# Soil Data Analysis and Crop Yield Prediction in Data Mining using R – Programming

**K. Samundeeswari, K. Srinivasan**

*Abstract: Data mining is better choices in emerging research filed- soil data analysis. crop yield prediction is an important issue for selecting the crop. earlier prediction of crop is done by the experience of farmer on a particular type of field and crop. predicting the crop is done by the farmer's experience based on the factors like soil types, climatic condition, seasons, and weather, rainfall and irrigation facilities. data mining techniques is the better choice for predicting the crop. the analysis of soil plays an important role in agricultural filed. soil fertility prediction is one of the very important factors in agriculture this research work implements to predict yield of crop, decision tree algorithm is used to find yield. the aim of this research to pinpoint the accuracy and to finding the yield of the crop using decision tree and c 4.5 algorithm is used to predict the yield of crop using r-programming and also to find range of magnesium found in the collected soil data set. this prediction will be very useful for the farmer to predict the crop yield for cultivation*

*Keywords: Crop Yield Prediction, Decision Tree Algorithm, C5.0, Data Mining, R-Programming.*

## I. INTRODUCTION

Data Mining is a process of extracting hidden information from a database and transforms it into useful information. Data Mining is used to find desired patterns from huge data sets and create useful classifications from the data sets. In agricultural problems Data Mining is widely used now a day. Data mining is a tools used to predict crop yield and Agricultural output

In India, it is common that farmers are not getting the expected crop yield due to some reasons. The agricultural yield is mainly depends on weather conditions. In this perspective, the farmers necessarily requires a appropriate advice to predict the future crop productivity and an study is to be made in order to help the farmers to capitalize on the crop production in their crops.. Agriculture is the main and important back-bone of Indian economy.

**K.Samundeeswari\***, Department of Computer Science, Govt. Arts College for Women, Krishnagiri - 635 001. Tamil Nadu, India. Email: samun.npgs18@gmail.com

**K.Srinivasan**, Department of Computer Science,, Periyar University Constituent College of Arts &Science Pennagaram, Dharmapuri - 636 803, Tamil Nadu, India. Email: vasanphdrk@gmail.com

The soil testing is a process that starts from the collected soil sample. The principle of soil testing is field is sampled using chemical analysis of the soil to find the true nutrient status. The soil is analyzed to physical properties and chemical properties such as EC, pH, along with micro nutrients N, P, K and macro nutrients Zn, B, Cu, Fe. The classification technique applied to soil data set, the results of classification is used to find the cultivation pattern. Classification method is applied to analyze the soil properties. The essential nutrient required for crop growths is determined during soil testing process. The aim of this research work is to find the accuracy of decision tree algorithm and C 4.5 algorithm to predicate the crop yield based on available data. Here we focus on decision tree algorithm – classification Method is used to for classify unidentified samples into identified samples.

## II. DATA MINING TECHNIQUES

This research work use agricultural raw soil data sets collected from district agricultural office at Krishnagiri. This soil data set is preprocessed, analyzed and implemented by using R- programming. Decision tree algorithm are used for given soil data set with the following steps,

1. Input - Agricultural data
2. Pre-processing – removing unwanted data or adding necessary information to build data mining easier.
3. Pre-processed data - input for implementation.
4. Creation of datasets - from available data sets (i.e) training data and testing data.
5. Decision tree algorithm and C4.5 algorithm is applied.

The soil dataset used for this research work are obtained for the year 2017 from few Taluks of Krishnagiri district of Tamil Nadu in India. The data collection is preliminarily carried out for all the Taluks of Krishnagiri district of Tamil Nadu in India. For soil optimum health Macronutrients and micronutrients combination is essential .A lack of any one micronutrients in the soil will limit growth, when even other nutrients are present in sufficient amounts.

### A. Decision Tree Algorithm

Decision Tree algorithm is supervised learning algorithm, it is used for solving classification and regression problems and also used to create a training data set which is used to predict crop yield inferred from prior data (training data).

