Forecasting Agricultural Crop Yield through Data Mining Methods

Sapparapu Meghavardhan

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract**—** *Agriculture is highly uncertain due to factors like geography, weather, and more. In India, farming plays a crucial role in the economy, with agriculture contributing significantly to the GDP. However, analyzing vast agricultural data for crop yield estimation is complex. Data mining offers a solution by extracting meaningful insights from raw agricultural data. This research aims to estimate crop yield using data mining techniques.*

Keywords: Statistical, production, estimation, tracking patterns.

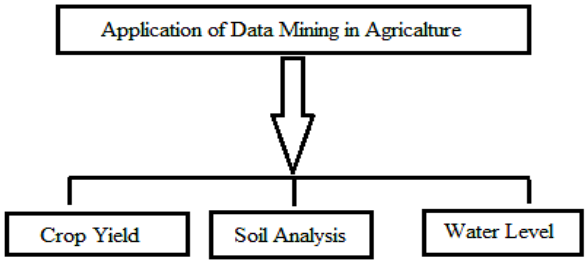
I. INTRODUCTION

India's diverse agriculture sector plays a crucial role in the economy, with crop yield projections being vital for food security. Factors like resource changes impact crop production. Crop forecasting is essential for policymaking and relies on efficient estimation techniques. Technological advancements have revolutionized agriculture globally since the 1990s, with intelligent agriculture decision systems integrating AI, database, and 3S technologies for improved decision-making

1.1 DM and Statistical Approaches for Production Estimation:

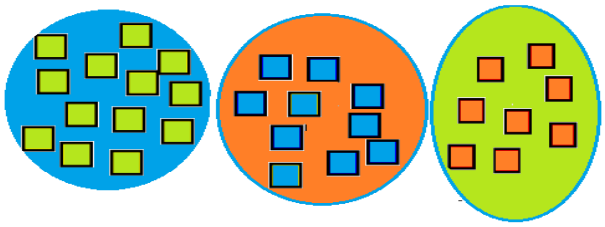
1.1.1 Tracking Patterns:

Data extraction is learning to recognize patterns in data set. This is usually to identify some of the intervals in your data that occur at regular intervals, or over time there is a flow and flow of a particular variable. Tracking and Classification methods are capable of handling a massive volume of data in data extraction. Classification is the method of data extraction to predict the category of agricultural data. The aim of tracking and classification techniques is to rigorously estimate the target class for each case in the data set of agriculture crop. Classification model could be used to find agricultural crop yield prediction as low, medium, or high-risk factors. Classification constructs a form used to predict agricultural crop segment labels to differentiate objects from different categories of anonymous objects.



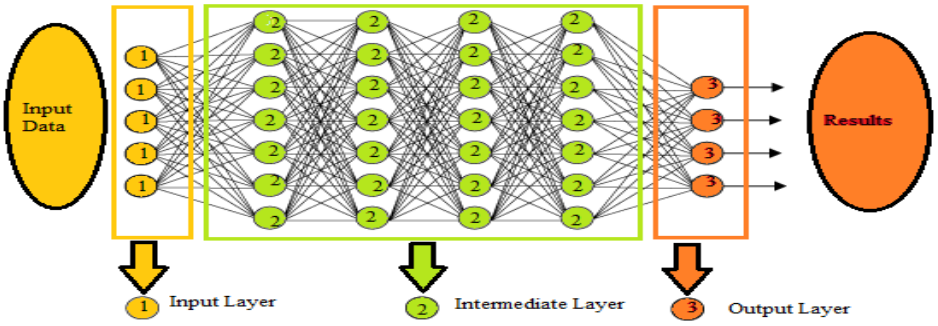
1.1.2 Cluster based Analysis (CBA):

CBA is most useful analysis to evolve the clusters of similar kind of crop with various type of parameter. Blocks are basically sub categories of agricultural data. Users understand the basic structure of data set. They are used as a standalone tool to gain insight into the distribution of data processing algorithms. Clustering is a group that consist same type objects. Simplification is achieved by representing the data in fewer clusters that require fine details.



