PSYCHOLOGY AND EDUCATION (2021) 58(2): 11796-11800
ISSN: 1553-6939

CROP YEILD PREDICTION USING ML

B.Venkateswarlu[1],Erugu Krishna[2],Dr.M.Bal Raju[3],Dr.M.Chandra Naik[4],
Assistant Professor[1,2],Professor[3,4],
Department of CSE,
Pallavi Engineering College[1,2,4],Swamy Vivekananda Institute of Technology[3],
Mail ID:bvenkat1109@gmail.com,Mail ID:krishna.cseit@gmail.com,Mail ID:drrajucse@gmail.com
Kuntloor(V),Hayathnagar(M),Hyderabad,R.R.Dist.-501505.

ABSTRACT

In India's economy, agriculture is by far the most significant industry, and it has the greatest impact on the country's gross domestic product (GDP). An estimated 50 percent of the country's workforce is employed in the industry, which accounts for around 18 percent of the country's Gross Domestic Product (GDP). People in India have been engaged in agriculture for a long time, but the results have never been satisfactory owing to a variety of variables that influence crop productivity at different times of the year in different regions. A high agricultural production is required to meet the demands of the world's approximately 1.2 billion people in order to ensure that they are met. All of the variables that influence crop output are directly related to soil type, precipitation, seed quality, and the existence or lack of technical infrastructure, to name a few. To meet the increased demand, new technologies are required, and farmers must use their resources effectively by embracing new technology rather than relying on inefficient farming practises. The purpose of this project is to demonstrate how to develop a crop production forecast system using Data Mining methods. The dataset pertaining to agriculture was the topic of the investigation. Several classifiers, including the J48, LWL, LAD Tree, and IBK are used to forecast it. The performance of each classifier is evaluated by comparing its performance to the others using the WEKA tools for enhancing Python with machine learning performance (python with machine learning). In order to evaluate total performance, it is necessary to include Accuracy factors such as linear regression, as well as the accuracy of Random forest and KNN classifiers, were employed in this study, and one of them was the accuracy of linear regression. The overall performance of the classifiers is then assessed by comparing their Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Relative Absolute Error (RAE) values to the values of Root Mean Squared Error (RMSE) obtained from the training data (RAE). As a result, the technique will perform more correctly as the number of errors lowers. Classifiers are evaluated based on how well they perform in classification by making comparisons with one another.

1. INTRODUCTION

The purpose of information extraction and forecasting is to identify patterns in huge data sets. Information extraction and forecasting is the process of analysing, extracting, and predicting crucial information in order to identify patterns in the data. When businesses want to translate raw data from their customers into information that can be utilisedto improve the efficiency and effectiveness of their operations, they employ this strategy. The pre-processing and alteration of data is a critical component of the Data Mining process, and it accounts for a large portion of the total time spent on it. This process begins with the selection of data and continues until patterns are uncovered that may be used to forecast crucial insights. The data is then analysed further. It is necessary to perform two jobs during the preprocessing stage: outlier identification and the detection of missing data. Transformation, on the other hand, is concerned with the establishment of a relationship between two or more separate parts. In this study, historical climatic and agricultural output data were mined and evaluated with software created expressly for this purpose, resulting in a large number of projections being generated. It is possible to make judgments that can help in the expansion of agricultural production when you have access to precise data. In order to lower the costs involved with making decisions about the soil and crop that will be planted in a field, it is critical to provide farmers with a Decision Support System (DSS).

The usage of this software system while analysing raw data, academic papers, or business models aids analysts in forecasting or detecting key information that may be utilised to analyse an issue and solve it through decision-making. Farmers are likely to gain from this strategy since it will aid them in making critical decisions that were previously done inefficiently or based on educated speculation. When it comes to creating the final version of the prediction system, the application of data mining techniques will be utilised. Previous research has demonstrated

