesis purpose. The Relative Humidity during day 1 to day 30 increases the stomata's of the groundnut crop and increases the photosynthesis capacity of the groundnut crop. The use of pesticide Choloropyrifos at the 45th day and application of Cypermethin 25% EC during the 60th day and use of Fungicide (Lustre) at the 70th day of controls the pest and fungal disease such as Tikka in the groundnut crop and increase the yield. The overall distribution of rainfall during the pre-flowering, peak flowering, post flowering and pod development phase helps to increase the yield of groundnut crop. The artificial irrigation applied to balance the water requirement of groundnut crop also plays a vital role in improving the yield of groundnut crop. Selecting the proper seed rate with appropriate seed treatment method also helps to improve the yield of groundnut crop. From this study, it is clear that the groundnut variety Virudhachalam2 is most suited to be cultivated in sandy soil and the groundnut variety JRL is most suited to be cultivated at Lateritic soil. The variety Gujarat7 is suited to be cultivated in both Sandy and Lateritic soil.

7. Conclusion

The proposed machine learning framework that contains two decision tree algorithms namely DTPCA and DTBE reduces the dimensionality of the real time dataset and built the decision tree. The association rules generated from the two decision trees serves as a recommender system for farmers and help them in decision making related to groundnut crop yield. This paper also suggests the various crop management strategies to be followed by the farmers to produce high yield. The performance of the two proposed algorithms were analyzed with the performance metrics and the results shows that the decision tree algorithm build using the combination of Principal Component Analysis performs better and produce results with higher accuracy. In future, the proposed algorithms are tested with various crop datasets and it will be recommended to apply in various domains such as Healthcare, Education etc.

REFERENCES

Karnasuta, S. (2021).Organic Farming Model of Paddy Rice Production with Environmental Efficiency in Thailand. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 3066-3074.

Kumar, M. S. (2021). Design And Development Of Automatic Robotic System For Vertical Hydroponic Farming Using Iot And Big Data Analysis. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 1597-1607.

Rajak, R. K., Pawar, A., Pendke, M., Shinde, P., Rathod, S., &Devare, A. (2017).Crop recommendation system to maximize crop yield using machine learning technique.International Research Journal of Engineering and Technology.4(12), 950-953.

Nathgosavi, V. (2021).A Survey on Crop Yield Prediction using Machine Learning.Turkish Journal of Computer and Mathematics Education (TURCOMAT).*12(*13), 2343-2347.

Gyamerah, S. A., Ngare, P., &Ikpe, D. (2020). Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov Kernel function. Agricultural and Forest Meteorology.280, 107808.

Rao, Haidi, Xianzhang Shi, AhoussouKouassiRodrigue, JuanjuanFeng, Yingchun Xia, Mohamed Elhoseny, Xiaohui Yuan, and LichuanGu. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. Applied Soft Computing, 74, 634-642.

Zhang, X., He, L., Zhang, J., Whiting, M. D., Karkee, M., & Zhang, Q. (2020). Determination of key canopy parameters for mass mechanical apple harvesting using supervised machine learning and principal component analysis (PCA).Biosystems Engineering, 193, 247-263.

Devi, K. Suganya, P. Srinivasan, and SivajiBandhopadhyay.(2020). H2K–A robust and optimum approach for detection and classification of groundnut leaf diseases.Computers and Electronics in Agriculture, 178, 105749.

Li, Z., Niu, B., Peng, F., Li, G., Yang, Z. and Wu, J. (2018). Classification of peanut images based on multi-features and SVM. IFAC-PapersOnLine. 51(17), 726-731.

Shah, V., & Shah, P. (2018).Groundnut Crop Yield Prediction Using Machine Learning Techniques. Int. J. Scient. Res. Comput. Sci. Eng. Inform. Technol, 3(5), 1093-1097.

Pathan, S. S. (2021). An Approach to Decision Tree Induction for Classification. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(12), 919-928.

Acknowledgements

This article has been written with the financial support of RUSA-Phase 2.0 grant sanctioned vide Letter No.F.24-51/2014-U, Policy (TNMulti-Gen), Dep.ofEdn.Govt. of India, Dt.09.10.2018.

AUTHOR'S PROFILE

S.Muthukumaran is a Research Scholar in the Department of Computer Science, Alagappa University, Karaikudi. His main Research Interest includes Machine Learning, Big Data Analytics, and Internet of Things. He has published 13 International Journals.

P.Geetha is working as an Associate Professor in the Department of Computer Science at Dr.UmayalRamanathan College for women, Karaikudi. Her Research Interest includes Data Mining, Network Security and Internet of Things. She has published 20 International Journals.

E. Ramaraj is working as the Professor and Head of the Department of Computer Science, Alagappa University, Karaikudi. He has very sound knowledge in many research fields especially in Data Mining, Network Security, Remote Sensing, and Big Data Analytics. He has published more than 100 international journals.