Decision Tree Algorithms and their Applicability in Agriculture for Classification

Abstract: Machine learning approaches has the advantage that in most of machine leaning algorithms data transformation is unnecessary, can handle missing predictor variables, success of prediction is not dependent on data meeting normality conditions or covariance homogeneity, variable selection is intrinsic to the methodology and provides good accuracy over the traditional methods. The decision tree is one of such machine learning algorithms which is capable of handling both complete and incomplete data, so it can be applied in the field of agriculture where such data occurs frequently. The algorithms which were considered for this study includes ID3 (Iterative Dichotomizer 3), Classification and Regression Tree (CART) and C4.5. This paper provides a detailed approach on the development of decision tree using its various algorithms. It is anticipated that this study will be a layman guide to all agriculture researchers towards enhancing awareness of the potential advantages of using decision tree in agriculture, and contributing to its wide applicability in the agriculture data. The secondary data of cotton genotypes was used for classifying the genotypes into two classes, and hold out method was used for cross validation for checking the performance of the algorithms.

Keywords: Decision tree, supervised classification, CART, ID3, C4.5

1. Introduction

Decision tree is an important classification model build by Hunt *et al.* (1966). It is a supervised machine learning technique which looks like an inverted tree, where each node represents a predictor variable. The link between the nodes represents a decision and each leaf node represents an outcome variable. The goal of the decision tree is to create a model that predicts the value of the target variable by learning simple decision rules inferred from the data feature. As we traverse down the tree, we must make decisions at each node, until a dead end is reached (Quinlan, 1986). The paths from root to leaf represent classification rules. So, decision trees use tree splitting logic until pure or somewhat pure leaf node classes are attained. Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable to solve any kind of problem in hand that is classification or regression. In our study we have discussed only the classification problem.

2. Methodology

The decision tree produces a sequence of if else rules that can be used to classification. The tree is build using the labeled (training) data and then this tree is used to classify the unlabeled (test) data. The next section provides a detailed insight to the development of a decision tree and the popular algorithms for its development, and the secondary data on cotton genotypes which will be used for classification

2.1 Structure of a Decision Tree

Before getting into the detail of algorithm, below are some terms that are mandatory to understand the decision tree. The figure 1 provides the better understanding of the basic terms and the structure of the decision tree.

- Root Node: It represents the entire training data which further gets divided into two or more homogeneous sets, the first split is performed at this point.
- Splitting: It is a process of dividing a node into two or more sub-nodes on the basis of some condition.
- Leaf Node (Terminal Node): Nodes that do not split further representing the final class of the outcome.

