A Comparitive Study of Clasification Liver Dysfunction with Machine Learning

Sattarpoom Thaiparnit[1], Narumol Chumuang[2] and Mahasak Ketcham[3]

[1]Faculty of Business Administration and Information Technology, Rajamanala University of Technology Suvarnabhumi, Thailand
[2]Department of Digital Media Technology, Muban Chombueng Rajabhat University, Thailand.
[3]Department of Management Information System, King Mongkut's University of Technology North Bangkok, Thailand.

sattarpoom@hotmail.com[1], sattarpoom.t@rmutsb.ac.th[1], lecho20@hotmail.com[2] and mahasak.k@it.kmutnb.ac.th[3]

***Abstract*— This article presents the comparison and comparison of data of patients with liver dysfunction. By collecting information on liver disease and collecting data for selection in data mining. The Liver Disorders Data Set (UCI Machine Learning Repository) was used to compare the 359 patients with liver disease. The classification consisted of 7 types of liver disease and divided into 2 classes, namely, those with normal liver function and those who did not. Abnormal liver. The result was that the data was sorted using the Rules Part accuracy rate with 64.62 %. The OneR rule technique was 58.21 %. The Tree Decision Stump technique was 60.16 %. Tree REPTree has 62.67 % and Tree Random forest technique is 75.76 %. The results of this study showed that the tree random sampling technique was used to extract data from the 359 samples. The sample was extracted with 75.76 %. Because of the comparison results, Tree Random forest provides the most accurate value.**

Keywords*— *Clasification, Comparitive, Liver, Dysfunction, Machine Learning

I. INTRODUCTION

Liver disease is the leading cause of death in Thailand. The health statistics show that the mortality rate for Hundreds of people with hepatitis has increased at a slower rate in the whole country in 2010-2014. 22.25 %, in 2012, 22.51 %, in 2013, 23.94 %, in 2014, 23.56 % [1] due to infection with hepatitis or other causes as in Fig.1. For example, large amounts of alcohol use narcotic drugs. Side effects from taking Include the immune system to destroy themselves. The liver damage caused by various illnesses. Chronic hepatitis may lead to abnormal liver function. Cirrhosis or risk of liver cancer and also affect the economy of the country. Because of liver disease and need to maintain a long time. Cause loss The main causes of liver disease are the daily lives of people in the consumption and consumption. Hygienic treatment From the previous study, it was found that most of the factors that were studied were internal factors. Lack of knowledge of alcoholism and alcohol consumption in long-term consecutively. Alcohol abuse can cause abnormalities in the use of protein, fat and carbohydrates in the liver. Cirrhosis and chronic cirrhosis are more common than cirrhosis. Cirrhosis is more common in men than in men, and can be caused by certain medications and chronic hepatitis, such as painkillers, paracetamol. Tetracycline antibiotic Some TB drugs when we eat a lot of it. It is one of the main causes of liver disease.

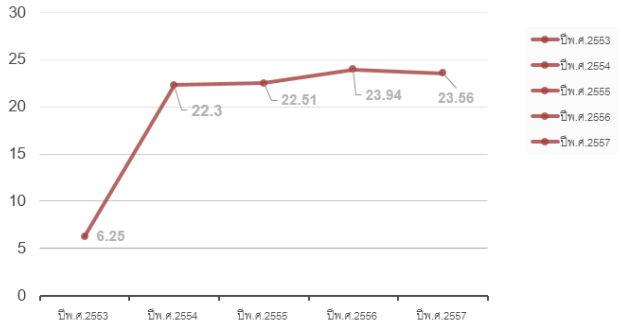

Fig. 1. Mortality rate per 100,000 population with liver disease.

For the above reasons, the development of information systems for the separation of patients with liver disease. Because there are currently limited data collection. There is a collection of factors that cause liver disease. By the organizer, the data is analyzed to compare the algorithm to get the most accurate results.

Therefore, the researcher uses data analysis with data mining techniques. For the screening of patients with liver dysfunction. Comparisons of algency performance were performed using 359 samples.

