Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)

N. Saravanan
Department of Computer Science
Periyar University, Salem, Tamil Nadu, India
saravanangacm@gmail.com

V. Gayathri
Pee Gee College of Arts and Science
Periyanahalli, Dharmapuri, Tamil Nadu, India
gayhar11@gmail.com

***Abstract* - We have been using the most popular algorithm J48 for classification of data. The J48 algorithm is used to classify different applications and perform accurate results of the classification. J48 algorithm is one of the best machine learning algorithms to examine the data categorically and continuously. When it is used for instance purpose, it occupies more memory space and depletes the performance and accuracy in classifying medical data. Our proposed method is to measure the improved performance and produce higher rate of accuracy. For this research, the dengue dataset was collected from various government hospitals in Krishnagiri District. To measure the entropy of information and to identify the dataset and to increase the accuracy of J48 algorithm, the entropy of J48 is modified with Kendall's Rank Correlation Coefficient algorithm (KNJ48) to improve the accuracy of classification and performance time. Thus, it is modified as Kendall's New Rank Correlation Coefficient J48 algorithm (KNJ48) for better performance.**

Key words - Data mining, Classification, Dengue, J48, Entropy, Kendall's Correlation J48 (KNJ48), WEKA

1. INTRODUCTION

1.1 Data Mining

Data Mining (DM) is a technique which is used to find, new hidden and useful patterns of knowledge from large databases. From statistics, artificial intelligence and data warehouses, it is very easy to design methods and procedures to classify the data for the use of real-world applications. DM concept is actually split of the knowledge discovery process. DM has become a current technology in existing research and for medical field applications. The data mining applications are applied to find the final result of a disease and it is one of the most inspiring works and a difficult task (Venkatesan, 2015).

1.2 Decision Tree

A decision tree be a flowchart-like tree structure, where each internal node represents a test happening an attribute, each branch represents an ending of the test, class label is represented by each leaf node or terminal node . Given each tuple the attribute value of the tuple are tested next to the decision tree. A path is traced beginning the root to a leaf node which holds the class prediction used for the tuple. It is simple to convert decision trees into classification rules. Decision tree learning uses a decision tree because a predictive model which maps observations on an item to conclusions about the item's object value. It is single of the predictive modeling approaches utilize in statistics, data mining and machine learning. Tree models where the object variable can take a finite set of value are called classification trees, inside this tree structure, leaves correspond to class labels and branches represent conjunction of features that lead to individuals class labels. Decision tree can be constructed moderately quick compare to other methods of classification.

SQL statements can be constructing from tree to can be used to access databases accuracy. Decision tree classifiers obtain like or better Accuracy when compare with other classification methods. A amount of data mining techniques have already been done on learning data mining to improve the performance of students like Regression, Genetic algorithm, Bays classification, k-means clustering, associate rules, prediction etc. Data mining techniques can be there use in educational field to improve our understanding of learning process to focus on identify,

extracting and evaluating variables linked to the learning process of students. Classification is one of the most regularly. The C4.5, ID3, CART, J48 decision tree are applied on the data of students to predict their performance. (Gaganjot Kaur, 2014)

1.3 Classification of Algorithms

Classification is one of the most essential data mining problems. The input is a dataset of training record, in which each record has got several attributes. Numerical attributes are attributes with numerical domains and categorical attributes are attributes with non-numerical domains. Besides, there is also a distinguished attribute called the Class label. This classification is intended at building a console model which can be utilized to predict the class label future, unlabeled records. There are many classification models are used and, I am using the technique called decision trees. (Nagaparameshwara chary, 2017).

1.4 Organization of Paper

This paper is divided into seven sections. Section 1, deals with the overall concept of the topic. Section 2, Review of Literature, discusses the main objective with reference to the authors who have done research on this topic. Section 3, deals with the existing methods in the technical field. Section 4, focuses on the proposed method Kendall's Rank Correlation Coefficient. Section 5, discusses about designing experiment, details of dataset and experimental outcomes. Section 6, analyses the results and the last section 7 concludes about the research.