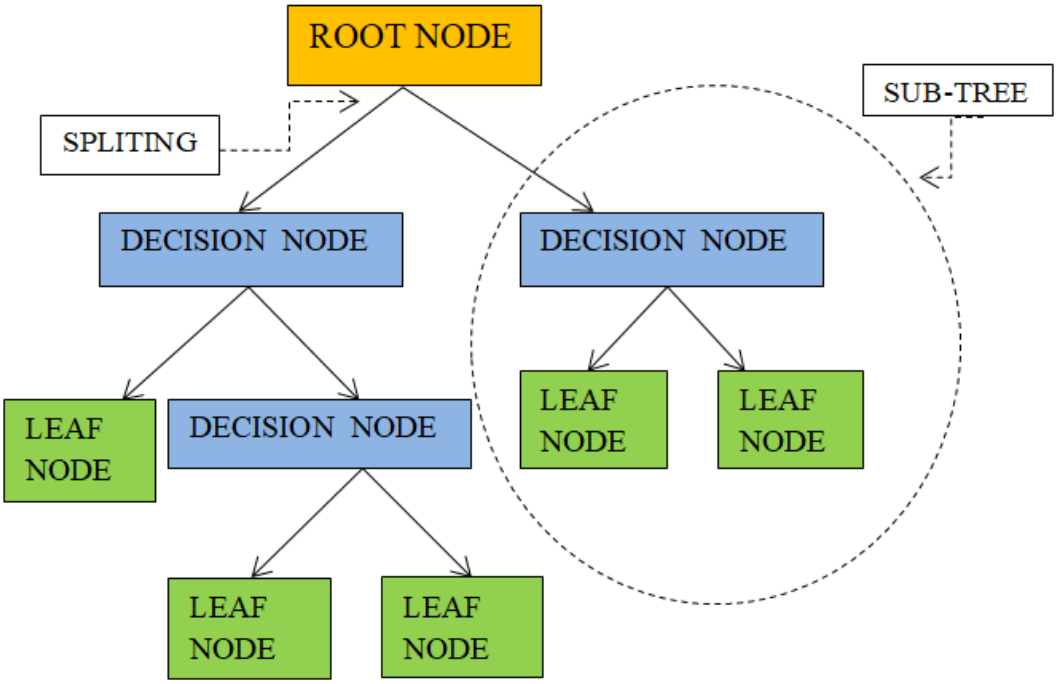

Figure 1 Decision tree structure

- Decision Node: When a sub-node splits into further sub-nodes, then it is called decision node. Each internal node represents a decision point that eventually leads to the prediction of the outcome.
- Sub-tree: It is a sub-section of entire tree.
- Branches: They are connections between the nodes, and are represented as arrows with a response such as yes or no.

2.2 Basic Decision Tree Algorithm

The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label. The below-mentioned steps represent the general workflow of a decision tree used for classification purposes.

(i) Create a root node out of all features of training data by using some feature splitting measure.

(ii) If labels of all observations in the node are of same class, then return that node as a leaf node named with that class label.

(iii) If list of attributes is empty then return that node as a leaf node named with the majority class in that node, and stop.

(iv) Apply feature splitting measure on remaining data and features for the finding next best-splitting attribute and label that node as a decision node.

(v) If the number of remaining observations is zero, then create a leaf node with a majority class in observations.

(vi) Repeat the above step until all nodes becomes pure node that is leaf node. The modeled decision tree can be used to predict the class of unknown observations.

2.3 Types of Algorithms for the Development of a Decision Tree

In decision tree one of the major challenges is the identification of the attribute for the root node in each level. There are various feature selection measures for nodes, and depending on that there are different algorithms for building a decision tree. The attribute having the best score for the selection metrics is selected as the splitting attribute for the given decision node. Following are the most commonly used algorithms:

• ID3 (Iterative Dichotomizer 3)

Iterative Dichotomizer 3 algorithm is one of the most effective algorithms used to build a Decision Tree. It uses the concept of entropy and information gain to generate a decision tree for a given set of data. It was developed by J. R. Quinlan (Quinlan, 1986). The algorithm creates a multi way tree, finding for each node the variable that will yield the largest information gain for the targets. Trees are grown to their maximum size and then a pruning step (Esposito *et al.,* 1997) is usually applied to improve the ability of the tree to generalize to unseen data. Using a decision tree, we can keep splitting until each leaf node has only a single training observation, which would be zero impurity. However, graphically it would have many small regions and it would be a case of over fitting. ID3 uses entropy and information gain as a metric for splitting the tree.

The word Entropy is borrowed from Thermodynamics which is a measure of variability or randomness. Shannon extended the thermodynamic entropy concept (Shannon, 1948) and introduced it into statistical studies and suggested the following formula for statistical entropy:

$$Entropy = -\sum_{i} p_i \log_2 p_i$$

This computes the entropy at a node by summing over all classes i and computing $p_i \log_2 p_i$, where p_i is the proportion of observations that belong to class i at a node. The lesser the entropy, the better it is. If the sample is completely homogeneous, then the entropy is zero and if the sample is an equally divided that is 50% each then it has entropy of one.

We want to make splits so that we have many observations of only one particular class and few observations of the other. In other words, the split should decrease the entropy. Higher entropy implies a mix of different classes and low entropy means that predominantly there is one class. Ideally, it is desired that nodes have low entropy, i.e. all observations at that node are definitely of one class. This low entropy is desirable at the leaf nodes to classify an observation. The figure 2 provides an easy understanding of homogeneity of classes with respect to entropy, where green and pink bar represents the sample for two classes.

