

A Novel Framework Crop Yield Prediction Using Data Mining

Ms.Juvanna I^a, Yuvesh Balaji V G^b, Sri Raam M A^c, Karthikeyan T^d

^{a,b,c,d} Department of Information Technology Hindustan Institute of Technology and Science Chennai, India

^a ijuvanna@hindustanuniv.ac.in, ^b yuveshmsd@gmail.com, ^c Sriraammadhavan777@gmail.com, ^d karthihawaz@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Modern agriculture has aided in the reduction of food costs as a percentage of profits. Estimating crop yield based on environment, soil, water, and crop parameters has been proposed as a potential research topic in modern agriculture. The prediction of crop yield is one of the most difficult issues in precision farming, and several models have already been suggested and tested. This is because a variety of factors such as climate, weather, soil, fertilizer and seed diversity affect crop yields. This issue requires the application of multiple datasets. Researchers have been unable to create a clear non-linear or linear relationship between raw data and crop yield values, and the performance of the derived characteristics is highly dependent on the production of those models. This means that it is not an easy task to predict yield, but rather a number of complex steps. Crop output models can now accurately estimate the actual output, but better output prediction results remain desirable. Data mining is an important tool for forecasting the yield of crops, with a focus on plants and what to do in the growth period. A number of algorithms for data mining were used for prediction studies on crop yields. For better prediction accuracy, an efficient data mining approach based on linear regression is proposed in this study.

Keywords: Agriculture, Crop yield, Data mining, Decision making, Prediction.

1. Introduction

Growing data development from sensors and aerial crop pictures, increasing agricultural production through intelligent systems, and government assistance for the adoption of modern farming practices are all driving the agricultural market forward. Agricultural organizations and farmers across the country are increasingly implementing data mining-enabled systems to raise farm productivity and gain a competitive edge in business activities [1][2]. Agriculture is a major industrial sector and the economy of the country depends on resilience in rural areas and factors such as global warming, precipitation, water levels and chemicals. Despite the fact that recent research has revealed statistical knowledge about agriculture, few studies have looked into crop prediction based on historical evidence. For agricultural productivity, crop yield forecasts are vital. Politicians rely upon reliable forecasts to take timely decisions on imports and exports to boost national food security. On the other hand, estimating the crop yields by various factors involved is very difficult [3][4][5].

Crop output is influenced by agro-climatic input parameters in general. However, agricultural input parameters differ from region to region, and gathering such data over a wider area is a difficult task. Climate data is also collected in various parts for each sq.m area. The resulting data sets are huge and can be used to forecast massive crops. Agricultural researchers are experimenting with different forecasting methodologies. Agricultural scientists have shown that the pro-pesticide state's policies have resulted in an alarmingly high use of pesticides. Also, there is a negative association between pesticide use and crop yield, according to the report. A data-mining algorithm using minimum incidents with limitations can be used to detect the laws of association between rainfall and climatic indices. During critical plant growth phases, a longer dry season or heavy plumage will drastically reduce crop yields[6].

Based on previous research, the most common applications of data mining in agriculture tend to fall into three broad categories:

- Robots in Agriculture

Companies create and customize intelligent drones to perform essential agricultural tasks including crop harvesting with greater strength and response than human staff.

- Monitoring of Crops and Soils

Businesses use computer vision and data mining techniques to analyze data obtained by robots and/or production tools systems in order to monitor crop and soil quality.

- Predictive Analytics

Data mining models are being developed to monitor and predict various environmental effects on agricultural production, such as climate.

Through a certain capacity of flexible thinking, a machine can interpret its conditions and manage measures to fix a specific objective related to that context, according to data mining theory. Data mining is when a computer improves its ability to solve environmental-related problems and goals by following a collection of protocols as the scientific nature of the data it collects grows. Simply put, as the machine receives more comparable data sets that can be grouped into specific protocols, its capacity to rationalize grows, allowing it to reliably determine a variety of outcomes [7][8].

Companies that produce data mining products or services like agricultural training information, drones and automated machines will make progress to make more practical applications in this sector to assist the world in dealing with issues of food production for the growing population [9]. As a result, the study's key contribution is

- i) To provide a detail of how data mining integrates with the agricultural domain.
- ii) To propose a data mining method for better crop prediction.