2. LITERATURE REVIEW

In general, all year after completion of rainy season a few epidemics are hitting to the people. Because the lack of living standards in slum areas nearly all of the people in those areas affected by epidemics. Fewer hygienist and poverty is the major classes of victims of epidemics used for these areas. The main factors for cause the epidemic are Poor hygienist, Rapid climate change, Drinking water contamination Unplanned sewage removal system etc. There are large amount water borne and mosquito-borne disease such as Cholera, Dengue fever, Malaria, Typhoid, Diarrhea, Jaundice, brain fever etc. Every rainy reasons, Hundreds of people be admit in the hospital due to some transmittable disease and a few of them were even failure their life. More types of epidemics finishing in this season. The major reasons for causing these epidemics are pollution of drinking water, non consistent change in climatic whether condition, mosquitoes because of water stagnation payable to unexpected sewage disposal system. Poor health of population in the slums resulted by poor quality and old age, Living conditions like personal hygienist, hygienist of the humanity, lack of health awareness or unsatisfactory knowledge in observing hygienist, per-capita income of the inhabitance etc are also some of the reasons behind this disaster.

Water-borne disease is, purely, any illness resulting from eating of or contact with water. Like food borne diseases, water-ingestion illness is also infections or intoxications. Organisms dependable for infections are mainly bacteria. These organisms generally occur in water contaminated by sewage (eg, especially bird and mammal excrement) or by infected persons or animals. Intoxications may well be chemical in nature (eg, copper, lead, insecticide poisonings) and generally occur as a result of metal leaching into water (from pipes or containers) and during the accidental spillage or seepage of chemicals into water supply. They can to happen during toxins produced by blue-green algae (cyanobacteria), eg, Anabaena, Microcystis or Oscillatoria. These organisms have cause even deaths in southern part of India during drinking pond water; Illnesses acquired through speak to with water are caused by bacteria.

The diabetes of the patients is calculated (Deepali Kharche, K. R. 2014) by use the decision tree within two phases: data pre-processing in which the attributes be identified and next be diabetes prediction model constructed with the help of using the decision tree method. Both the phases are implemented use WEKA data mining tool. The performance evaluation of Decision Tree Algorithms and Artificial Neural Network (Dr. Anjali B Raut, A. A. 2017) on health data was performed going on the base of parameter like kappa statistics, mean absolute error, relative

squared error, time toward model and mean-squared error. On the basis of results it have been examine to Decision Tree Algorithms perform improved than the Artificial Neural Network. Hypertension have been predict next to generating (Gaganjot Kaur, A. C. 2014) J48 plus Naive Bayesian classifiers in WEKA. The in general accuracy is around 83%. A slight development of ensemble five J48 classifier is see over clean Naive Bayesian and J-48 in sensitivity, accuracy and F-measure. Rough set tools be able to decrease the ensemble of five member to three except there is substantial growth of sensitivity. (Kameswara Rao N K, 2014)

Classification accuracy is usually measured by determining the percentage of rows positioned in a proper class. This disregards the fact that there can be also a cost associated with an improper assignment to the incorrect class. This perhaps should also determine. (Tina, 2013).

There are so many classification algorithms including J48 and we compared it with each other. The worth mentioning improvement in the classification accuracy, while implementing J48 is identified for the cognitive behavior and cognitive load. (Jayasimman, 2015).

Machine Learning Techniques (MLT) is applied to predict the medical datasets at an early stage to protect human life. Much of medical datasets are available in various data store that are used in the real world application. To group and predict symptoms in medical data, different data mining techniques had been used by various researchers in different time. In this system the popular predictive algorithms apply various algorithms including J48 to ensemble hybrid model by combining individual techniques/methods into one in order to increase the performance and accuracy. (Minyechil Alehegn, 2017).

It is compared that the performance of the datasets using the various classification techniques with evaluation principle as accuracy and implementation time. Its examined that performance of classification techniques differ with datasets. Factors which affect the classifier's performance are Dataset, Number of instances and attributes and Type of attributes. The updatable J48, has come out with better results with other datasets utilized in the comparison. (Deepali Kharche, 2014).

The decision trees produced by J48 can be utilized for classification. At every node of the tree, J48 chooses the attribute of the data that most effectively splits its arrangement of tests into subsets improved in one class or the other. The splitting criterion is the standardized information gain (on contrast in entropy). The attribute with the highest worthy standardized information gain is making on the decision. The J48 algorithm at that point recurs on the smaller sub lists. Also using Generalized Sequential Pattern mining algorithm, we have used it for predicting medical dataset and to improve the performance. (Dr. Anjali B Raut, 2017).

