Comparative Analysis of Random Forest and Caret Algorithm for Prediction of Crop Yield

Aishwarya R
CS&E department
AIT college, Chikkamagalur

Amulya H O
CS&E department
AIT college, Chikkamagalur

Nidhi Dixith H R
CS&E department
AIT college, Chikkamagalur

Spandana K M
CS&E Department
AIT college, Chikkamagalur

Dr.Pushpa Ravikumar
Professor and head, CS&E department,
AIT college, Chikkamagalur

***Abstract*: Crop yield of a country has a direct impact on the economy of that country as well as on its overall development. It plays a very important role in determining every aspect that effects the growth of a country i.e., from insufficient yield to the economic trading of crops. So to understand the various factors that affects the crop yield as well as in order to predict the yield of the crop based on these factors we make use of R tool and various packages that are available in it, such as randomForest and caret. In this approach, we consider the important factors that has the direct effect on the crop yield for analysis and prediction. For example, the basic factors on which the yield of crops depends are rainfall, temperature, humidity, soil characteristics, fertilizers, wind speed, sunshine and also on the farming practices. These factors varies for various crops as well as for various locations. In this context we consider rice, the most likely grown crop for analysis and prediction. The factors are considered with respect to rice for better understanding and also for easy evaluation. With the help of R tool we determine the error in the prediction through which the deviation can be understood easily, for example, low rainfall due to drought in some areas can be a reason for low yield when all the other factors are in favour. We can also determine the main factors that has a great influence on the yield.**

Keyword: R tool, randomForest, crop yield, prediction, analysis.

I. INTRODUCTION

Rice is one of the dominant crop that is grown in most of the countries and India stands second in the contribution of rice production next to China. Each crop requires various kinds of external and internal factors. For example, some crop requires rainfall less than 1000mm while some other may require more than 1500mm. Some crop may require less humidity while some other may require more humidity. So in order to understand the efficiency of this analysis and the prediction, we consider rice for the analysis and the same technique can be applied for other crops by collecting the data of the factors that has the influence on that particular crop.

In context to Rice we consider the factors shown in Table 1.1 to understand the range of values that has an implicit effect on its yield. The deviation from these range may lead to deviation from the expected crop yield. So in order to predict the yield of rice the data are collected from past circumstances which gives a clear cut idea of how the variance of the attributes has a direct impact on the yield. These data are collected and stored in an organized way and the evaluation is done on them to understand the influence of these data, so that the future analysis can be carried based on these references.

II. LITERATURE SURVEY

In this section, related literature about rice yield analysis, Random forest, crop yield prediction strategies will be reviewed and discussed.

A. Rice yield analysis

Rice is a more adversely grown crop whose yield varies for various location and it is also based on the climatic condition of that region. Rice is the seed of a monocot plant and it is the important staple food for large part of human population. It is also found that rice is the third widely grown cereal grain next to maize and wheat. Determining better techniques for the rice yield would help to assist the farmers and other stake holders in order to make better decisions which has been determined in [6].

B. Random forest

In order to predict the crop yield more accurately random forest is used which is a standard and supervised machine learning algorithm. Random forest provides a various kinds of algorithms to make better understanding of the results that has to be obtained. It helps to plot multi-dimensional scaling for proximity matrix, extract a single tree in the randomForest as explained in [7].

C. Crop yield prediction

Yield of a crop mainly depends on the function on the resources as well as on the conditions around it. The factors such as temperature, humidity, rainfall, sunshine plays a very important role in determining the yield of a crop. In addition to making use of the historic data for analysis the future prediction can also be done using the algorithm. These predictions helps to understand the necessary steps that has to be taken in order to increase the yield of a crop [5].

III. METHODOLOGY

The complete analysis is carried out based on following steps:

i. Read data and partitioning data.
ii. Classification using random forest.
iii. Confusion matrix and statistics.

Crop Yield	High	Medium	Low
Rainfall(in mm)	1400-1800	1000-1400	<1000
Temperature(in 'C)	21-37	16-24	<15
Humidity (in %)	60-80	40-60	<40
Soil pH	6-7	4-6 0r 7-8	<4 or >8
Nitrogen (in kg)	50	40-50	<40
Phosphate(in kg)	25	16-24	<15
Potash(in kg)	25	16-24	<15
Sunshine(in hours)	>300	200-300	<200