*Retrieval Number: C8683019320/2020©BEIESP*
*DOI: 10.35940/ijitee.C8683.019320*
*Journal Website: www.ijitee.org*

1857

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

# Soil Data Analysis and Crop Yield Prediction in Data Mining using R – Programming

The decision tree algorithm is used to solve the problem, by using tree representation.

## B. C 4.5 algorithm

C4.5- Decision Tree algorithm is the popular and effectively used classifier for crop yield in soil data set classification in present study. The decision trees generated by C 4.5 is used for classification C4.5 builds decision trees from training data using the concept of information entropy.

## III. EXPERIMENT AND RESULTS

The experiment is done on the real world soil data sets obtain from district agricultural office at Krishnagiri District in Tamil Nadu. Soil data sets taken in this work have enough amounts of macronutrients and micronutrients. The soil data sets used in this work consists of 249 tuples with 24 attributes the tuples of soil dataset define the availability of nutrients(N,P,K), pH, Ec, fertilizer recommendation(N,P$_2$O$_5$,K$_2$O,urea,super phosphate, muriate of potash) micronutrientscontents (Fe,Zn,Mn,Cu) and MicroNutrient Recommendation(FeSo4,ZnSo4,MnSo4,CuSo4) in soil.

**Attributes and description of input variables**

### TABLE 1: Input Variables.

| Attributes | Description |
|---|---|
| **Macronutrients** | |
| Ca | Calcium |
| Mg | Megnisium |
| N | Nitrogen |
| P | Phosphorus |
| K | Potassium |
| SO4 | Sulphur |
| **Fertilizer recommended** | |
| P2O5 | Phosphorus Pentoxide |
| K2O | Potassium Oxide |
| FeSo4 | Iron Sulphate |
| Zn SO4 | Zinc Sulphate |
| Mn | Magamese Sulphate |
| CuSo4 | Copper Sulphte |

### TABLE 2: Input Variables – Macronutrients.

| Attributes | Description |
|---|---|
| PH | potential hydro gen |
| EC | Electrical Conductivity |
| **Micronutrients** | |
| Fe | Iron |
| B | Boran |
| Mn | Maganese |
| Zn | Zinc |
| Cu | Copper |

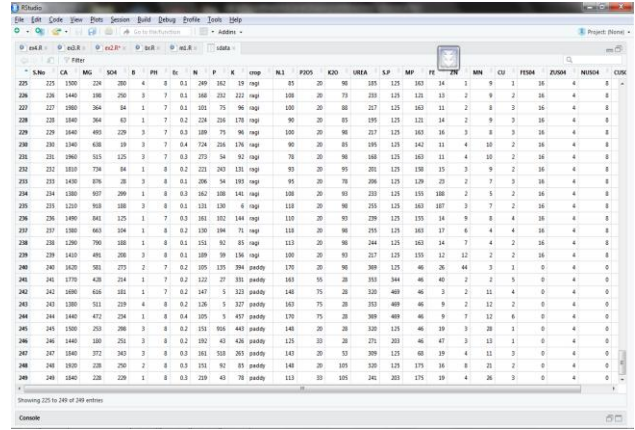The collected dataset has organized in Excel Sheet in CSV file format which shows in Figure 1



**Figure 1: Soil dataset in CSV format**

The summary of the soil dataset is used to find the accuracy of the algorithm, mean, median, minimum, maximum, first quartile and third quartile values are calculated for all the twenty four input variables.
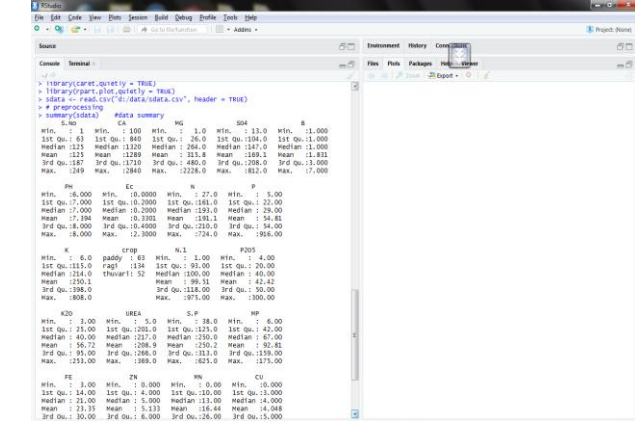


