Crop Prediction Under Disease (grass grub insect) Condition Using Hybrid Approach of Weighted K-Mean and Evolutionary Techniques.

[1] Manpreet Kaur Research Scholar, Department of Computer Applications, Guru Kashi University, Talwandi Sabo, PB, India

[2] Dr. Dinesh Kumar Associate Professor, Department of CSE, Guru Kashi University, Talwandi Sabo, PB, India

Abstract: Data analytics is the main focusing point for different fields. Agriculture field is the new entrant to the data analytics. It specifically picks the data related to agriculture different parameters and evaluates the parameters with different machine learning algorithms. In proposed technique the agriculture disease prediction based on different parameters has been evaluated. These parameters are related to patient different aspects like previous years produce, diseases etc. The proposed technique for the prediction is weighted k-mean and logistic regression. The proposed technique is showing better results in terms of accuracy, precision and recall. The accuracy improvement is around 1.67%, Recall is improved by 1.15% and precision is improved by 3.15%.

Keywords: Prediction, Logistic regression, k-mean

1. INTRODUCTION

Data analytics is picking and becoming natural choice for various different organizations to be dependent on. The big data analytics is used for different types of data processing and analysis. These analyzed data will be presented for having various strategic decisions. The quality of the decisions will be improved if the decisions are based on some authentic information. There are various advantages for the data analytics approach.

High quality: the quality of the decisions will be improved by involving different types of information. The quality of the decisions will directly be dependent on how well the information will be provided to the system.

Authentic: Data analytics will provide authentic information which will be useful for quality decision. There are various types of techniques which are used for generating the quality of the data. This will be authentic data will helps in taking authentic decisions.

Timely: Data analytics can produce quality information in timely manner. There are various types of fast processing algorithms which can process the data with higher level of authenticity and timely manner.

Agriculture field is one new addition to the whole scenario. Where large amount of patient related data will be used for prediction purpose. Patients previous data collected at different sources provide the way for prediction model to predict the disease for the patients. This will also enhance the system capacity to

have preventive medicines. The diagnosis for the different types of diseases can be based on prediction model. This will increase the accuracy of the diagnosis. These are automated machines which will continuously collect the data and generate the analysis.

There are various machine learning techniques which are used for the prediction purpose. Majority of the techniques are based on classification, will classify the whole data into four different classes one is the true positive, second is the true negative, false positive, false negative. The requires system accuracy is measured on the basis of total of true positive and true negatives [1].

These machine leaning techniques are

- **KNN**: It is the k-nearest neighbor. It identifies the centroid value of all the parameters. Each parameter distance value is measured from the centroid value. These distance values are kept in the descending order of distances. K numbers of elements are kept in one set and remaining in another. This technique is having higher level of accuracy for the predict purpose.
- **SVM**: is the support vector machine is based on supervised learning, where the total dataset will be sub divided into two classes. It is two class problem. It identifies the space vector line in the state space.
- **Decision tree**: It is the approach where total set will be having different conditions which classify the entities into multiple subsets based on the condition. There are two classes that can be prepared to have two subsequent classes of the dataset with tree of conditions. The proposed approach is having higher accuracy for classification.

2. LITERATURE SURVEY

[1] Mackay J et. al(2004): Author in this paper has proposed a technique which helps in prediction of the heart stoke for the patient. The proposed technique author has applied is for the large dataset having different data items related to patients. These patients related data is regarding different parameters which will helps in identifying the heart stoke prediction. The proposed SVM based technique is used which will generate the classes of the whole dataset. The proposed technique has achieved the success rate of 92%.

[2] Vasighi Mahdi et. al (2013): author in this paper has proposed a technique based on genetic based approach. It classify the dataset based on different features of the blood plasma. The large number of patients data will be having creation of the classes which generate the positive results for the prediction. The proposed genetic based approach has achieved the success rate of 93% for the prediction.

[3] Amin Mohammed Shafennor, et al. (2019): Author in this paper has proposed a data mining based approach for the heart rate prediction. The proposed approach is having higher level of accuracy of the prediction for the heart rate prediction. The proposed approach of decision tree will be used for

classification of the big data related to the patient different parameters. The proposed approach has achieved th accuracy of 95% . The proposed approach is having success in terms of achieving the final results.

