IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

Performance Study on Diabetic Disease Prediction Using Classification Techniques

P. Hema[1*], K. Palanivel[2]

[1*]Dept of Computer Science, AVC College, Mayiladuthurai, India
[2]Dept of Computer Science, AVC College, Mayiladuthurai, India

[*]*Corresponding Author: hemapalanivel@gmail.com*

Available online at: www.ijcseonline.org

Received: 14/Jan/2018, Revised: 23/Jan/2018, Accepted: 11/Feb/2018, Published: 28/Feb/2018

***Abstract*—** Data mining techniques can be used by Health organizations to identify the diseases like heart, tumor, diabetic, liver and thyroid disease using symptoms as parameters. Diabetic disorder is one of the growing diseases worldwide currently faced by people because of modified life style. Valuable data can be observed from application of knowledge mining techniques in the fitness care system particularly in Diabetic Disease. In this direction, this research paper studies the performance of three classifier algorithms available, namely JRip, PART and Random Tree using WEKA tool and proposed a new algorithm Weighted Classifier to classify the data a diabetic data set. The objective of this research is to classify data, assist the people by extracting useful knowledge from classified data and identify the efficient algorithm to best prediction of disease. From the experimental analysis, it is concluded that weighted Classifier is the effective algorithm for classification accuracy. The result will help doctors in a diagnosis process.

***Keywords*—** Data mining, Diabetes disease, JRip, PART, Random Tree, Weighted Classifier, classification, WEKA tool.

I. INTRODUCTION

Diabetic is a very risky disorder. The diabetic infection has a lot of side effects, like kidney problems, eye problems and more difficulties. Normal symptoms of diabetes are Weight loss, vomiting and increased urination. Disease prediction performs essential role in data mining. Data mining is used extremely in the field of medicine to predict unwellness such as heart, lung cancer, diabetic, thyroid etc., the prime goal of this paper is to experiment the data from a diabetic data set classification approach to predict class effectively in each status in data. The most important contributions of this paper are:

- To extract helpful classified accuracy for prediction of diabetic diseases.
- Comparison of variety of data mining algorithms on diabetic dataset.
- Identify the most overall performance algorithmic software for prediction of diseases.

In this paper we have utilized a diabetic data set for classification technique. The steps include collection of data set for the essential the accuracy, classification and then comparison of results. The dataset has been used to classify the following diabetic attribute based on Blood pressure, Age, BMI, 2-hrs insulin, Plasma glucose. Though data mining has several different algorithms to analyze data but analysis using all the methods is impossible. In this paper we have performed the analysis using PART, JRip (RIPPER), Weighted Classifier, Random tree algorithms by using Explorer of WEKA Tool. The reminder of this paper is presented as follows. Section II lists on top of each other work. Section III expresses some fundamental concept of classification algorithms. Section IV describes experimental results of the classification algorithms for diabetic data set. Finally, Section V the conclusion of this research work.

II. RELATED WORK

Yasodha P. and Kannan M. performed analysis of a population of diabetic patient database using weka tool. They have classified the data and then outputs were compared by using Bayes Network, REP Tree, J48 and Random Tree algorithms. Finally the results conclude that these algorithms help to determine and identify the stage or state in which a of disease like diabetes is in by entering patients daily glucose rate and insulin dosages thereby predicting and consulting the patients for their next insulin dosage [1]. Vijayarani S. and Sudha S. have compared the analysis of classification function techniques for heart disease prediction. Classification was done using algorithms such as Logistic, Multilayer Perception and Sequential Minimal Optimization algorithms for predicting heart disease. In this classification