With this detailed introduction, Section II describes the role of data mining in modern agriculture, Section III shows proposed system model, followed by Experimental result in Section IV and Conclusion in Section V.

2. Related Work

Agriculture is very important in the country's global economy. The agricultural sector is continuously under pressure to increase crop productivity and produce more crops as a result of population growth. Data mining is a popular technology today that can be applied to the modern agricultural sector. In agriculture, data mining aids in the production of more stable crops. The principle developed by Arthur Samuel in machine learning research is now used in modern agriculture. Artificial techniques are being used in the agricultural sector to increase precision and solve problems [10][11][12].

Precision agriculture is an organizational culture that uses machine learning systems to maximize harvest productivity and precision. Predictive analytics employs data mining technology to aid in the diagnosis of plant pests, insect infestations, and slow crop growth on farms. AI systems will control and track weeds when deciding the herbicides to apply within the right buffer, reducing herbicide overuse and tolerance. Farmers use predictive results to determine agricultural precision by designing probability distributions for seasonal forecasting [13].

These models will look months ahead of time and use collected data to provide farmers with clear season forecasts, ideal planting times, and locations of the best crop varieties. Agricultural data mining technologies will then boost farm management by basing predictions on predicted weather conditions during the coming season. Knowing when to sow the seed at the right time can mean the difference between a fruitful year and a failed crop. Another important aspect of crop management is yield mapping and prediction, which helps to balance supply and demand. Using remote sensing technologies and weather conditions, data mining models can be developed for specific crops in a given region to forecast precise yield estimates. Governments often employ such yield estimation innovations, reducing or eliminating the need for expensive crop cutting studies [14][15][16].

In emerging markets, the gap between capital and education is also widening. Many farmers lack access to telephones, training, and the skills necessary to evaluate the data that is available. Recommendations must be provided that not only promote the appreciation of these farmers, but also recommend the measures that must be taken to achieve high returns[17][18]. It's critical to take coordinated steps to address the pressing need to link rural areas, as well as to work with policymakers and technology firms to reduce the cost of data collection equipment and software. As a result, better data mining approaches for agricultural initiatives are needed.

3. System Model

The current framework used a multiple linear regression method and a density-based clustering approach for estimating harvest yield analyses.

A) Multiple Linear Regression

The technique for modeling the straight-linear relationship between a dependent variable and one or more independent variables is multiple linear regression (MLR). The variable dependent is often called a prediction or an independent variable called a predictor.

B) Density-based Clustering Technique

The main concept of cluster technology based on density is that at least a minimum number of points are found in the vicinity of a particular unit distance for each point of the cluster. In other words, there should be a threshold of density in the vicinity.

C) Drawbacks

- There is no large volume of dataset implemented.
- Feature extraction is not handled well.
- Less accurate of prediction crop yield.

All critical parameters required for high crop yield are included in the crop yield forecast. This increases the outcome of the agricultural yield rating. Inputs are considered for all important parameters. One of the most common problems with the prediction method is that almost all of the necessary parameters that must be considered for an exact prediction are not taken into account. Farm data such as crop varieties, crop year, location, and seasonal parameters such as Khrif, rabbi, and summer crops make up the crop information base. The knowledge-based also includes regions, district information, and ecological parameters such as extreme and lowest temperature values, as well as average precipitation. A related input module which collects data from the farmer needs to be used for the plant return model. Crop name, land area, crop year, and prediction tons are all included in the input module. The model for feature selection is in charge offset.

Crop specifics are used to select an associate attribute. To predict the yield, a crop yield prediction model was used. The data is then sent to a classification rule for similar grouping content following the selection of the feature. Crop growth with climatic data and crop parameters can be forecast. The performance of classification of crop information as regards the name of the crop, the season and the total yield data is then predicted by law.