Table 1.1: Attributes and range of values that influences rice yield

A. Reading and partitioning the data

Data is stored in a file with the crop yield as a major attribute. The data is read from the file and viewed in the R environment as shown in the figure 2.1. The data can also be used to classify and determine the number of objects that belong to a particular category as shown in figure 2.2. During partitioning the data is grouped into 2 divisions. One partition specifies the objects that is to be evaluated while the other refers to the set of objects with respect to which the test is carried out. In this analysis the partition is done by 70/30. Where the 70% of the objects are used for training and the rest 30% is used for testing. The analysis is carried out on small set of data which is the test data set and based on the result of these data an estimation is done for the training data set.

```
'data.frame':   500 obs. of  9 variables:
 $ CropYield  : Factor w/ 3 levels "high","low","med": 1 3 2 1 2 1
2 2 3 1 ...
 $ Rainfall   : int  1254 1278 1787 1265 1676 1356 578 563 1345
1786 ...
 $ Temperature: int  28 20 13 23 11 29 13 12 17 35 ...
 $ Humidity   : int  66 53 34 77 23 58 36 25 41 67 ...
 $ SoilpH     : num  4.2 4.9 3.6 6.1 3.3 6.2 3.2 8.9 5.5 6.9 ...
 $ Nitrogen   : int  52 43 31 54 33 58 23 32 40 58 ...
 $ Phosphate  : int  23 16 3 28 14 29 14 13 18 29 ...
 $ Potash     : int  25 21 12 29 14 28 13 12 24 28 ...
 $ Sunshine   : int  290 210 50 456 132 367 20 123 256 343 ...
```

Fig 2.1: Data read from data file and displayed on the R environment.

```
High  low  med
 184  164  152
```

Fig 2.2: Classification of objects based on the attribute values of training data set.

B. Classification using random forest

In this module we make use of the randomforest package in order to determine the classification of objects and also to find out the OOB (Out Of Bag) estimate of the error rate. It is more likely to give the accuracy in the measurement as it shows the deviation, which in turn explains the objects that do not satisfy the required conditions. For example, sometimes the crop yield can be high even when the rainfall in below the expected range because of some reasons like climate fluctuation or sometimes there might be a change in the soil pH or any other external factors. In such situations these objects has to be discarded from the analysis. The error has to be very small to carry out the analysis otherwise the probability of prediction going wrong will be high. The OOB estimation and the confusion matrix is as shown in figure 3.1.

```
   OOB estimate of  error rate: 0.84%
Confusion matrix:
     high low med class.error
high  135   1   0 0.007352941
low     0 121   1 0.008196721
med     0   1  98 0.010101010
```

Fig 3.1: OOB estimation and confusion matrix obtained using randomForest.

C. Confusion matrix and statistics

In this stage we can obtain the reference matrix and also statistics for both the training as well as test data set. It provides the accuracy and various other details such as sensitivity in measurement, specificity etc. Figure 4.1 shows the reference matrix and the statistics of training data set. A similar matrix and statistics can also be obtained for a test data set as shown in figure 4.2. The reference matrix provides the objects number that satisfies the required condition. In case of training data set it can be observed that there are 136 objects that satisfies the condition of factor values required for high yield. Similarly, there are 122 objects that provides result for low yield and 99 objects which gives an average yield.

```
                Reference
Prediction high low med
     high  136   0   0
     low     0 122   0
     med     0   0  99

Statistics by Class:
                     Class: high  Class: low  Class: med
Sensitivity                1.000      1.0000      1.0000
Specificity                1.000      1.0000      1.0000
Pos Pred Value             1.000      1.0000      1.0000
Neg Pred Value             1.000      1.0000      1.0000
Prevalence                 0.381      0.3417      0.2773
Detection Rate             0.381      0.3417      0.2773
Detection Prevalence       0.381      0.3417      0.2773
Balanced Accuracy          1.000      1.0000      1.0000
```

Fig 4.1: Reference matrix and statistics of training data set

In case of the test matrix very low number of data set is considered and hence evaluation is easy to carry out. In this data set there are about 48 objects that gives a high yield and about 42 objects with low yield and 53 with average. Sensitivity can be obtained which gives us the approximation in the accuracy. As the error rate is

relatively very small for this huge data set, it can be seen that each high low and med class has a sensitivity equal to 1. It shows that the value that has been considered is about 100% accurate. If the deviation is quite more, then refining of the data set has to be done in order to avoid future risks such as very high deviation in the prediction or fluctuation in the analysis that has been carried out.

```
          Reference
Prediction high low med
     high   48   0   0
     low     0  42   0
     med     0   0  53

Statistics by Class:

                     Class: high Class: low Class: med
Sensitivity               1.0000     1.0000     1.0000
Specificity               1.0000     1.0000     1.0000
Pos Pred Value            1.0000     1.0000     1.0000
Neg Pred Value            1.0000     1.0000     1.0000
Prevalence                0.3357     0.2937     0.3706
Detection Rate            0.3357     0.2937     0.3706
Detection Prevalence      0.3357     0.2937     0.3706
Balanced Accuracy         1.0000     1.0000     1.0000
```