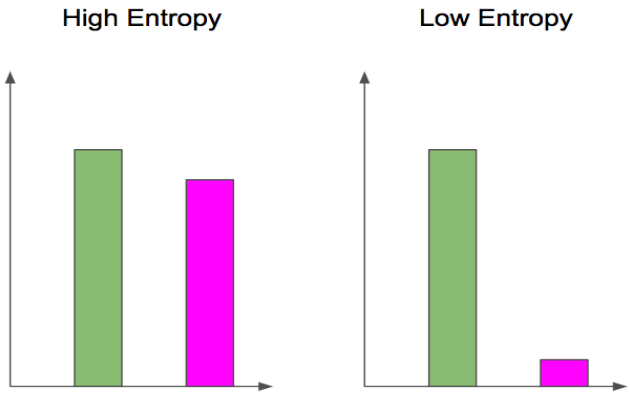

Figure 2: Entropy of impure and pure nodes

The information gain is the decrease in entropy after a dataset is split on a feature. Constructing a decision tree is all about finding feature that returns the highest information gain. Less impure node requires less information to describe it. And, more impure node requires more information. The feature with the highest information gain is used to split the data at the root node.

Steps to calculate entropy for splits:

i. Calculate entropy of parent node

ii. Calculate entropy of each individual node of split and calculate weighted average of all sub-nodes available in split.

The information gain from entropy can be obtained as

$$Information\ Gain = Entropy(Parent\ node) - \{Weighted\ average * Entropy(Child\ node)\}$$

- **CART (Classification and Regression Tree)**

Classification and regression tree is a non-parametric decision tree algorithm that produces either classification or regression trees (Breiman *et al*., 1984), depending on whether the dependent variable is categorical or continuous, respectively. It uses Gini impurity as a metric for splitting the tree. Gini impurity is a measure of the homogeneity (or purity) of the nodes. If all data points at one node are of same class, then that node is considered pure and has the smallest value for Gini impurity. So, by minimizing the Gini impurity the decision tree finds the features that separate the data best.

$$Gini\ Impurity = 1 - \sum_{i=1}^{c} (p_i)^2$$

where, p_i is the probability of class i in a node.

- **C4.5**

It is the successor of ID3 and was also developed by J. R. Quinlan (1993). This algorithm removed the restriction that features must be categorical. C4.5 algorithm uses information gain ratio as the metric for splitting the trees. The algorithm recursively classifies data until it has been categorized as perfectly as possible. It performs error based pruning that is a part of tree is removed and accuracy is checked repetitively. This algorithm is the most popular algorithm amongst all existing algorithms of decision tree. It is different from CART in terms of how the features are selected for a node. It uses information gain ratio for feature selection instead of Gini impurity.

Soon after the development of entropy mathematicians realized that information gain is biased toward multi-valued attributes. It favors the attributes that have large number of distinct values. To solve this issue, gain ratio was developed which is more reliable than information gain. This overcomes the bias in information gain. It is simply normalization of the information gain. The gain ratio is defined as:

$$Gain\ Ratio = \frac{Information\ Gain}{Entropy}$$

2.4 Agriculture data

The secondary data of cotton genotypes was taken from an experiment conducted by Central Institute for Cotton Research, Sirsa during kharif season of 2018-19. The data was split into two classes with class of low yield genotypes having 214 observations and the class of high yield genotypes having 151 observations. The variable yield was a categorical variable with "LOW" and "HIGH" yield categories, and there were other seven independent variables (Boll weight, plant height, plant width, number of bolls/plant, leaf shape, number of sympodia and monopodia) which were continuous, discrete and categorical. The holdout method in 80:20 ratios was used for cross validation.

3. Results and Discussion

The CART and C4.5 algorithms were used to develop the decision tree. The ID3 algorithm was not used since it can handle only continuous variables and our data had mix of continuous, discrete and categorical variables. The algorithms discussed in the previous section can be summarized as:

Table 1: Comparison between the algorithms of decision tree

S. No.	Algorithm	Feature Selection Measure	Independent Variable	Split on each node
1.	ID3	Information Gain	Continuous	Multiple
2.	CART	Gini Impurity	Categorical/Continuous	Binary
3.	C4.5	Information Gain Ratio	Categorical/Continuous	Multiple

CART Algorithm

A decision tree of 7 nodes and depth 4 was obtained using this algorithm on the training data. The variable number of bolls becomes the root node based on its minimum Gini index. The decision tree keeps on splitting until all nodes which are left are the leaf nodes.