the relevance of data mining approaches in the agriculture industry, and the current findings corroborate this. It is necessary to use this data collection, as well as the dataset that has been mentioned previously. Several distinct algorithms were used in the course of the investigation and analysis. There were a variety of algorithms employed in the study, including classification methods like J48 and LAD Tree, and Lazy Learner approaches like IBK and LWL, all of which are detailed fully in detail in the WEKA programme, which was used to conduct the research. It is used to explain the confusion 2 matrix for each classifier in WEKA by using the experimental phases, as well as the confusion 1 matrix for that classifier, to describe the confusion 2 matrix for each classifier. The concepts of accuracy and sensitivity (specificity and sensitivity), as well as the concepts of performance metrics and their general concepts, such as root mean square error (RMSE), mean absolute error (MAE), root mean square error (RAE), and root mean square error (RMSE), are discussed in this chapter (accuracy). There are three metrics that are used to evaluate the performance of each classifier: the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Relative Absolute Error (RAE) (RAE). A statistic known as the Root Mean Squared Error (RMSE) is the most often used statistic in computer science (RAE). Precision is taken into consideration, as can be seen in the final result, which also contains recommendations for additional study. Precision is taken into consideration The purpose of this study was to evaluate the performance of machine learning-based crop production prediction systems to the performance of other types of algorithmic crop production prediction systems, which we accomplished via the use of artificial intelligence.

2. Literature survey

It is the technique for doing a literature review that is the most critical phase in the software development process. Before the tool can be developed, several aspects must be taken into consideration, including time limits, budgetary concerns, and the general strength of the business. Once these requirements have been met, the following ten steps will be utilised to identify which operating system and programming language will be used to construct the tool in order to finish the project. Once these phases have been completed, the project will be completed. The programmers will require a significant amount of assistance from other sources in order to accomplish their task once they begin working on the instrument. For example, older programmers may provide guidance, as can books and websites, among other sources of information and assistance. When designing the system that will be used to create the suggested system from large amounts of data in order to extract some pattern from it, the prior considerations are taken into consideration. This method is used by organisations to transform raw data from their customers into information that can be utilised to improve the efficiency and effectiveness of their business. It is the pre-processing and modification of data that constitutes a significant element of the Data Mining process. This process begins with the selection of data and continues until patterns are discovered that may be utilised to predict critical insights. In preprocessing, two tasks are involved: the identification of outliers and the detection of missing data. In contrast, transformation is concerned with the connection between distinct elements.

By combining historical climatic and agricultural production data with data mining techniques and statistical analysis, it is possible to make a wide variety of forecasts based on the information acquired. This knowledge may then be used to agricultural yields in order to increase them. Providing farmers with a Decision Support System (DSS) is important in order to reduce the expenditures associated with making decisions regarding the soil and crop that will be planted in the field. Using this system, farmers could save time and money because it would assist them in making critical decisions that were previously made using inefficient, simplistic approaches or by guessing." It is a software system that aids analysts in predicting or identifying meaningful information from raw

datasets, papers, or business models in order to assess an issue and resolve it through decision-making. The application of data mining techniques will be used to create the implementation of the prediction system in its final form. This study is a continuation of the previous research and is divided into several portions, each of which was completed by a different researcher in the same field. It is organised in the following ways: Detailed explanation of the goal of this research may be found in the section on machine learning. The dataset described in the section on machine learning must be used in order to execute the experiments detailed in this part. Section V goes into further detail about the methodology used to do the analysis, which includes classification approaches such as J48 and LAD Tree, as well as lazy learner algorithms like as IBK and LWL. Section V also goes into greater detail about how to conduct the study. The study is being carried out with the help of the WEKA tool. Section VI describes the WEKA experimental methodology, whereas Section VII illustrates the confusion matrix for each of the classifiers employed in the experiments. Section VI also describes the WEKA experimental methodologies. This portion of the text provides an in-depth explanation of performance metrics, as well as their broad implications in general. There are definitions for the phrases root mean square error, mean absolute error, root mean square error, sensitivity, specificity, and accuracy. There are also definitions for the terms sensitivity, specificity, and accuracy. Definitions are provided for the words sensitivity, specificity, and accuracy, among other things. Each classifier's performance is evaluated using three metrics: the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Relative Absolute Error (RAE). The Root Mean Squared Error (RMSE) is the most commonly used statistic (RAE). Accuracy is also explored in depth towards the conclusion, as well as in the perspective of future study. Conclusion:

