Agricultural Data Modeling and Yield Forecasting using Data Mining Techniques

Rithesh Pakkala P.[1], Akhila Thejaswi R[2]

[1,2]Assistant Professor, Department of Information Science of Engineering, Sahyadri College of Engineering & Management, Mangaluru, Karnataka, India

---***---

Abstract - *Agriculture encompasses a great impact on the economy of developing countries. The Global change in climatic conditions and the cost of investment in agriculture are major obstacle for small-holder farmers. The proposed work intends to design a predictive model that provides a cultivation plan for the farmers to high yield of paddy crop using data mining techniques. Data mining techniques extract hidden knowledge through data analysis, unlike statistical approaches. The dataset is collected from the agricultural department. K- Means clustering and various classifiers like Support Vector Machine, Naïve Bayes are applied to meteorological and agronomic data for the paddy crop. The performance of various classifiers are validated and compared. The result of the work is the accurate prediction of crop yield. The final rules extracted by this work are useful for farmers to make proactive and knowledge-driven decisions before harvest.*

***Key Words*: Data mining, Predictive Model, K - Means Clustering, Support Vector Machine, Naïve Bayes classifiers**

1. INTRODUCTION

Data mining is the process of analysing various hidden patterns of data according to different views for categorization into required information. This data is been collected and gathered from common area, such as agriculture department, for efficient analysis, data mining algorithms, improving business decision making and other information requirements to ultimately reduce the costs and increase revenue.

Data mining technique is intended in extracting the hidden, useful and interesting patterns from raw data. Data mining tools predict future trends and behaviours, allowing businesses to make proactive knowledge based decision. Data mining accompanies the use of complicated statistics, analysis tools to find previously unknown, suitable unseen structure and interaction in the huge dataset. It helps to develop a predictive model that provides a cultivation plan for farmers to get high yield of paddy crops.

Descriptive data mining tasks featurize the general properties of the data in the database while predictive data mining is used to predict explicit values based on patterns determined from known results. Prediction also involves usage of some fields or variables in the database to predict unknown or future values of other variables that are of concern. As far as data mining technique are concerned in most of cases, predictive data mining approaches are been used. Predictive data mining technique is used to predict future crop, pesticides, weather forecasting and fertilizers to be used, revenue to be generated and so on.

Forecasting crop productivity is one of the scientific techniques of predicting crop yield before harvest. Data mining techniques like clustering and classification are performed in order to maximize the crop yield prediction. A final prediction model is developed and implemented, that protects farmers from agricultural risks by providing a framework that helps them in scientific decision making in agriculture.

Using this predictive model, farmers can plan the cultivation process well in advance. To prevent loss, farmers can identify suitable combinations of varying factors like seed quality, rainfall, temperature and sowing procedure. It is a scientific model that provides suitable cultivation plans to farmers in accordance with the changing agronomic factors. Paddy is a pivotal crop in south India. Yield of paddy crop depends on various meteorological and agronomic factors such as seed quality, rainfall, temperature and sowing procedure. In order to evaluate the relationship between these factors and crop yield and to identify the input variables effecting the output of paddy crop, a realtime data set is collected from farmers cultivating paddy is used in this research.

Raw agricultural data are pre-processed and only the necessary factors are established by filtering. The major data mining techniques used in this research are K-means clustering and classifiers such as Support Vector Machine, Naïve Bayes. Performances of the above are compared based on classier accuracy measures. The final knowledge regarding the cultivation plan is discovered, evaluated, and presented. The result of the desire models will help agribusiness associations in equip agriculturists with necessary information as to which factors add to high yield.

2. RELATED WORK

This section describes the various works carried out in the relevant fields.

Jharna Majumdar *et al.*[1] proposed a data mining model which is applied on agriculture dataset using different clustering algorithms such as DBSCAN, PAM and CLARA.

