IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12742

ANALYSIS OF AGRICULTURAL DATA USING DATA MINING TECHNIQUES

G.Ramya

M.Phil, Research Scholar, H.H.The Rajah's College (Autonomous), Pudukkottai, India

Abstract: Agriculture is undoubtedly the largest livelihood provider in India and contributes a significant figure to the economy of our Country. The technological factors affecting the crop production includes practices used and also managerial decisions. So, predicting the crop yield prior to its harvest would help farmers to take appropriate steps. We attempt to resolve the issue by building a user-friendly prediction system. The results of the prediction are suggested to the farmer such that suitable changes can be made to improve the produce. There are different techniques or algorithms which help to predict crop yield. By analyzing all the parameters like location, soil nutrients, pH value, rainfall, moisture a potential solution can be obtained to overcome the situation faced by farmers. This paper focuses on the analysis of the agriculture data and finding optimal yield to provide an insight before the actual crop production using data mining techniques and Machine Learning algorithms.

Keywords*:* Data mining, Random forest regression, Decision Tree regression, GDP.

I.INTRODUCTION

Today, India is one of the main makers across the world in the farming area. Horticulture is the broadest monetary area and assumes a remarkable part in the financial piece of India. Horticulture is an unconventional business crop creation which is impacted by numerous background and monetary factors. Andhra Pradesh, fundamentally being an agro-Based economy offers over 29% of the Gross domestic product as against 17% in the nation's Gross domestic product. Periodical guidance to the ranchers either as far as improved farming procedures or headways in factors influencing the creation of harvests may fortify the state in thehorticulture sector. Yield forecast is one among the rural progressions.

Because of these sorts of developments farming is driving the interest of present-day man. In the past ranchers used to anticipate their yield from past encounters. Digitalization in cultivating gives mindfulness about the development of the yields at the perfect time and at the perfect spot even to youthful ranchers. These sorts of headways need the utilization of information analytics. This is one such framework that can be utilized to address yield forecasts.

II.HOW DATA MINING IS USED IN AGRICULTURE SECTOR

Data mining techniques are used in performing several activities in the agricultural sector such as pest identification, detection and classification and prediction of crop diseases. It can also be used in yield prediction, input management (planning of irrigation and pesticides), fertilizer suggestion and predicting soil. In a world full of data , data mining is the computational process for discovering new patterns[3]. Data mining techniques provide a major advantage in agriculture for detection and prediction for optimizing the pesticides. Techniques for agriculture related activities provide a lot of information. The yield of agriculture primarily depends on diseases, pests, weather conditions, planning of various crops for the harvest productivity are the results.

Crop production for reliable and timely requirements for various decisions for agriculture marketing. Predictions are very useful for agriculture data. For instance, by applying data mining techniques, the government can fully benefit from data about farmers' buying patterns and also to achieve a superior understanding of their land to achieve more profit on the farmer's part.

Data mining techniques followed in two ways[4]:

1) Descriptive data mining.

2) Predictive data mining.

Descriptive data mining tasks characterize the final properties of the info within the database while predictive data mining is employed to predict the direct values supported patterns determined from known results. Prediction involves using some variables or fields within the database to predict unknown or future values of other variables of interest.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12742

III.PROPOSED SYSTEM

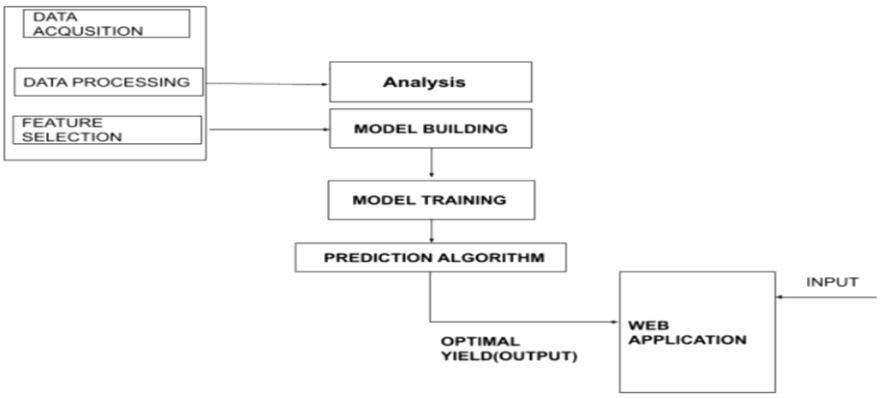
The main objectives of proposed work is to analyse the agricultural parameters using data mining algorithms and predict the yield.Inour proposed work, agriculture data has been collected from various sources which include: Dataset in agricultural sector[5] ,Crop wise agriculture data:[6], Soil data of different districts:[7]

