Plant Disease Prediction using Machine Learning Algorithms

G. Prem Rishi Kranth
UG Student
Koneru Lakshmaiah Education Foundation
Guntur (A.P), 522502

M. Hema Lalitha
UG Student
Koneru Lakshmaiah Education Foundation
Guntur (A.P), 522502

Laharika Basava
UG Student
Koneru Lakshmaiah Education Foundation
Guntur (A.P), 522502

Anjali Mathur, PhD
Associate Professor
Koneru Lakshmaiah Education Foundation
Guntur (A.P), 522502

ABSTRACT
Machine learning is the one of the branch in Artificial Intelligence to work automatically or give the instructions to a particular system to perform a action. The goal of machine Learning is to understand the structure of the data and fit that data into models that can be understood and utilized by the people. The proposed research work is for analysis of various machine algorithms applying on plant disease prediction. A plant shows some visible effects of disease, as a response to the pathogen. The visible features such as shape, size, dryness, wilting, are very helpful to recognize the plant condition. The research paper deals with all such features and apply various machine learning technologies to find out the output. The research work deals with decision tree, Naive Bayes theorem, artificial neural network and k-mean clustering and random forest algorithms. Disease development depends on three conditions-host plants susceptible to disease, favorable environment and viable pathogen. The presence of all three conditions is must for a disease to occur.

Keywords
Naive Bayes, Artificial neural network, random forest, k-means clustering

1. INTRODUCTION
Machine Learning behaves like self-learning concept which will work without any interruption of a human. Now a day's self-driving cars, hand-writing recognition, Stock market are some of the examples of Machine Learning concepts. Machine learning will be able to predict the future based on the past or historical data. A computer program is said to be learned from experience E with respect to some clause of task T and performance measure P, if its performance on T as measured by P improves with experience E. Machine learning broadly uses three major learning algorithms Supervised learning, Unsupervised learning, Reinforcement learning. Machine learning can be used in each and every routine task performed by human being. The research work deals with plant disease prediction with the help of machine learning

A plant disease is a physiological abnormality. Once a plant suffers from any diseases it shows up certain symptoms. symptoms are the outward changes in the physical appearance that are gradually developed and can be witnessed by naked eyes. Illustrations of symptoms are wilt leaf spots, rots, cankers and many more. The visible effects of disease can broadly categorize in following types: -

Wilting, it is loss of turgor pressure in a plant leading to temporary or permanent drooping of leaves, shoots, or entire plants from lack of water or infection by different pathogens.

Spot, is a definite, localized, round to regular lesion, often with a border of a different color, characterized as to location (leaf spot, fruit spot) and color (brown spot, black spot);

Powdery mildew, is a fungal disease that affects a wide range of plants. Infected plants display white powdery spots on the leaves and stems. As the disease progresses, the spots get larger and denser.

Galls, these are abnormal growths that occur on leaves, twigs, or branches. They may be simple lumps or complicated structures, plain brown or brightly colored.

Dryness, after normal aging process generally leaf's get dry and fall down from the tree, but at other times drying of leaves may be a symptom of fungal attacks.

In plant disease diagnosis, data provided is small and some of the values are missing that will require imputation of values we will replace all the null values with -1

The proposed research work applies the concept of ensemble learning, that is implemented through machine learning algorithms. After implementation the result is compare to get the model has the highest accuracy.

2. LITERATURE SURVEY
In 2011, an innovative approach was presented[1] to automatically grade the disease on plant leaves. According to that, plant pathologists mainly rely on naked eye prediction and a disease scoring scale to grade the disease. That leads some problems associated with manual grading This manual grading is not only time consuming but also not feasible. Hence an image processing-based approach to automatically grade the disease spread on plant leaves by employing Fuzzy Logic had been proposed. The results are proved to be accurate and satisfactory in contrast with manual diseases are inevitable in plants. The proposed methodology aims to model a promising disease grading system for plant leaves. The system was divided into the following steps: (1) Image acquisition (2) Image Pre-processing (3) Color image segmentation (4) Calculating AT and AD (5) Disease grading by Fuzzy Logic.