System Analysis

3.1 Existing system:

Since they would not be suggested if there was no reason for them to be, all recommended meals must have a high level of stability in order to be recommended. Various studies carried out in the past have indicated that data mining methodologies may be employed to aid in the development of a successful information system. As a result, it will be feasible to minimise the amount of human involvement required to handle difficult agricultural challenges. In order to assess agricultural data, a variety of different processes were used, and the findings were then compared in order to establish which strategy was the most effective overall. Precision agriculture was achieved through the use of the ensemble model, and the use of data mining methods was advised for the application of the Crop Selection Method (CSM), which offers information on the planting sequence of crops. Different data mining techniques and how they can be applied in the agriculture sector were investigated by the research team, which also analysed agriculture datasets for prediction of rice yield in the humid subtropical climate zone and tropical wet and dry climate zone in India, as well as worked on the implementation of the Crop Selection Method (CSM). This study examined the use of data mining techniques such as Support Vector Machine (SVM), Multiple Linear Regression (MLR), Adobos, and Modified Non Linear Regression on agricultural datasets for the purpose of comparing different algorithms. Factors such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Accuracy, or factors such as Average Rainfall in a year were considered when comparing the algorithms in this study. Throughout [19], the authors treat precision agriculture as a broad agricultural practise, not as a subset of precision agriculture. According to a study published in [20], the collecting of remote sensing data and the subsequent use of that data to develop indices for assessing agricultural productivity were investigated, with the results published in [21]. Some of the disadvantages of classification algorithms such as

KNN, Bayesian Network, and Decision Tree are listed below:

This classification approach's performance is lower to that of other classification approaches, for the following reasons:

2. A lower degree of accuracy has been applied to the data.

3.2 Proposed system:

As a result of the quantity of Creatinine present in the body, it has been hypothesised that we could be able to produce meal suggestions for people. On the contrary, the research described above is carried out mostly through long-term clinical trials that exclusively propose diets for certain disorders, and it seldom explores the association between nutritional elements and diseases using data mining techniques.

1. As a result of the adjustment, the overall performance has been enhanced significantly.

Second, they are less cost-effective than alternative solutions.

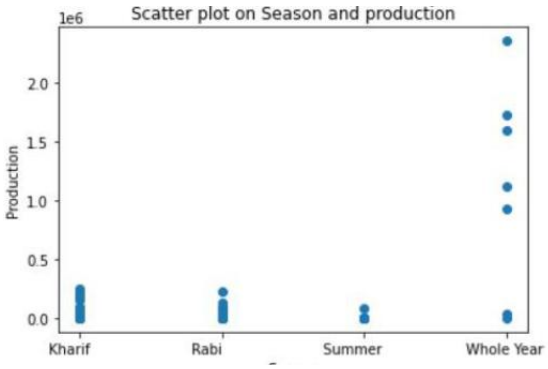
4. Results:

	District_Name	Crop_Year	Season	Crop	Area	Production
0	AURANGABAD	2014	Kharif	Arhar/Tur	32900	8700.0
1	AURANGABAD	2014	Kharif	Bajra	45300	24200.0
2	AURANGABAD	2014	Kharif	Cotton(lint)	422600	244300.0
3	AURANGABAD	2014	Kharif	Groundnut	4100	900.0
4	AURANGABAD	2014	Kharif	Jowar	3700	1700.0

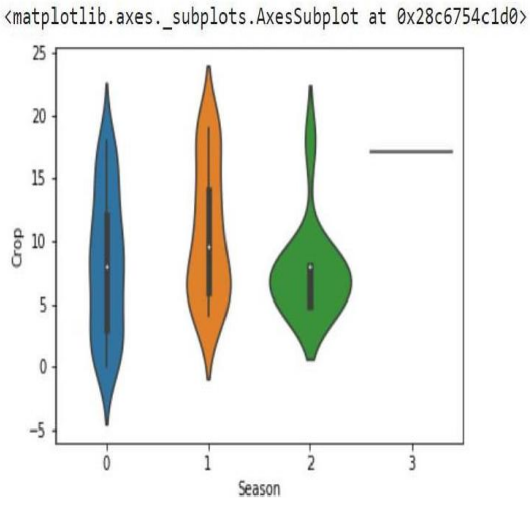
Data main to identify dataset

	Crop_Year	Area	Production
count	151.0	151.000000	1.350000e+02
mean	2014.0	38315.238411	7.848001e+04
std	0.0	74309.750744	3.074922e+05
min	2014.0	100.000000	1.000000e+02
25%	2014.0	1300.000000	6.000000e+02
50%	2014.0	6200.000000	4.200000e+03
75%	2014.0	41450.000000	2.495000e+04
max	2014.0	422600.000000	2.354600e+06

Data Description of dataset



Scatter Plot graph



Violin plot graph

Conclusion:

In order to limit the number of mistakes made by various classifiers to a bare minimum, it is necessary to trim the dataset further and lower the confidence factor. The lesser the amount of errors, the more accurate the analysis will be, and the reverse is true. In terms of accuracy, IBK achieves the highest level of precision, whereas LAD Tree achieves the lowest level. In order to facilitate early prediction and decision-making, farmer-friendly information may be assembled from the information gained through the research of the machine learning algorithm currently under development in a manner that is user-friendly. As a result of this information, the proportion of losses and bad output will decrease, as the management of the entire process may be performed through the use of real-time data. It is possible to make use of real-time weather and soil datasets in the future, either by collecting them manually with equipment or by acquiring them from reputable websites such as ndiaagristat.com or india.gov.in/data-portal-India in the near future, but this will be more difficult in the short term. We may combine several classifiers to form a single model, which is referred to as an Ensemble, and then modify the model as needed to improve its performance. Via the use of this strategy, we are able to achieve levels of performance that would be difficult to achieve through the use of a single method alone. Additionally, the nature of the data collection has an influence on the analysis; as a result, more cleaned and pre-processed data may be used to get better results in the study, if this is possible. Using machine learning methods such as random forest, linear regression, svm, and KNN algorithms based on results to analysis, weka library in python was integrated with machine learning to increase the performance of the executable on data set to forecast the outcomes.

REFERENCES

1. Mucherino A, Papajorgji P, Pardalos PM: A survey of data miningtechniques applied to agriculture. Oper Res. 2009, 9 (2): 121-140.

2. Cunningham, S. J., and Holmes, G. (1999). Developing innovativeapplications in agriculture using data mining. In the Proceedings of theSoutheast Asia Regional Computer Confederation Conference,1999.

3. N.Gandhi and L.J. Armstrong, "Applying data mining techniques to predictyield of rice in Humid Subtropical Climatic Zone of India", Proceedings ofthe 10th INDIACom-2016, 3rd 2016 IEEE International Conference onComputing for Sustainable Global Development, New Delhi, India, 16th to18th March 2016.

4. N. Gandhi and L. Armstrong, "Rice Crop Yield forecasting of Tropical Wetand Dry climatic zone of India using data mining techniques", IEEEInternational Conference on Advances in Computer Applications (ICACA),pp. 357-363, 2016.

5. N.Gandhi and L.J. Armstrong, "A review of the application of data miningtechniques for decision making in agriculture", 2016 2nd InternationalConference on Contemporary Computing and Informatics(ic3i).

6. S.Pudumalar, E. Ramanujam, R. HarineRajashreeń, C. Kavyań, T.Kiruthikań, J. Nishań, 'Crop Recommendation System for PrecisionAgriculture', 2016 IEEE Eighth International Conference on AdvancedComputing (ICoAC).

7. Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh (2015), 'CropSelection Method to Maximize Crop Yield Rate using Machine LearningTechnique', International Conference on Smart Technologies and60anagement for Computing, Communication, Controls, Energy andMaterials (ICSTM).

8. WEKA 3: Data Mining Software in Java, Machine Learning Group at theUniversity of Waikato, Official Website:http://www.cs.waikato.ac.nz/ml/weka/index.html, accessed on 12nd October2017.

9. D. Ramesh and B. Vardhan, "Analysis of crop yield prediction using datamining techniques", International Journal of Research in Engineering andTechnology, vol. 4, no. 1, pp. 47-473, 2015.10. Umid Kumar Dey, Abdullah HasanMasud, Mohammed NazimUddin,"Rice yield prediction model using data mining", International Conferenceon Electrical, Computer and Communication Engineering (ECCE), February16-18, 2017, Cox's Bazar, Bangladesh.

11. A. Ahamed, N. Mahmood, N. Hossain, M. Kabir, K. Das, F. Rahman, R.Rahman, "Applying data mining techniques to predict annual yield of majorcrops and recommend planting different crops in different districts inBangladesh", 16th IEEE/ACIS International Conference on SoftwareEngineering, Artificial Intelligence, Networking and Parallel/DistributedComputing (SNPD), pp. 1-6, 2015.

12. A.B. Mankar and M.S. Burange, "Data Mining-an evolutionary view ofagriculture", ", International Journal of Application or Innovation inEngineering and Management, Vol 3, No 2 pp .102-105, March 201413. JharnaMajumdar, SnehaNaraseeyappa and ShilpaAnkalaki, "Analysis ofagriculture data using data mining techniques: plication of big data",Jouranl of Big Day, Springer Open.