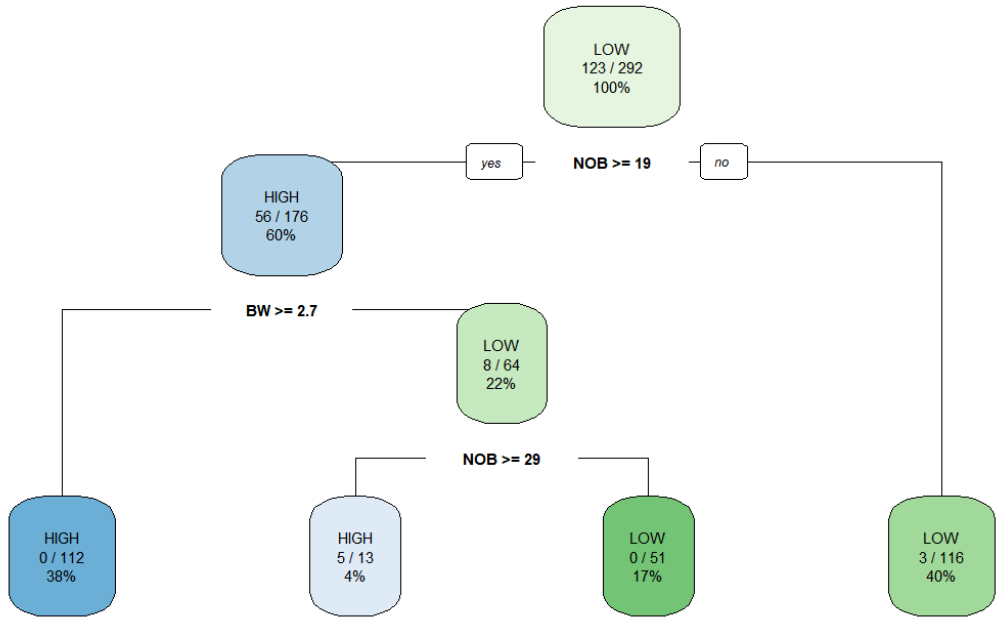

Figure 3: Decision tree developed from CART algorithm

The decision tree was then provided with the test data. The accuracy of the CART algorithm in classifying the unlabeled genotypes was 93.15%. The figure 4 represents the confusion matrix of CART algorithm with five misclassifications from the class of low yield genotypes.