[4] Nahar J et. al(2013): Author in this paper has proposed approach of expert system for classification. There are higher end entities which are used for generating more than one classes. The decision algorithm is used in association to the expert system. It will help in having genuine intelligence in the system for prediction. The proposed approach has achieved the accuracy of 97% over the multiple datasets. The prediction for the heart attack is having higher level of system automation.

[5] Guru Niti et. al(2017): Author in this paper has proposed a technique based on neural network for two different layers. There are different classes created using multiple layers. The proposed system is achieving the accuracy of 95% for the dataset related to IIT Delhi. The Proposed approach has created a supervised learning layered based mechanism. It will classify the whole system in four different classes. The proposed system is ready for further enhancing the results for given datasets.

3. METHODOLOGY

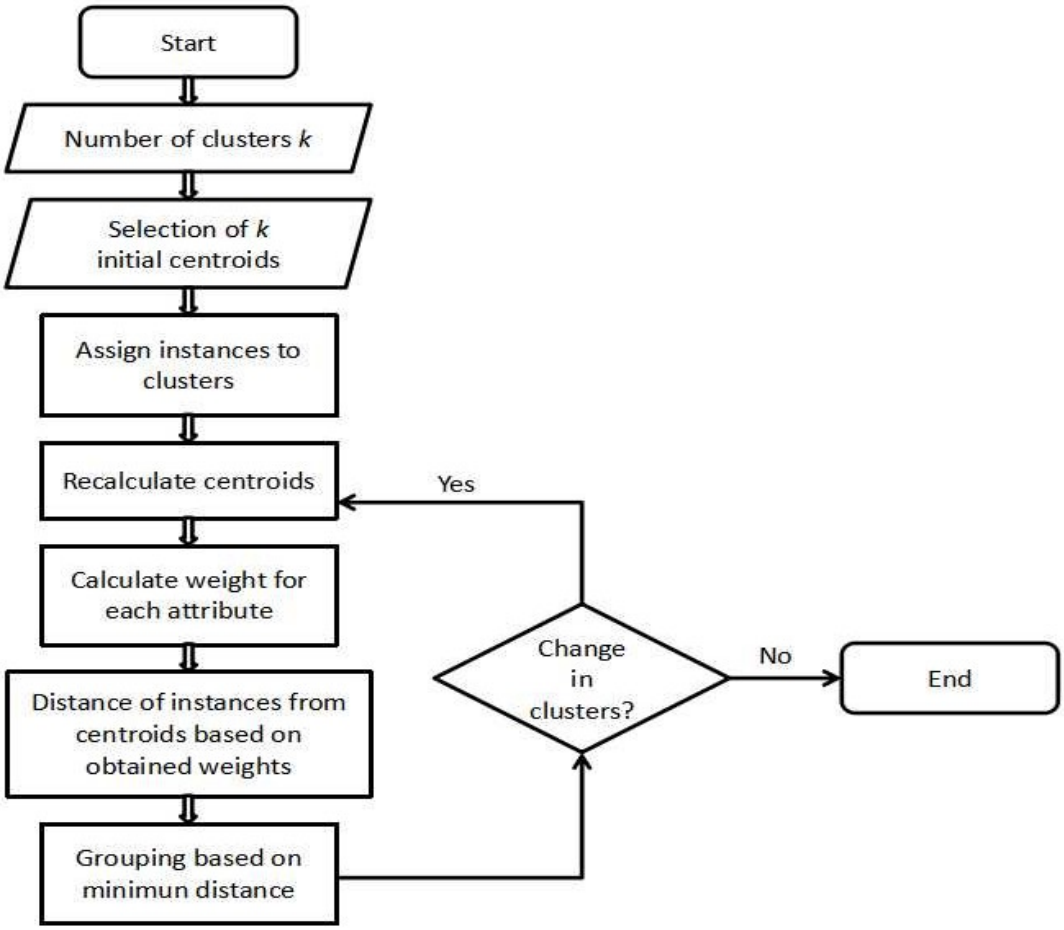


Fig. 1 Methodology

4. RESULTS

The proposed technique is a hybrid approach of weighted k-mean and logistic regression for prediction of the agriculture disease. The proposed technique has been applied on the dataset having different parameters. The proposed technique predicts for the disease whether the crop will be having disease or not. The accuracy of the prediction is compared with various classification techniques. The proposed hybrid technique performs with better accuracy, recall and precision.