comparison logistic algorithm trained out to be best classifier for heart disease having more accuracy and least error rate [2]. Karthikeyani.v et al. restorative information helps the specialists to analyze distinctive examples in the information set. The examples found in information sets might be utilized for grouping, expectation and determination of the sicknesses [3]. Tirunagari et al. Applied SOM to cluster heterogeneous diabetes data. They were able to reduce the dimensionality of the data and demonstrate the similarities between patients by placing them in groups using the U-matrix. As a result, the profiles of patients who need self care management were grouped clearly and easily were identified. In another study, SOM to recognize the behavior of self care based on survey data collected from type I diabetic patients. The visualization result improved understanding pattern of various behaviors as well as detecting patients who need to adjust their lifestyle [4]. Nikita Sigh and Alka Jindal have concluded that SVM is better classifier as compared to KNN and Bayesian. Accuracy of SVM is about 84.62%. KNN found the nearest neighbourhood automatically. It represented by the graph each vertices having object. Bayesian based on the probability classification which gives the sample data belongs to a class [5]. Adidela et al.,presented the type of diabetes by using Fuzzy ID3 method. The author uses the system for predicting the disease from data set as it initially clusters the data and applies the classification algorithms on clustered data. The author presented a combination of classification method where they developed EM algorithm for clustering and fuzzy ID3 algorithm to attain decision tree for each cluster [6]. Durairaj et al Neural Networks are one of the soft computing techniques that can be used to make predictions on medical data. Neural Networks are known as the Universal predictors. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The Artificial Neural Networks (ANNs) based system can effectively applied for high blood pressure risk prediction. This improved model separates the dataset into either one of the two groups. The earlier detection using soft computing techniques help the physicians to reduce the probability of getting severe of the disease. The data set chosen for classification and experimental simulation is based on Pima Indian Diabetic Set from (UCI) Repository of Machine Learning databases. In this paper, a detailed survey is conducted on the application of different soft computing techniques for the prediction of diabetes. This survey is aimed to identify and propose an effective technique for earlier prediction of the disease [7].

III. METHODOLOGY

This paper has analyzed four different classification algorithms namely JRip (RIPPER), PART, Random Tree and Weighted Classifier Tree to predict which of the algorithm is most suitable for predicting the diabetic disease. The data set to be given as input to the WEKA tool should be formatted to the Attribute –Relation-File Format (ARFF). WEKA is a very good data mining tool for the users to classify the accuracy in the field of bioinformatics. Classification techniques are more appropriate for predicting or relating data sets with binary or nominal categories. In data mining tool classification deals with identifying problem by observing characteristics of disease amongst patients and diagnose or predict which algorithm shows best performance on the basis of WEKA`s statistical output.

Software	*Dataset*	*Weka Data Ming Technique*	*Classification Algorithm*	*Operating System*	*Data set file format*
WEKA	Diabetic	Explorer	JRip PART Random Tree Weighted Classifier	Windows 8.1	ARFF

Table 1. WEKA data mining technique using different algorithm

JRip:

Jrip (RIPPER) is one of the most popular algorithms. It has classes that are examined in increasing size. It also includes set of rules for class is generated using reduced error Jrip (RIPPER). Proceed by treating examples of judgments made in training data as a class and finding rules that covers all the members of the class. Then it proceeds to the next class and repeats the same action, repetition is done until all classes have been covered [8].

PART:

PART is a separate-and-conquer rule leaner. The algorithm producing sets of rules called "decision lists" which are planned set of rules. A new data is compared to each rule in the list in turn and the item is assigned the class of the first matching rule. PART builds a partial c4.5 decision tree in each iteration and makes the "best" leaf into a rule [9].

Random Tree:

A random tree is a collection of tree predictor that is called forest. It can deal with both classification and regression problem. The classification works as follows; the random trees classifier takes the input feature vector, classifies it with every tree in the forest and outputs the class label that received the majority of "votes". In case of a regression, the classifier response is the average of the responses overall the trees in the forest. All trees are trained with the same parameters but on different training sets.

Weighted Classifier:

A Weighted Classifier is a promising methodology in information mining that uses the association rule discovery techniques to build classification systems, otherwise called weighted classifiers. It is an unsupervised learning wherever no class attribute is engaged with finding the association rule. On the opposite, classification is a supervised learning wherever class attribute is concerned within the construction of the classifier and is used to predict the information unknown sample. Weighted Classifier could be a recent and rewarding technique that integrates association rule mining and classification to a model for prediction and achieves more accuracy. Weighted classifiers are particularly suitable applications where most accuracy is desired to a model for prediction. Different Techniques which might be utilized are apriori algorithm, eclat algorithm, FP-growth algorithm.

