International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

Comparison of Different Techniques to Predict Disease in Agriculture Production - A Review

Manpreet Kaur
Research Scholar, Dept of Comp. Applications
Guru Kashi University, Talwandi Sabo,
PB, India

Dr. Dinesh Kumar
Associate Professor, Dept of Engg and Technology, Guru Kashi University, Talwandi Sabo,
PB, India

***Abstract*- With the increase in agricultural produce suffering from different types of disease, early detection and prediction of disease is the major area of concern. Analyzing the data can help in improving the quality of decision making and help the clinician's to monitor the high risk area and provide specialized treatments. India is the agriculture based economy. Good production will leads to well being of all the persons from bottom to top. In India like country, where due to the poor people who belongs to agriculture cannot take up the detection process by their own. They are dependent on the government. They want to have collective action for whole area by the Government. So at the larger scale there requires detection of the disease for early corrective action. Using predictive analysis techniques can help in solving this problem. Existing literature provides various models that can be helpful but a combination of cluster and class based approach has yielded better results.**

Keyword: Classification; Clustering; Prediction

I. INTRODUCTION

With the ever-increasing data in every field, there is a need to utilize the obtained information in a productive manner focus is on usage and advantages of various technologies and their applications, they can result in refined data analysis and better overall predictions. These technologies and platforms can provide provisions for improved solutions, better strategies and improve the process of decision making.

Data mining is an emerging technology that is designed to efficiently handle big data and provide useful implications in various fields. The term "Data mining" was first coined in 1990's and is often considered as a synonym of Knowledge discovery in databases. But data mining is actually a crucial part of KDD where the details analyzed to obtain connections and find patterns within the data and using them predictions can be made or the recent trends can be uncovered and analyzed whereas KDD is the overall process of knowledge extraction. These tasks are broadly classified into supervised, unsupervised and reinforcement learning. Supervised learning refers to process of training the model using the data which is already labeled, that is, data is already classified correctly. In unsupervised learning, the model is not trained because the instances are neither labeled nor classified initially. The model itself classifies the data based on the patterns observed and their similarities. In reinforcement learning, the machine learns from the feedback of the previous input and outputs.

Various search engines and social networking sites collect huge amounts of data and by using different data mining techniques, they try to find the hidden patterns within the data. Data mining is widely used in diverse areas such as banking, e-commerce, health and medicine, genetics, education, stock exchange and various other fields. The data is available in structured, semi-structured and unstructured format which is initially processed and then analyzed using different techniques. These techniques are broadly classified into two main categories, namely, predictive and descriptive. The predictive data mining techniques focus on understanding the future, making predictions using the available datasets whereas the descriptive techniques summarize and analyze the past data and properties to make it useful in predicting new ones.

Various fields like Marketing, Education, Banking Sector, Bio-Informatics and Healthcare Agriculture make use of these techniques. But nowadays, a lot of focus is given to the healthcare sector to improve the process of decision making and provide better facilities to the patients. The data for analyzing the health data can be collected from various sources like paper records, x-rays, etc and can be stored in electronic health records. The accumulated data can then be processed and analyzed. The predictive analysis considers the various symptoms, their effects on health and can help in early detection and prevention of these diseases. Different data mining techniques have been implemented to obtain the results regarding breast cancer, heart disease and disease to agriculture to analyze the current status and make future predictions regarding the occurrence of disease, early detection and preventable patient deaths.

II. LITERATURE SURVEY

Nahato et al. proposed a combination of fuzzy sets and extreme learning machine for classification of three different datasets regarding heart disease and disease to agriculture. In this hybrid approach features of the dataset were divided into fuzzy sets and then extreme learning machine was used to classify them. Three different datasets were used to implement the proposed algorithm, namely, Cleveland Heart Disease Dataset (CHD), Stat Log Heart Disease Dataset (SHD) and Pima Indian Disease to agriculture dataset (PID). The algorithm was run by varying the number of hidden layer neurons and the ones that produced the most efficient results were selected. This model performed better in terms of accuracy and training time. The accuracy of the classifier for CHD, SHD and disease to agriculture data came out to be 73.77%, 94.44% and 92.54%, respectively.

Roxana et al. focused on providing easy and accessible method for diagnosis of disease to agriculture. Ensemble perceptron algorithm which is a combination of ensemble learning algorithm and perceptron algorithm is proposed. This approach was validated using three different datasets and results showed that the value of AUC increased from 0.72 to 0.75.