4.1 Performance parameters

4.1.1 Accuracy

The accuracy of the prediction is the total number of true positives and true negatives. The accuracy of the result can be evaluated using calculating the confusion matrix. It includes four parameters true positive, true negative, false positive, false negative. The accuracy can be calculated by

Accuracy=(True positive+ true negative)/(true positive+ true negative+ false positive+ false positive)

4.1.2 Precision

It is the measure of total predicted positives and actual total positive

Precision=predicted positive/ total positive

4.1.3 Recall

It is the classification based on imbalance.

Recall= (true positive)/ (true positive +false negative)

The parameters will helps in measuring the performance of the proposed technique in comparison of the base technique. These parameters are based on measuring the different types of values for example true positive, true negatives, false positive, false negatives in the single confusion matrix. Accuracy, Recall, and precision are calculated from the above mentioned parameters.

4.2 Implementation configuration

Parameter Name	**Value**
Dataset	agriculture
Format	CSV
Testing set size	30%
Training set size	70%
Dataset columns count	9
Actual given result attribute	Predicted_value

These parameters are set based on the requirements. Few parameters always remain fixed. But few of the parameters values can be vary for enhancing the problem solutions.

4.3 Implementation Tool

The implementation tool is taken as R. It is the modern era tool specially used for running various machine learning algorithms. There are different fields in the current times where machine learning algorithms are used for prediction purpose. These algorithms are classification algorithms classify the whole dataset values into different subset based on prescribed conditions. R tool is used for data analysis and for the statistical evaluation on the large datasets.

4.4 Libraries required for prediction

S. No	Library name
1	Readr
2	CaTools
3	ISLR
4	ggplot2
5	Caret
6	ROCR
7	Devtools
8	plot3D
9	Plotly

These libraries are required for the purpose of classification with the hybrid approach of weighted k-mean and the logistic regression. These libraries are the standard libraries can be downloaded from the R official website.

4.5 Dataset

The data related to different parameters for the crop disease prediction is taken from open repository of UCI. It is the open source provide the datasets for educational and analysis purpose. The given dataset is having 8 parameters and one actual predicted value. The dataset is having different parameters which affect the crop disease rate. User can predict the crop disease by using machine learning technique. These techniques are having higher level of performance, because they will classify the whole dataset into smaller classes. In his crop disease prediction the dataset is to be classified into two subsets, one is for those crops which can have disease and second is for those crops which do not have crop disease.

4.6 Confusion matrix for Logistic regression

	1	2
0	44	108
1	55	24

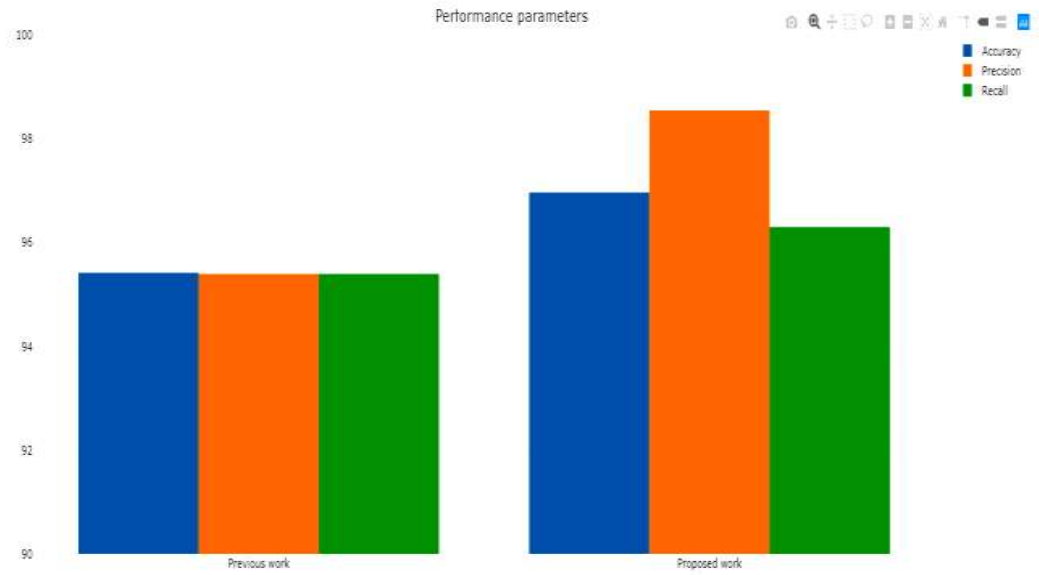
This shows the confusion matrix for logistic regression