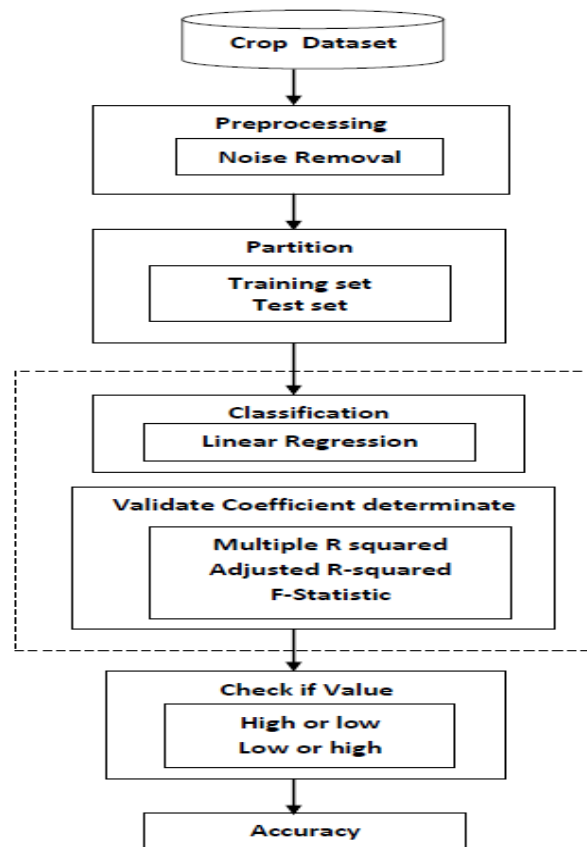


Fig 3.1 Proposed Framework

C) Procedure

Step 1: Data extraction from the Agricultural Department's online repository.

Step 2: Applying pre-processing to clean up results.

Step 3: For training and testing, standard cross validation is used.

Step 4: Compute the results for each classifier individually.

Step 5: Based on the output measure such as precision, choose the top three classifiers and compose a majority voting-based ensemble.

Phase 6: Measure the yield based on the accuracy ranking.

D) Data Collection

The first step in the machine learning pipeline is to collect data for training the data mining model. Data mining systems can only be as successful as the data they've been trained on.

E) Preprocessing

Once the data has been collected as datasets from the Twitter source, it must be transferred to the classifier. Before the study, the classifier cleans the dataset by extracting redundant data such as stop words and emoticons to ensure that non-textual material is detected and deleted.

F) Classification

The data mining approach employs classification algorithms that use input training data to predict the likelihood or probability that the data that follows will fall into one of the predetermined categories.

G) Advantages

1 It increases and verifies the accuracy of yield predictions, which are helpful to farmers in predicting the yield of a particular crop.

2 In the following work, the crop yield forecast will be compared to all available data, and appropriate methods to improve the effectiveness of the proposed technique will also be taken into account.

4. Experimental Results

The SSPTS library was used to evaluate all agricultural experimental data sets. IBM SPSS is a research environment that provides a wide range of data and text analysis, as well as a strong set of statistical algorithms for classification, processing, and association rules. It has been discovered that using DM techniques, more successful techniques for solving complex agricultural issues can be created. The obtained results were tested and analyzed using IBM SPSS statistical tools.

A) Tracking Patterns

Data extraction learns to recognize patterns in a data set. This usually refers to some of the time in your data that occurs at regular intervals, or over time there is a flow and flow of certain variations. Tracking and classification methods are able to manage large amounts of data in data extraction. Separation is a method of extracting data to predict the category of agricultural data.

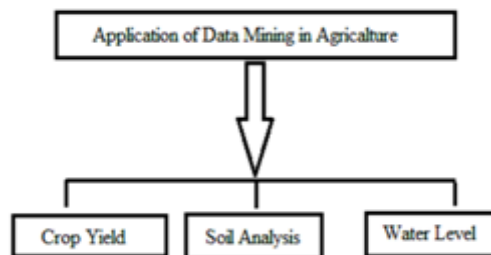


Fig 4.1 Tracking pattern

B) Cluster based Analysis (CBA)

CBA analysis is very useful for converting clusters of similar plant species to different parameters. Blocks are sub-categories for agricultural data. Users understand the basic structure of preset data. They are used as a standalone tool to gain insight into the distribution of data processing algorithms. Compilation by a group that contains objects of the same type. Simplification is achieved by representing data in a few groups that require good detail.

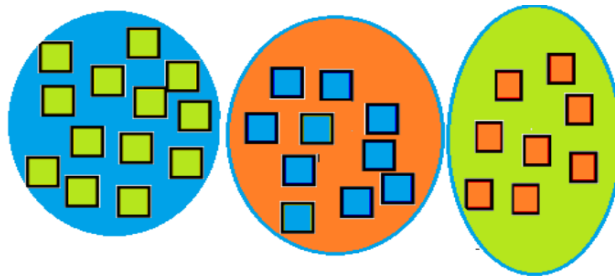


Fig 4.1 Cluster based analysis

C) Association Rules for yield Estimation (ARYE)

ARYE has a large number of applications and is widely used to help determine the yield of foreign yields and integration into agricultural datasets. The mineral law of the organization describes how often events occur together. Through the application of the Mining Association Act, we can see some interesting connections between the various varieties in large-scale agricultural production data.