Lekha et al. implemented one dimensional convolution neural network to detect disease to agriculture. The input for this model is the human breath signals obtained using an array of MOS sensors. Initially, the features are reduced and then based on the optimal set of features; the signals acquired were fed to the neural network implemented in MATLAB environment. The results showed a reduction in mean square error and a better overall performance of the model.

Deepika et al. presented a comparative study of five different algorithms on two datasets, namely, breast cancer and disease to agriculture. The algorithms used were SMO, Naive Bayes, MLP, J48 and REP tree. The preprocessed data was classified using the algorithms and several evaluation metrics were used to measure the performance of each algorithm. According to the results obtained, J48 is best suitable for breast cancer prediction and SMO for disease to agriculture prediction.

Meng et al. presented a comparative analysis of three different models for prediction of disease to agriculture and pre disease to agriculture. Based on 12 different features and one outcome variable, the models, based on logistic regression, ANN, decision tree, were implemented. Results showed that the highest accuracy of 77.87% was obtained using decision tree. The logistic regression model attained an accuracy of 76.13% whereas ANN model had the least accuracy of 73.23%.

Deepti et al. compared three different machine learning algorithms for prediction of disease to agriculture using the PIMA Indian disease to agriculture dataset. The preprocessed data is classified using naïve bayes, decision tree, SVM. Experimental results showed that naïve bayes achieved the highest accuracy of 76.30%. This work can be extended for prediction of various other diseases.

Archana et al. presented a comparative study of the performance of k-means clustering using distance metrics, namely, Euclidean, Manhattan and Minkowski distance. Results showed that performance of k-means clustering is affected by the selection of distance metrics. It is also concluded that Euclidean distance metric provides the best result as compared to others whereas Manhattan distance metric is the worst.

Vaishali et al. Implemented multi objective Evolutionary fuzzy algorithm for classification of PIMA Indian disease to agriculture dataset. Initially, before classification genetic feature selection is performed to remove the redundant features which help in improving the accuracy and speed of the classifier. The performance of classification algorithm with and without using feature selection was compared and results showed that the combination of genetic feature selection and MOE fuzzy classifier is better than the others.

Sun et al. Used a combination of logistic regression and random forest algorithm to select differently expressed genes of breast cancer on micro array dataset. The prediction accuracy rates were analyzed by varying the threshold value. Top 20 genes were recognized that are expected to influence the development of breast cancer and a maximum accuracy obtained was 95.57%.

Isaac et al. compared the accuracy obtained by implementing k-means algorithm and decision tree for the diagnosis of breast cancer. Based on 15 attributes the model was trained and used for predicting the disease. Results showed that both the techniques resulted in high accuracy but statistical results show that k-means algorithm has higher performance than decision tree.

Han Wu et al. proposed a model that implemented k-means algorithm and logistic regression using WEKA toolkit to predict type-2 disease to agriculture. The main aim was to improve the accuracy and implement the model using various datasets. This model produced satisfying results and was less time consuming. Accuracy of the proposed model was 3.04% more than the existing ones. The performance of the model was evaluated using two other datasets as well.

III. COMPARATIVE ANALYSIS

Table 1. Comparative Analysis of Existing System

Research Paper	**Publication**	**Techniques Used**	**Results / Observations**	**Limitations / Future Scope**
Research on Logistic Regression Algorithm of Breast Cancer Diagnosis Data by Machine Learning	International Conference on Robots & Intelligent System, IEEE 2018	Logistic Regression	Quick and efficient results. This is based on identifying the best solutions based on prediction.	Analyze using different feature combinations. They may also encourage researcher farmer-advisor-stakeholder interaction, and thus create enabling environment for cooperation for further research around these ILTER sites.
An Efficient Mixed Model for Screening Differentially Expressed Genes of Breast Cancer Based on LR-RF	IEEE/ACM Transactions on Computational Biology and Bio-informatics, 2018	Logistic Regression-Random Forest	Improvement inaccuracy and speed of the screening of cancer-causing genes	Explore LR-RF to gather knowledge and methods to identify disease-related genes of breast cancer. SOM and the crop residue management. The results obtained by data mining are in line with previous studies and enhance our knowledge about the driving forces of primary