IV. EXPERIMENTAL RESULTS

In this part to analysis the result in predict diabetic disease data set. The datasets are collected from UCI Repository. A diabetic data set contains 768 instances and 09 attributes. In this research work accuracy, error rate and performance measures are calculated by using the performance factors to determine the best algorithm for the diabetic dataset.

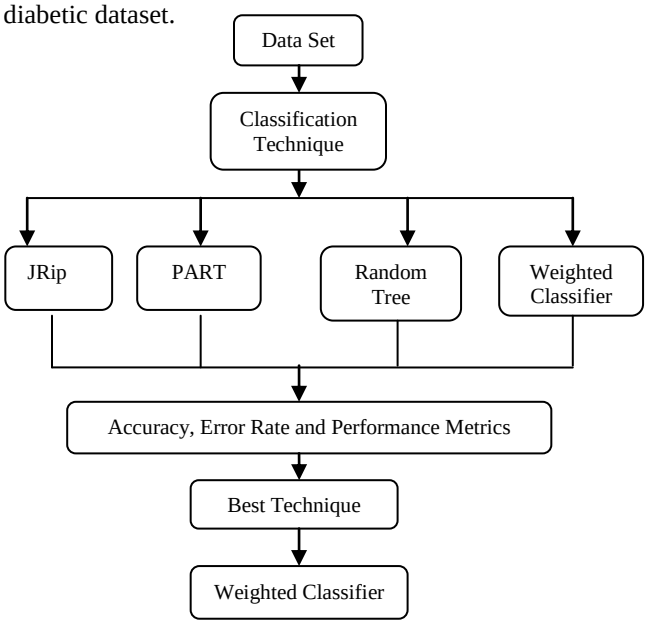

Figure 1. Working Architecture of proposed work

The data mining technique are utilized by us to predict diabetic Disorder. Predictions have been done by us using weka data mining tool for classification and accuracy by applying kind of different algorithms approaches. The interfaces of weka utilized as a part of this paper are the accompanying:

It initial pre-processes the data and then filters the data. Users can then load the data file in CSV (Comma Separated Value) format and then analyze the classification accuracy result by selecting the following algorithms using 10 cross validation. JRip, PART, Random Tree and Weighted Classifier algorithm Figure 2 shows the interface of explorer when using diabetic dataset is opened using CSV file along with its graphical view

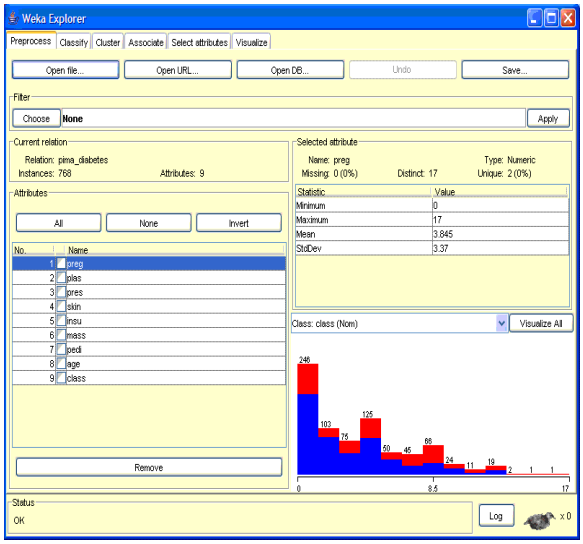

Figure 2. Screenshot view of CSV Diabetic Dataset File open in Explorer interface

Table 2. Comparison of accuracy measures for the classification algorithm using diabetic datasets.