Clustering is considered as an unsupervised classification process. Clustering techniques can be divided into Partitioning clustering, Hierarchical clustering, Density based methods, Grid-based methods and Model based clustering methods. Clustering methods are contrasted using quality metrics. According to the analysis of clustering quality metrics, DBSCAN gives the better clustering quality when compared to PAM and CLARA, CLARA gives the better clustering quality than the PAM. At the end comparison through different factors which includes Root Mean Squared Error(RMSE), Mean Absolute Error (MAE), etc.

A predictive model that provides a cultivation plan for farmers to get high yield of paddy crops using data mining techniques is proposed by Anitha Arumugam[2]. Data is collected from farmers cultivating paddy along the Thamirabarani river basin. K-means and various decision tree classifiers are applied. The performance of various classifiers is validated and compared. In this research, K-means clustering is integrated with decision tree classifiers in order to improve classification accuracy. Even with worst-case performance using decision stump, integration of k-means clustering has improved the accuracy from 63.5.

Shruti Mishra *et al*. [3] describes the use of data mining in crop yield prediction. Classification is a technique in data mining that assigns item in a collection to target categories of classes. Different classifiers are used namely J48, LWL, LAD Tree and IBK for prediction and then the performance of each is compared using WEKA tool. The classifiers are compared with the values of accuracy, root mean squared error (RMSE), mean absolute error (MAE) and relative absolute error (RAE). Lesser the value of error, more accurate the algorithm will work.

Akanksha Verma *et al*. [4] proposed trim forecast framework utilizing fuzzy bunching strategies and neural system. In the proposed approach initially the raw data set is taken and then clustering and classification will be performed in order to get the results in MATLAB. To perform clustering, Fuzzy c means clustering approach is taken. Fuzzy C Means clustering is a precised learning algorithm that provides less error rate probability and arranges the data in the hierarchical manner. The framework encourages ranchers to do right things at opportune time. The model will help the ranchers in expanding their profitability by choosing the proper harvest for their land, soil temperature, humidity and few other conditions.

Pooja M C *et al*. [5] designed a model to forecast crop yield using data mining technique J48of C4.5 algorithm. A model is built which uses past information like soil type, soil pH, ESP, EC of a particular region to give better crop yield estimation for that region. This model can be used to select the most excellent crops for the region and also its yield there by improving the values and gain of farming also. This aids farmers to decide on the crop they would like to plant for the forthcoming year. Prediction will help the associated industries for planning the logistics of their business.

Rupinder Singh and Gurpreeth Singh[6] proposed a model using decision tree algorithm to predict accurate yield of crop. The work done is based on detecting the influence of rainfall and relative humidity on wheat crop yield. Decision tree results indicated that rainfall and relative humidity has more influence over wheat crop yield during vegetative period as compared to reproduction and maturation period. Rules generated from decision tree analysis will help the users to predict the conditions responsible for variable wheat crop yield under given meteorological parameters.

Ramesh Babu Palepu and Rajesh Reddy Muley[7] used data mining for forecasting the future trends of agricultural processes. This paper presents about the role of data mining in perspective of soil analysis in the field of agriculture and also confers about several data mining techniques and their related work by different authors in context to domain of soil analysis. The data mining techniques are of very up-to-the-minute in the area of soil analysis. In this model fuzzy algorithm are applied for managing crops, K-means algorithm used for classify the soils and Support Vector Machine technique applied to predict the crop yield.

U. Kumar Dey *et al*.[8] in their study analyzed crop yield prediction by using Support Vector Machine (SVM), Multiple Linear Regression (MLR), AdaBoost and Modified Non Linear Regression for the regions in Bangladesh. Rice is divided into three categories: Aman, Aus and Boro. Prediction is done during the aforementioned seasons. The different training techniques were judged based on the Root Means Square Error(RMSE) and Mean Absolute Error(MAE) values.

R.Sujatha and Dr P. Isakki[9] proposed the model to predict crop yield using classification techniques. The paper describes how improving agriculture efficient by prophesying and improves yields by previous agriculture information. It also used to select a best crop by farmer, to plant depending on the weather situation and provides required information to prefer the suitable season to do excellence farming.