1.1.3 Association Rules for yield Estimation (ARYE)

ARYE has a large number of applications and it is widely used to help find outcrop yield estimation and correlations in agricultural datasets. An association rule mining describes how frequently events have occurred together. With the use of Mining Association rule, we can identify some interesting interconnection between different varieties in a large amount of agricultural crop production data.

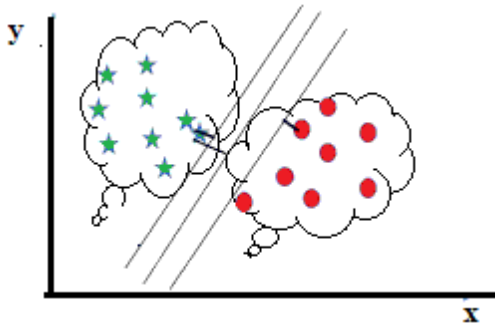


1.1.4 Artificial Neural Network (ANN)

ANN is very useful algorithm in forecasting agricultural crop yield using different crop performance factors Biological nerve organization is the basis of ANN. Agriculture, farming method is very multifarious since it deals with the large data situation which comes from large number of factors. It consists of set of interconnected neurons. Connection between neurons is links and having weight associated with it.

1.1.5 Support Vector Machine

SVM algorithm is help to forecast the category of crop based on number of factors of crop and soil of crop yield estimation SVM is the specific discriminatory selection of various factor of crop by the excellent interval. SVM is an automated learning algorithm under supervision that can also be used for classification or regression challenges.



1.1.6 Multiple Linear Regressions (MLR)

LRA expression is derived from the statistical regression model based on which the agricultural crop yield is estimated. Basically, MLR is used to find linear relationship between a dependent variable and one or more independent variables. For formulation consider a polynomial of the nth degree.

$$y = a_1 + a_2x^1 + a_3x^2 + a_4x^3 + ---- a_nx^n$$

II. LITERATURE REVIEW

2.1 Big Data Analysis Technology Application in Agricultural Intelligence Decision System

Ji-chun Zhao Jian-xin Guo et. Al – 2019

The rapid development of big data technology provides a new technical means for the research and development of agricultural intelligent decision system. It can effectively improve the processing speed and accuracy of the agricultural intelligent decision

system, and can provide guidance for agricultural production. The application of big data analysis technology and artificial intelligence technology in the agricultural intelligent decision system is the next development direction

2.2 Analysis of Crop Yield Prediction Using Data Mining

Ramesh, B Vishnu Vardhan et. al – 2019

Initially the statistical model Multiple Linear Regression technique is applied on existing data. The results so obtained were verified and analyzed using the Data Mining technique namely Density-based clustering technique. In this procedure the results of two methods were compared according to the specific region i.e., East Godavari district of Andhra Pradesh in India. Similar process was adopted for all the districts of Andhra Pradesh to improve and authenticate the validity of yield prediction which are useful for the farmers of Andhra Pradesh for the prediction of a specific crop. In the subsequent work a comparison of the crop yield prediction can be made with the entire set of existing available data and will be dedicated to suitable approaches for improving the efficiency of the proposed technique.

2.3 Survey on Data Mining Techniques in Agriculture

M.C.S.Geetha et. al – 2020

Agriculture, especially in developing nations like India, benefits greatly from information technology. Data mining plays a key role in decision-making for agricultural issues. This paper compiles various authors' work on data mining applications in agriculture, providing researchers with insights into current techniques and applications.

2.4 Data Mining in Agriculture: A Review

K Raghuveer, M J Yogesh, Shwetha S – 2020

According to him the decision trees suggested there exists a correlation between climatic factors and soybean crop productivity and these variables influence on the soybean crop productivity were confirmed from the rule accuracy and Bayesian classification. The salient conclusions of the present study are:

a. The decision tree analysis indicated that the productivity of soybean crop was mostly influenced by Relative humidity followed by temperature and rainfall.

b. The decision tree from the present study is fast to execute and much to be desired as representations of knowledge interpretations.

c. The rules formed from the decision tree are helpful in predicting the conditions responsible for the high or low soybean crop productivity under given climatic parameters.