In 2014, an survey report was published[2], based on different classification techniques that could be used for plant leaf disease classification. A classification technique deals with classifying each pattern in one of the distinct classes. A classification is a technique where leaf is classified based on its different morphological features. There are so many classification techniques such as k-Nearest Neighbor Classifier, Probabilistic Neural Network, Genetic Algorithm, Support Vector Machine, and Principal Component Analysis,

Artificial neural network, Fuzzy logic. Selecting a classification method is always a difficult task because the quality of result can vary for different input data. Plant leaf disease classifications have wide applications in various fields such as in biological research, in Agriculture etc. The paper provides an overview of different classification techniques used for plant leaf disease classification.

In 2012, an article was published[3] having a detailed description on definition of disease, types of diseases, symptoms and causes of most commonly observed plant diseases

One article was published by Michigan University[4] regarding the threats caused due to diseases. Various conditions for disease development had been discussed there. An overview of major disease-causing organisms and the effect of diseases caused by them was given.

3. EXPERIMENTAL ANALYSIS

The machine learning deals with classification, the research work is used to classify healthy and unhealthy plants. Our work is based on morphological features of the plant leaf. The techniques represented in this paper are

1. Decision tree
2. K-mean clustering
3. Naive Bayes
4. Random forest
5. Artificial neural network

3.1 Decision Tree

The algorithm contains Predefined target variables. It is a tree in which each branch node represent a choice between a number of alternatives and each leaf node represents a decision. The decision trees takes input as object or situation described by set f properties & output as Yes / No.

1. An Entropy[5] H(S) is a measure of the amount of uncertainty in the (data) set S (i.e. entropy characterizes the (data) set S).

   $H(S)=\sum_{x\epsilon X} -p(x)log_2\, P(x)$

2. Where, S – The current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm), X – Set of classes in S, p(x) – The proportion of the number of elements in class x to the number of elements in set S.

   When H(S)=0, the set S is perfectly classified (i.e. all elements in S are of the same class).

   In ID3 algorithm, entropy is calculated for each remaining attribute. The attribute with the **smallest** entropy is used to split the set S on this iteration. The higher the entropy, the higher the potential to improve the classification here.

3. Information gain[5] IG(A) is the measure of the difference in entropy from before to after the set S is split on an attribute A. In other words, how much uncertainty in S was reduced after splitting set S on attribute A.

   $IG(A,S)=H(S)-\sum_{t\epsilon T} p(t)\, H(t)$

   Where, H(S) – Entropy of set S T – The subsets created from splitting set S by attribute A such that. $S = \cup_{t\epsilon T}\, t$

   p(t) – The proportion of the number of elements in t to the number of elements in set S, H(t) – Entropy of subset t.

   In ID3, information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the largest information gain is used to split the set S on this iteration.

The plant disease prediction based on decision trees experimental setup was performed in R language and follows results were obtained, in figure-1 and Figure-2.