Monali Paul *et al*. [10] proposed the model to predict crop yield using two classification techniques i.e., K-Nearest Neighbor and Naive Bayes algorithms. These algorithms are applied to the soil dataset which is taken from the soil testing laboratory Jabalpur, M.P. There accuracy is obtained by evaluating the datasets. Classification of soil into low, medium and high categories are done by adopting data mining techniques in order to predict the crop yield using available dataset. This study can help the soil analysts and farmers to decide farmers to decide sowing in which land may result in better crop production.

Researchers namely Ramesh and Vishnu Vardhan [11] are analysed the agriculture data for the years 1965–2009 in the district East Godavari of Andhra Pradesh, India. Rain fall data is clustered into 4 clusters by adopting the K means clustering method. Multiple linear regression (MLR) is one of the data mining technique that is used to model the linear relationship existing between a dependent variable and one or more independent variables. The dependent variable is rainfall and independent variables are year, area of sowing, production. Purpose of the work is to obtain many suitable data models that achieve high accuracy in terms of yield prediction capabilities.

Geraldin B Dela Cruz *et al*. [12] focus of this study is to implement an efficient data mining mechanism based on the combination of Principal Component Analysis (PCA) as a pre-processing method and a modified Genetic Algorithm (GA) as the learning algorithm, in order to reduce computational cost and time by keeping a number of features as discriminating and small as possible. In doing so, generating agricultural crops classification models is efficient and characterization is improved. The Principle Component Analysis-Genetic Algorithm(PCA-GA) data mining mechanism will be implemented for agricultural crops dataset to identify key attribute combinations and characteristics that determine crop performance.

Anshu Bharadwaj *et al*. [13] in their research attempt to extend the boundaries of discretization and to evaluate its effect on other machine learning techniques for classification, support vector machines with and without discretization of the datasets. On comparing the results obtained by the algorithms which was previously mentioned, it was inferred that Discretization based Support Vector Machine (D-SVM) produced the model with highest accuracy for all the datasets. The results clearly indicate that the accuracies of discretization based SVM are better when compared to that of the classification accuracy without SVM of the same datasets when they were classified without getting discretized.

Shruthi Ramdas and RPakkala discussed about how Web mining makes use of data mining classification techniques to automatically discover Web documents and services, extract information from Web resources, and uncover general patterns on the Web for analyzing visitor activities in network-based systems in [14]. The objective of this study is to classify the users into interested or non interested for particular website based on their access pattern using web server log's by making use of decision tree classification technique.

3. PROPOSED METHODOLOGY

The main objectives of the proposed work are as follows:

- To perform classification using different algorithms like Support Vector Machine, Naïve Bayes
- Predict the crop yield.

The agricultural dataset involving attributes such as seed quality, rainfall, sowing procedure and temperature is collected from different sources. The raw data collected is pre-processed using pre-processing techniques which removes the outliers and irrelevant data. This is followed by data mining techniques such as clustering and classification techniques in order to determine the yield of the crop. The proposed framework is shown below.

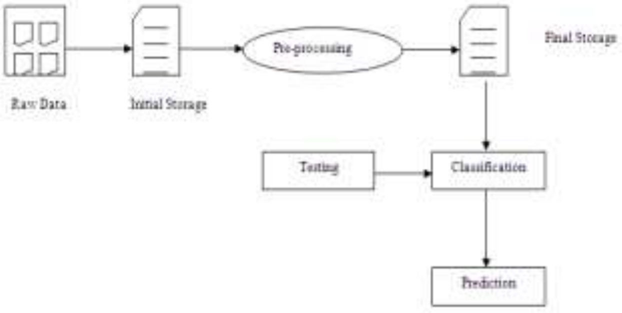

Fig -1: Schematic representation of Proposed Framework

Pre-processing is a technique to improve the quality of data to be presented to the mining process. In the proposed system data pre-processing is dine in 3 major ways: (i) data cleaning (ii) attribute selection (iii) transformation.