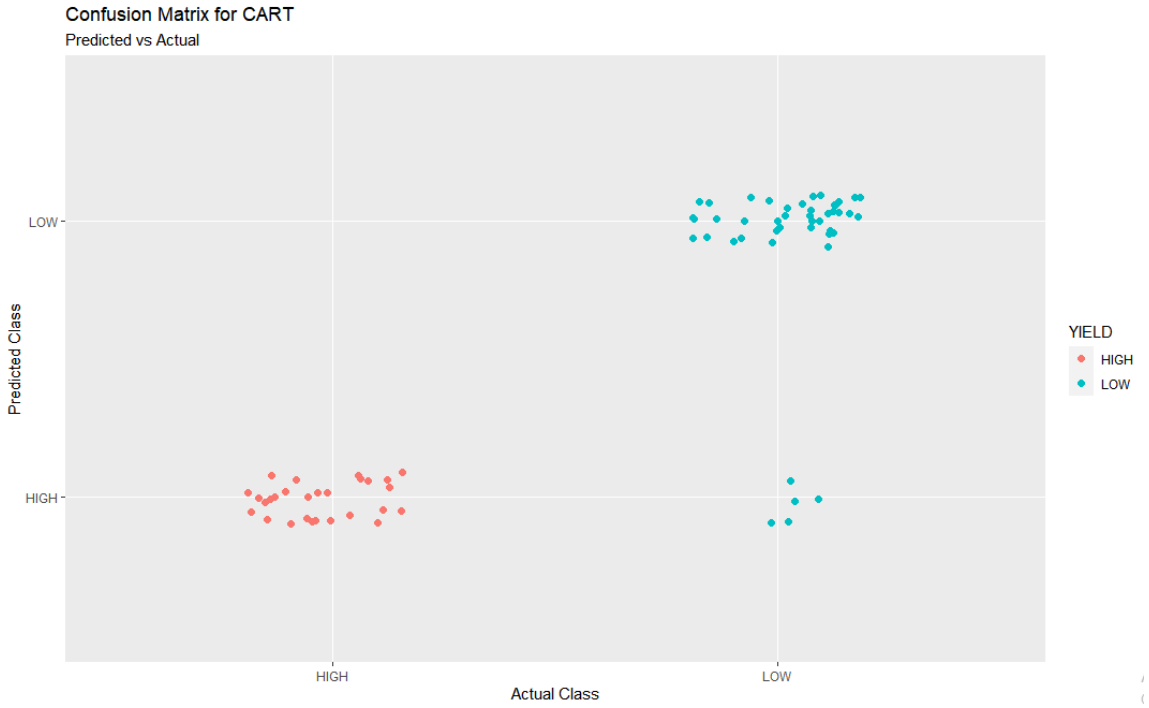

Figure 4: Confusion matrix for CART algorithm

Over fitting is very common problem in decision trees, which implies that a decision tree may be perform good with the training data but it fails in classification with the test data (Ying, 2019). This can be countered with “Pruning”, which is done by reducing the size of tree. Pruning of the decision tree was done and was then provided with the same test data. It was found that pruning has reduced the accuracy and total numbers of misclassified observations were increased. So, we kept the original decision tree of seven nodes and four depths.