Fig 4.2: Reference matrix and statistics of test data set

IV. RESULTS AND DISCUSSION

A. *Determining Error rate*

In this step a graph is plotted in order to determine the error in the data that has been considered. This might have taken place due to mistyping or due to fault during collection of data or may be because of other reasons such as variation in the factors. In this plot an estimation can be clearly made about the error. If the error is too small then it can be neglected and the further processing can be carried out, otherwise steps must be taken to reduce these errors in order to avoid future problems. The plot for the data set considered for rice is as shown in figure 5.1.The straight line at the error rate of about 0.01 shows that there is a very small error which becomes constant after the 15th tree. So hence it is not expected to create much problem in the analysis. Any variation in the estimation of error such as finding a large error rate as the number of tree grows intends to create a high variation in the estimation of crop yield.

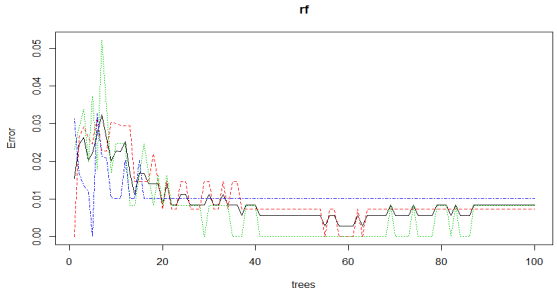

Fig 5.1: Plot of error versus number of trees.

B. *Tuning processes*

This step is carried out in order to estimate the algorithm that is efficient for the analysis. This yields a graph which provides a standard for determining the value at which the error rate is very less. This helps in refining the algorithms that are provided for the analysis and selecting the value that provides least error. Fig 6.1 shows the result of using tuning to determine the value with less error.

```
mtry = 2  OOB error = 19.42414
Searching left ...
mtry = 4        OOB error = 18.67981
0.03831976 0.05
Searching right ...
mtry = 1        OOB error = 22.84844
-0.176291 0.05
```

Fig 6.1: Result obtained for tuning process

A plot can be obtained for the same result for a better understanding of the variation in error based on the mtry value as shown in figure 6.2.We can observe that the OOB error goes on decreasing as the mtry value increases. For mtry value 1 it is at the peak and it becomes relatively less when the mtry value is 2 and it finally decreases when the mtry value increases to 4. The OOB error has a range of approximately 19-23 which refers to the objects that has a deviation in the value from that of the expected range.

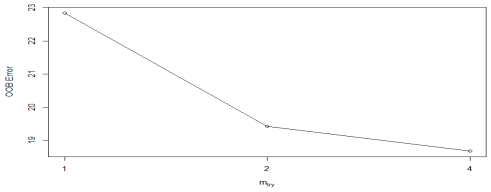

Fig 6.2: Plot for Error estimation based on tuning method

C. *Determining number of nodes for the tree*

The optimum frequency of the tree is obtained when the tree size becomes 10. This means that the number of nodes is more when the tree size reaches 10. A graph can be plotted which shows the number of nodes for different tree size with respect to the frequency.

A maximum number of nodes can be obtained when the tree size ranges from 10 to 12. This provides a reference to the way in which the classification is carried out based on the values in the data set. The graph representing the number of nodes of the trees is as shown in the figure 7.1.

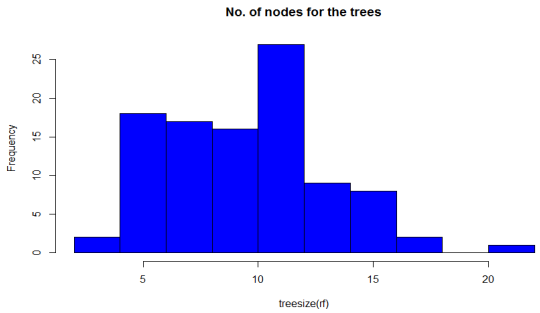

Fig 7.1: Bar chart representing the number of nodes of a tree based on the tree size.

D. Identifying variable importance

In this step the importance of the factors affecting the crop yield can be determined with respect to the mean decrease accuracy as well as the mean decrease in gini coefficient as shown in figure 8.1. It provides the idea about the factors that has a direct impact on the measurement of the crop yield and how each variable contributes to homogeneity of nodes and leaves.