2.1 Objective

The objective of this research is to show the improved J48 classification algorithm by modifying the entropy and using Kendall's Rank Correlation Coefficient as KNJ48 for developing accuracy and save the time to get the expected results in an accurate manner.

3. EXISTING METHOD

3.1 J48 Algorithm

Quinlan's C4.5 algorithm actualizes J48 to create a trimmed C4.5 decision tree. The every aspect of the information is to split into minor subsets to base on a decision. J48 look at the standardized data gain that really the results the split the information by choosing an attribute. To summarize, the attribute extreme standardized data gained is utilized. The minor subsets are returned by the algorithm. The split strategies stop if a subset has a place with a similar class in all the instances. J48 develops a decision node utilizing the expected estimations of the class. J48 decision tree can deal with particular characteristics, lost or missing attribute estimations of the data and varying attribute costs. Here accuracy can be expanded by pruning (Venkatesan, 2015).

The Algorithm

Stage 1: The leaf is labeled with a similar class if the instances belong to similar class.

Stage 2: For each attribute, the potential data will be figured and the gain in the data will be taken from the test on the attribute.

Stage 3: Finally the best attribute will be chosen depending upon the current selection parameter.

3.2. Limitations of J48 Algorithm

Despite the fact that J48 one of the well known algorithms, there are a few shortcomings of this algorithm. A few limitations of J48 are discussed below.

3.2.1. Empty Branches

Constructing tree with significant value is one of the important steps for rule generation by J48 algorithm. In our research, we have come out with many nodes with zero values or very close to that. But, these values don't contribute to create or help to create any class for classification task. Instead it makes the tree wider and still complicating. (Prerna Kapoor, 2015).

3.2.2 Insignificant Branches

Number of chosen distinct attributes produces the same number of potential division to build a decision tree. But the fact is, not all of them are significant for classification task. These least important branches not only decrease the usability of decision trees but also bring on the problem of over fitting. (Srishti Taneja, 2014)

3.2.3 Over Fitting

Over fitting happens when algorithm display gets information with exceptional attributes. This causes many fragmentations in the process distribution. Statistically unimportant nodes with least examples are known as fragmentations. Usually J48 algorithm builds trees and grows its branches 'just deep enough to perfectly classify the training examples'. This approach performs better with noise free data. But most of the time this strategy over fits the training examples with noisy data. At present there are two strategies which are widely used to bypass this over fitting in decision tree learning. (SAGAR, 2015) Those are:

- If tree grows taller, stop it from growing before it reaches the maximum point of accurate classification of the training data.
- Let the tree to over-fit the training data then post-prune tree.

Yet, nothing of those is perfect solution of this problem. So we have proposed two tools to minimize the input space of data in this research. The first tool is Entropy of Information Theory and the second is Correlation Coefficient. In this experimentation, we have examined dengue medical data. The particulars of the datasets explanation are provided Java based machine learning tool WEKA which is used to perform the research.

4. PROPOSED METHOD

The primary focus of our research is to reduce the input space of data file, turn back the processing time and increase the percentage of classification accuracy. From doing so, we propose widely used measurement of Information Theory the Entropy. Entropy looks for the average uncertainty of collection of information. We have applied it to come out with the central point of the data file. After obtaining the central point, the Correlation Coefficient is used to select significant attributes in the data files. After that we have implemented J48 algorithm with Kendall's Rank Correlation coefficient method. There are brief discussions on the Entropy with Kendall's Rank Correlation Coefficient and new KNJ48 algorithm. Figure 1 is the structure of the proposed methodology.