2.5 Applying Data Mining Technique to Predict Annual Yield of Major Crops

Praful S. Tayde, Balkrishna K. Patil, Rajesh A. Auti

Climate and other environmental changes impact on agricultural economy of any country. Production of crop normally depends on factors like biology, climate, economy and geography, this factor leads to impacts on agriculture. By applying different methodologies and techniques on yields of crops, it is possible to obtain information about which help to government organization for making good decisions and applying different policies, also it’s helpful for farmers to make better plan to increase the production.

Problem Statement: The Existing system used the method namely Multiple Linear Regression technique and Data Mining method namely Density-based clustering technique were taken up for the estimation of crop yield analysis.

2.6 Multiple Linear Regression

A Multiple Regression Model involves multiple predictor variables to predict a dependent variable. Multiple Linear Regression (MLR) is a common method for modeling the linear relationship between a dependent variable and several independent variables. MLR, based on least squares, is widely used in climatology and crop yield prediction. In this model, Production is predicted using seven predictors: Year, Rainfall, Area of Sowing, Yield, and three types of Fertilizers (Nitrogen, Phosphorous, and Potassium.

2.7 Density-based Clustering Technique

The primary idea of Density-based clustering techniques is that, for each point of a cluster, the neighborhood of a given unit distance contains at least a minimum number of points. In other words, the density in the neighborhood should reach some threshold. However, this idea is based on the assumption that the clusters are in the spherical or regular shapes. These methods group the objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighborhood of data objects. In these approaches, a given cluster continues to grow as long as the number of objects in the neighborhood which exceeds some parameter. This is considered to be different from the idea in partitioning algorithms that use iterative relocation of points that give a certain number of clusters.

Disadvantages

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

III. PROPOSED WORK

This research aims to estimate crop yield using data mining techniques, considering various essential parameters for accurate prediction. Parameters such as crop varieties, crop year, area, and seasonal factors like Khrif, rabbi, and summer crops are included in the crop information database. The prediction model involves an input module for farmer input and a feature selection model to select relevant attributes. Climate data and crop parameters are used to predict crop growth, and prediction rules are applied to classify crop particulars in terms of crop name, season, and total yield

Advantages

- It improves and authenticate the validity of yield prediction which are useful for the farmers for the prediction of a specific crop.
- In the subsequent work a comparison of the crop yield prediction can be made with the entire set of existing available data and will be dedicated to suitable approaches for improving the efficiency of the proposed technique

IV. METHODOLOGY

- Step-1: Extraction of datasets through an online repository of Agricultural department.
- Step-2: Application of pre-processing for data cleaning.
- Step-3: Standard cross validation is applied for training and testing.
- Step-4: Computation of results for all individual classifiers.
- Step-5: Select top-3 classifiers based on the performance measure such as accuracy and compose majority voting-based ensemble.
- Step-6: Predict the yield with accuracy score

4.1 Data Collection

Collecting data for training the ML model is the basic step in the machine learning pipeline. The predictions made by ML systems can only be as good as the data on which they have been trained. Following are some of the problems that can arise in data collection:

Data refers to land use statistics, which is an important element for planning and developing policy formulation in agriculture. The Agricultural data set used for this kind of study is freely available at the website of directorate of economics & statistics of India [www.data.gov.in].For this study, we have selected Crop (Arhar, Barley, Maize, Potato).We have collected data from year 1997 to year 2013 that cover the different parameter of agricultural crop production that are Year, Season, Land area, production, yield, area under irrigation

- Inaccurate data. The collected data could be unrelated to the problem statement.

- Missing data. Sub-data could be missing. That could take the form of empty values in columns or missing images for some class of prediction.
- Data imbalance. Some classes or categories in the data may have a disproportionately high or low number of corresponding samples. As a result, they risk being under-represented in the model.
- Data bias. Depending on how the data, subjects and labels themselves are chosen, the model could propagate inherent biases on gender, politics, age or region, for example. Data bias is difficult to detect and remove.

4.2 Pre-Processing

Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.