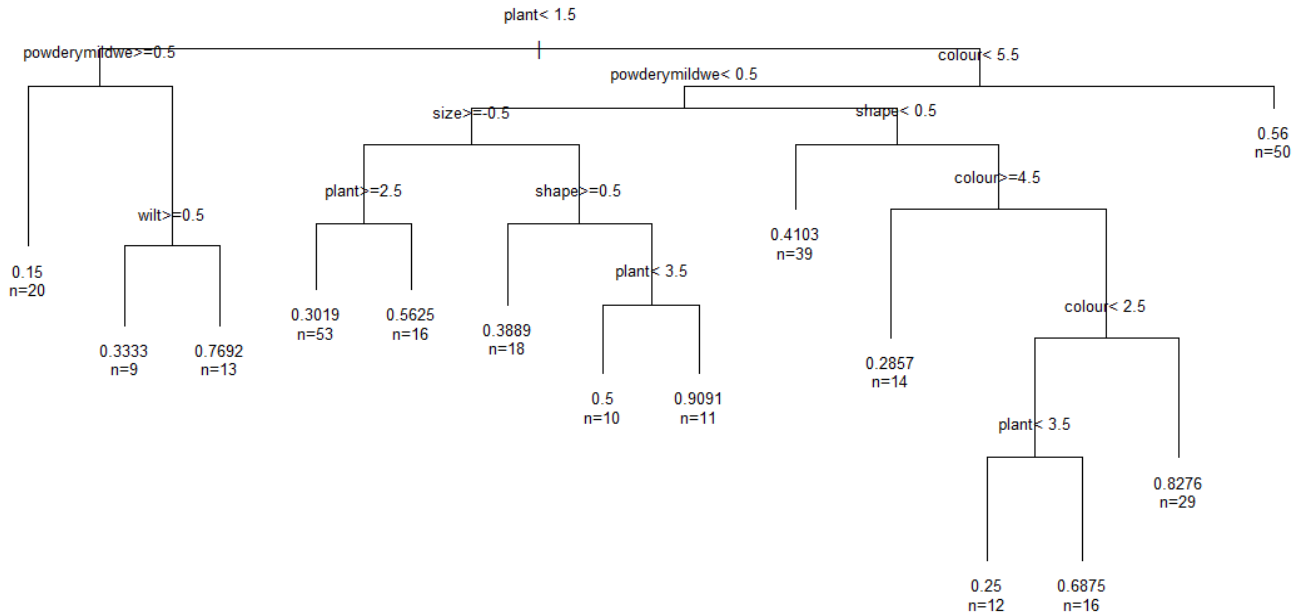

Figure 1: Decision Tree based on Symptoms

```
Overall Statistics

               Accuracy : 0.9782609
                 95% CI : (0.9236693, 0.9973564)
    No Information Rate : 0.1956522
    P-Value [Acc > NIR] : < 0.00000000000000022204

                  Kappa : 0.9738413
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: APPLE  Class: BLUE BERRY Class: CHERRY Class: POTATO
Sensitivity             1.0000000         0.9411765     1.0000000     1.0000000
Specificity             1.0000000         1.0000000     1.0000000     0.9870130
Pos Pred Value          1.0000000         1.0000000     1.0000000     0.9375000
Neg Pred Value          1.0000000         0.9868421     1.0000000     1.0000000
Prevalence              0.1304348         0.1847826     0.1956522     0.1630435
Detection Rate          0.1304348         0.1739130     0.1956522     0.1630435
Detection Prevalence    0.1304348         0.1739130     0.1956522     0.1739130
Balanced Accuracy       1.0000000         0.9705882     1.0000000     0.9935065
                     Class: STRAW BERRY Class: TOMATO
Sensitivity                   0.9333333     1.0000000
Specificity                   1.0000000     0.9870130
Pos Pred Value                1.0000000     0.9375000
Neg Pred Value                0.9871795     1.0000000
Prevalence                    0.1630435     0.1630435
Detection Rate                0.1521739     0.1630435
Detection Prevalence          0.1521739     0.1739130
Balanced Accuracy             0.9666667     0.9935065
```

Fig 2: Decision tree accuracy

The decision tree has been produced (Figure 1) according to the root node selected. Here the root node is HEALTHY. Figure 2 shows the accuracy of the decision tree.

3.2 K-Means Clustering [8]:

It is for partitioning where each cluster center is represented by the mean value of objects in cluster. The input is data set and number of clusters. The output is set of k clusters. The algorithm randomly selects k objects as initial clusters for the remaining objects, an object is assigned to cluster to which it is most similar based on equidistance between cluster object and cluster centers. K-means algorithm improves by computing new mean values or centroids all the objects are again reassigned using the updated means.