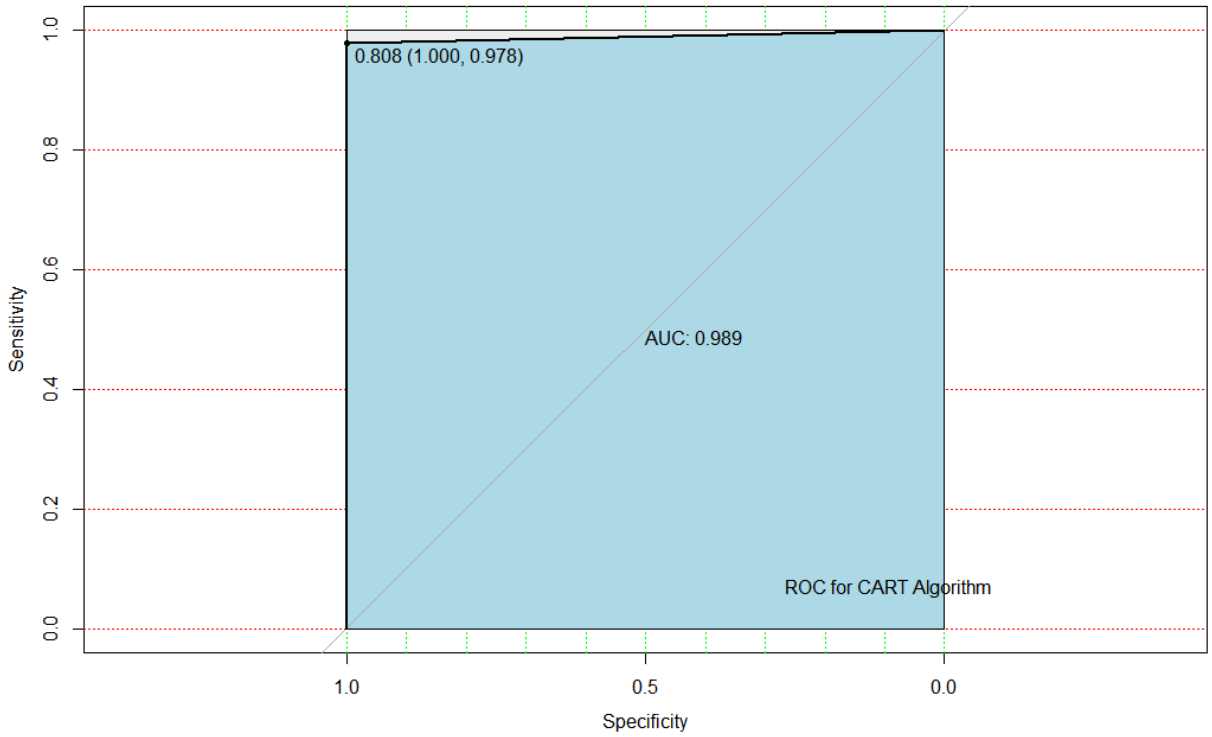

Figure 5: ROC curve for CART algorithm

It is clear from the confusion matrix that all the observations from “HIGH” class are correctly classified so the sensitivity of the model was 100%. The figure 5 is plot between sensitivity and specificity of the model and it shows the area under curve (AUC) and receiver operating characteristic (ROC). It is visual representation of the performance of a classifier at various threshold settings and it incorporates specificity and sensitivity in one figure (Bowers & Zhou, 2019). Higher the value of AUC, better is the model in distinguishing the classes. The value of AUC for CART algorithm 0.989 which is close to 1.

C4.5 Algorithm

The decision tree developed from C4.5 algorithm has 13 number of nodes, which is almost double of the CART algorithm. Since this tree is bigger and has more nodes so its accuracy must be more than the previous algorithm. The figure 6 shows the decision tree from C4.5 algorithm. In this tree also the number of bolls variable is the root node, since it had minimum information gain ratio among all the variables.

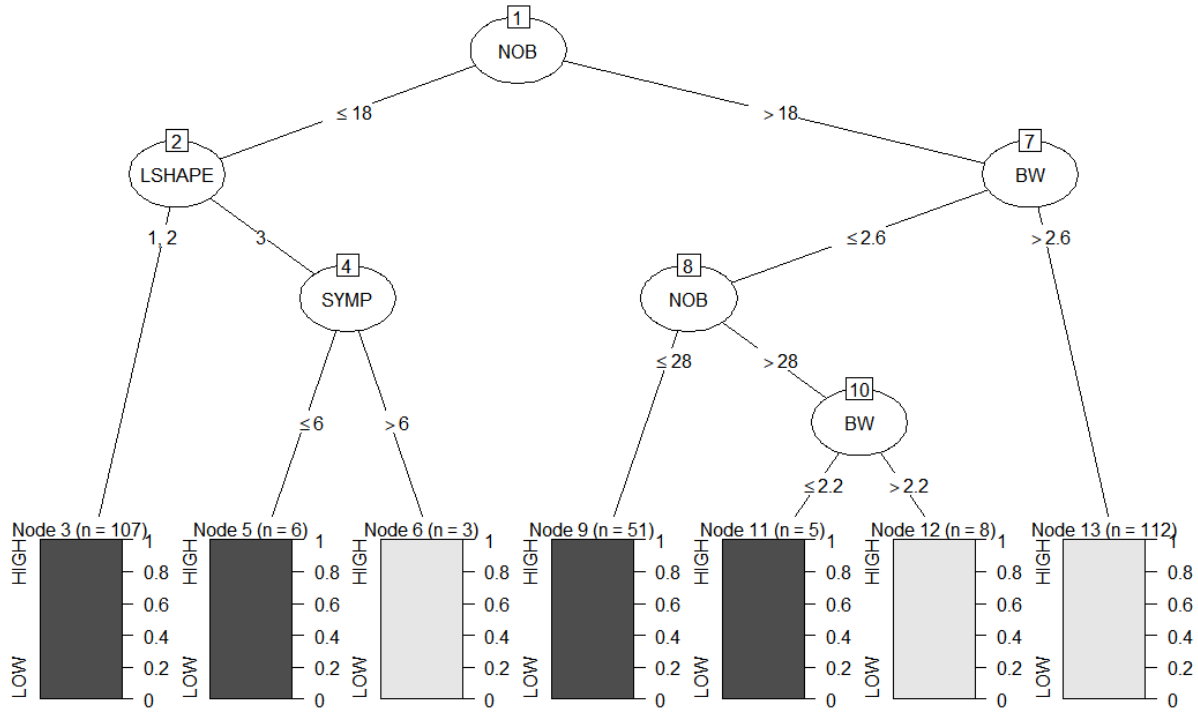

Figure 6: Decision tree developed from C4.5 algorithm

Test data was provided to the tree and the confusion matrix in the figure 7 shows the number of misclassifications. Four observation from low yield class were misclassified into high yield class. Since all the observations from high yield class are correctly classified, so the sensitivity is 1 for this tree.