II. LITTERATURE REVIEWS

A. About Liver Disease

Hepatitis is an infection in the viral hepatitis group. It is divided into two groups: contact hepatitis, food and water contaminated with the virus, hepatitis A and E, and hepatitis. Currently, the Bureau of Epidemiology has been reported to monitor only Hepatitis B and Hepatitis. Because in the year. 2012 has canceled the surveillance of hepatitis C and good. Hepatitis in Thailand over the last 10 years In 2005-2005, there was a reported high rate of 15.32 deaths per 100,000 population and the lowest In 2014, the rate of illness was 12.06. In 2014, 7,385 cases of hepatitis were reported. The rate of sickness was 12.06 per 100,000 population, classified as Hepatitis A 445, 5.68% Hepatitis B 6,283 (80.19%) E 31 cases of hepatitis (0.40%) and non-hepatitis Font pathogens 1,076 cases (13.73%) have been reported deaths, including 3 cases of hepatitis incidence rate was 0.04 percent. 1 death from hepatitis B and 2 non-pathogenic hepatitis, female to male ratio [2].

978-1-5386-8458-0/18/$31.00 ©2018 IEEE

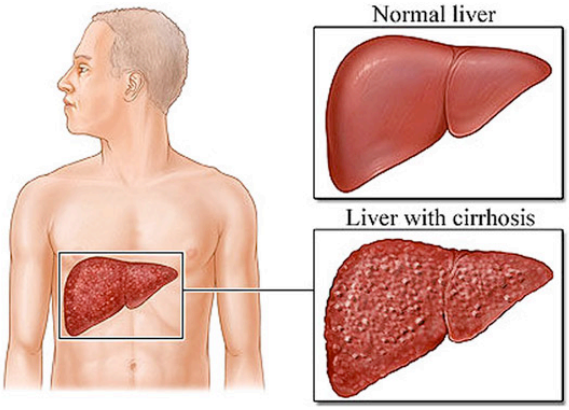

Fig.2. Comparison of normal and non-normal liver characteristics.

B. *Tree Based*

The decision tree is a data tree that generates predictive models that resemble trees. Rules are created to make decisions. Supervised Learning is the ability to create a classification model from a set of predefined data sets called Training Set and automatically predict groups of items that have not been categorized [3],[4]. The tree structure consists of a node and a branch. Each node is represented by a feature of the data set that is learned and tested. Each branch of the tree displays the results in the test and the node. Leaf Node displays user-defined classes. The selection of attributes for tree nodes is based on the information gain calculation. Considering attributes with low information or entropy, this means that the attribute has high classification capability [5].

C. *Random Forest*

The algorithm is a type of unpaved or unpredictable decision tree algorithm, which is generated from the training data, randomly selected data samples and data attributes. Create a decision tree, which contains an unmodified sample, to be used in a tree decision test. This method is called bagging. The independent results of each decision tree are taken as the result of the most votes. The Random Forest algorithm does not require OOB data is used to test the decision tree [6].

D. *OneR Rule*

OneR, short for "One Rule", [7] is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor, we construct a frequency table for each predictor against the target. It has been shown that OneR produces rules only slightly less accurate than state-of-the-art classification algorithms while producing rules that are simple for humans to interpret.

Predictors contribution simply, the total error calculated from the frequency tables is the measure of each predictor contribution. A low total error means a higher contribution to the predictability of the model. Model evaluation the following confusion matrix shows significant predictability power. OneR does not generate score or probability, which means evaluation charts are not applicable.

E. *Related Research*

A. Stojanova, et al. [8] presented data mining techniques. Helps to classify the VARK learning model by using a questionnaire that contains the student's general information and the VARK learning model and ask questions from undergraduate students. Data from the questionnaire will be used to classify data in three ways: Bayes method, Decision tree method, Rule-Bases, Association Rules Support, NBTree. From the experiment, it was found that the decision tree method gave the highest accuracy at 82.78%. The NBTree algorithm gave the accuracy of 81.78%, followed by the rule-based method.