**Figure 2: Summary of Soil dataset**

## A. Density of Mg

Density of Mg is calculated to find amount of Mg present in soil, Mg is one of the macronutrients present in soil, it is used to find the efficiency and deficiency of the soil for cultivation of crop.
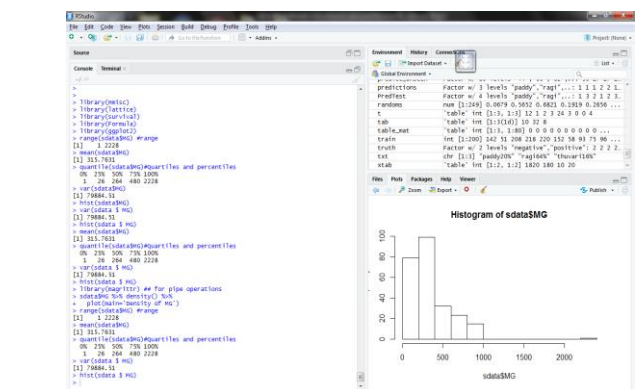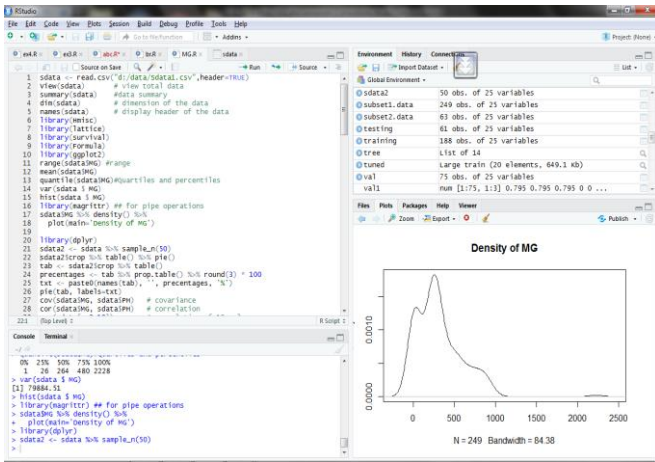