4.7 Confusion matrix with weighted k-mean

	0	1
0	111	237
1	136	53

This table shows the confusion matrix for classification accuracy. There are four values for the whole training set. These parameters are true positive, true negative, false positive, false negative. All these parameters will denote the values shown in the matrix.

4.8 Comparison of parameters



4.9 Table for comparison

Parameter Name	**Base**	**Proposed**	**Improvement**
Accuracy	95.42	96.97	1.67%
Recall	95.4	96.30	1.15%
Precision	95.4	98.55	3.15%

5. CONCLUSION

Data mining and various machine learning algorithms which are used for prediction purpose are giving higher level of accuracy. Different authors are applying the machine learning based techniques for prediction in the field of crop health. Data analytics is the main focusing point for different fields. Agriculture field is the new entrant to the data analytics. It specifically picks the data related to patient different parameters and evaluates the parameters with different machine learning algorithms. In proposed

technique the crop disease prediction based on different parameters has been evaluated. These parameters are related to crops different aspects like previous years diseases etc. The proposed technique for the prediction is k-mean and logistic regression. The proposed technique is showing better results in terms of accuracy, precision and recall. The accuracy improvement is around 1.67%, Recall is improved by 1.15% and precision is improved by 3.15%.

6. FUTURE WORK

The proposed technique for the prediction is based on the k-mean and logistic regression. This technique has been applied on to the standard dataset of crop disease parameters. In future the hybrid technique can be applied onto the different datasets available on the disease prediction

REFERENCES

[1] Mackay J, Mensah G. Atlas of heart disease and stroke. Nonserial Publication; 2004.

[2] Vasighi Mahdi, Ali Zahraei, Bagheri Saeed, Vafaeimanesh Jamshid. Diagnosis of coronary heart disease based on Hnmr spectra of human blood plasma using genetic algorithm-based feature selection. Wiley Online Library; 2013. p. 318–22.

[3] Amin Mohammed Shafennor, et al. Identification of Significant features and data mining techniques in predicting heart disease. Telematics Inf 2019:82–93.

[4] Nahar J, Imam T, Tickle KS, Chen YPP. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert Syst Appl 2013;40(1):96–104.

[5] Guru Niti, Dahiya Anil, NavinRajpal. Decision support system for heart disease diagnosis using neural network, Delhi Business Review. 2007;8(1). January-June.

[6] Detrano Robert. Cleveland heart disease database. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation; 1989.

[7] Patil SB, Kumaraswamy YS. Extraction of significant patterns from heart disease warehouses for heart attack prediction. Int. J. Comput. Sci. Netw. Secur(IJCSNS) 2009;9(2):228–35.

[8] Chauhan Shraddha, Aeri Bani T. The rising incidence of cardiovascular diseases in India: assessing its economic impact. J. Prev. Cardiol. 2015;4(4):735–40.

[9] Vanisree K, JyothiSingaraju. Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. Int J Comput Appl April 2011;19(6). (0975 8887).

[10] Verma L, Srivastava S, Negi PC. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst 2016;40(7):1–7.

[11] Liu Xiao, Wang Xiaoli, Su Qiang, Zhang Mo, Zhu Yanhong, Wang Qiugen, Wang Qian. A hybrid classification system for heart disease diagnosis based on the RFRS method. Comput. Math. Methods Med. 2017;2017:1–11.

[12] Xing Yanwei, Wang Jie, Yonghong Gao Zhihong Zhao. Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease. Convergence Information Technology. 2007. p. 868–72.