Recognition of Liver Disease Based on Feature Choice Approach: An Experimental Review

Suchithra K[1], Anjan Babu G[2]

Dept of Computer Science, SV University, Tirupati

***Abstract*—** *Feature choice strategy is utilized for creating an ideal number of features to be utilized for a specific undertaking like arrangement. The target of the dispensing with measure is to decrease the size of the info features set and simultaneously to hold the class oppressive data. This paper proposes and assesses another element determination calculation utilizing data hypothesis which is the common data (MI) between mixes of info features and the class rather than shared data between solitary information include and the class for both consistent esteemed and discrete-esteemed features. Features class MI has been utilized to choose a subset of features dependent on its importance. The investigation is done on the Bupa Liver problem dataset taken from the University of California at Irvine Machine Learning Data Repository. The dataset contains a colossal volume of feature estimations which are diminished using MI based segment assurance methodology. The dataset contains an immense rundown of abilities which is reduced using a further developed segment decision methodology named as covering procedure. They chose subset of features then, at that point go through a pre-processing step to present a consistency in the conveyance of information. Since Classification by means of Regression is perceived to have the advantage of giving a striking execution in grouping stage. The arrangement exhibitions have been discovered promising when contrasted and characterizations performed utilizing typical classifiers and with utilizing common data.*

I. INTRODUCTION

The liver is completely inspected to be one of the central organs in any living body with fundamental limits like getting ready additional things, making mixtures, and taking out drained tissues or cells [4]. We can stay alive a couple of days if our liver shuts down. The liver is the greatest glandular organ of the body. It weighs around 3 lb (1.36 kg). It is blushing brown in concealing and is disengaged into four projections of conflicting size and shape [10]. The liver lies on the right 50% of the stomach melancholy under the stomach. Blood is passed on to the liver through two colossal vessels called the hepatic stock course and the passageway vein. The heptic vein passes on oxygen-rich blood from the aorta (a huge vessel in the heart). The section vein passes on blood containing handled food from the little stomach related parcel. These veins parcel in the liver again and again, finishing off with little vessels. Each thin prompts a lobule. Liver tissue is made out of thousands of lobules, and each lobule is involved hepatic cells, the crucial metabolic cells of the liver [1] [2]. This paper depicts Excessive usage of alcohol can cause an exceptional or persevering disturbance of the liver and may even harm various organs in the body, alcohol incited liver contamination remains a huge issue.

Right when the liver gets feeble, it may have various certified results. Liver ailment (furthermore called hepatic disease) is a wide term depicting any single number of ailments impacting the liver. Many are joined by jaundice achieved by extended levels of bilirubin in the structure. The bilirubin results from the detachment of the hemoglobin of dead red platelets; usually, the liver wipes out bilirubin from the blood and releases it through bile.

II. FEATURE SELECTION

Feature choice has been broadly examined and utilized by the AI and information mining local area. In this specific circumstance, an element, likewise called trait or variable, addresses a property of an interaction or framework than has been measured or built from the first info variables. The objective of feature determination is to choose the littlest feature subset given a specific speculation blunder, or alternatively finding the best component subset with k features, that yields the minimum speculation mistake. Extra destinations of feature determination are as per the following: (I) further develop the speculation execution regarding the model assembled using the entire arrangement of highlights, (ii) give a more vigorous speculation and a quicker reaction with inconspicuous information, and (iii)achieve a superior and less difficult comprehension of the process that produces the information [6] [7]. Highlight choice methods are typically ordered in three fundamental gatherings: covering, inserted, and channel techniques [5]. Coverings use the induction learning calculation as a

component of the capacity assessing highlight subsets [9]. The presentation is typically measured in terms of the grouping rate got on a testing set, i.e., the classifier is utilized as a black box for surveying feature subsets. Albeit these methods might accomplish a good generalization, the computational expense of preparing the classifier a combinatorial number of times becomes prohibitive for high-dimensional datasets

2.1 Mutual Information (MI)

The MI which is a proportion of the reliance between the arbitrary factors is consistently symmetric and non-negative. It is zero if and just if the factors are free. The shared information [8] between two discrete arbitrary factors U= (u1, u2, uk) and V= (v1, v2, vd) is characterized as

$$I\ (U,V) = \ \Sigma_u \Sigma_v \mathrm{P}\ (\mathrm{u},\ \mathrm{v})\ \log \frac{p(u,v)}{p(u)p(v)}$$

Where U= (u1, u2, UK) and V= (v1, v2, vd) are the upsides of the discrete factors U and V individually. P (u, v) is a joint thickness work and p (u) and p (v) are the peripheral thickness capacities.