Data cleaning is a method of replacing incomplete, inconsistent, and noisy data. Some attributes have null data or missing data. To obtain high quality knowledge, cleaning is done by eliminating null values. Attributes that contribute more to mining process are identified in this process. Attribute selection is done by eliminating irrelevant and redundant attributes. The process of converting the data to the form suitable for mining task is called transformation. Attributes with numeric values are converted to categorical attributes.

Basically data mining techniques are used in order to predict the crop yield. The farmer gives different attributes or parameters as an input, and later by using support vector machine the possible yield of crop is been predicted in terms of low, medium or high. The output of support vector machine is been directed to the Naïve Bayes which gives the accuracy of prediction.

Classification is a supervised learning process by which data objects are grouped into classes of known labels. Classification algorithms uses classifiers to classify a group of similar objects under one type and when a new object is introduced, prediction is made so as to put that object into one of the class. It involves two phases: learning phase and classification phase. In the learning phase, training data are analyzed and a classifier model is built. In the classification phase, the test data are used to estimate the accuracy of classification.

The support vector machines(SVM) are supervised learning models with associated learning algorithm that examines the data that have been used for the purpose of classification and clustering analysis. A set of training examples are given which includes the dataset, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifiers. An SVM model is a representation of the given examples as points in space, mapped in order that the examples of the separate categories are divided by a clear gap that is as wide enough. New examples are then mapped and adapted in such a way that the same space and predicted value to belong to a category based on which side of the gap they will appear.

Naive Bayes classifiers are a community of simple probabilistic classifiers in data mining that are based on applying Bayes theorem with intensive and strong independence considerations that are existing between their characteristics. Naive Bayes classifiers are highly scalable, highly changeable, requiring a number of parameters that are linear in the number of variables (features/predictors) defined in a learning problem. Maximum likelihood training can be done by evaluating a closed form expression, which takes linear time, rather than by any other cost inefficient iterative approximation that are used for other types of classifiers.

Psuedo Code:

```
1. Preprocess the raw data and prepare it for classification
2. Input the preprocessed data for Support Vector Machine
      if the dataset is separable
            then create a hyperplane using equation $ax+b=0$
      else
            then create a hyperplane using equation $y(ax+b)>=1$
3. Use the classes formed from hyperplane for further classification
4. Input the data for classification
5. Classify it with SVM into different classes
6. Predict the output
```

4. RESULTS

From the raw data collected from various agricultural departments, a pre-processing operation is carried out. The classification method identifies suitable classes such as low depicted as 0, medium depicted as 1 and high depicted as 2.