W. Jitsakul, P. Meesad and S. Sodsee [9] presented the classification analysis. To study the stability of the four basic classification algorithms, the TREE BASED RULE BASED association rules support vector machine and three test texts from www.imdb.com, www.yelp.com and www.amazon.com. Data as of November 11, 2016) to study the stability of the algorithm. The analysis of functional characteristics curve (ROC) and paired-t test were presented to the stability of the algorithm studied. The results show that. Tree-based algorithms, such as Random Forest, show stability, message classification, and other algorithms. The mean ROC> 0.80 and the difference between the mean and the test mean were 10 and the experimental data was 0.

P. Khakham, N. Chumuang and M. Ketcham [10] proposed a novel application for the recognition of Isan Dhamma characters. Their algorithm does not require any complicated method and performs word recognition on the whole image holistically, departing from the character based recognition systems of the past. Functional trees (FTs) classifier is the models at the center of this algorithm are trained. The functional trees build univariate decision tree consists of two phases. In the first phase to construct a large decision tree. In the second phase this tree is pruned back. The algorithm to grow the tree follows the standard divide and conquer approach. The local of ten features for Isan Dhamma characters are used. The experiments illustrate the accuracy is 82.33%.

III. Prepare Your Paper Before Styling

An algorithm for classification of patients with liver disorders. The process is as follows Fig.3:

A. *Pre-processing dataset*

The data set used in the study is liver disorders data set from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Liver+Disorders) to compare the patient classification. The number of 359 liver diseases, consisting of mcv (mean corpuscular volume), alkaline phosphatase, sgpt (alanine aminotransferase), sgt (aspartate aminotransferase), gamma gt (gamma-glutamyl transpeptidase), drinks (number of half-pint equivalents of alcoholic beverages drunk per day), selector (field created by the BUPA researchers to split the data into train / test sets), and divided into two classes: those with normal liver and Information that is not a malfunction of the liver.

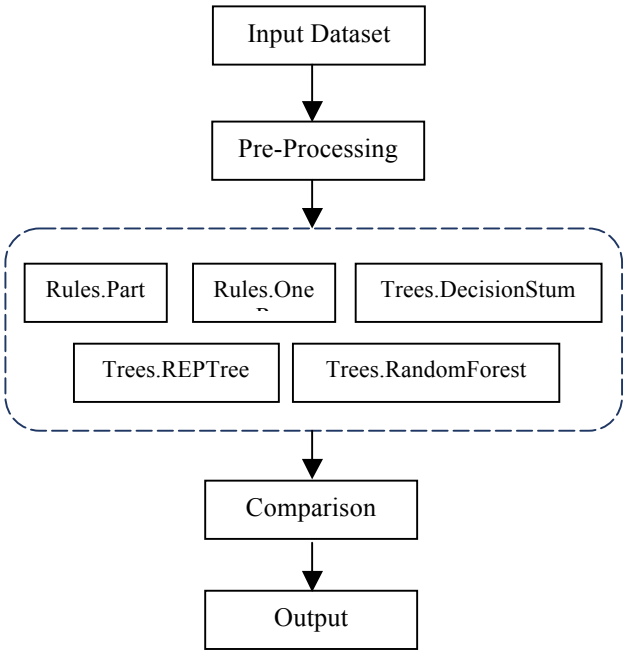

Fig. 3. The overview of system

The number of 359 samples consisted of 7 attributes, including mean blood cell volume (mcv), enzyme activity in the liver (alphos), blood test for liver enzyme (sgpt), Enzymes generated from liver damage (sgot), enzyme levels in the liver (gamma gt), daily drinks (drinks), types of selectors. The information is of an example. Details are as follows:

TABLE I. SAMPLE DATA FROM A TOTAL OF 359 SAMPLES

mcv	alkphos	sgpt	sgot	gammagt	drinks	selector
85	92	45	27	31	0	n
85	64	59	32	23	0	a
86	54	33	16	54	0	a
91	78	34	24	36	0	a
87	70	12	28	10	0	a
98	55	13	17	17	0	a
88	62	20	17	15	0.5	n
88	67	21	11	11	0.5	n
92	54	22	20	16	0.5	n
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Attribute namely “selector” has two class by using nominal *n* is normally and *a* is abnormally. The information in this study is divided into two classes, from the actinomycetes, including liver dysfunction and abnormalities of the liver.