In this proposed system , we mainly focussed on Andhra Pradesh State in India. As the state has two major rivers flowing , it has a diversity in factors useful for agriculture at district level. Periodical data about the crop , soil and water a particular region is the major focus of this study.The final dataset has been tabulated as in table-1:

Sno	Feature	Description
1	Year	The year in which the crop will be cultivated. Generally the upcoming year
2	Season	One among Kharif,Rabi and Whole Year.
3	Crop	Name of the crop
4	District	Name of the district
5	pH Level	This describes the nature of the soil
6	Nitrogen	Amount of nitrogen present
7	Potassium	Amount of potassium present
8	Phosphorus	Amount of phosphorus present
9	Rainfall	Expected rainfall in millimeters
10	Area	Area of field in hectares

Description of Input data

The below diagram depicts the system architecture of our proposed system. Our whole system can be divided into 2 modules as a whole i.e., one model predicts the optimal yield and the other model analyses the patterns in the dataset. The operation of these models as a whole is specified clearly in the below diagram.



The blueprint of the proposed system

IV.METHODS

In the implementation of this yield prediction system Regression Analysis is used.Regression Analysis is considered as one of the oldest,and widely used multivariate analysis techniques in the social sciences. Unlike others regression stands as an example of dependence analysis in which the variables are treated asymmetrically. In regression

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023
DOI: 10.17148/IJARCCE.2023.12742

analysis, the object is to obtain a prediction of one variable, based on given the values of the others[8].Random Forest and Decision Tree algorithms are generally used in classification problems but these can also be used in regression problems as well.

A. Decision Tree Regression

The Decision Tree algorithm comes under supervised machine learning techniques. A decision tree arrives at an outcome by asking a series of questions to the input data , each question narrows down the possible outcomes until the model gets enough potential to make a unique prediction[9]. The order of the questions as well as their contents are being determined by the model.All the questions that are raised have their answer as either true or false.

B. Random Forest Regression

Random Forest algorithm comes under the family of ensemble algorithms. This is also a supervised learning algorithm. This can be implemented in classification and regression as well. Random forest algorithm basically works on Decision Tree principle by constructing a number of decision trees having different sets of hyper-parameters for tuning and training on different subsets of data[10].

V.EXPERIMENTAL RESULT

PERFORMANCE ANALYSIS

Experimental data in science and engineering is data produced by a measurement, test method, experimental design, or quasi-experimental design. Experimental data can be reproduced by a variety of different investigators and mathematical analysis may be performed on these data.

Cross Validation Score

Cross-validation may be a statistical procedure that is used to estimate the skill of machine learning models. It is commonly utilized in applied machine learning to match and choose a model for a given predictive modeling problem because it is easy to know, easy to implement, and leads to skill estimates that generally have a lower bias than other methods. It is also known as a resampling procedure used to evaluate machine learning models on a limited data sample. Cross-validation gives a more accurate measure of model quality, which is especially importantif you are making a lot of modeling decisions. Sometimes it takes longer to run because it estimates multiple models. It is a popular method because it is simple to understand and it generally results in a less biased or less optimistic estimate of the model skill than other methods,such as a simple train/test split.

Cross-validation of Random forest and Decision Tree Regression can be seen in below The number of splits are 5 and test size is 0.2 which means 20 records out of every100 records are taken for the test.

Once the data has been pre-processed, it's time to train the model. First, select the algorithm that most closely aligns with the machine learning task to be performed. Because the predicted value is a numerically continuous value, the task is regression. One of the regression algorithms implemented by ML.NET is the StochasticDualCoordinateAscentCoordinator algorithm. To train the model with cross-validation use the CrossValidate method.

Performance Measures

The end users of prediction tools should be able to understand how evaluation is done and howto interpret the results. Six main performance evaluation measures are introduced.

These include

- Sensitivity
- Specificity
- Positive predictive value
- Negative predictive value
- Accuracy and
- Matthews correlation coefficient.