In this strategy, a shared data measure is utilized to ascertain the data acquire among highlights just as among highlight and class ascribes. Utilizing a voracious way, each time we pick an element from the list of capabilities actually left one that gives most extreme data about the class quality with least excess.

III. METHODOLOGY

The data might contain overabundance and pointless attributes; there is a need to dispose of these characteristics without lessening the exactness using a part assurance technique. Dimensionality decline in Liver issue dataset insightful model involves the going with progresses:

- To scale the data and to isolate the features from the first dataset using Mutual Information (MI).
- Create getting ready and testing dataset.
- Apply order through Regression systems to the readiness set.
- Generate the judicious model.
- Evaluate model using testing dataset.
- Compare execution among the features and without incorporate decision procedures.

3.1 Classification through Regression

The arrangement in this investigation dependent on order by means of relapse, and Classification through grouping. The grouping through relapse is an arrangement technique that can change issues into relapse capacities [3]. This technique joins the standards of the choice tree calculation and straight relapse on a few sub-trees (leaves) that are assembled. This technique has two fundamental advances in particular [11]:

1. Settle on a conventional choice tree, by amplifying the division of boundaries/ascribes and their varieties as per the objective/yield esteems. In settling on a choice tree, it should be determined the deviation decrease standard.
2. Managing the choice tree (pruning) on a few potential sub-trees, and filling it with the relapse work (straight model) as needs be, as a rule on leaves.

IV. EXPLORATORY RESULTS

This part gives results and related discussion on data driven examination of Bupa Liver issue dataset was accumulated from UCI storehouse [8]. This investigation work was executed using WEKA. WEKA is made by investigators at the University of Waikato in New Zealand. The item is written in the Java language and contains a GUI for speaking with data records. WEKA also gives the graphical UI of the customer and gives various workplaces. WEKA is a state of the art office for making AI (ML) strategies and their application to genuine data mining issues. These records were masterminded into two classes, contains 846 cases and 19 credits. The investigations were performed considering 593 models which suggest 70% of

the total models were planning data and 30% were attempting data. The Bupa Liver problem dataset data are displayed in the table-1. The nitty gritty data of the dataset is displayed in the figure-1.

TABLE 1
DATASET INFORMATION

S.No	Name of the Dataset	No. of Attributes	No. of Instances	No. of Classes
1	Bupa dataset	7	345	1 : 145
				2 : 200

