An Overview on Data Mining Techniques for Rice Yield Prediction on Clustered Region Of Kerala

Joslin Joy
MCA Student, School of CSA
REVA University, Bangalore
joslinjoy4@gmail.com
Pinaka Pani. R
Assistant Professor
School of CSA
REVA University, Bangalore
rppani.mca@gmail.com

----------**ABSTRACT**----------

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining extract knowledge from the historical data. Rice production depends on climate, type of the soil, geography, biology and economy. Several factors are having an impact of the production of rice, especially the availability of rainfall in riverbeds near to the paddy fields. For low lying region like Kerala, predicting supply of rice is critical due to varying climate over the last couple of decades. Excessive conversion of paddy fields to households and industry especially in Kuttanad (rice bowl of Kerala).Hence to overcome these issues several statistical methods are being used. Applying such methodologies and techniques on historical yield of crops, it is possible to acquire the data and information which can be helpful to farmers and cultivators to improve the production of rice and helps farmers and government sectors to make favorable decisions and policies that lead to increase the production of rice. Here my focus is on literature study on application of data mining techniques to extract information from the rice yield data of previous years to estimate rice yield.

Key words: **SOM (self-organizing maps, k- Clustering, decision tree, linear regression, rice yield**

I. INTRODUCTION

Rice is the most important food crop in Kerala. The climate in Kerala is suitable for the cultivation of rice. This led to the people in Kerala to be more devoted in cultivating rice. There are many varieties of crop used to cultivate rice in Kerala[1]. Some are hybrid varieties which give more yield than normal type. In Kuttanad there are 3 major crop varieties, e.g. Jyoti (12-85), Jaya, Uma (H1), in which Jaya is normal crop and other two are hybrid. These are the 3 crops which is very suitable for the soil in this clustered area [2]. There are many more crops can be cultivated, but farmers used to grow these three crops because of the high yield, good quality and comparing the less expensive crops. There are three main rice growing seasons in Kerala.[2] First season is Virappu season/ Autumn season/First crop season, which starts in April-May and extends up to September-October; and the second is Mundakan season/winter season/second crop season, which starts in September-October and extends up to December-January and the last season is Puncha season/summer season/ third crop season, which starts in December- January and extends up to March- April. In Kerala, winter crop (Mundakan) has been greater than the other two crops both in terms of area as well as production.

II. RELATED WORK

There has been couple of studies related to irrigation, temperature, moisture in soil, geography, effect of climate changes & natural calamities in rice yield [5]. These studies helps and support a lot to research about the current topic, which is been affected by all the factors mentioned above. For good cultivation of rice, it needs rainfall in a particular centimeter and less or more than that centimeter will definitely affect the rice growth. As per the studies in this clustered region is a low land region and the 80% of this region's district is covered by coastal regions. Kuttanad has the lowest altitude in India and is one of the few places in the world where farming is carried on around 1.2 or 3.0 meters below sea level. The farming is mainly on bio saline. FAO has declared the Kuttanad farming system as a Globally Important Agricultural Heritage Systems (GIAHS) [3]. The studies say that the geographical area is mostly covered by water bodies, such as Kerala's major rivers, the Pamba, Meenachil, Achankovil and Manimala.

Farmers of Kuttanad have developed and mastered the spectacular technique of below sea level cultivation over 150 year ago. They made this system unique as it contributes remarkably well to the conservation of biodiversity and ecosystem services including several livelihood services for local communities. This case study has been done in 2012 in geographical coverage of Asia and the Pacific by M S Swami Nathan Research foundation and supported by government of Kerala, Food and Agriculture Organization of the United Nations (FAO)[3].