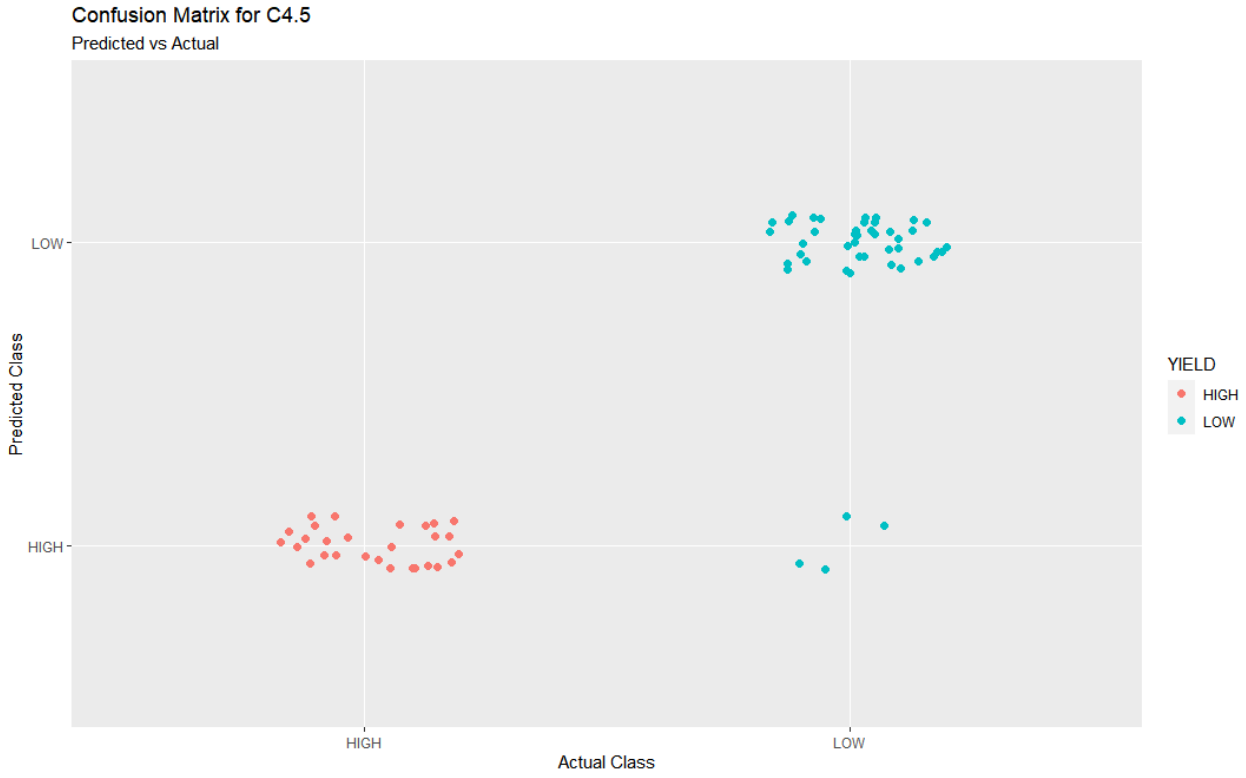

Figure 7: Confusion matrix for C4.5 algorithm

The ROC curve for the decision tree of C4.5 algorithm is presented in the figure below, which is almost similar to that of CART algorithm with same value of AUC.