▪

Accuracy: Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positives and false Negatives are almost the same. Therefore, you must look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.

Accuracy = TP+TN/TP+FP+FN+TN

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12742

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes, and the value of predicted class is also yes. E.g. if the actual class value indicates that this passenger survived, and the predicted class tells you the same thing.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no, and value of predicted class is also no. E.g. If the actual class says this passenger did not survive and the predicted class tells you the same thing.

False Positives (FP) – When actual class is no and predicted class is yes. E.g. if the actual class says this passenger did not survive but the predicted class tells you that this passenger will survive.

False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if the actual class value indicates that this passenger survived, and the predicted class tells you that the passenger will die.

Performance Metrics

To evaluate how good our regression model is, we can use the following metrics:

- **R-squared:** indicate how many variables compared to the total variables the model predicted. R-squared does not take into consideration any biases that might be present in the data. Therefore, a good model might have a low R-squared value, or a model thatdoes not fit the data might have a high R-squared value.
- **Average error:** the numerical difference between the predicted value and the actual value.
- **Mean Square Error (MSE)**: good to use if you have a lot of outliers in the data.
- **Median error:** the average of all differences between the predicted and the actual values.
- **Average absolute error:** like the average error, only you use the absolute value of the difference to balance out the outliers in the data.
- **Median absolute error** represents the average of the absolute differences between prediction and actual observation. All individual differences have equal weight, and big outliers can therefore affect the final evaluation of the model.

Mean Absolute Error

Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on all instances in the test set. Each prediction error is the difference between the true value and the predicted value for the instance.

Mean Square Error

The mean squared error (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the"errors") and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It's called the mean squared error

Root Means Square Error

Root mean squared error (RMSE) is the square root of the mean of the square of all of the errors. The use of RMSE is very common, and it is considered an excellent general-purpose error metricfor predictions. RMSE is a good measure of accuracy, but only to compare prediction errors of different models or model configurations for a particular variable and not between variables, as itis scale-dependent.

The models are tested using the above metrics and their results are compared manually. The below figure 8.2 shows Comparison of performance metrics on Random Forest and Decision Tree Regression

Experimental Results

The best fit model for our system has been found out through the above-mentioned model's comparison. So now let's take some sample data and analyze how our model is performing with respect to that data.

The below Table-8.1 shows a sample of data points along with sample id, it's actual rating and predicted rating by the model.

Test Case Number	Actual Value	Predicted Value
01	18000	18000
02	7100	7100
03	9400	9400
04	7100	7310
05	500	500