III. WORKFLOW OF THE SYSTEM

Indian Bureau of Statistics (IBS) is responsible for storing various statistical data sets, such agricultural record, population census, and employment statics[5]. However, most of distributed in the form of yearbooks. In order to work with data-sets one has to be manually copied and then entered into digital device such as a computer. In this way the data of rice yield has been collected from 10-20 years. It is very easy to collect data of climate, rainfall from these data-base the

studies will give an accurate results. These climate details will be easily available in Indian weather forecasting department records. In each year there are some fixed features/factors that will be considered. They are rainfall, temperature, humidity, moisture, wind, PH of soil etc. Each year's worth of rice yield in this clustered area contains 3 rice types namely Jyoti (12-85), Uma (h1) and Jaya. There are many other varieties that has been tried to cultivate, unfortunately didn't found a good yield from those crop varieties. And gradually these were not been used by the farmers because there were no expected results. Those varieties are kochuvithu, myna, and vykkatharan [4]. These are non-hybrid varieties that was used before 10 years for rice cultivation. There yield amount became less as compared to hybrid crops, thus hybrid crops are having more resistance to pests and they need less amount of fertilizers and less time to take harvesting than compared with non-hybrid crops.

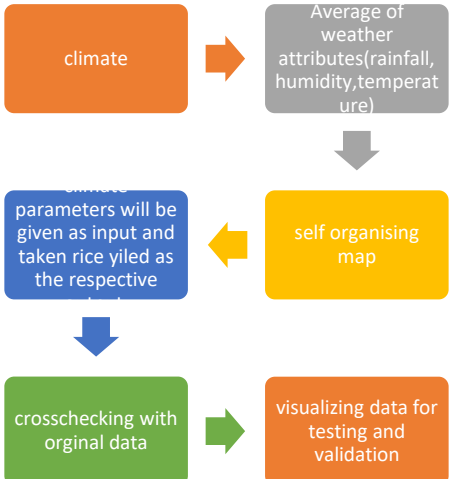

Fig: 1 Workflow of data.

IV. MOTIVATION

Paddy cultivation in kerala has witnessed a steady decline since the year 1980s.The sharp fall in the area under paddy cultivation as well as in the quantity of rice produced in the state has had important implications for Kerala's economic, ecological and social development over the last 5 years, however there have been commendable signs of a revival in rice production in among paddy cultivators in Kuttanad a region that is referred to as the "rice bowl" of kerala[4]. This field report, which is based partly on interviews with farmers, government officials and leaders of mass organizations in kerala. The government of kerala is supporting the rice cultivation in a great extend where the seeds for cultivation is completely given free to the farmers and the pesticides and other supportive measures is given in subsidy rate. Thus government is motivating the farmers in kerala for rice cultivation. Rice is the staple food of the people of kerala, and traditionally the cultivation of rice has occupied pride of place in the agrarian economy of the state. The lush green of paddy fields is one of the most captivating features Kerala's landscape[6].

V. DATA MINING TECHNIQUES:

1. Preprocessing

In the above workflow (fig.1) the data is preprocessed. One year is parted into 3 seasons according to the climate and other geographical features. The time periods are from April-September, October-December, and January –April. For all these seasons the data of each attribute (temp, humidity, rainfall, wind) is been taken and monthly average is calculated from this data [9]. This procedure is followed in our clustered area (Kuttanad) in last 5 years. In each season the yield is calculated and the average is taken for finding the total yield. The result of the preprocessing will be taken as the input for the next process. The approximate result of the fig.1 is tabulated below:

Season	**April-Sept**	**Oct-Dec**	**Jan-Mar**
Attributes			
Temperature	36˚C	27˚C	32˚C
Wind	17km/h	14km/h	13km/h
Humidity	98%	56%	70%
Rainfall	18.4mm	2.7mm	3mm
Total yield	7.03lk tones	9.53lk tones	13.63lk tones

2. Clustering

In my study I have taken a clustered region called "Kuttanad" which is lying in three districts of kerala which are Alappuzha, Thrissur and Kottayam. In this the major portion of Kuttanad lies in Alappuzha district and this district is covered with most of the area with water bodies such as rivers, lakes and coastal areas. To finding the yield of the rice production in this area I have taken some attributes that can help in assuming the results, thus the historical data that defines the value of these attributes will give a clear picture of the next harvest [9]. By this we can easily predict the yield rate will increase or decrease.

Cluster 1: This cluster evaluates based on the attributes such as temperature, rainfall, humidity and availability of sunlight. These are the climatic attributes that is considered in our study. The percentage of similarity in the values of our attributes will indicate the final result.

Cluster 2: This type is based on the pH of the soil and the salinity. This is very important to know about the soil and water in the clustered area is having an acidic or alkaline soil and water. Because according to this only the fertilizers can be used in the paddy filed.

