www.ijaceeonline.com ISSN: 2456 - 3935
**International Journal of Advances in Computer and Electronics Engineering**
Volume 3, Issue 12, December 2018, pp. 1–6

# Survey on Crop Yield Prediction Using Data Mining Techniques

Prashant Govardhan
Assistant Professor, Department of Computer Science and Engineering,
Priyadarshini Institute of Engineering and Technology, Maharashtra, India
Email: er.phgovardhan@gmail.com

Rasika Korde
UG Scholar, Department of Computer Science and Engineering,
Priyadarshini Institute of Engineering and Technology, Maharashtra, India
Email: rasikakorde8997@gmai.com

Rashi Lanjewar
UG Scholar, Department of Computer Science and Engineering,
Priyadarshini Institute of Engineering and Technology, Maharashtra, India
Email: rashilanjewar111@gmail.com

**Abstract:** *Food plays a very important role in every living being lives. Farmers are the backbone to Yield crop in human favorable habitat. It's not mandatory that every farmer knows the best method for cultivation of crops. The technologies have gone beyond the trivial methods that farmers were using earlier. Such software is boon in the production of agriculture and the fields are data mining, big data, data science. These fields help us to analyze and predict and provide best results. In the domain of the data mining, we have enormous algorithms to find the accuracy of the equation precisely. It also tries to give the nearer best as well as the accurate percentage to find the answers through the dataset. In totality, such kind of works will help us to yield the good crop which maximizes the production and never the less helps to reduce the unwanted loss in agriculture*.

**Keyword:** *Crop; Yield; Climate; Season; Data Science; Data Cleaning;*

## 1. INTRODUCTION

Agriculture is a backbone to the Indian economy. India is the second highest crop producing country in the world. The agriculture is very important aspect of the Indian occupation, the Indian economy mostly include the agricultural production. Nearly 2/3 of the Indian population directly depends upon the agriculture for its livelihood. As the time passed the methods of agriculture has changed. Many new technologies came into existence day by day today in the world of science and technology we have many new ideas and methodology for the agriculture There are several crops which are being produced every year. In the small village's agriculture is the highest occupation of the people.

The production of crops depends on the weather, temperature, soil and many other factors which are directly proportional in the crop production. There is different season in which crops are produced, some are kharif, rabbi, summer, winter, and some crops can be produced whole year. Different crops are produced in different seasons. The crops which are produced whole year can probably give more production. Mostly the agriculture is done in the village.

The literacy rate among the farmers are less, hence it is bit difficult for them to have the knowledge about the modern agriculture which are followed in the foreign countries. The farmers yield their crops from their experience. But sometimes due to lack of knowledge and poor prediction of weather amongst the farmers the production of the crops gets reduced. To create awareness among the farmers for the better production of crops the new technique is introduce using the data mining. Data mining is the field of data science which helps us to predict the changes from the large amount of historic data. Such large amount of data which could not be easily understood and studied in small amount of period and to avoid this the data mining concept is used, which filters the unwanted data and get the useful data and from it so that we could analyze the future trends of agricultural process.

In the process, data mining provides multiple algorithms to work with and get the best possible results out of it. The methods help us to find the optimal accuracy by applying the different algorithms simulta-

**Cite this paper:**

Prashant Govardhan, Rasika Korde, Rashi Lanjewar, "Survey on Crop Yield Prediction Using Data Mining Techniques", International Journal of Advances in Computer and Electronics Engineering, Vol. 3, No. 12, pp. 1-6, December 2018.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 12, December 2018, pp. 1–6

neously and comparing their results by confusion matrix. Confusion matrix helps us to find the best accurate results of the algorithms, we have various aspects to find the error percentage, lesser the error more will be the accuracy of the algorithm by which we can judge which one of the algorithms is best to follow. It is seen that different algorithms have different capacity to work with the quantity of datasets, algorithm like SVM (Support Vector Machine) works with the smaller number of datasets where as J48 gives the best possible results even if with the quantity of lakhs of rows in the datasets. Algorithms such as LAD Tree, neural networks, JWL also give the average results. By obtaining such results, it will be a great bliss in the field of agriculture to cope up with the losses farmers face.