				productivity in arable systems.
A new variant of Fuzzy K-Nearest Neighbor using Interval Type-2 Fuzzy Logic	International Conference on Fuzzy Systems, IEEE 2018	Fuzzy logic, K-Nearest Neighbor	Better classification rate. Classify the dataset into multiple classes based on features set.	Make changes to fuzzy inference system and study their effects on classification rate. Data mining analyses of the experimental data were carried out in order to investigate trends in the productivity data novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease.
Development of a Hybrid Neuro - Fuzzy System as Diagnostic Tool for Type 2 Diabetes Mellitus	6th Iranian Joint Congress on Fuzzy and Intelligent Systems, IEEE 2018	Fuzzy Logic, Neural Networks	Hybrid model overcomes the performance of individual techniques. It hybridize multiple techniques to generate single process framework	Examine more hybrid approaches to improve the efficiency. Feature selection was done using genetic algorithm and Fuzzy logic was used to classify the data values which resulted in increased performance
Diagnosis Prognosis and Prevention of Breast Cancer Based on Present Scenario of Human Life	International Conference on Communication information and Computing Technology, IEEE 2018	K-Means algorithm, Decision Tree	Better diagnosis and prevention of breast cancer, suggests test to get a clear understanding of the illness and further treatment required	Refine the attributes to study their effect on accuracy. In this hybrid approach features of the dataset were divided into fuzzy sets and then extreme learning machine was used to classify them.
Prediction of Pre diabetes using Fuzzy Logic based Association	Second International Conference on	Fuzzy Logic based Associative Classification	Overcomes the problem of boundary value misinterpretati	Explore other applications to under its performance and scalability
Classification	Inventive Communication and Computational Technologies, IEEE 2018		on, predicts the outliers	
Type 2 Diabetes Mellitus Prediction Model Based on Data Mining	Informatics in Medicine Unlocked, Elsevier 2018	Improved K-Means algorithm, Logistic Regression algorithm	Higher accuracy of prediction, applicable on various datasets, less time consuming and maximum retention of original data	They have performed the process of data prediction for the small geographical area. But it will be very difficult for generates the hypothesis based model.
A Hybrid Prediction Model for Type 2Diabetes Using K-Means and Decision Tree	8th IEEE International Conference on Software Engineering and Service Science, IEEE 2017	K-Means, Decision Tree J48	Improved accuracy and promising results. The result of the prediction has shown the improvement in prediction for the results.	Use different type of data to check the behavior of proposed model and see behavior on multiclass classification problems
Analysis and Prediction of Breast cancer and Diabetes Disease Datasets using Data Mining Classification Techniques	International Conference on Intelligent Sustainable Systems, IEEE 2017	Naive Bayes SMO, REP Tree, J48 and MLP algorithms	J48 works best for cancer dataset and for diabetes SMO is better than other algorithms	Implement these algorithms using various other datasets and analyze their performance
Real-Time Non-Invasive Detection and Classification of Diabetes Using Modified Convolution Neural Network	IEEE journal of biomedical and health informatics, IEEE 2018	Modified Convolution Neural Network	Reduced computational cost and mean square errors and optimizes the overall performance of the classifier. Because the process involves the process framework for size reduction.	Implement the proposed algorithm with suitable enhancements from breath signals obtained from other gas sensors and to include more data samples.
Genetic algorithm based feature selection and MOE fuzzy classification algorithm on PIMA Indians disease to agriculture dataset	International Conference on Computing Networking and Informatics, IEEE 2017	Genetical agorithm, Multi Objective Evolutionary fuzzy classier	High feature reduction rate and improved performance. Reduction in the features set will enhance the results and reduces the confusion for the accuracy.	Analyze the effect of outliers and missing data on feature selection. Using data mining techniques to model primary productivity from international long-term ecological research