Datasets	Algorithm	Correctly Classified	Incorrectly Classified
Diabetic	JRip	584	184
	PART	578	190
	Random Tree	523	245
	Weighted Classifier	**592**	**176**

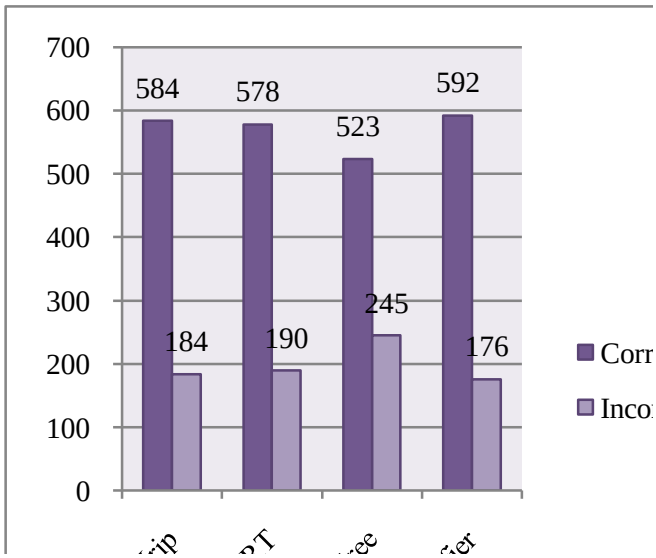

Figure 3. Comparison of accuracy measures for the classification algorithm using diabetic datasets

From the results (Table 2) it is inferred that for the diabetic dataset the proposed algorithm performs well as compared to Jrip (RIPPER), PART and Random Tree. The Proposed algorithm (Weighted Classifier algorithm) gives more Correctly Instances compared to others.

Table 3. Comparison of Error rate measures for the classification algorithm using diabetic datasets

Algorithms	***MAE***	***RMSE***	***RAE***	***RRSE***	***Kappa Statistic***
JRip	0.341	0.423	0.752	0.889	0.453
PART	0.310	0.414	0.682	0.870	0.439
Random Tree	0.319	0.564	0.701	1.184	0.303
Weighted Classifier	***0.308***	***0.405***	***0.674***	***0.936***	***0.416***

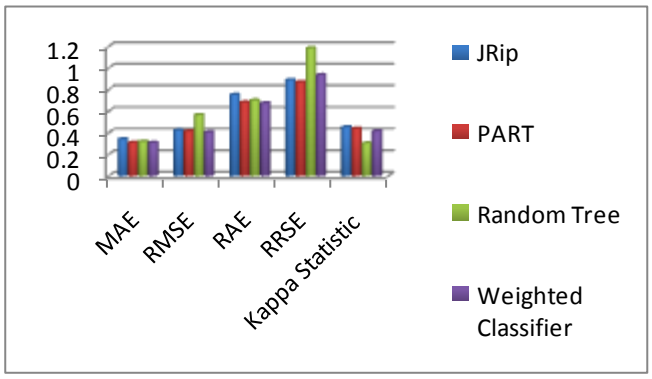

Figure 4. Comparison of Error rate measures for the classification algorithm using diabetic datasets.

From the results (Table 3) it is inferred that the diabetic datasets, the Error Rate for proposed algorithm (Weighted Classifier algorithm) is less compared to others. From the experimental results that the proposed algorithm the parameter Root Relative Squared Error (RRSE) increases, Mean Absolute Error (MAE), Root Mean Square Error(RMSE) and Relative Absolute Error (RAE) value decreases and the Kappa value fluctuates. For the Random Tree, JRip and PART algorithm Kappa, RAE, RMSE, RRSE and MAE fluctuates.

Table 4. Comparison of Performance Measures for the classification algorithms.