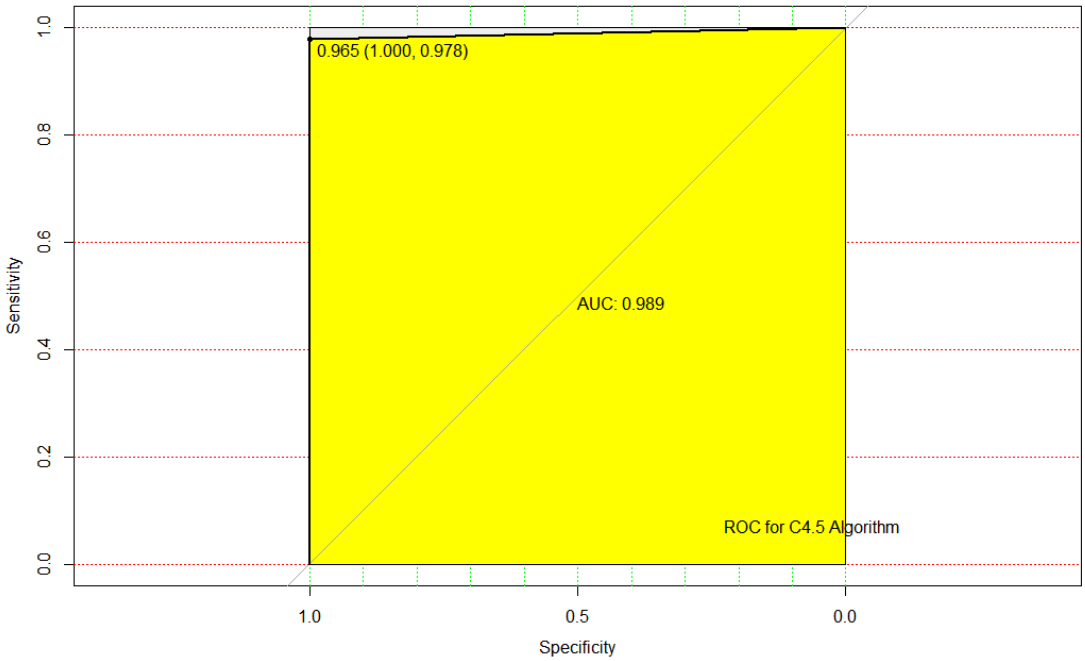

Figure 8: ROC curve for C4.5 algorithm

The accuracy is not always an appropriate measure for validating a model (Deepak & Ameer, 2019), and it should be considered along with other performance measures. So, other performance measures were also calculated and their values for both the algorithms are given in the table 2.

Table 2: Comparative measures of CART and C4.5

Measure	**CART**	**C4.5**
Accuracy	0.9315	0.9452
95% CI	(0.8474, 0.9774)	(0.8656, 0.9849)
Error rate	0.0685	0.0548
Kappa	0.8599	0.8872
Specificity	0.8889	0.9111
Sensitivity	1.0000	1.0000
Balanced accuracy	0.9444	0.9556

*Positive class: HIGH

By comparing the performance measures of the two algorithms, we can deduce that the decision tree made from C4.5 algorithm performed best in classifying the genotypes. It can be considered in future for such type of classification problems in agriculture. The algorithms were implemented using the R software (version 4.1.2) by R Core Team (2021) and the above performance measures (table 2) were also calculated using various packages in R.

4. Conclusion

In this paper, we have given the detailed methodology for the development of the decision tree algorithms and its implementation in the field of agriculture. The algorithms for development of decision tree included ID3, C4.5 and CART. The efficiency of the decision tree algorithms was analyzed based on their accuracy and other performance measures. The C4.5 algorithm performed best in classifying the cotton genotypes. The future scope on this includes the comparison of decision tree algorithms with other machine learning algorithms with respect to classification problem in agriculture data.

5. Reference

- Bowers, A. J. & Zhou, X. (2019). Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *Journal of Education for Students Placed at Risk*, 24(1), 20–46.
- Breiman, L., Friedman, J. H., Stone, C. J. & Olshen, R. A. (1984) Classification and regression trees. *Chapman & Hall.*
- Deepak, S. & Ameer, P.M. (2019). Brain tumor classification using deep CNN features via transfer learning. *Computers in Biology and Medicine*, 111, 1-8.
- Esposito, F., Malerba, D. & Semeraro, G. (1997) A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), 476–91.
- Hunt, E., Marin, J. & Stone, P. (1966) Experiments in induction. *Academic Press, New York.*
- Quinlan J. R. (1986) Induction of decision trees. *Machine Learning*, 1(1), 81–106.
- Quinlan J. R. (1993) C4.5: Programs for machine learning. *Morgan Kaufmann Publishers, United States*.
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379-423.
- Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*. 1168(2), 1-6.