## 2. LITERATURE SURVEY

Agriculture not only provides live but also gives bread and butter to many people, if one is affected them then others affect too. There are many reasons to make crop cultivation strong. To make it happen as good as possible many researches has been implemented and many more are still going on. such kind of work was done by Pune students in which they have used the WEKA tool and applied the datasets from kaggle.com which was in CSV format [1]. The initial step of the research was data cleaning. Because the works is being processed in WEKA Tool there is not much human effort, the software itself has the capability of pre-processing, classification, regression, clustering and visualization [1]. There are inbuilt algorithms and classifications. and analyzed the crop yield prediction but with different kind of crops [1]. For the pre-processing DSS is used for the datasets, further for analysis part different algorithms are used namely J48, LAD tree, LWL AND IBK [1]. After executing these algorithms to find the accurate algorithm error methods are used namely RMSE, MAE, and RAE [1]. To describe the performance of the classifiers confusion matrix is used and hence gave us the results that LAD has the lowest accuracy whereas IBK gave the highest accuracy.

Similarly, in Bangladesh a research has been done regarding rice yield predictions that too of different kinds, since Bangladesh is a country with a blessed weather condition favorable for the rice cultivation [2]. The research was done with the data mining concepts in which they have considered the tuples of rainfall, humidity, area and yield for the acquire datasets and applied the various equations like MLR, AdaBoost, SVMR and MNR for the values. Further they have checked the accuracy of the values by applying error techniques of RMSE, MSE and MAE and lastly concluded that MNR is the best equation found among the three and not only this it has given the highest R- square value [2].

In brief survey the author seeks for the techniques in data mining which will be helpful for the researchers to get the details of the techniques [3]. They have classified the task into two main data mining categories – Descriptive and predictive. From their study they have analyzed that descriptive data mining approaches are used more as compare to predictive and found out the data mining techniques called classification, clustering, association rule mining and regression and multiple number of classification techniques for the discovery of knowledge base namely- Rule Based Classifiers, Bayesian Networks, Decision Tree, Nearest Neighbor, Artificial Neural Network, SVM, Fuzzy Logic, Rough set, Genetic algorithms [3]. Hence lastly stated that these data mining techniques will be fruitful for the farmers and will help them to cope up from the calamity [3].

The prediction techniques were also implemented in the fruits crop like grapes which are also used for the winemaking. Such works was done in New Zealand for making the grape wine with two major aspects of temperature difference in different months and area of land in yards [4]. They stated that daily whether, grapevine phenology and yield indicators are directly connected to each other, they got the daily whether updated from the nearby meteorology station. Their main work was done in the algorithms of Neural Networks and calculations were obtained and the calculation part was done in chi-square methods to find the results among the inter related factors [4]. The conclusion was found that the temperature and other terms related to it like wind-speed, humidity, precipitation played a major role for the quality and cultivation of the grape wine.

India is a country where rice consumption is 40%, due to fitful climatic conditions crops are affected [5]. Hence the study has been conducted in the Western Australia aiming the prediction of rice production via Neural Networks in various districts of Maharashtra State considering the frameworks of temperature, precipitation, area, the analysis was done in the Kharif season for the time period of 1998-2002 [5]. The process was run in the WEKA tool in which Perceptron Neural Network a multilayer thing was developed [5]. To validate the data set the cross validation was used. The results were obtained by MLP and further the results were analyzed for the accuracy by confusion matrix. The researcher lastly inferred that the ANN is one of the possibilities for the linear regression techniques and also that ANN is more accurate as compared to other techniques used by them [5].

The effective supervised filter-based feature selection algorithm using rough set theory paper published by Rubul Kumar Bania assistant professor in computer application department of North Eastern hill University has stated in this paper that, the data is generally represented by high dimensional vector in many areas. feature selection is the important aspect in data mining [6]. Using the feature selection algo-

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 12, December 2018, pp. 1–6

rithm, we filter the data and select the required attribute in the data. with the help of feature selection, the accuracy of data increases [6]. The rough set theory is used, it is based on the two important concepts, an upper and a lower approximation of a set [6]. The lower approximation is a description of the domain object which are known with certainty to belong to the subset of interest whereas the upper approximation is a description of the object which possibly belong to the subset. here the datasets are taken from the UCI repository of machine learning, they have used the WEKA tool for the classification process [6]. It is an open source Java based machine learning workbench. The JRip and J48 are used for the classification task. J48 is a decision tree-based classifier and JRip is a rule-based classifier [6].They have analyzed that pattern selection is still challenging in feature selection ,also they have proposed the improvement in QR algorithm with different stopping criterion[6].