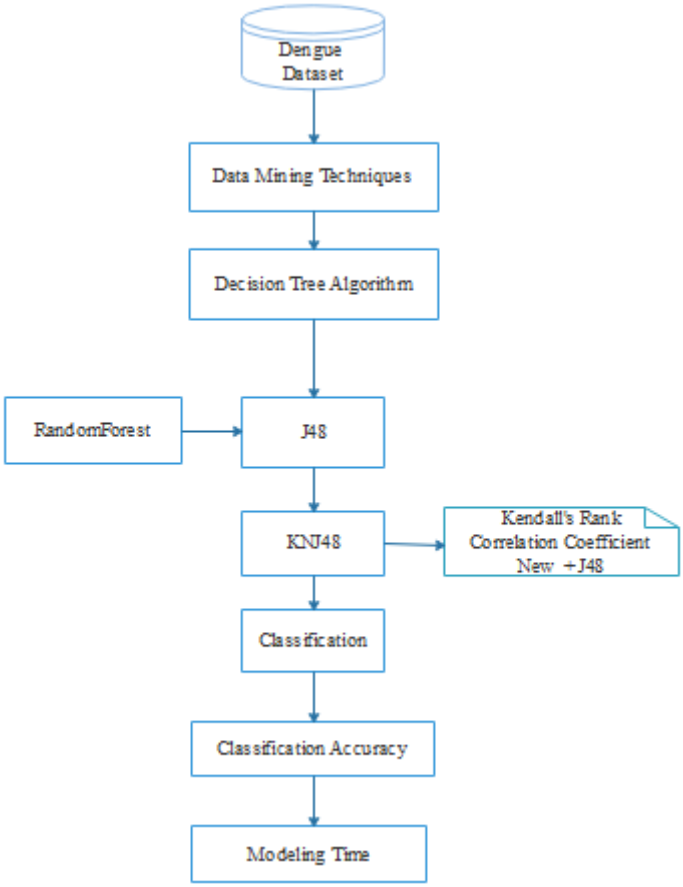

Figure 1. Methodology

4.1. Entropy

Information theory is a popularly used theme for computer scientists, cognitive scientists, data miners, statisticians, biologists and engineers. In information theory, entropy measures the uncertainty among unstoppable variables in a dataset. The idea of entropy of random variables has already been developed. Also the beginnings of information theory and of the modern age of Ergodic theory have also been introduced. Entropy and information pertaining to that offers the long term behavior of random processes that are beneficial to analyse data. The behavior of random process is also an important factor for implementing the coding for information theory. Entropy is a measurement of moderate uncertainty of collection of information when we are not aware of the result of a data source. It means that it's a measurement of how much data we lack. This also points out the average amount of data we will obtain from the result of a data source. (Kalpesh Adhatrao, 2013). The equation 1 stands for measuring information theory of entropy.

Let X is an attribute, p is each element and j is position of each element of X then algorithm for entropy is evaluated using the equation 1.

$$H(X) = \sum_{j=1}^{k} P_j \ log_2 \frac{1}{p_j} = -\sum_{j=1}^{k} P_j \ log_2 p_j \qquad (1)$$

Larger value H(X) indicates that attribute X is more random. On the other hand, attribute with smaller H(X) value implies less random i.e. this attribute is more significant for the data mining. The value of the entropy attains its minimum 0, when all other pj's are 0. The value reaches its maximum log2 k, when all pj's are equal to 1/k.

4.2. Correlation Coefficient

Correlation coefficient is one of the chief statistical tools to examine sets of variables and find out their relationship. So that user would be able to come out with decisions on the basis of available information by correlation coefficients. Hence, it secures millions even billions of dollars for businessman, saves much for researchers and scale down exertion for many other professional in various fields. Researchers have toiled on this tool to develop its efficiency by introducing various ways of calculation. Out of various correlation coefficients, we have selected the most popular one which is Kendall's Rank Correlation Coefficients. In the following sections we have discussed briefly about it. (GANG KOU, 2012)

4.2.1. Kendall's Rank Correlation Coefficients

Kendall's Rank Correlation Coefficient also makes use of nonparametric system for correlation measure. When other correlations are calculated from variables' rank rather Kendall's Rank Correlation Coefficient is related to probability calculation. Kendall's Rank Correlation Coefficient is denoted with the Greek Letter τ (tau). Kendall-tau makes use of concordant or discordant values. The range of value of Kendall's Rank Correlation Coefficient is -1 to +1. The equation 2 stands for measurement of average.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (2)$$

The Kendall τ coefficient is defined as: n_c = Number of concordant pairs, n_d= Number of discordant pairs, n= Number of pairs values

Let X and Y are the pairs of measured and estimated inhibitory activity. Kendall tau coefficient is defined as, nc is concordant value, nd is discordant value and n is a total number of instances (Mohammed M Mazid, 2013).