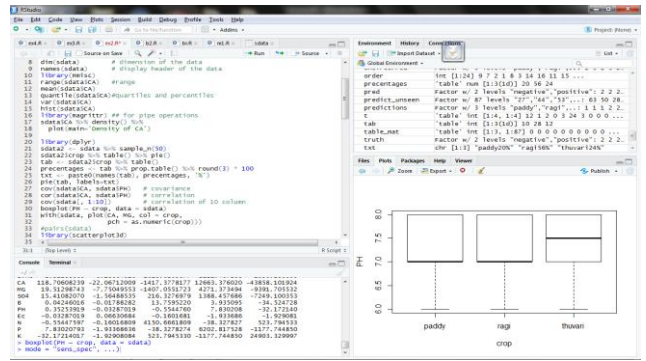
**Figure 3: Histogram of Mg**

**Figure 4 : Density of Mg**



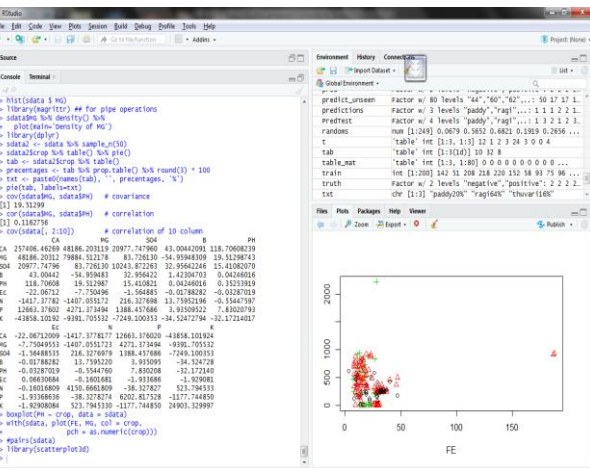**FIGURE 5: SATTER PLOT DIAGRAM OF FE**

### B. Pie chart of crops

Pie chart show the pictorial representation of the crop grown and also we can easily identify which crop gives the high yield in that particular area according the available micro nutrient and macro nutrients and what are the nutrients are recommended to give the high yield of a particular crop in that particular area can be easily identified using the pie chart from the give data set.
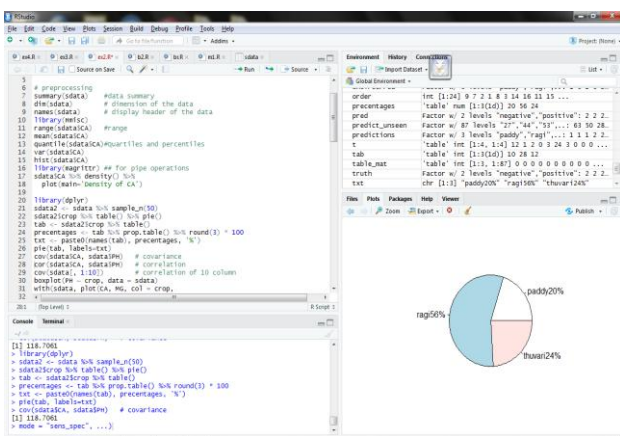


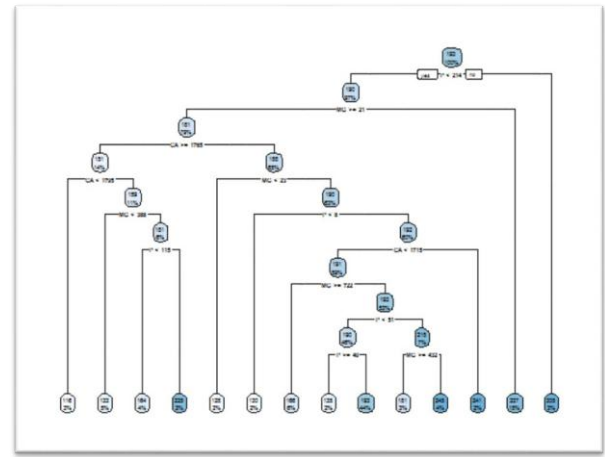**Figure 6: Pie chart of crops with percentage**



**Figure 7 : Boxplot diagram of crops**



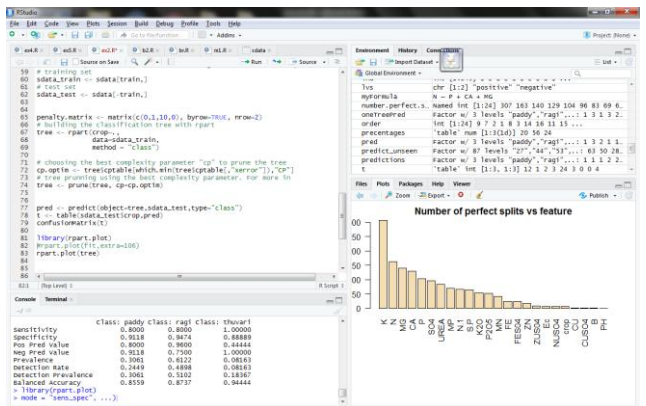**Figure 8 : Rplot Diagram of the Nutrients**



**Figure 9: Diagram of nutrients**

## IV. RESULTS AND DISCUSSION

If you This paper mainly focuses on finding the accuracy of decision tree and C4.5 algorithm is used to find the high yield of the crop from the available sample soil dataset collected from the Krishngiri district in Tamil Nadu .The results are listed below

1. Density of Mg - calculated to find the accuracy soil data set using decision tree C4.5 Algorithm.
2. Histogram of Mg - to identify the range of values present from 249 soil dataset.
3. Pie chart of crop, pie chart of crop with percentage – to identify the high yield of the crop.