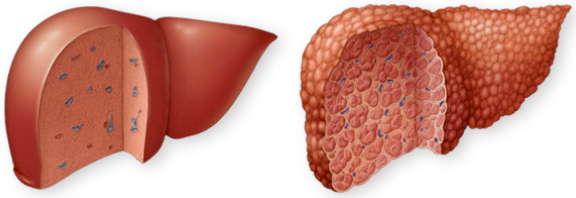

Fig. 4. Normal liver function vs. abnormal liver.

B. *Setting*

In the trial, the researchers used seven data sets, including mcv (mean corpuscular volume), alkaline phosphatase, sgpt (alanine aminotransferase), sgt (aspartate aminotransferase), gamma gt (gamma-glutamyl transpeptidase) Drinks (number of half-pint equivalents of alcoholic beverages drunk per day), selectors 359 samples were divided into data into train / test sets. The five folds cross validation for more reliable performance testing By dividing the training data into 5 parts with the same number. After that, the performance of the model was five times as follows:

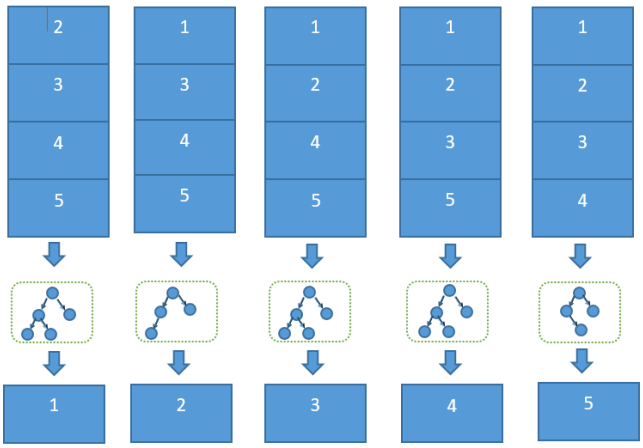

Fig. 5. The five folds cross validation experiments

Round 1 uses data sections 2,3,4 and 5 to create model and predict data.

Round 2 uses data sections 1,3,4 and 5 to create model and predict data.

Round 3 uses data sections 1,2,4 and 5 to create model and predict data.

Round 4 uses data sections 1,2,3 and 5 to create model and predict data.

Round 5 uses data sections 1,2,3 and 4 to create model and predict data.