Algorithm	TP Rate	FP Rate	Precision	Re-Call	F-MES
Jrip	0.719	0.281	0.738	0.719	0.726
PART	0.713	0.574	0.728	0.713	0.719
Random Tree	0.653	0.347	0.651	0.653	0.651
Weighted Classifier	***0.642***	***0.288***	***0.699***	***0.651***	***0.648***

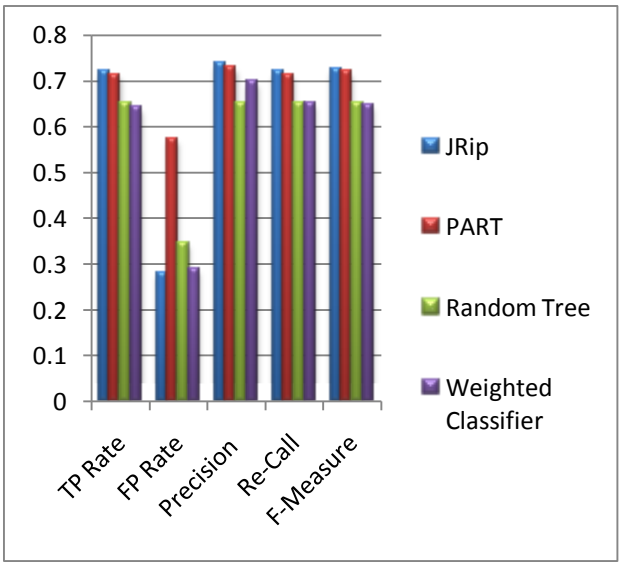

Figure 5. Comparisons of Performance Measures for the classification algorithms.

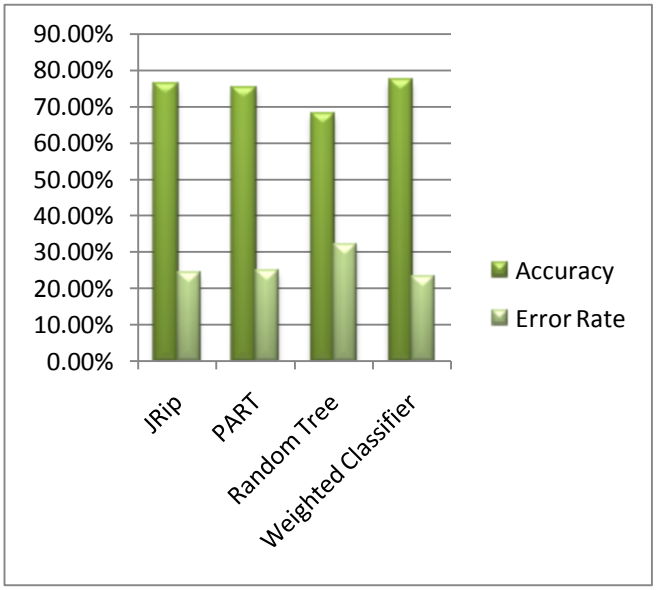

Figure 6. Diagram of Accuracy and Error Rate of Classification Algorithms

The results of following analysis on the dataset are clearly given by the table 3 and 4. Table 4 listed the True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, Re-Call and F-Measure to analyses the classifier.

Table 5. Accuracy and Error Rate Analysis for classification algorithms

Algorithm	***Accuracy***	***Error Rate***
JRip	76.04%	23.91%
PART	75.26%	24.73%
Random Tree	68.09%	31.90%
Weighted Classifier	***77.10%***	***22.91%***

Table 5 shows the accuracy and error rate of the algorithm compared with the all other algorithms in data mining. It clearly shown the proposed algorithm (Weighted Classifier) will be better than the JRip, PART and Random Tree algorithm clearly.

V. CONCLUSION AND FUTURE WORK

In this research, the accuracy of three existing algorithms namely JRip, PART, Random Tree and a proposed algorithm namely Weighted Classifier is evaluated using an experiment conducted using WEKA tool and the diabetic dataset. This research work focuses on four algorithms and studies the performance of these algorithms using diabetic data set to predict the diabetic disease using symptoms. Out of the four algorithms, Weighted Classifier gives more accuracy and less Error rate. In future using Weighted Classifier algorithm can be studied for accuracy in different datasets and different tools. In future to handle the vagueness in data, the evolutionary algorithms like fuzzy or rough sets can be adopted.

REFERENCES

[1] P. Yasodha , M. Kannan M, "*Analysis of Population of Diabetic Patient Database in WEKA Tool*", International Journal of Science and Engineering Research, VoL.2 Issue.5, 2011.