4. Calculated the mean, median, variance, correlation and quartiles – to find the accuracy of the soil dataset.
5. Box plot diagram of crop .
6. Scatter Diagram of FE.
7. Rplot Diagram of the Nutrients.

**Table 3: Crop yield**

| S.No | Crop | Yield in % |
|------|--------|------|
| 1 | Ragi | 48% |
| 2 | Thuvari | 28% |
| 3 | Paddy | 24% |

## V. CONCLUSION

Data mining is one of the emerging fields of technology now-a-days used in agricultural field. the Balanced Accuracy of Decision Tree Algorithm is 0.788333 and the Crop Yield prediction are Ragi – 48 %,Paddy -28 % and Thuvari - 24%.using sample soil data set collected from district agricultural office at Krishnagiri district in Tamil Nadu, the sample soil data set taken it from the dimension of 249 tuples and 24 attributes. From the experiment and results the accuracy of Decision Tree Algorithm is .78 and high yield predicted Crop is Ragi. This paper focused only on three crops to find the crop that produce high yield, from sample soil dataset collected from Krishnagiri district, but in future planned to take at least 6 to 8 crops to find the high yield of the Crop in Krishnagiri district. And also this work can be extended, to find the accuracy of j48 algorithm and feature selection algorithms - applied predictive model is used to predict the crop yield and accuracy.

## REFERENCES

1. N Hemageetha and G.M. Nasira ,"Analysis of the Soil Data Using Classification Techniques forAgricultural Purpose" International Journal of Computer Sciences and Engineering, 4(6) (2016).
2. S.Hari Ganesh and Mrs. Jayasudha," Data Mining Technique to Predict the Accuracy of the Soil Fertility",  International Journal of Computer Science and Mobile Computing,  330-334 (7) (2015).
3. P. Kanjana Devi and S.Shenbagavadivu" Enhanced Crop Yield Prediction and Soil Data Analysis Using Data Mining", International Journal of Modern Computer Science 4(6)(2016).
4. R. M. Kumar, "Crop Selection Method to Maximize Crop Yield Rate Using Machine Learning Technique", International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, 1 (1) (2009).
5. E. Manjula and S. Djodiltachoumy, "A Model for Prediction of Crop Yield", International Journal of Computational Intelligence and Informatics, 6(4)(2017).
6. V. Rajeswari and K. Arunesh "Analysing Soil Data using Data Mining Classification Techniques", Indian Journal of Science and Technology, 9(19),(2016).
7. K.Samundeeswari and K.Srinivasan," Crop Yield Prediction And Soil Data Analysis Using Data Mining Techniques InKrishnagiri District" International Journal of Computer Science and Engineering(JCSE), 49-55, 6(8)(2018).
8. [8].K.Samundeeswari and K.Srinivasan," Data Mining Techniques In AgriculturePrediction Of Soil Fertility" International Journal of Scientific & Engineering Research (IJSER), 8(4)(2017).
9. .Miss.Snehal,S.D. "Agricultural Crop Yield Prediction Using Artificial", International Journal of Innovative Research in Electrical,Electronic, 1 (1) (2014).
10. https://krishnagiri.nic.in/about-district/history.

## AUTHORS PROFILE

**K. Samundeeswari,** Department of Computer Science, Govt. Arts College for Women, KrishnagirI, Tamil Nadu, India

**K.Srinivasan,** Department of Computer Science, Periyar University Constituent College of Arts &Science, Pennagaram, Dharmapuri, Tamil Nadu, India