Text pre-processing is an essential a part of any NLP method and the significance of the NLP pre-processing are

- To minimize indexing (or knowledge) records dimension of the textual content records
    1. Stop words bills 20-30% of total phrase counts in a special textual content record
    2. Stemming may just diminish indexing size as much as forty- 50%
- To make stronger the efficiency and effectiveness of the IR method
    1. Stop words aren't valuable for shopping or textual content mining
    2. Stemming used for matching the similar words in a text record

Tokenization: Tokenization is the process of breaking a circulate of textual content into phrases, phrases, symbols, or different significant factors called tokens. The aim of the tokenization is the exploration of the phrases in a sentence. The list of tokens turns into input for further processing akin to parsing or textual content mining. Tokenization is valuable both in linguistics (where it's a form of textual content segmentation), and in laptop science, the place it forms a part of lexical analysis. Textual knowledge is simplest a block of characters at the starting. All strategies in know-how retrieval require the words of the data set. For that reason, the requirement for a parser is a tokenization of records. This might be sound trivial because the text is already saved in computing device-readable codecs. However, some problems are nonetheless left, like the removing of punctuation marks. Different characters like brackets, hyphens, and so on require processing as well.

Stop word Removal: Stop phrases are very more often than not used fashioned phrases like ‘and’, ‘are’, ‘this’ etc. They don't seem to be useful in classification of records. So, they must be removed. However, the development of such stop phrases record is problematic and inconsistent between textual sources. This process also reduces the text knowledge and improves the approach performance. Each textual content report offers with these phrases which are not vital for text mining applications.

Stemming and Lemmatization: The aim of both stemming as well as lemmatization is to scale down inflectional types & mostly derivationally associated varieties of a phrase to a fashioned base kind. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of accomplishing this goal accurately more often than not, and quite often involves the removal of derivational affixes. Lemmatization often refers to doing matters competently with the usage of a vocabulary and morphological analysis of phrases, in most cases aiming to eliminate inflectional endings only and to come back the base or dictionary type of a word, which is often called the lemma.

4.3 Classification

Classification is defined as the process of recognition, understanding, and grouping of objects and ideas into preset categories with the help of these pre-categorized training datasets, classification in machine learning programs leverage a wide range of algorithms to classify future datasets into respective and relevant categories. Classification algorithms used in machine learning utilize input training data for the purpose of predicting the likelihood or probability that the data that follows will fall into one of the predetermined categories. Our proposed model involves six classification models such that Support Vector Machine, K-Nearest Neighbour, Decision Tree, Random Forest Tree, Logistic Regression, Naïve Bayes

To evaluate the accuracy of our classifier model, we need some accuracy measures. The following methods are used to see how well our classifiers are predicting:

- Holdout Method: It is one of the most common methods of evaluating the accuracy of our classifiers
- In this method, we divide the data into two sets: a Training set and a Testing set. The training set is shown to our model, and the model learns from the data in it. The data in the testing set is withheld from the model, and after the model is trained, the testing set is used to test its accuracy.
- The training set will have both the features and the corresponding label, but the testing set will only have the features and the model will have to predict the corresponding label.

The predicted labels are then compared to the actual labels and accuracy is found out seeing how many labels the model got right.

Bias and Variance: Bias is the difference between our actual and predicted values. Bias are the simple assumptions that our model makes about our data to be able to predict on new data. It directly corresponds to the patterns found in our data. When the Bias is high, assumptions made by our model are too basic, the model can't capture the important features of our data, this is called underfitting.We can define variance as the model's sensitivity to fluctuations in the data. Our model may learn from noise. This will cause our model to consider trivial features as important. When the Variance is high, our model will capture all the features of the data given to it, will tune itself to the data, and predict on it very well but new data may not have the exact same features and the model won't be able to predict on it very well. We call this Overfitting.

Precision and Recall: Precision is used to calculate the model's ability to classify values correctly. It is given by dividing the number of correctly classified data points by the total number of classified data points for that class label.

Where:

TP = True Positives, when our model correctly classifies the data point to the class it belongs to.

FP = False Positives, when the model falsely classifies the data point.

Recall is used to calculate the ability of the mode to predict positive values. But, "How often does the model predict the correct positive values?". This is calculated by the ratio of true positives and the total number of actual positive values

4.4 Evaluation

All agricultural experimental data set were analyzed using SSPS library. IBM SPSS offers variety of data, text analysis in research environment that has the good collection of statistical algorithms for Classification, processing and association rules It has been observed that of more effective techniques that can be developed to find the solution of complex agricultural issue using DM techniques. The obtained results were verified and analyzed through statistical software IBM SPSS package.