5. EXPERIMENTAL DESIGN

Performance of our experiment design; we have computed entropy with MODIFIED WEKA programming tools. We select the attribute with least entropy value. We propose that attribute as the central attribute of the database. Then we find out Kendall's rank correlation coefficient depending upon the central attribute using MODIFIED WEKA. Finally, we have applied J48 algorithm with MODIFIED WEKA. MODIFIED WEKA presents different types of test options to classify data files such as user training set, supplied test set, cross-validation and percentage split. We choose 10 fold cross-validation data.

5. 1 Data Set Description

5.1.1 Dataset

In this research work, fifteen attributes are used namely pid, pname, sex, age, tname, bgroup, neutrophil, wbc, rbc, platlet, ast, alt, hb, sodium and urban. The data are collected from 512 dengue affected patients at various Government Hospitals in Krishnagiri District.

5.1.2 Data Preparation

Table 1 is dengue patients' data collected from various Government hospitals in Krishnagiri Distirct.

Table 1. Attributes, Description their Possible Values

Attributes	Description	Possible Values
p_id	Patient ID	1-512
p_name	Patient Name	Patient Name
Sex	Patient Sex	Male , Female
Age	Age of Patient	Patient Age
t_name	Patient Taluk Name	Bargur, denkanikottai, Hosur, thally

Attributes	Description	Possible Values
		Krishnagiri, Pochampalli, Shoolagiri,
b_group	Patient Bloood Group	A+,A-,B+,B-,AB+,AB-,O+,O-
neutrophil	Low(0-1.5), Normal(1.5-8), High(8-above)	Normal, Low, High
Wbc	Low(0-3.8), Average(3.8-10.58), High (10.58-above)	Low,Average,Hight,
Rbc	Poor(0-4.23), Average(4.23-5.59), Good(5.59-above)	Poor,Average,Good
Platlet	Low(0-141), Average(141-316), High(316-above)	Low, Average, High
Ast	Worst(0-blow), Average(0-40), High(40-above)	Worst, Average, High,
Alt	Worst(0-blow), Average(0-40), Normal(40-above)	Worst, Average, Normal
Hb	Low(0-11), Average(11-16), Normal (16-above)	Low, Average, Normal
Sodium	Bad(0-135), Average(135-145), Good(145-above)	Bad, Average, Good
Urban	Urban	Yes, No

5.2 Experimental Outcome

Table 2 has the analysis of AST based classification from RandomForest, J48, and KNJ48 algorithm. Correctly classified on RandomForest, J48 and KNJ48 algorithm and the values are 87.1610%, 88.8672%, 91.6750% respectively. The KNJ48 algorithm is an efficient classification of accuracy than the original J48 and RandomForest algorithm on dengue dataset. The figure 2 describes AST Based Classification.

Table 2. AST Based Classification

Algorithm	Correctly Classified	Incorrectly Classified
RandomForest	87.1610	12.8390
J48	88.8672	11.1328
KNJ48	91.6750	8.3250