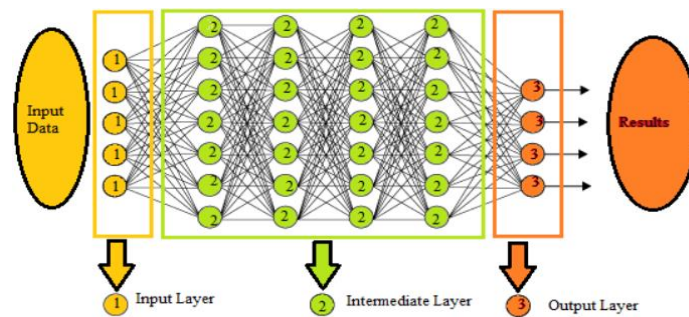


Fig 4.2 Association rules for yield estimation.

D) Performance Metrics

The real effectiveness passes through the procedure with it – only once it includes precision, moreover recognition as accuracy. The final results, means performance of the classification could be higher by using the bagging approach as an optimization strategy within the process of the category. The performance of crop prediction is measured using the following performance metrics.

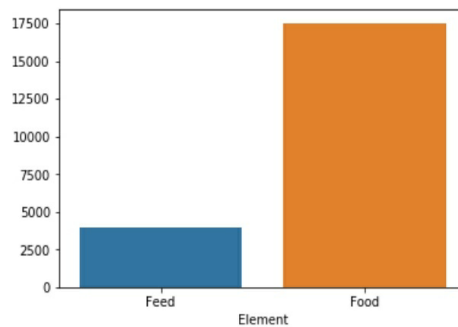


Fig 4.3 Feed and Food graph

E) Accuracy

Historically, the exactness rate was the most common statistical indicator used. In unbalanced data sets the reliability no longer is an appropriate measure, because it does not distinguish between the numbers of instances of properly categorized classes. It can thus lead to misconceptions that a 90 percent precision classifier in a set of data with an IR value of 9 is incorrect if all examples are classified as negative.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}.$$

Where

TP- True Positive

TN- True Negative

FP - False Positive

FN - False Negative

The accuracy level is 95%. Using the existing algorithms of KNN and ANN, the accuracy level is 80 and 85 respectively.

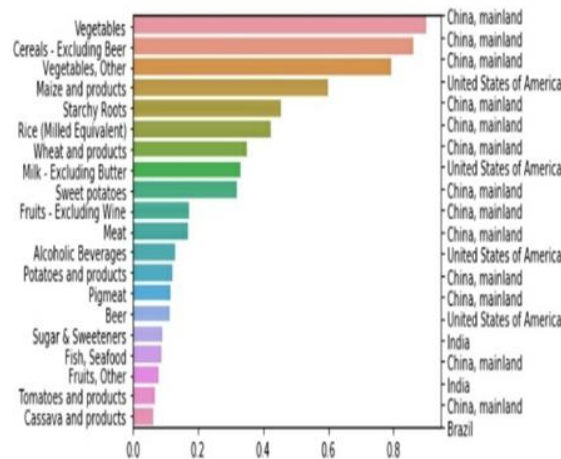


Fig 4.4 Crop yield prediction

The above graph represents the annual crop yield performed in selective land. As per the graph the left alignment represents the crop section and the right alignment represents the countries with the yield prediction. From the above graph, it is evident that cultivating vegetables in China might lead to a higher yield. Hence, cultivating vegetables in China will be profitable for the upcoming years (based on the graph prediction). But, the graph conveys that the cultivation of tomatoes and products might lead to a lower yield, hence results in lesser profit. Hence, for successive years tomatoes and products can be cultivated at a lower scale, such that the loss might be lighter, if the crops of tomatoes face any of the natural disasters (heavy rain etc.).