The paper Performance Analysis of Feature Selection Algorithm for Educational Data Mining is based on feature selection used in the data mining for the analysis of performance for educational purpose [7]. Feature Selection is very dynamic and productive field and research area of machine learning and data mining [7]. The main goal of feature selection is to choose a subset by eliminating nonpredictive data. Furthermore, it increases the predictive accuracy and reduces the complexity of learned results [7]. Feature Selection techniques can be classified in to three groups: filter, wrapper, and embedded models. Filter method depends upon general characteristics of training data, this method is done on pre-processing stage and not dependent on a learning algorithm [7]. Wrapper method uses learning algorithms to evaluate the features. Embedded methods are specific to some given learning algorithms, and these methods are performed on training process of classifiers.

The main aim of this paper is to apply different feature selection algorithm on the datasets and check which works better. In this paper the datasets are taken from the Kaggle.com and Weka tool is used to perform all the work of this experiment [7]. Weka is an open source data mining tool. In this research work six FS algorithm Cfs Subset Eval, Chi Squared Attribute Eval, Filtered Attribute Eval, Gain Ratio Attribute Eval, Principal Components, and Relief Attribute Eval are evaluated. The classification algorithm Bayes Net (BN), Naïve Bayes (NB), Naïve Bayes Updateable (NBU), MLP, Simple Logistic (SL), SMO, Decision Table (DT), Jrip, OneR, OneR, Decision Stump (DS), J48, Random Forest (RF), Random Tree (RT), REP tree (RepT) are evaluated through the educational data set. among all these available feature selection methods, principal components have shown better results by using it with Random Forest classifier. This study has also shown that MLP classifier performed slightly better than other classifiers on student data set [7]. In their input datasets Random Forest gave the better results among all the classifiers and found there is not major change in the feature selection algorithm in the WEKA tool [7].

In the paper predicting fault- Prone software Module using feature selection and classification through data Mining Algorithms the author has stated about the feature selection and classification through data mining algorithm [8]. They have studied on the software defect detection system using the supervised machine learning techniques. They have used seven datasets (CM1, JM1, MW1, KC3, PC1, PC2, PC3 and PC4) and they have categorized them in two classes namely defective and normal. Feature selection, is the method of deciding on a subset of important features for building reliable learning models [8]. It makes training and utilizing a classifier more efficient by reducing the size of the effective training set. Moreover, feature selection often increases classification accuracy by removing noise features. They didn't used any specific feature selection method, they have used random tree classification algorithm for the classification of the data. This research work was specifically centered on investigating the performance of the classification techniques on categorizing the nature of software modules in publicly available datasets [8]. According to their research the random tree classification algorithm gave 100 percent results on their available datasets [8].

In the paper web document clustering using similarity measures the proposed work tells that there are lots of data being extracted daily from the internet, there are many people who surfs the data every now and then, the data stored in the database is in huge amount every data is not useful to the user every time so the data which is useful should be provided to the user, so for this purpose the clustering is important . using clustering the data which is useful is being extracted from the data. In this paper the web document clustering is done. The text processing plays an important role in the information retrieval [9].

The goal of clustering is to determine the intrinsic grouping in a set of unlabelled data. An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. A document is usually represented as a vector in which each component indicates the value of the corresponding feature in the document [9]. The authors found that study of similarity measure for clustering is initially motivated by a research on automated text categorization. The application of document clustering to information retrieval has been motivated by the potential effectiveness gains postulated by the cluster hypothesis [9].

The paper survey on data mining techniques in

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 12, December 2018, pp. 1–6

Agriculture, this paper discusses about the role of data mining in perspective of agriculture field and also confers about several data mining techniques .it also discusses on the different data mining applications in solving the different agricultural problems [10]. This paper provides a survey on different data mining techniques used in Agriculture such as Artificial neural networks, k-nearest neighbor, decision tree Bayesian network, fuzzy set, support vector machine and k- means. The data can be analyzed in a relational database, a data warehouse, a web server log or simple text file [10]. analysis of data in effective way requires understanding of appropriate techniques of data mining. The association rule mining helps to search unseen or desired pattern among the large amount of data. This method is used to find relationship between the different items. The different association rule mining algorithm are Apriori Algorithm (AA), Partition, Dynamic Hashing and Pruning (DHP), Dynamic Itemset Counting (DIC), FP Growth (FPG), SEAR, Spear, Eclat and Declat, MaxEclat.