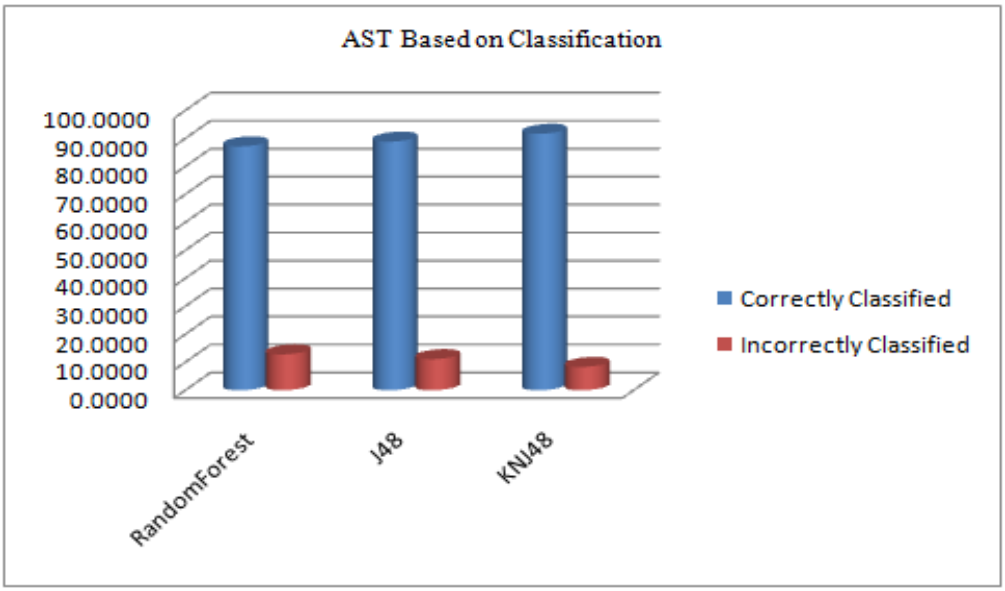

Figure 2. AST Based Classification

Table 3 has the analysis of AST Based Classification of accuracy from RandomForest, J48, and KNJ48 algorithm. The KNJ48 algorithm is an efficient classification of accuracy and the consumption of time is lower than the original J48 and RandomForest algorithm on dengue dataset. The figure 3 AST Based Classification.

Table 3. AST Based Classification

Algorithm	TPR	FPR	Precision	Recall	F-Measure
RandomForest	1.0000	0.0000	0.8711	1.0000	0.9311
J48	1.0000	0.0000	0.8887	1.0000	0.9411
KNJ48	1.0000	0.0000	0.9160	1.0000	0.9562

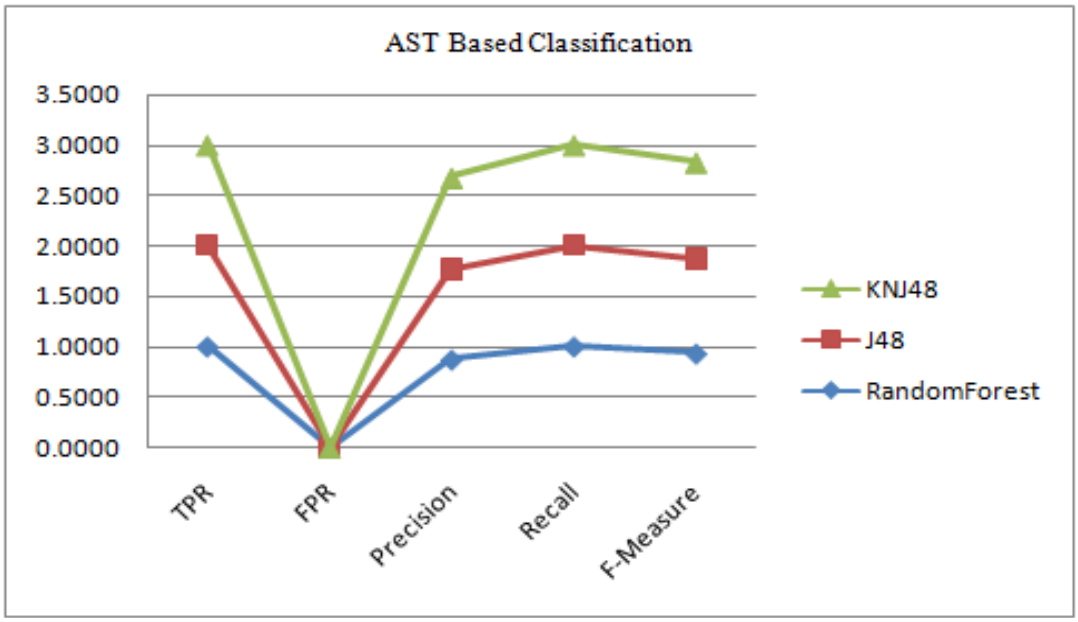

Figure 3. AST Based Classification

Table 4 has the analysis of WBC based classification from RandomForest, J48, and KNJ48 algorithm. Correctly classified on RandomForest, J48 and KNJ48 algorithm and the values are 73.1200%, 74.6094% 75.3100% respectively. The KNJ48 algorithm is an efficient classification of accuracy and the consumption of time is lower than the original J48 and RandomForest algorithm on dengue dataset. The figure 4 is the description of WBC Based Classification.

Table 4. WBC Based Classification

Algorithm	Correctly Classified	Incorrectly Classified
RandomForest	73.1200	26.8800
J48	74.6094	25.3906
KNJ48	75.3100	24.6900

Table 5 has the analysis is WBC based classification of accuracy from RandomForest, J48, and KNJ48 algorithm. The KNJ48 algorithm is an efficient classification of accuracy and the consumption of time is lower than the original J48 and RandomForest algorithm on dengue dataset. Figure 5 is the description of WBC based classification.