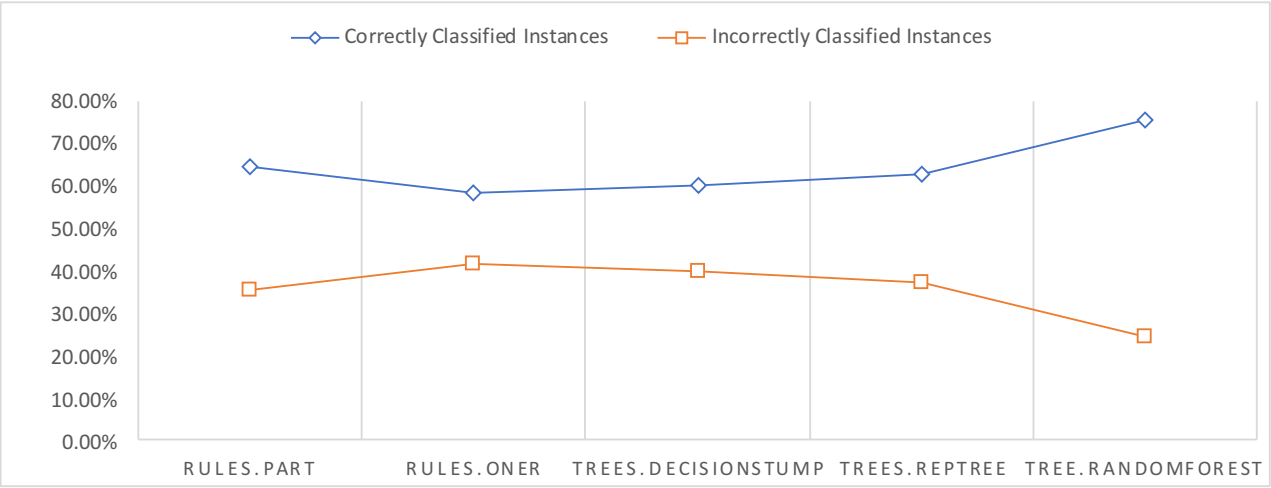

Fig. 6. The comparison of the liver disorders data set

IV. THE EVALUATION AND RESULT

To evaluate the classified , the hypothesis of the testing is setting 5-folds cross validation using the rules.Part and tree.RandomForest. Results from all 359 samples. Data extraction using Rules Part has a value of 64.62 %. The OneR rule has 58.21 % accuracy. The Tree Decision Stump technique is 60.16 % accurate. The Tree REPTree technique is accurate. 62.67 % and Tree Random forest technique was 75.76 %. The detail shown in Table II.

TABLE II. THE CLASSIFIER RESULT FROM DIFFERENT ALGORITHM

Classifier	Corrected rate	Error rate
Rules.Part	64.62 %	35.37 %
Rules.OneR	58.21 %	41.78 %
Trees.DecisionStump	60.16 %	39.83 %
Trees.REPTree	62.67 %	37.32 %
Trees.RandomForest	75.76 %	24.23 %

From Table II, Correctly Classified Instances The hypothesis of the test is a 5-fold cross validation method using different algorithms. There are different discrepancies and differences. Of the 359 samples, Tree.RandomForest Correctly Classified Instances is 75.766 %, which is the most accurate value.

CONCLUSION

For the purpose of this paper, a comparative study was conducted to separate the data of patients with liver dysfunction. The Liver Disorders data set (UCI Machine Learning Repository) was used to compare the 359 records with liver disease. The classification consisted of 7 types of liver disease and divided into 2 classes, namely, those with normal liver function and abnormal liver. The experiment was a 5 folds cross validation method. Random Forest had the accuracy of 75.76 % with the highest accuracy. It can be concluded that the Random Forest is the most effective. So, the forecast model. A comparative study of the data for randomized data analysis using random forest was performed.

REFERENCES

[1] R. T. Ribeiro, R. T. Marinho and J. M. Sanches, "Classification and Staging of Chronic Liver Disease From Multimodal Data," in IEEE Transactions on Biomedical Engineering, vol. 60, no. 5, pp. 1336-1344, May 2013. doi: 10.1109/TBME.2012.2235438

[2] M. P. Andre et al., "Accurate diagnosis of nonalcoholic fatty liver disease in human participants via quantitative ultrasound," 2014 IEEE International Ultrasonics Symposium, Chicago, IL, 2014, pp. 2375-2377. doi: 10.1109/ULTSYM.2014.0592

[3] Ganmaa, D., Li, X. M., Wang, J., Qin, L. Q., Wang, P. Y., & Sato, A. (2002). Incidence and mortality of testicular and prostatic cancers in relation to world dietary practices. International Journal of Cancer, 98(2), 262-267.

[4] Quadri, M. M., & Kalyankar, N. V. (2010). Drop out feature of student data for academic performance using decision tree techniques. Global Journal of Computer Science and Technology.

[5] Lee, C., & Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. Information processing & management, 42(1), 155-165.

[6] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[7] S. K. Solanki and J. T. Patel, "A Survey on Association Rule Mining," 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Haryana, 2015, pp. 212-216.

[8] A. Stojanova, N. Stojkovikj, M. Kocaleva, B. Zlatanovska and C. Martinovska-Bande, "Application of VARK learning model on “Data structures and algorithms” course," 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, 2017, pp. 613-620.

[9] W. Jitsakul, P. Meesad and S. Sodsee, "Enhancing comment Feedback classification using text classifiers with word centrality measures," 2017 2nd International Conference on Information Technology (INCIT), Nakhonpathom, 2017, pp. 1-5.

[10] P. Khakham, N. Chumuang and M. Ketcham, "Isan Dhamma Handwritten Characters Recognition System by Using Functional Trees Classifier," 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Bangkok, 2015, pp. 606-612.