				(ILTER) agricultural experiments in Austria
Machine Learning Based Prediction of Depression among Type 2 Diabetic Patients	12thInternationalConference on Intelligent Systems and Knowledge Engineering, IEEE 2017	SVM K-MEAN, F-C MEAN, Probabilistic Neural Network	SVM classier generates more precise results than the others. SVM based classifiers will be to sub divide the whole training and testing set into two classes. One is the predicted positive and other is predictive negative.	Implement other learning methods for higher accuracy and optimize them. But the result should be applied for different environment conditions and soil conditions.
Disease to Agriculture Disease Prediction Using Data Mining	International Conference on Information, Embedded and CommunicationSystems, IEEE2017	Naive Bayes, K-nearest neighbor	Large database, improved accuracy Agriculture based prediction will help in forecasting the total amount of produce for the period of time.	Improvise the algorithms to further improve the efficiency. There are different classification techniques produces week results in the terms of classifiers.
Predictive Analysis Using Hybrid Clustering in Disease to Agriculture Diagnosis	Recent Developments in Control, Automation & Power Engineering, IEEE 2017	K*-means clustering, genetical algorithm, SVM	Increase in the accuracy, improved sensitivity and the positive predicted value metrics Hybrid technique for the prediction of the results will helps in having prediction for the results.	Explore different machine learning techniques combinations, replace missing values using a better technique. Hybrid approach features of the dataset were divided into fuzzy sets and then extreme learning machine was used to classify them.
Disease to Agriculture Disease Diagnosis Method based on Feature Extraction using K-SVM	Int J Adv Computer Science Applications, 2017	K-means algorithm, SVM	Hybrid approach enhanced the performance and produced accurate results. It is the hybrid approach for agriculture produce amount. Planning can be fine tuned by having early prediction system.	Integrate and optimization technique for further enhancement. Varying the number of hidden layer neurons and the ones that produced the most efficient results were selected. This model performed better in terms of accuracy and training
				time.
Hybrid Approach using fuzzy sets and Extreme Learning Machine for Classifying Clinical Datasets	Informatics in Medicine Unlocked, Elsevier 2016	Fuzzy logic, Extreme learning machine	Performance of proposed work is competent to existing work. The result of the option for the new hybrid based technique has enhance the results.	Explore Hybrid FELM with bio-inspired optimization techniques. Were divided into fuzzy sets and then extreme learning machine was used to classify them. Three different datasets were used to implement the proposed algorithm,
Disease to Agriculture Prediction Using Ensemble Perceptron Algorithm	9th International Conference on Computational Intelligence and Communication Networks (CICN), 2017	Ensemble learning, Perceptron Algorithm	The value of AUC increased and the execution time was almost the same compared to Perceptron Algorithm	They have performed the process of data prediction for the small geographical area. But it will be very difficult for generates the hypothesis based model.
Comparison of three Data Mining models for Predicting Disease to Agriculture or Pre Disease to Agriculture by risk factors	The Kaohsiung journal of medical sciences, 2013	Logistic regression, ANN, decision tree	The highest accuracy was obtained using decision tree whereas ANN model had the least accuracy	Author has proposed a system for the single type environment like temperature and soil condition. But the result should be applied for different environment conditions and soil conditions.
Prediction of Disease to Agriculture using Classification Algorithms	Procedia computer science, 2018	Naive Bayes, Decision Tree, SVM	Naive Bayes achieved the highest accuracy	Prediction of various other diseases using these techniques and automation of disease to agriculture analysis based on this approach
A Survey on Medical Diagnosis of disease to agriculture Using Machine Learning Techniques	Recent Developments in Machine Learning and Data Analytics, 2019	Decision Tree, Artificial Neural Network, Random Forest, K-Nearest Neighbor,	Logistic regression provides the best accuracy as compared to other algorithms	Identification of Type 1 and Type 2 disease to agriculture using a single classifier

		Naive Bayes, Logistic Regression and Support Vector Machine		
Performance Analysis of Classifier Models to Predict Disease to Agriculture Mellitus	Procedia Computer Science, 2015	KNN, Random Forest, J48 Decision Tree and SVM	In case of original dataset, J48 performs better whereas for preprocessed data, accuracy improved for all the classifiers but Random Forest and KNN (k=1) were the best classifiers.	Analyze the performance of the classifiers for prediction of other diseases

IV. CONCLUSION

With the increasing demand of predictive analysis in the medical field, the aim of this paper is to study an efficient model for diagnosing and predicting the occurrence of disease to agriculture based on several parameters. The proposed model is a combination of cluster and class based techniques. Though, k-means clustering is an efficient technique to obtain clusters based on similarity between the instances but it is sensitive towards the selection of initial centroids. Decision tree algorithm in data mining is used for predicting soil fertility. By using clustering techniques based on Partitioning Algorithms and Hierarchical Algorithm, the land utilization for agriculture and non-agriculture areas for the past ten years has been determined. As early into the growing season as possible, a farmer is always concerned with how much yield of his crop. In the past, this yield prediction has been relied on farmer's experience for particular yield, crops and climatic conditions. However, this knowledge might also be available, but not exactly for the small scale. Accurate data which can collect in seasons using a multitude of seasons. In this paper, study a median based approach for the selection of initial centers to reduce the effect of outliers and further enhance the performance of the classifier. Further, k-means and weighted k-means have been used for clustering and classification of instances is studied using logistic regression. Researcher's research showed that about 80% and 84% of the original data is retained after k-means and weighted k-means clustering, respectively, which is fed to the classifier. Accuracy obtained for classification using k-means and weighted k-means in combination with logistic regression is 96.97% and 97.84%, respectively. The model using weighted k-means performs slightly better than the one using k-means for clustering. Further, the risk associated with disease and non diseased agricultural is analyzed using the results of the classifier.