Cluster 3: This cluster focused on the irrigation. As we discussed earlier our region is having a very good water resource that will help in cultivation of rice. But also there is another problem that is it is nearby coastal area and the salt water is not good for the rice cultivation. When the phenomenon such as low tide and high tide happens there will be chances of water in the sea to come into the river and other resources and then to paddy filed. This is one of the great threat in Kuttanad.

Cluster 4: This is based on the three major crops that we selected for the study which is cultivated in Kuttanad. These major crops are Uma, Jaya and Jyoti. There are many other

varieties of crops which are cultivated by farmers, but the reason why we choose these crops specifically is because of the time span that it takes is different and these crops will give great yield in less amount of fertilizers and are having great immunity which will resist against the pests.

In clustering self-organizing maps was used over K-means to cluster the year based average dataset. The main benefit of SOM over K-means is that SOM has low classification error than K-means. This helps to classify the data collected into correct cluster [10]. The other benefit is it reduces the dimension of its input data, but k-means cannot do that. Also studies says that the end clustered result of SOM is more stable than the k-means.

3. Classification

The result of the above four clusters will be taken for the classification. According to the historical data of last year's rice yield will be taken to classify the coming year yield. Here the main process is our clustered region Kuttanad is arranged against the input attributes i.e. min temperature, max temperature, rainfall, humidity, pH, sunshine, wind. The output we get from this will the yield of the sample crops we have taken for study (Jaya, Uma and Jyoti)[2].

The test is done for all the environment variables as the input and yield of each Jyoti, Jaya and Uma as output. This is represented in the above table.

a) Linear regression: This is a statistical measure that can be used to determine the strength of the relationship between one dependent variable and a series of other changing variables known as independent variables. If independent variable contains multiple input attributes like in my research such as max temperature, min temperature, rainfall, humidity, pH etc. linear regression provides the model for the relationship between a scalar variable one or more explanatory variable.

In the linear regression the 70% of data was used for training and left 30% is used for testing. The cross validation is done by taking all the errors of the 3 crops yield values in past 5 years.

b) Non- linear regression:

In this errors from different type of nonlinear regressions are compared. This training, testing and cross-validation errors will be recorded.

In this 70% of data is used to train each and left 30% of data is used for testing.

c) Regression Tree:

The corresponding training, testing and cross validation error for each crops in our clustered region is been calculated. 70% of data is been taken for training sample and the rest data is used for testing. And for the cross validation the all crops value is taken for finding the error [9].

VI. COMPARISON OF METHODS

The final result will be obtained from the end value of each model by comparing the value. For this we will be having an expected value and the original value i.e. our final result of each model is been taken and compared with the expected output and the most approximate value is been taken as the final result.

1) Cluster type 1: weather attributes
2) Cluster type 2: pH of the soil and salinity
3) Cluster type 3: irrigation and availability of rain.
4) Cluster type 4: sample crops taken for the prediction of yield.

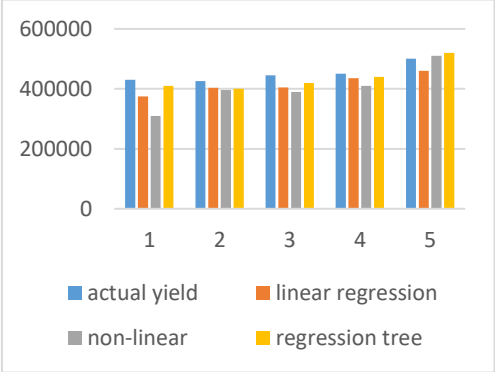

Figure1: actual vs predicted Yield of crop Jyoti according to historical data [7],[8]