Given a set of observations ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$arg_S min \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 = arg_S min \sum_{i=1}^{k} |S_i| \, var S_i$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$arg_S min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x,y \in S_i} ||x - y||^2$$

The equivalence can be deduced from identity $\sum_{x \in S_i} ||x - \mu_i||^2 = \sum_{x \neq y \in S_i} (x - \mu_{i)}(\mu_i - y)$. Because the total variance is constant, this is also equivalent to maximizing the sum of squared deviations between points in *different* clusters (between-cluster sum of squares, BCSS),[1] which follows easily from the law of total variance.

The experimental results is using k-mean cluster are obtained as follows

```
K-means clustering with 6 clusters of sizes 50, 60, 58, 50, 50, 42

Cluster means:
        [,1]
1  0.8156449
2  1.4009457
3 -0.9402573
4 -0.3549566
5  0.2303442
6 -1.5255581

Clustering vector:
  [1] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
 [43] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [85] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[127] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[169] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 1 1 1 1 1 1 1 1 1
[211] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2
[253] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[295] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 6.162976e-29 1.449532e-28 6.434147e-30 5.546678e-30 3.466674e-31 0.000000e+00
 (between_SS / total_SS = 100.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Figure 3: K-Means

3.3 Naive Bayes[8]

Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability and the probabilities of observing various data.

$$P(h/D)=(P(D/h)*P(h)$$

- A concept learning algorithm considers a finite hypothesis space H defined over an instance space X. The task is to learn the target concept C:X--->[0,1], The learner gets a set of training examples. Brute force bays concept learning algorithm finds the maximum a posterior hypothesis.

R programming implementation of naïve-bayes classifier for plant disease prediction is as follows

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.5299539 0.4700461

Conditional probabilities:
   plant
Y        [,1]     [,2]
  0 3.521739 1.773824
  1 3.745098 1.669092

   colour
Y        [,1]     [,2]
  0 3.626087 1.893946
  1 3.813725 2.018383

   spots
Y        [,1]      [,2]
  0 0.4869565 0.5020173
  1 0.4313725 0.4977137

   dry
Y        [,1]      [,2]
  0 0.5043478 0.5021692
  1 0.4705882 0.5015991

   size
Y         [,1]      [,2]
  0 0.13913043 0.8044499
  1 0.07843137 0.7795184
```

Figure 4: Naive Bayes

Naïve Bayes is performed, and every result is obtained by using the numerical data and doing the hypothesis.

3.4 Random Forest[9]

In this algorithm there are two steps, in the first stage a random forest creation is produced and in the second step the values

that are produced from step one are used and predictions are made. Here select m features from n features where m<<n.

Among the m features calculate the node d which has the best split point. Later split the node into child node and repeat the above steps until k number of nodes has been reached. Build the forest by repeating all the above steps z times to produce z number of tree.