V. REFERENCES

[1] Bhatia, K. and Syal, R., 2017, October. Predictive analysis using hybrid clustering in disease to agriculture diagnosis. In 2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE) (pp. 447-452). IEEE.

[2] Bora, D., 2019, Big Data Analytics in Healthcare: A critical analysis. In: Big Data Analytics for Intelligent Healthcare Management. [online] Mara Conner. Available at https://www.sciencedirect.com/book/9780128181461/big-data-analytics-for-intelligent-healthcare-management.

[3] Chen, W., Chen, S., Zhang, H. and Wu, T., 2017, November. A hybrid prediction model for type 2 disease to agriculture using K-means and decision tree. In 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 386-390). IEEE.

[4] Choudhury, A. and Gupta, D., 2019, A Survey on Medical Diagnosis of disease to agriculture Using Machine Learning Techniques. In Recent Developments in Machine Learning and Data Analytics (pp. 67-78). Springer,

[5] Singapore. de Amorim, R.C., 2016, A survey on feature weighting based K-Means algorithms. Journal of Classification, 33(2), pp.210-242.

[6] Fatemidokht, H. and Rafsanjani, M.K., 2018, February. Development of a hybrid neuro-fuzzy system as a diagnostic tool for Type 2 disease to agriculture Mellitus. In 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) (pp. 54-56). IEEE.

[7] India Today., 2018, Disease to Agriculture Epidemic: 98 million people in India may have type 2 disease to agriculture by 2030. [online] Available at: https://www.indiatoday.in/education-today/gk-current-affairs/story/98-million-indians-dia betes-2030-prevention-1394158-2018-11-22.

[8] Isaac, L.D. and Suresh kumar, C., 2018, February. Diagnosis prognosis and prevention of breast cancer based on present scenario of human life. In 2018 International Conference on Communication information and Computing Technology (ICCICT) (pp. 1-7). IEEE.

[9] M.S., Ghindawi, I.W. and Mhawi, D.E., 2018, An Accurate disease to agriculture Prediction System Based on K-means Clustering and Proposed Classification Approach. International Journal of Applied Engineering Research, 13(6), pp.4038-4041. Kaggle.com. (2016). PIMA Indians disease to agriculture Database. [online] Available at: https://www.kaggle.com/uciml/pima-indians- disease to agriculture -database.

[10] Kandhasamy, J.P. and Balamurali, S., 2015, Performance analysis of classifier models to predict disease to Agriculture Mellitus. Procedia Computer Science, 47, pp.45-51. 43

[11] Kaur, S. and Kalra, S., 2016, August. Disease prediction using hybrid K-means and support vector machine. In 2016 1st India International Conference on Information Processing (IICIP)(pp. 1-6). IEEE.

[12] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., 2017, Machine learning and data mining methods in disease to agriculture research. Computational and structural biotechnology journal, 15, pp.104-116.

[13] Khalil, R.M. and Al-Jumaily, A., 2017, November. Machine learning based prediction of depression among type 2 diabetic patients. In 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) (pp. 1-5). IEEE.

[14] Lekha, S. and Suchetha, M., 2017, Real-time non-invasive detection and classification of disease to agriculture using modified convolution neural network. IEEE journal of biomedical and health informatics, 22(5), pp.1630-1636.

[15] Liu, L., 2018, May. Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning. In 2018 International Conference on Robots & Intelligent System (ICRIS) (pp. 157-160). IEEE.

[16] Mehta, N. and Pandit, A., 2018, Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics, 114, pp.57-65.

[17] Melin, P., Ramirez, E. and Prado-Arechiga, G., 2018, July. A new variant of Fuzzy K-Nearest Neighbor using Interval Type-2 Fuzzy Logic. In 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-7). IEEE.

[18] Meng, X.H., Huang, Y.X., Rao, D.P., Zhang, Q. and Liu, Q., 2013, Comparison of three Data Mining models for Predicting Disease to Agriculture or Pre Disease to Agriculture by risk factors. The Kaohsiung journal of medical sciences, 29(2), pp.93-99.

[19] Mirshahvalad, R. and Zanjani, N.A., 2017, September. Disease to Agriculture Prediction using Ensemble perceptron algorithm. In 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 190-194). IEEE.

[20] Nahato, K.B., Nehemiah, K.H. and Kannan, A., 2016, Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets. Informatics in Medicine Unlocked, 2, pp.1-11.