DOI: 10.17148/IJARCCE.2023.12742

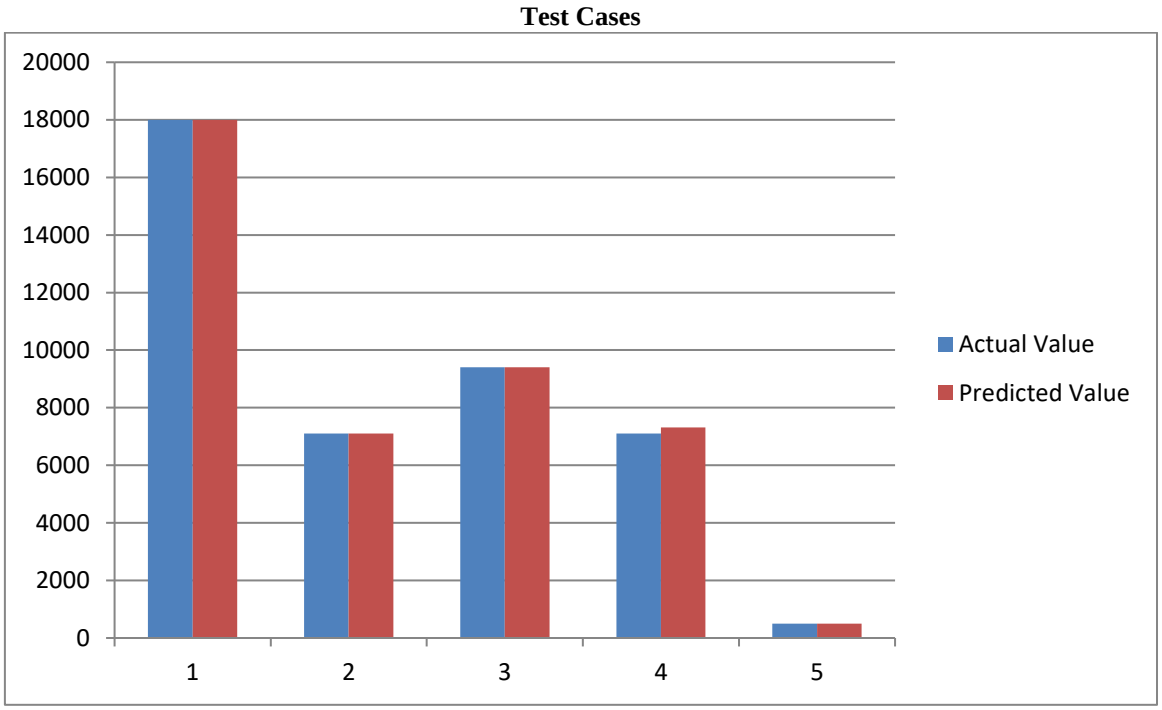


Fig 8.3 Test Cases

In the above table, the performance of the system is compared for a sample of 5 data points. The predictions are made by the Random Forest Regressor Model as it is our best fit model. We can see for our sample that predictions production value do not deviate / vary much from the actual production value.

VI.CONCLUSION

Both Decision tree regression and Random Forest regression techniques are implemented on the input data to assess the best performance yielding method. These methods are compared using performance metrics. According to the analyses of metrics both the algorithms work well , but Random Forest regression gives a better accuracy score on test data than Decision tree regression. The proposed work can also be extended to analyse the climatic conditions and other factors for the crop and to increase the crop production.

REFERENCES

[1]. Veenadhari S, Misra B, Singh CD. Data mining techniques for redicting crop productivity—A review article. In: IJCST. 2011; 2(1).
[2]. Gleaso CP. Large area yield estimation/forecasting using plant process models.paper presentation at the winter meeting American society of agricultural engineers palmer house, Chicago, Illinois. 1982; 14–17
[3]. Majumdar J, Ankalaki S. Comparison of clustering algorithms using quality metrics with invariant features extracted from plant leaves. In: Paper presented at international conference on computational science and engineering. 2016.
[4]. Jain A, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31(3):264–323.
[5]. Jain AK, Dubes RC. Algorithms for clustering data. New Jersey: Prentice Hall; 1988.
[6]. Berkhin P. A survey of clustering data mining technique. In: Kogan J, Nicholas C, Teboulle M, editors. Grouping multi- dimensional data. Berlin: Springer; 2006. p. 25–72.
[7]. Han J, Kamber M. Data mining: concepts and techniques. Massachusetts: Morgan Kaufmann Publishers; 2001.
[8]. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Paper presented at International conference on knowledge discovery and data mining. 1996
[9]. Ramesh D, Vishnu Vardhan B. Data mining techniques and applications to agricultural yield data. In: International journal of advanced research in computer and communication engineering. 2013; 2(9).
[10]. MotiurRahman M, Haq N, Rahman RM. Application of data mining tools for rice yield prediction on clustered regions of Bangladesh. IEEE. 2014;2014:8–13.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023
DOI: 10.17148/IJARCCE.2023.12742

[11]. Verheyen K, Adrianens M, Hermy S Deckers. High resolution continuous soil classification using morphological soil profile descriptions. Geoderma. 2001;101:31–48.
[12]. Gonzalez-Sanchez Alberto, Frausto-Solis Juan, Ojeda-Bustamante W. Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res. 2014;12(2):313–28.
[13]. Pantazi XE, Moshou D, Alexandridis T, Mouazen AM. Wheat yield prediction using machine learning and advanced sensing techniques. Comput Electron Agric. 2016;121:57–65.
[14]. Veenadhari S, Misra B, Singh D. Machine learning approach for forecasting crop yield based on climatic parameters. In: Paper presented at international conference on computer communication and informatics (ICCCI-2014), Coim- batore. 2014.
[15]. Rahmah N, Sitanggang IS. Determination of optimal epsilon (Eps) value on DBSCAN algorithm to clustering data on peatland hotspots in Sumatra. IOP conference series: earth and environmental. Science. 2016;31:012012.