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. It is a process in which a model learns to predict a class label from a set of training data which can then be used to predict discrete class labels on new samples [10]. To maximize the predictive accuracy obtained by the classification model when classifying examples in the test set unseen during training is one of the major goals of classification algorithm. The different classification techniques for discovering knowledge are Rule Based Classifiers, Bayesian Networks (BN), Decision Tree (DT), Nearest Neighbour (NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Rough Sets, Fuzzy Logic, Genetic Algorithms [10]. In clustering, the focus is on finding a partition of data records into clusters such that the points within each cluster are close to one another.

Clustering groups the data instances into subsets in such a manner that similar instances are assembled together, while dissimilar instances belong to diverse groups. The different clustering methods are Hierarchical Methods (HM), Partitioning Methods (PM), Density-based Methods (DBM), Model-based Clustering Methods (MBCM), Grid-based Methods and Soft-computing Methods [fuzzy, neural network based], Squared Error—Based Clustering (Vector Quantization), network data and Clustering graph [10]. Regression is learning a function that maps a data item to a real-valued prediction variable. The different applications of regression are predicting the amount of biomass present in a forest, estimating the probability of patient will survive or not on the set of his diagnostic tests, predicting consumer demand for a new product. The methods for prediction are Nonlinear Regression (NLR) and Linear Regression (LR) [10]. The author studied various papers as her research and winded up that it is useful for researchers to get information of current scenario of data mining techniques and applications in context to agriculture field [10].

The paper survey on approaches to feature selection states that, the feature selection the very important aspect in the data mining and machine learning to decrease the dimensionality of the data and increase the performance of an algorithm, such as a classification and clustering algorithm. There are various techniques of feature selection [11]. There are many features in the dataset from which many of them are of no use or unnecessary, so to use the important data an make the work easy the feature selection algorithms are used. A variety of search techniques have been applied to feature selection, such as complete search, greedy search, heuristic search, and random search. Research on feature selection started around 1990, but it has become popular since 2007, when the number of features in many areas became relatively large [11]. Based on the evaluation criteria, feature selection algorithms are generally classified into two categories: 1) filter approaches and 2) wrapper approaches. Their main difference is that wrapper approaches include a classification/learning algorithm in the feature subset evaluation step. The classification algorithm is used as a “black box” by a wrapper to evaluate the goodness (i.e., the classification performance) of the selected features. A filter feature selection process is independent of any classification algorithm. Filter algorithms are often computationally less expensive and more general than wrapper algorithms [11].The author have centered his research in the EC algorithm of feature selection which has recently gained importance in the large scale feature selection task [11].

Due to explosive growth of accessing information from the web, efficient access and exploration of information are needed critically. The Text processing plays an important role in information retrieval, data mining, and web search. Text mining attempts to discover new, previously unknown information by applying techniques from data mining [12]. Clustering, one of the traditional data mining techniques is an unsupervised learning paradigm where clustering methods try to identify inherent groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intra-cluster similarity and low inter-cluster similarity [12]. In this paper different techniques for similarity measure has been analysed for the improvement of accuracy ,precision and computational speed [12].

## 3. CONCLUSION

As per previous studies we have analyzed that there are different crops which are produced in different regions and they produce different amount of pro-

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 12, December 2018, pp. 1–6

duction from the same amount of yield. As India is the fast-developing country, so it is very important to make strong backbone for country's development. In India there are 29 states each state includes several numbers of districts in it. With such big geographical region, it finds too much difficult to predict crop production scenario because everywhere the climate is different. Due to which in every district different crop production is done from which some gives the more production and some gives less. To avoid unbearable loss and to increase the production of the crop yield the data mining technique can be used.

The data mining helps us in major aspect to accomplish our desired expectations related to the work of mining. Researchers have worked and analyzed the different algorithm and technique in favor to expand the production. Some of the researchers have found their results by using WEKA tool. Algorithm such as J48, SVM, and neural network have resulted that J48 is the highest accurate algorithm till now. In future we can have such algorithm whose accuracies are above average and by working with such kind of algorithm we can get the best possible results for reduction of crop loss and for the maximization of crop production.

## 4. ACKNOWLEDGEMENT

We would like to express our sincere thanks to Board of Management and Principal, Priyadarshini Institute of Engineering and Technology, Nagpur, Maharashtra, India for providing opportunity to carried out our work in very loving and peaceful environment. We are also thankful to Dr. P. S. Prasad, Head of Department, Computer Science and Engineering Department at Priyadarshini Institute of Engineering and Technology, Nagpur, Maharashtra, India for his constant support and guidance. Finally, we would like to thank all those who extended their direct or indirect support for carrying out our research work.