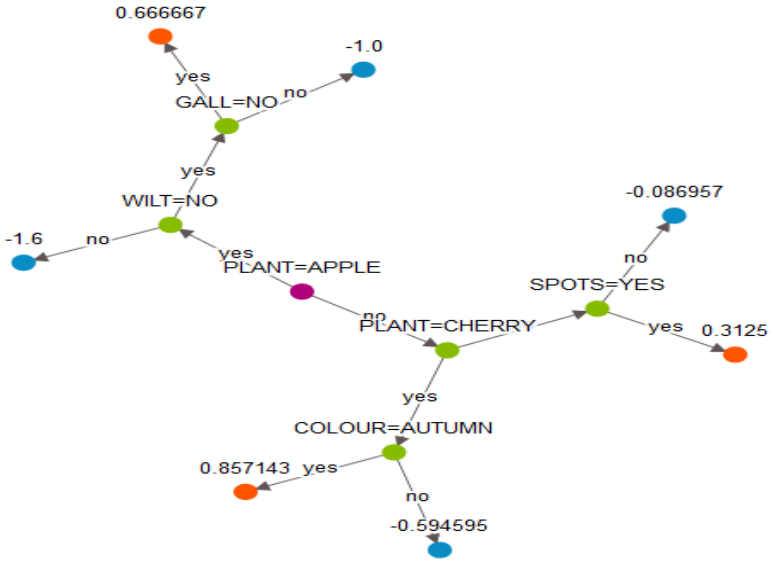

Figure 5: Random Forest Prediction 1

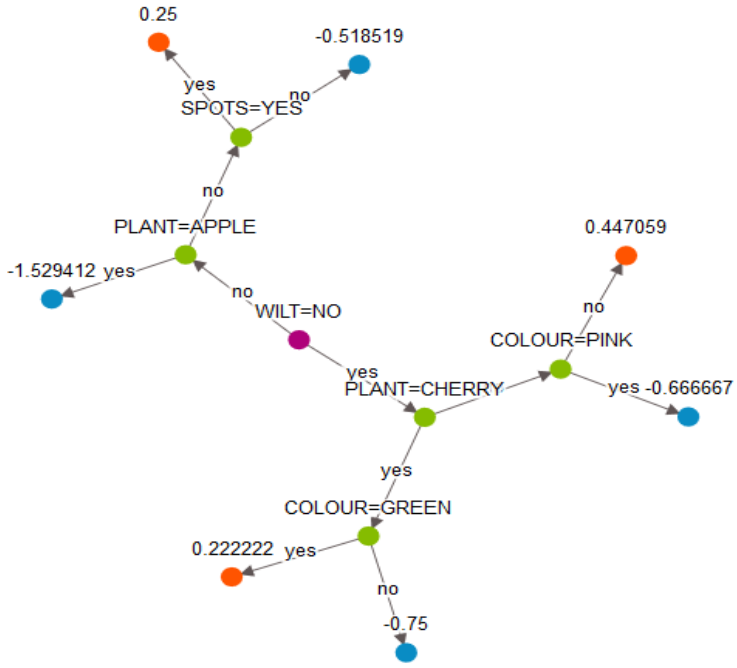

Figure 6: Random Forest Prediction 2

In Figure 5, Figure 6 different trees has been generated with the help of Random Forest algorithm and final prediction is made on overall dataset

3.5 Neural Network[9]

Let y be the correct output, and f(x) the output function of the network.

Error: E =y-f(x) , Update weights: w_j<-$x_j\alpha$+w_j

For the total output: O_i=g($\sum_j w_j, a_{ij}$)

At first initialize the weights and threshold and later perform the following steps over the input and desired output dj.

Then calculate the actual output

$$y_j(t) = f|w(t).x_j|$$

$$=f|w_0(t)x_{j,0} + w_1(t)x_{j,1} + w_2(t)x_{j,2} + \cdots + w_n(t)x_{j,n}|$$

Then later weights are updated

$$w_i(t+1) = w_i(t) + r.(d_j - y_j(t))x_{j,i}$$

Here in the artificial neural network the output has same variable named HEALTHY and the inputs varies. The input in Figure 7 is plant, color, spots. But Figure 8 has four different inputs color, gall , powdery mildew, wilt.