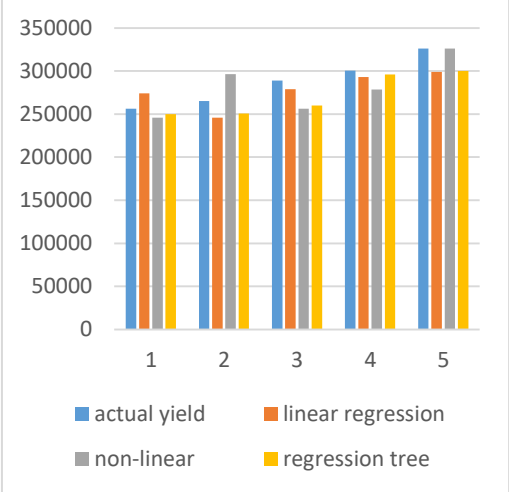

Figure 2: actual vs predicted Yield of crop Uma according to historical data [7],[8]

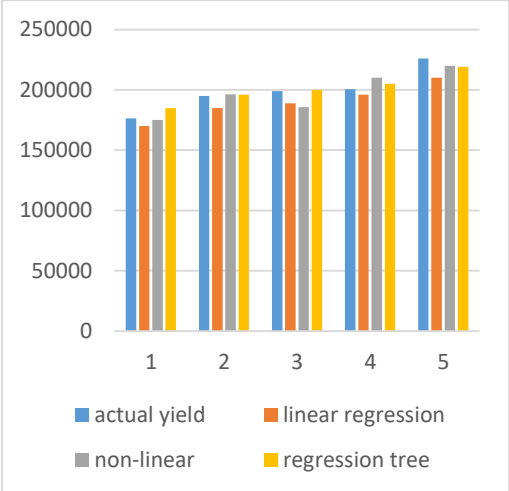

Figure1: actual vs predicted Yield of crop Jyoti according to historical data (In hectares)[7],[8]

VII. RESULT ANALYSIS

The final result is obtained by comparing all the models. The average of all the models which is very near to the actual value is taken. Comparison of all models with the actual value is tabulated below:

Clustered Region						
Test	Jyoti		Uma		Jaya	
	AY	PY	AY	PY	AY	PY
1	4.3	3.6	2.5	2.5	1.7	1.7
2	4.2	3.9	2.6	2.6	1.9	1.8
3	4.4	4.1	2.8	2.7	1.9	1.9
4	4.5	4.2	3.0	2.8	2.0	2.0
5	5.0	4.9	3.2	3.0	2.2	2.1

AR- Actual yield, PY-Predicted Yield

VIII. CONCLUSION AND FUTURE WORKS

In the research I clearly found that the accurate prediction of different species of crop yields across the clustered region that could help a lot of farmers, cultivators and government agriculture sector. According to the results the yield is getting increased in every year. And the predicted value is very near approximate that will definitely help the farmers to take preventive measures against the factors that can become a cause for less production of crops. The major crops that I have taken as sample are Jaya, Jyoti and Uma. In this the historical data says that farmers can earn more profit by using the crop Jyoti than others, because in each year the yield is getting increased in a good quantity. In the case of other crops also yield is varying but, comparing with Jyoti yield is less. In the overall study it says that the crops farmers using now a days are giving good yield and can get more yield in next year also by using the same crop varieties.

References

[1]. Status paper on rice in kerala (http://www.rkmp.co.in)

[2]. Crop wise analysis –economic review 2016, state planning board (spb.kerala.gov.in)

[3]. RAS /Paddy cultivation in kerala. (ras.org.in)

[4]. Paddy cultivation in kerala- Christ university journal

[5]. Indian bureau of statistics, Annual report on agriculture 2016-2017 –department of agriculture

[6]. International rice research institute/ India (IRRI)

[7]. Fernando Bacao, et al., "Self-organizing Maps as Substitutes for Clustering," Springer Computational Science – ICCS 2005

[8]. Teuvo Kohonen . "The Self- Organizing Map", Proceedings of IEEE, Vol.78, No.9, pp.1464- 1480, 1990.[Online]Available: http://www.eicstes.org/EICSTES_PDF/PAPERS/The%20SelfOrganizing%20Map%20(Kohonen).pdf

[9]. Shanmuganathan, S.et al., "Data Mining Techniques for Modelling the Influence of Daily Extreme Weather Conditions on Grapevine, Wine Quality and Perennial Crop Yield," IEEE Second International Conference on Computational Intelligence, Communication Systems and Networks(CICSyN),pp90-95,2010.

[10]http://www.mathworks.com/help/stats/kmeans.html, accessed last on20th November, 2014.