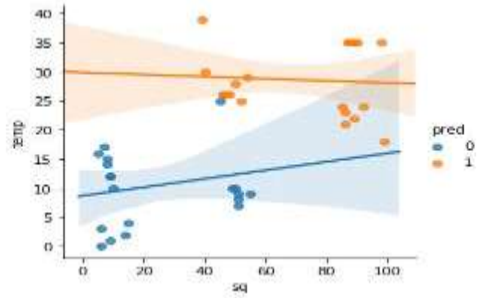

Fig -2: Classification of Data based on Temperature and Seed Quality

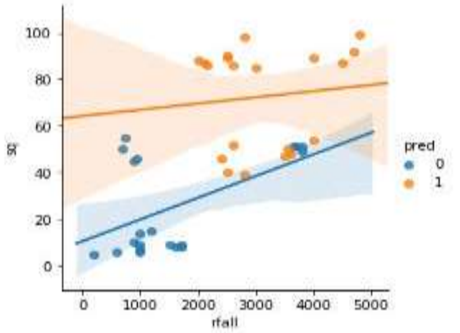

Fig -3: Clustering of Data based on Rainfall and Seed Quality

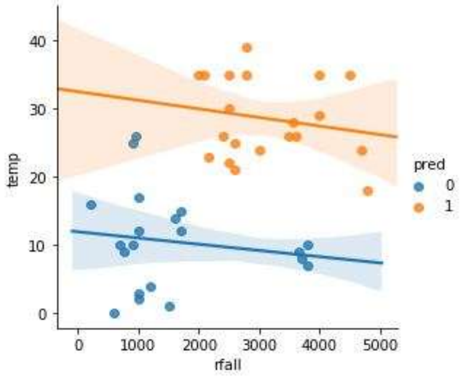

Fig -4: Classification of Data based on Temperature and Rainfall

The attribute which affects the crop yield are taken on different axes on the graph and their respective crop yield result is shown in Figure 2, 3 and 4. Figure 2 depicts the graph which is plotted by taking seed quality along the horizontal plane and temperature along the vertical plane. Figure 3 depicts the graph which is plotted by taking seed quality along the horizontal plane and rainfall along the vertical plane. Figure 4 depicts the graph which is plotted by taking rainfall along the horizontal plane and temperature along the vertical plane.

5. CONCLUSION

The major outcome of this work is the prediction of the rice crop yield based on the input given by the farmers. There is also an accurate extraction of hidden knowledge about a cultivation plan involving major agronomic and meteorological factors. This knowledge thus obtained is useful for getting high yield of rice crop. In the project, since there is the usage of both clustering as well as classification the accuracy attained is high. This accuracy is been measured by using Naïve Bayes algorithm. The result shows that different kinds of estimations that can be done on a crop.

REFERENCES

[1] Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki, “Analysis of Agriculture Data using Data Mining Techniques: Application of Big Data”, Journal of Big Data, Springer, 2017.

[2] Anitha Arumugam, ”A Predictive Modeling Approach for Improving Paddy Crop Productivity using Data Mining Techniques“, Turkish Journal of Electrical Engineering Computer Sciences 2017.

[3] Shruti Mishra, Priyanka Paygude, Snehal Chaudhary and Sonali Idate, “Use of Data Mining in Crop Yield Prediction”, IEEE International Conference on Inventive Systems and Control (ICISC 2018).

[4] Akanksha Verma, Aman Jatain and Shalini Bajaj, “Crop Yield Prediction of Wheat using Fuzzy C Means Clustering and Neural Network”, International Journal of Applied Engineering Research 2017.

[5] Pooja M C, Sangeetha M, Shreyaswi J Salian, Veena Kamath and Mithun Naik, “Implementation of Crop Yield Forecasting using Data Mining”, International Research Journal of Engineering and Technology 2017.

[6] Rupinder Singh and Gurpreeth Singh, “Wheat Crop Yield Assessment using Decision Tree Algorithms”, International Journal of Advanced Research in Computer Science 2017.

[7] Ramesh Babu Palepu and Rajesh Reddy Muley, “An Analysis of Agricultural Soils by using Data Mining Techniques”, International Journal of Engineering Science and Computing 2017.

[8] Umid Kumar Dey, Abdulla Hassan Masud and Mohammed Nazim Uddin, “Rice Yield Prediction Model using Data Mining”, International Conference on Electrical, Computer and Communication Engineering (ECCE) 2017.

[9] R.Sujatha and Dr.P.Isakki, “A Study on Crop Yield Forecasting using Classification Techniques”, IEEE International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), 2016.

[10] Monali Paul, Santosh K. Vishwakarma and Ashok Verma, “Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach”, IEEE International Conference on Computational Intelligence and Communication Networks (CICN), 2015.

[11] D Ramesh and B Vishnu Vardhan, “Analysis of Crop Yield Prediction using Data Mining Techniques”, International Journal of Research in Engineering and Technology 2015.

[12] Geraldin B Dela Cruz, Bobby D Gerardo, Bartolome T, “Agricultural Crops Classification Models Based on PCA-GA Implementation in Data Mining”, International Journal of Modeling and Optimization 2014.

[13] Anshu Bharadwaj, Shashi Dahiya and Shashi Dahiya, ”Discretization based Support Vector Machine(D-SVM) for Classification of Agricultural Datasets”, International Journal of Computer Applications 2014.

[14] Shruthi Ramdas, Rithesh Pakkala P., Akhila Thejaswi R, “Determination and Classification of InterestingVisitors of Websites using Web Logs”, International Journal of Computer Science and Mobile Computing, Vol.5, Issue.1, January-2016, pg.01-09.