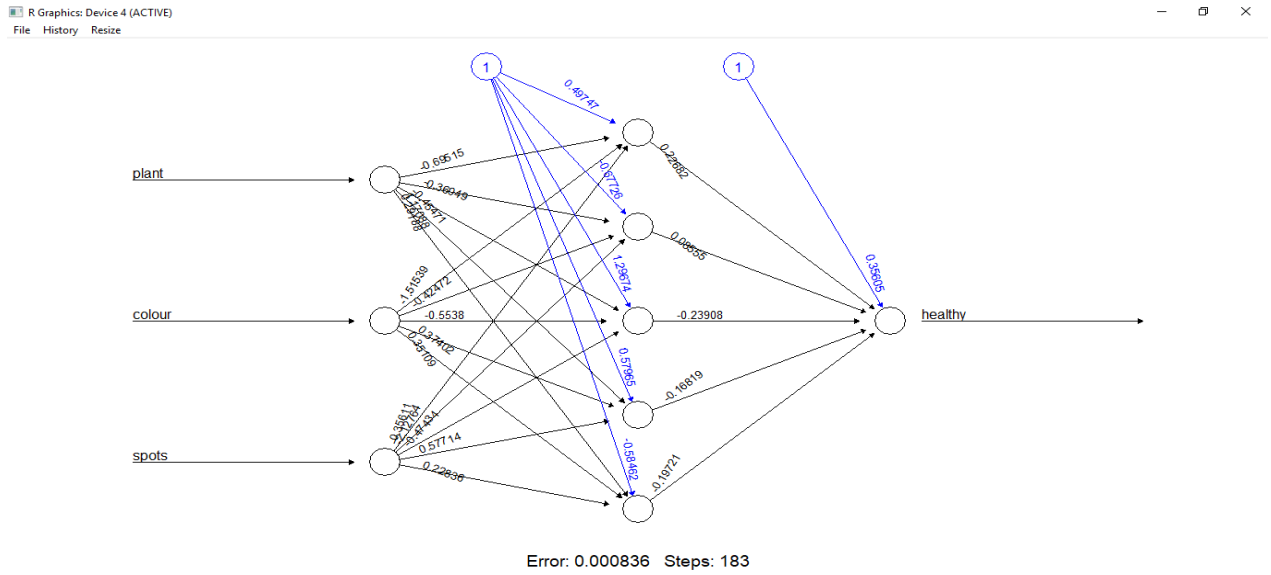

Figure 7: Neural Network with three parameters

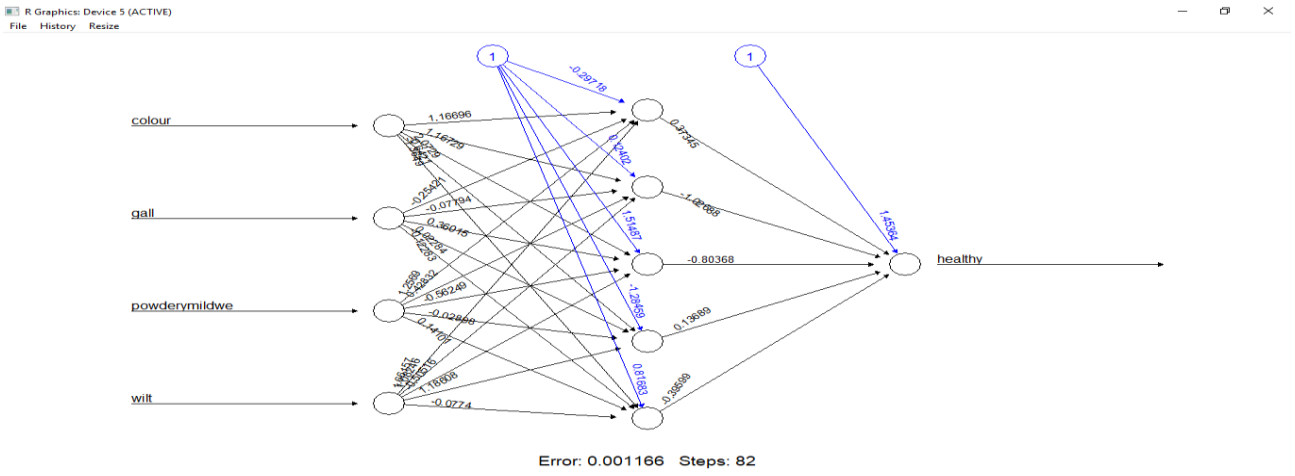

Figure 8: Neural Network with four parameters

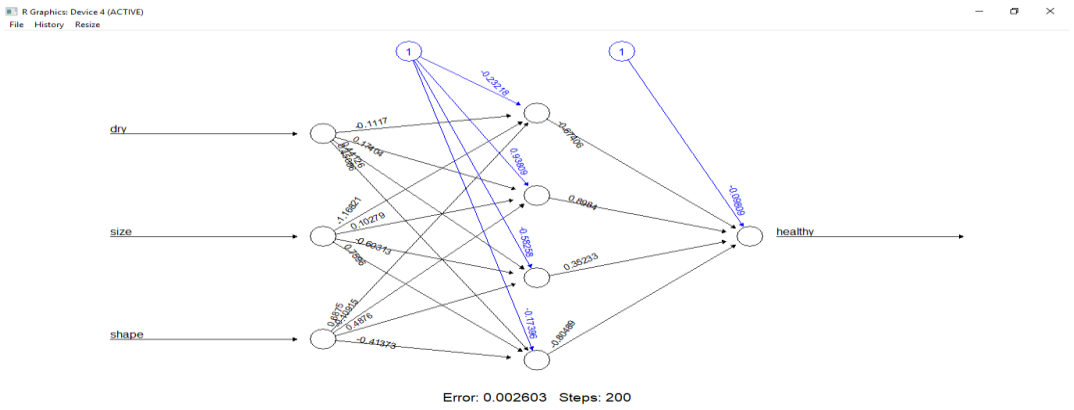

Figure 9: Error Calculation

Here how many steps are needed and also the error while calculating the path is shown.

The error does not rise with number of inputs this can be proved with the Figure 8 and Figure 9.

4. CONCLUSION

In the above information about prediction of plant disease a dataset is taken with 11 attributes and 310 rows. With the data different techniques like decision tree, Naive Bayes, Neural network and different plots like box plot, bar plot are performed. By performing the techniques it is concluded that,

the accuracy of given data in Decision tree, Naive Bayes, SVM , Neural Network we got accuracies as show in Table 1.

Table 1: Result Analysis

Methods	Accuracy
Decision tree	89.97-97.83
Naive Bayes	83-.98-86.76
Neural Network	89.93

Here some statistical tests are performed to find out possible output predictions with the help of some inbuilt dataset and compared with the techniques with the help of the data set. While techniques are compared, other techniques may be better or poor. But the accuracy differs from each sample taken from dataset.