V. CONCLUSION

Accurate prediction of crop yields across districts in India is crucial for farmers. Yield estimation models are used in Precision Agriculture to increase production and inform government decisions on imports. Regression approaches, particularly linear regression, showed acceptable accuracy in predicting yields using factors like year, crop, area, production, climate, and soil characteristics. Linear regression models can be recommended for similar agricultural conditions. This system predicts crop yields for sugarcane, cotton, and turmeric and can be adapted for other crops like wheat and rice.

REFERENCES

[1] Zhao, J.C. and Guo, J.X., 2018, April. Big data analysis technology application in agricultural intelligence decision system. In *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* (pp. 209-212). IEEE.

[2] Elsheikh, Ranya, Abdul Rashid B. Mohamed Shariff, Fazel Amiri, Noordin B. Ahmad, Siva Kumar Balasundram, and Mohd Amin Mohd Soom. "Agriculture Land Suitability Evaluator (ALSE): A decision and planning support tool for tropical and subtropical crops." *Computers and electronics in agriculture* 93 (2013): 98-110.

[3] Conţiu, Ştefan, and Adrian Groza. "Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning." *Expert Systems with Applications* 64 (2016): 269-286.

[4] Parry, Martin L., Cynthia Rosenzweig, Ana Iglesias, Matthew Livermore, and Gunther Fischer. "Effects of climate change on global food production under SRES emissions and socio-economic scenarios." *Global environmental change* 14, no. 1 (2004): 53-67.

[5] Jeguirim, M., Goddard, M.L., Tamosiunas, A., Berrich-Betouche, E., Azzaz, A.A., Praspaliauskas, M. and Jellali, S., 2020. Olive mill wastewater: From a pollutant to green fuels, agricultural water source and bio-fertilizer. Biofuel production. *Renewable Energy*, *149*, pp.716-724.

[6] Khot, Lav R., Sindhuja Sankaran, Joe Mari Maja, Reza Ehsani, and Edmund W. Schuster. "Applications of nanomaterials in agricultural production and crop protection: a review." *Crop protection* 35 (2012): 64-70.

[7] Taghizadeh-Mehrjardi, Ruhollah, Kamal Nabiollahi, Leila Rasoli, Ruth Kerry, and Thomas Scholten. "Land suitability assessment and agricultural production sustainability using machine learning models." *Agronomy* 10, no. 4 (2020): 573.

[8] Zhang, Shurui, Shuo Wang, Lingran Yuan, Xiaoguang Liu, and Binlei Gong. "The impact of epidemics on agricultural production and forecast of COVID-19." *China Agricultural Economic Review* (2020).

[9] Owen, Darren, A. Prysor Williams, Gareth Wyn Griffith, and Paul JA Withers. "Use of commercial bio-inoculants to increase agricultural production through improved phosphrous acquisition." *Applied Soil Ecology* 86 (2015): 41-54.

[10] Tripathi, M.K. and Maktedar, D.D., 2020. A role of computer vision in fruits and vegetables among various horticulture products of agriculture fields: A survey. *Information Processing in Agriculture*, *7*(2), pp.183-203.

[11] Zhao, Ji-chun, and Jian-xin Guo. "Big data analysis technology application in agricultural intelligence decision system." In *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 209-212. IEEE, 2018.

[12] Ramesh, D., and B. Vishnu Vardhan. "Analysis of crop yield prediction using data mining techniques." *International Journal of research in engineering and technology* 4, no. 1 (2015): 47-473.

[13] Sabareeswaran, D., and A. Edwin Robert. "A survey on data mining techniques in agriculture." *Indian J. Innov. Dev.* 5 (2016): 8.

[14] Issad, Hassina Ait, Rachida Aoudjit, and Joel JPC Rodrigues. "A comprehensive review of Data Mining techniques in smart agriculture." *Engineering in Agriculture, Environment and Food* 12, no. 4 (2019): 511-525.

[15] Tayde, Praful S., Balkrishna K. Patil, and Rajesh A. Auti. "Applying Data Mining Technique to Predict Annual Yield of Major Crops." *Int. J* 2, no. 2 (2017).