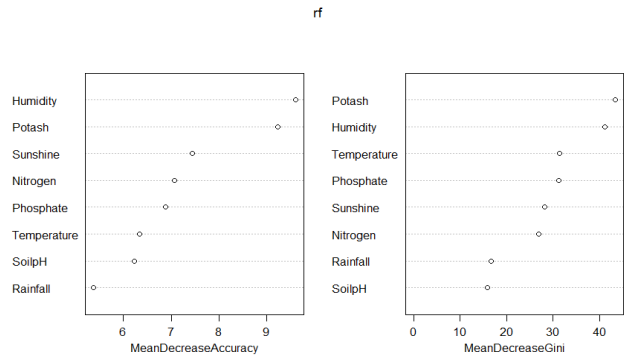

Fig 8.1: Plot to determine the influence of factors on the crop yield.

In the graph (8.1) it is clear that the humidity plays a very important role in the accuracy when compared to other attributes while in case of mean decrease in gini coefficient, potash has a high value. In case of mean decrease in accuracy rainfall has the least effect which shows that there is a wide change in the rainfall range for the same class. This may be due to the variation in climate from one location to another location or may be because of some unexpected condition such as cyclone or drought. In case of gini coefficient soil pH has the least position which shows that purity of split for this variable is comparatively very less.

E. Determining partial dependency plot

In this stage analysis is made on each factors that affect the yield of the crop. It helps to understand the fact that some factors have a direct impact on the yield while some may have deviation. Figure 9.1 shows that the yield is high when the fertilizer used for the rice is having nitrogen content greater than 50kg. But in figure 9.2 it can be observed that it is little difficult to obtain the particular range required for high crop yield based on rainfall as variation can be seen, however a peak value is reached when the rainfall is about 1400mm .

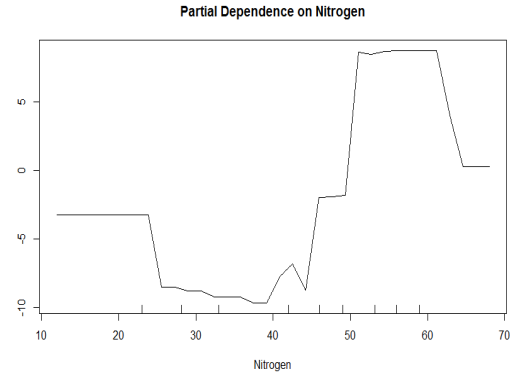

Fig 9.1: Plot of partial dependence of crop yield on nitrogen

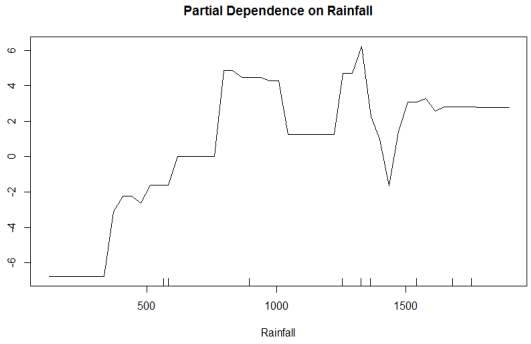

Fig 9.2: Plot of partial dependence of crop yield on rainfall

Hence various factors may have various kind of effect on the yield of a crop. Some may have a direct influence on it while the other may not have much impact. Hence an analysis for each factor can be carried in the similar way and final conclusion can be obtained about the factors and their range of values during which a high yield can be expected.

In case of phosphate and potash high yield can be expected when the fertilizer used contains phosphate and potash with minimum amount of 25kg each. The plot for these two variables appears almost similar as shown in figure 9.3

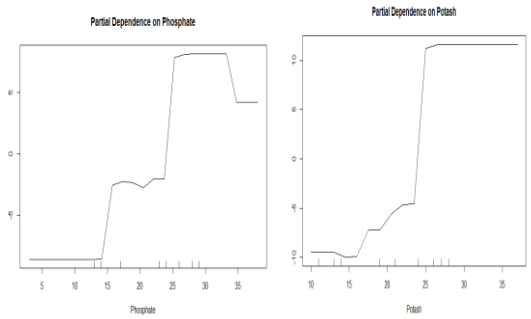

Fig 9.3: Plot of partial dependency of two variables that provides high yield in the same range.

F. Extraction of single tree

A data of a single tree can be extracted and an analysis can be carried on based on the values. For example if the status value is set to 1 then it implies that there exists a sub tree for that particular node. If the value is -1 then it shows that

the node is a terminal node and it can be observed that the left daughter and right daughter of such node would be set to 0 as shown in figure 10.1.