Future work will involve in image processing and spreading the usage of the model by training it for plant disease recognition on wider land areas, combining aerial photos of captured by drones and convolution neural networks for object detection. And the work may be on the shadowing techniques which is produces best results. By extending this research, there is a hope to achieve a valuable impact on sustainable development, affecting crop quality for future generations.

5. REFERENCES

[1] Sanjeev S Sannakki, Vijay S Rajpurohit, V B Nargund, Arun Kumar R, Prema S Yallur. 2011. Leaf Disease Grading by Machine Vision and Fuzzy Logic. Gogte Institute of Technology

[2] Savita N. Ghaiwat, Parul Arora. 2014. Detection and Classification of Plant Leaf Diseases Using Image processing Techniques:

[3] Signs and symptoms of plant disease: Is it fungal, viral or bacterial?, Michigan State University Extension.

[4] Kimberly Leonberger, Kelly Jackson and Robbie Smith, Nicole Ward Gauthier. Plant Diseases. University of Kentucky College of Agriculture, Food and Environment.

[5] Online reference- Wikipedia wiki/ID3_algorithm, wiki/K-means_clustering,

[6] Online reference- Wikipedia wiki/Perceptron

[7] Tom M.Mitchell .1997.”Machine Learning”,McGraw Hill.

[8] Ethem Alpaydin. 2010. “Introduction to Machine Learning”, The MIT Press

[9] Stephen Marsland. 2009. “Machine Learning an Algorithmic Perspection”, CRC Press.