## REFERENCES

[1] Shruti Mishra, Priyanka Paygude, Snehal Chaudhary, Sonali Idate, (2018)," Use of Data Mining in Crop yield Prediction", Proceedings of the second International Conference on Inventive Systems and Control (ICISC), pp.971-1-5386-0807-4.

[2] Md. Abdul Rashid Sarker, Khor shed Alam, Jeff Gow, (2012)," Exploring the relationship between climate change and rice yield in Bangladesh: An analysis of time services data", Agricultural system 112, pp. 11-16.

[3] Hetal Patel, Dharmendra Patel, (2014)," A brief survey of data mining techniques applied to agricultural data", International journal of Computer Applications, vol.95-no.9, pp.0975-8887.

[4] Subana Shanmuganathan and Philip sallis, Ajit Narayan, "Data mining technique for modelling the influence of daily Extreme weather conditions on Grapevine, wine quality and perennial crop yield",2010 Second International Conference on Computational Intelligence, Communication Systems and Networks, pp.978-0-7695-4158-7.

[5] Niketa Gandhi, Owaiz Petkar, Leisa J. Armstrong. (2018)," Rice Crop Yield Prediction Using Artificial Neural Networks", 2016 IEEE International Conference on Technological Innovations in ICT For Agriculture and Rural Development (TIAR 2016), pp. 978-1-5090-0615-1.

[6] Rubul Kumar Bania, (2017)" An Effective Supervised Filter based Feature Selection Algorithm using Rough Set Theory", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017).

[7] Maryam Zaffar, Manzoor Ahmad Hashmani, K.S. Savita (2017),"Performance analysis of Feature Selection Algorithm for Educational Data Mining" 2017 IEEE Conference on Big Data and Analytics (ICBDA) , pp.978-1-5386-0790-9.

[8] Dr. R. Geetha Ramani, S. Vinodh Kumar, Shomona Gracia Jacob, (2012)" Predicting Fault –Prone Software Modules Using Feature Selection and Classification through Data Mining Algorithms", 2012 IEEE International Conference on Computational Intelligence and Computing Research, pp. 978-1-4673-1344-5.

[9] P. H. Govardhan, K. P. Wagh , P. N. Chatur (2014)," Web Document Clustering using Proposed Similarity Measure " International Journal of Computer Applications, *pp.0975-8887.*

[10] M.C.S Geeta. (2015),"A Survey on Data Mining Techniques in Agriculture" International journal of Inovative research in computer and communication Engineering, vol. . 3, Issue 2, pp. 2320-9798.

[11] Prashant Govardhan, Prakash Prasad(2017)," A Survey on Approaches to Feature Selection" International Journal of Advanced Engineering, Management and Science (IJAEMS),issue-2, pp.2454-1311.

[12] P. H. Govardhan, Prof. K . P. Wagh, Dr. P.N. Chatur (2013), "Survey on Similarity Measure for Clustering" International Journal of Advanced Research in Computer Science and Software Engineering, vol 3,issue 2, pp.2277-128X

### Authors Biography

**Mr. Prashant Govardhan,** in 2014 he joined Department of Computer Science and Engineering at Priyadarshini Institute of Engineering and Technology, Nagpur, Maharashtra, India as an Assistant professor. He received his B. E. and M. Tech. in Computer Science and Engineering from Sant Gadge Baba Amravati University, Amravati. His research work interest includes Data Mining, Data Science, Distributed Systems and Machine Learning. His work has been documented in many publications. Currently he is working on Data Mining techniques for crop yield prediction.

**Miss. Rasika Korde**, perceiving Under Graduate Course of Engineering in Computer Science and Engineering at Priyadarshini Institute of Engineering and Technology Nagpur, Maharashtra, India. She Received Diploma in Computer Engineering from Maharashtra State

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 12, December 2018, pp. 1–6

Board of Technical Education, Mumbai, India. Currently she is working on Data Mining Technique for Crop Yield prediction.

**Miss. Rashi Lanjewar,** perceiving Under Graduate Course of Engineering computer science and engineering from Priyadarshini institute of engineering and technology Nagpur Maharashtra. Ongoing she is active in the data mining techniques in the crop yield production.

**Cite this paper:**

Prashant Govardhan, Rasika Korde, Rashi Lanjewar, "Survey on Crop Yield Prediction Using Data Mining Techniques", International Journal of Advances in Computer and Electronics Engineering, Vol. 3, No. 12, pp. 1-6, December 2018.