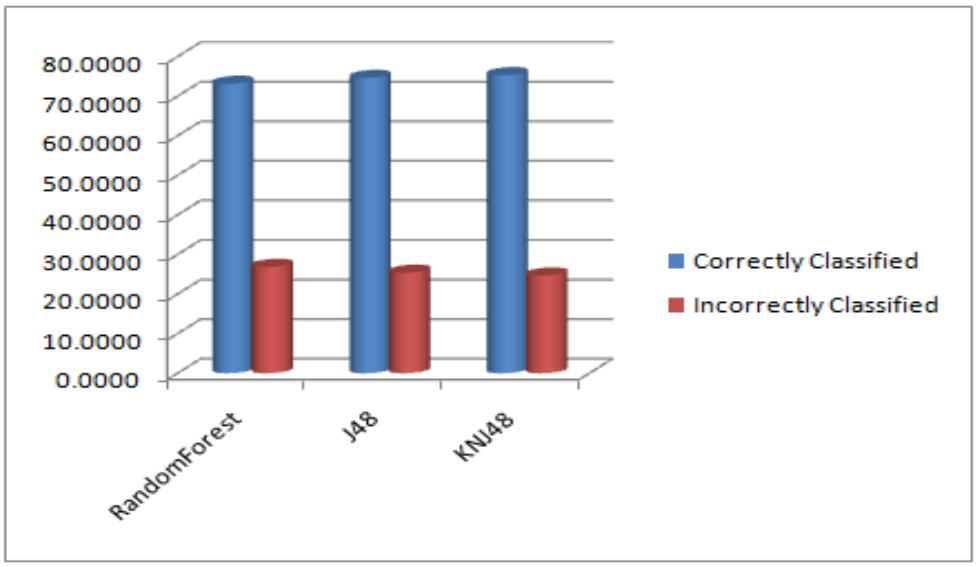

Figure 4. WBC Based Classification

Table: 5 WBC Based Classification

Algorithm	TPR	FPR	Precision	Recall	F-Measure
RandomForest	1.0000	0.0000	0.7305	1.0000	0.8442
J48	1.0000	0.0000	0.7461	1.0000	0.8546
KNJ48	0.9747	0.0253	0.7705	0.9747	0.8606