5. Conclusion

Farmers in India will benefit from precise forecasts of different crop yields in different districts. In precision farming, yield estimate models are used to enhance yield production in order to meet demand, and to advise the government on crop yield estimates of imports from Trichy, Tamilnadu, in order to avoid overlapping. During this project, the regression method was put to the test in terms of yield prediction. The data was used to create model inputs. While linear regression algorithms provided reasonable estimation accuracy, higher prognostic power could be obtained by including alternative variables such as environment, agricultural practices, and soil characteristics in the model growth, such as year, crop, area, and output (in tons). For Ecuadorian conditions, a linear regression model can be suggested. There are no yield prognostic models for any crop. Crop yields (sugarcane, cotton, and turmeric) are expected to be at their maximum levels using this proposed method. In future, this model can be reformulated by alternative crop valuations to determine methods of yield growth and land management in excellent potential crops such as wheat and rice..

References

1. Arumugam, Suresh & Pugalendhi, Ganeshkumar & Marimuthu, Ramalatha. (2018). Prediction of major crop yields of Tamilnadu using K-means and Modified KNN. 88-93. 10.1109/CESYS. 2018.8723956.
2. Balducci, Fabrizio & Impedovo, Donato & Pirlo, Giuseppe. (2018). Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement. Machines. 6. 38. 10.3390/machines6030038.
3. Bejo, Siti & Mustaffha, Samihah & Ishak, Wan & Wan Ismail, Wan Ishak. (2014). Application of Artificial Neural Network in Predicting Crop Yield: A Review. Journal of Food Science and Engineering. 4. 1-9.
4. Chlingaryan, Anna & Sukkarieh, Salah & Whelan, Brett. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture. 151. 61-69. 10.1016/j.compag.2018.05.012.
5. Crane-Droesch, Andrew. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters. 13. 10.1088/1748-9326/aae159.

6. Didar, Mostafa & Jahan, Nusrat & Shams, Maleeha & Chowdhury, Labib & Siddique, Shahnewaz. (2020). A Deep Gaussian Process for Forecasting Crop Yield and Time Series Analysis of Precipitation Based in Munshiganj, Bangladesh. 10.1109/IGARSS39084.2020.9323423.
7. Gandhi, Niketa & Petkar, Owaiz & Armstrong, Leisa & Tripathy, Amiya. (2016). Rice crop yield prediction in India using support vector machines. 1-5. 10.1109/JCSSE.2016.7748856.
8. González-Sánchez, Alberto & Frausto-Solis, Juan & Ojeda, Waldo. (2014). Predictive ability of machine learning methods for massive crop yield prediction. SPANISH JOURNAL OF AGRICULTURAL RESEARCH. 10.5424/sjar/2014122-4439.
9. Jahan, Raunak. (2018). Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land soils. International Journal for Research in Applied Science and Engineering Technology. 6. 189-193. 10.22214/ijraset.2018.5030.
10. Kim, Nari & Ha, Kyung-Ja & Park, No-Wook & Cho, Jaeil & Hong, Sungwook & Lee, Yang-Won. (2019). A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015. ISPRS International Journal of Geo-Information. 8. 240. 10.3390/ijgi8050240.
11. Kuwata, Kentaro & Shibasaki, Ryosuke. (2015). Estimating crop yields with deep learning and remotely sensed data. 858-861. 10.1109/IGARSS.2015.7325900.
12. Lobell, David & Burke, Marshall. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology - AGR FOREST METEOROL. 150. 1443-1452. 10.1016/j.agrformet.2010.07.008.
13. N, Mahendra. (2020). Crop Prediction using Machine Learning Approaches. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS080029.
14. Sanjudharan, M S Minu & Dharrsan, Vigash & Immanuel, Christo. (2020). Crop Yield Prediction Using Machine Learning. 9. 98. 10.37896/aj9.4/012.
15. Shastry, K Aditya & Sanjay, H. & Deshmukh, Abhijeeth. (2016). A Parameter Based Customized Artificial Neural Network Model for Crop Yield Prediction. Journal of Artificial Intelligence. 9. 23-32. 10.3923/jai.2016.23.32.
16. Suraparaju, Veenadhari & Mishra, Bharat & Singh, Cd. (2011). Soybean Productivity Modelling using Decision Tree Algorithms. International Journal of Computer Applications. 27. 975-8887. 10.5120/3314-4549.
17. Veera, Sellam & Poovammal, E.. (2016). Prediction of Crop Yield using Regression Analysis. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016 /v9i38/ 91714..