Split variable can also be obtained which shows the variable or the attribute with respect to which the splitting of the tree has been carried. With respect to the tree that has been considered in figure 10.1 we can observe that the first splitting of the tree has occurred by the splitting variable 'temperature', with a split point of about 15.5. It can also be observed that the terminal nodes with status value -1 has no split variable and split point and it is set to not applicable i.e., <NA>. It can also be observed that the prediction can be made only for the terminal nodes. This data about the tree can be made used to obtain the category of the yield that has been expected, i.e.., whether the yield is high, low or at an average level.

left	right	split var	split point	status	prediction
1	2	3 Temperature	15.500	1	<NA>
2	4	5 Rainfall	777.000	1	<NA>
3	6	7 Humidity	59.500	1	<NA>
4	8	9 Nitrogen	55.000	1	<NA>
5	10	11 SoilpH	5.085	1	<NA>
6	12	13 Rainfall	1645.000	1	<NA>
7	0	0 <NA>	0.000	-1	high
8	0	0 <NA>	0.000	-1	low
9	0	0 <NA>	0.000	-1	high
10	0	0 <NA>	0.000	-1	low
11	14	15 Humidity	49.500	1	<NA>
12	16	17 Temperature	21.500	1	<NA>
13	18	19 Nitrogen	40.500	1	<NA>
14	20	21 Rainfall	1427.500	1	<NA>
15	0	0 <NA>	0.000	-1	high
16	0	0 <NA>	0.000	-1	med
17	22	23 Nitrogen	34.500	1	<NA>
18	0	0 <NA>	0.000	-1	low
19	0	0 <NA>	0.000	-1	high
20	24	25 Sunshine	236.000	1	<NA>
21	0	0 <NA>	0.000	-1	low
22	0	0 <NA>	0.000	-1	low
23	26	27 Humidity	53.500	1	<NA>
24	0	0 <NA>	0.000	-1	low
25	0	0 <NA>	0.000	-1	high
26	0	0 <NA>	0.000	-1	med
27	28	29 Rainfall	1365.500	1	<NA>
28	0	0 <NA>	0.000	-1	high
29	0	0 <NA>	0.000	-1	med

Fig 10.1: Data extracted for the first tree

G. Multidimensional scaling plot of proximity matric

Here we can observe the cluster of objects which depicts a particular class. We can make use of this observation to determine the variance in the data set.

For example, an object that is expected to produce a high yield might have produced a low yield due to some difference in the external or the internal factors. These are nothing but the error that are estimated in the previous stages. However, these deviations are neglected as they are small in number when compared to the data set that has been used for the analysis.

In the figure 11.1 it can be observed that the cluster of objects expected to have low yield are scattered between the low and medium yield range and it can also be observed that there is no much variation in the high yield except for only one object and the medium yield is expected to be comparatively more deviating as it is spread across low as well as high yield. So, it can be stated that more error rate can be expected in the medium yield analysis

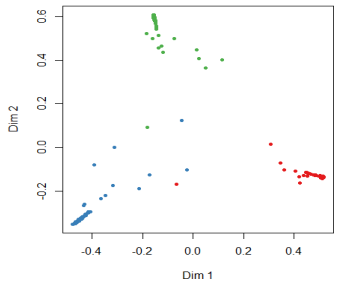

Fig 11.1: Cluster of objects showing the distribution in various classes

V. CONCLUSION

From the analysis and the prediction strategies that has been carried out for rice we come to know about the fact that each factor has its own influence on the yield of the rice and a small deviation in one variable may result in a wide change in the analysis. As mentioned earlier the same strategy can be applied for various other crops in order to determine their yield. The only difference is that the factors and the values are considered with respect to the crop about which an analysis is to be made. This helps in understanding the various factors that affect the crop yield and taking corrective measures in order to increase the yield of the crop so that it finally results in the overall development of a country as well as it satisfies one of the basic needs for living.

REFERENCES:

[1] Crop yield predictions- High yield statistical model for intra season forecasts applied to corn in the US.
[2] Rice crop yield forecasting using Random Forest Algorithm
[3] Package 'random forest'.
[4] S. Bejo, S. Mustaffha and W. Ismail, "Application of artificial neural network in predicting crop yield: A review", Journal of Food Science and Engineering, vol. 4, pp.1-9, 2014.
[5] M. S. Rasmussen, Operational yield forecast using AVHRR NDVI data.
[6] D. B. Lobell, M. B. Burke, on the use of statistical models to predict611 crop yield responses to climate change, Agric.
[7] K. Dzotsi, B. Basso, J. Jones, Development, uncertainty and sensitivity analysis.
[8] F. A. Vogel, G. A. Bange, et al., Understanding USDA crop forecasts.