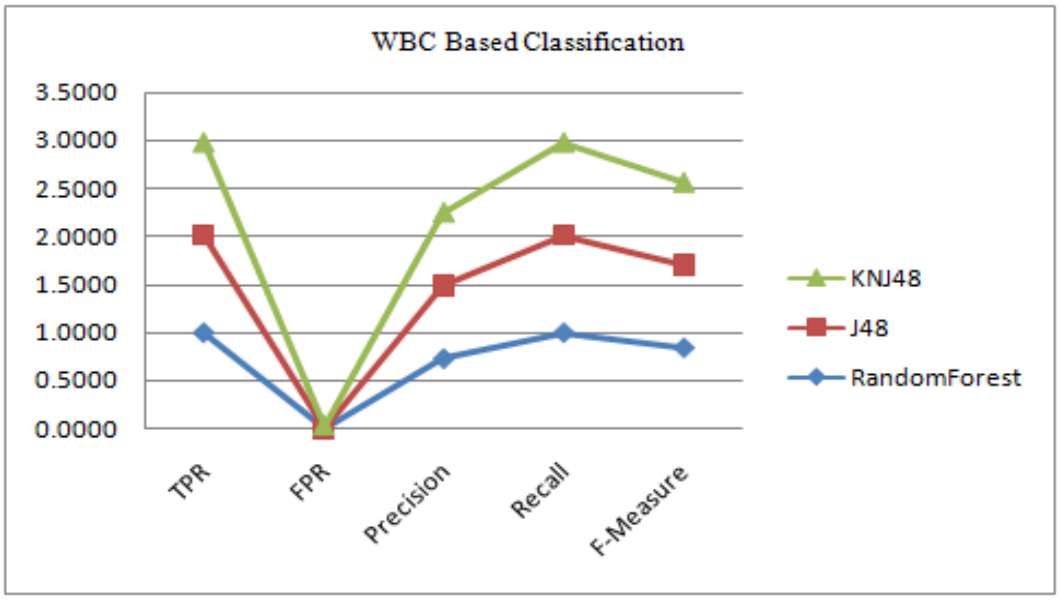

Figure 5. WBC Based Classification

6. RESULT ANALYSIS

The original J48 algorithm first classified on the WEKA tool and then changed the entropy on the J48 algorithm using MODIFIED WEKA on entropy Kendall's rank correlation coefficient and we get KNJ48. KNJ48 is Kendall's new modified J48 algorithm and we used this new algorithm with dengue data set. The performance of KNJ48 is

higher in performing accurate results. After correctly classifying the results are RandomForest, J48 and KNJ48 algorithm and the values are 87.1610%, 88.8672%, 91.6750% is correspondingly. The KNJ48 algorithm is an efficient classification method for dengue dataset and performs good accuracy rate and working for lower time compare to other algorithm J48 and RandomForest algorithm

Table 5. Comparison of original RandomForet, J48 and Modified Kendall's New J48

Data Field	RandomForest		J48		Modified New J48 (Kendall's)	
	Modeling Time	Accuracy	Modeling Time	Accuracy	Modeling Time	Accuracy
WBC	1.05	0.7305 %	0.11	0.7461 %	0.07	0.7559 %
AST	0.87	0.8711 %	0.09	0.8887 %	0.08	0.9162 %

7. CONCLUSION

Data mining gives a grouping of techniques to extract hidden pattern beginning the healthcare industry. This proposed study given an overview of data mining techniques similar to classification algorithms and tools like WEKA tool. This research work proposed to the improved J48 classification algorithm which means developing accuracy and to save the time to get the expected results in an accurate manner. As it was already discussed regarding the modified J48 to become KNJ48 and its higher performance, this shows greater performance on the dengue dataset and the highest classification accuracy rate is 91.62%. The improved J48 provides better classifications than the original J48. The performance of KNJ48 is faster than J48 and in future it can be used to find accurate classifications of different medical datasets.

REFERENCES

Deepali Kharche, K. R. (2014). Comparison Of Different Datasets Sing Various Classification Techniques With Modified WEKA. International Journal of Computer Science and Mobile Computing, IJCSMC, 389 – 393.

Dr. Anjali B Raut, A. A. (2017). Students Performance Prediction Using Decision Tree Technique. International Journal of Computational Intelligence Research, 1735-1741.

Gaganjot Kaur, A. C. (2014). Improved J48 Classification Algorithm forthe Prediction of Diabetes. International Journal of Computer Applications, 22.

Gang Kou, Y. L. (2012). Evaluation Of Classification Algorithms Using Mcdm And Rank Correlation. International Journal of Information Technology & Decision Making .

Jayasimman, L. (2015). Performance Accuracy of Classification Algorithms for Web Learning System. International Journal of Computer Applications .

Kalpesh Adhatrao, A. G. (2013). Predicting Students performance using id3 and c4.5 classification algorithms. International Journal of Data Mining & Knowledge Management Process (IJDKP) .

Kameswara Rao N K, D. V. (2014). Classification Rules Using Decision Tree for Dengue Disease. International Journal of Research in Computer and Communication Technology, 340-343.

Minyechil Alehegn, R. J. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology (IRJET) .

Mohammed M Mazid, A. B. (2013). Improved C4.5 Algorithm for Rule-Based Classification. Recent Advances In Artificial Intelligence, Knowledge Engineering And Data Bases .

Nagaparameshwara chary, S. D. (2017). A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining. International Conference on Innovative Applications in Engineering and Information Technology(ICIAEIT-2017) .

Prerna Kapoor, R. R. (2015). Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning. International Journal of Engineering Research and General Science .

Sagar, N. S. (2015). A Comparative Study of Classification Techniques in Data Mining Algorithms. Oriental Journal Of Computer Science & Technology , 13-19.

Srishti Taneja. (2014). Implementation of Novel Algorithm (SPruning Algorithm). IOSR Journal of Computer Engineering (IOSR-JCE), 57-65.

Tina, R. P. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications .

Venkatesan, E. V. (2015). Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. Indian Journal of Science and Technology .