[2] S. Vijayarani , S. Sudha, "*Comparative Analysis of Classification Function Techniques for Heart Disease Prediction*", International Journal of Innovative Research in Computer and Communication Engineering, Vol.1, Issue.3, pp.735-741, 2013.

[3] Dr.V.Karthikeyani , I.Parvin Begum "Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction Of Diabetes Disease" International Journal on Computer Science and Engineering (IJCSE) Vol. 5 No. 03 Mar 2013 205-210

[4] S. Tirunagari, N. Poh, K.Aliabadi, D.Windridge & D.Cooke, "*Patient level analytics using self-organising maps: A case study*

on Type-1 Diabetes self-care survey responses". In Computational Intelligence and Data Mining (CIDM), IEEE Symposium on pp. 304-309, 2014.

[5] N. Singh, A. Jindal, *"A Segmentation Method and Comparison of Classification Methods for Thyroid Ultrasound Images",* International Journal of Computer Applications, Vol.50, Issue.11, 2012.

[6] D.R.Adidela , D.G.Lavanya ,S.G. Jaya,A.R. Allam , "*Application of fuzzy ID3 to predict diabetes*". International Journal Advance Computer Mathematics Science, Vol.3, Issue.4, pp.541, *2012.*

[7] M. Durairaj ,G. Kalaiselvi , *" Prediction Of Diabetes Using Soft Computing Techniques- A Survey",* International Journal of Scientific & Technology Research, Vol.4, Issue.3, 2015.

[8] *A.* Rajput, *R.P.Aharwal*, M. *Dubey, S.P. saxena "J48 and JRIP Rules for E-Governance Data" I*nternational *J*ournal of *C*omputer *S*cience and *S*ecurity, Vol.5, Issue.2, pp.201-207, *2011*.

[9] E. Frank, Ian H. Witten, *"Generating Accurate Rule Sets Without Global Optimization*". In Fifteenth International Conference on Machine Learning, pp.144-151, 1998.

[10] M. H. Danham, S.Sridhar," *Data mining, Introductory and Advanced Topics*", Pearson education, 1st ed., 2006.

[11] R. Remco, Bouckaert, E. Frank, M. Hall, R. kirkby, P.Reutemann, A.Seewald, D. Scuse, "*WEKA Manual for Version 3-7-5*", 2011.

[12] Dr. V. Karthikeyini, I. Parvin Begum," *Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction Of Diabetes Disease*", International Journal on Computer Science and Engineering (IJCSE), Vol.5 Issue.3, 2013.

[13] P.P.Dhakate, S. Patil, K. Rajeswari, D.Abin "*Preprocessing and Classification in WEKA Using Different Classifier*", International Journal of Engineering Research and Applications, Vol.4, Issue.8, pp.91-93, 2014

[14] I.Parvin begum, V. Karthikeyini, K. Tajuddin, I. Shahina Begum, "*Comparative of data mining classification algorithm (CDMCA) in Diabetes Disease Prediction*", International journal of Computer Applications, Vol.60, Issue.12, pp. 26-31, 2012.

[15] K. Rajesh , V. Sangeetha, "*Application of data mining methods and techniques for diabetes diagnosis*" . International Journal of Engineering and Innovative Technology (IJEIT), Vol.2,Issue.3, pp.224.

Authors Profile

P.Hema has completed her MSc Computer Science degree at Bharathidasan University (CDE), Trichy. Currently she is doing M.Phil in computer Science at A.V.C College(Auto), Mayiladuthurai. She is doing research in the area of datamining.

Dr.K.Palanivel received his M.Sc. (Computer Science) degree from Bharathidasan University, M.Phil. (Computer Science) degree from Manonmaniam Sundaranar University and Ph.D. degree from Bharathidasan University. He is currently working as Associate Professor in the Department of Computer Science at AVC College (Autonomous), Mayiladuthurai. He has published many research papers in international journals. His research area includes Human Computer Interaction, Machine Learning, Recommender systems and Data mining.