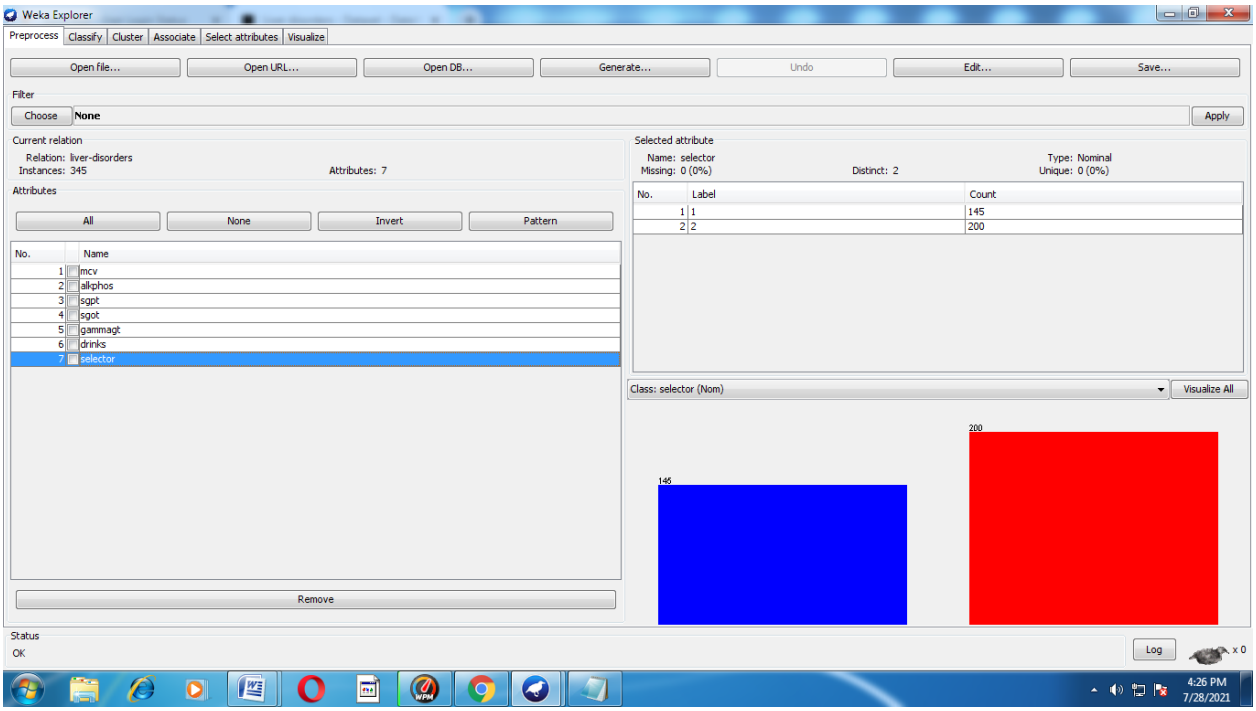

FIGURE 1: Bupa Liver disorder dataset

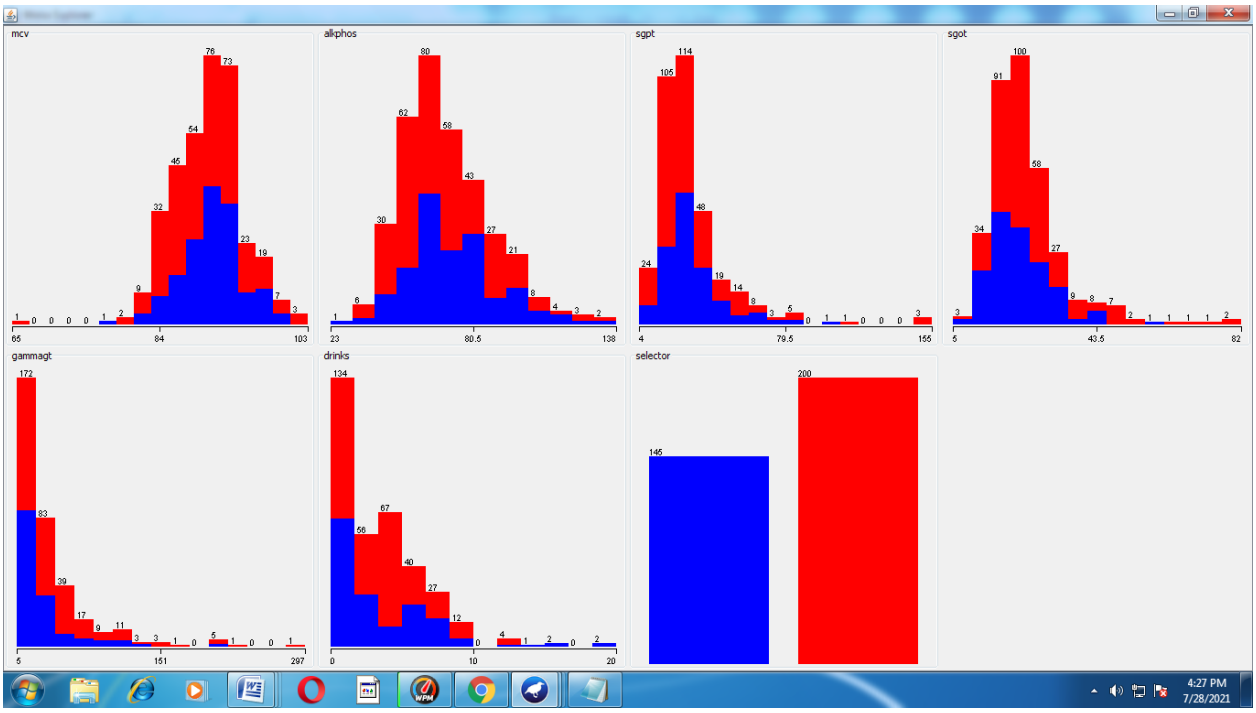

FIGURE 2: Statistical summary of the dataset

4.1 Results

In the first stage Classification via Regression algorithm was trained on the original set of features was used in the experiment. In the second stage we implement the MI algorithm for obtaining the adequate number of features to identify the features selected. The results that we got for Classification via Regression without feature selection and with feature selection are shown below in the table-2 and same as shown in the figure-3 with their corresponding values.

TABLE 2
PERFORMANCE OF CLASSIFIERS

Algorithm	Accuracy	precision	Recall
Classification Via Regression without MI	70	69	70
Classification Via Regression with MI	72.4	72	72

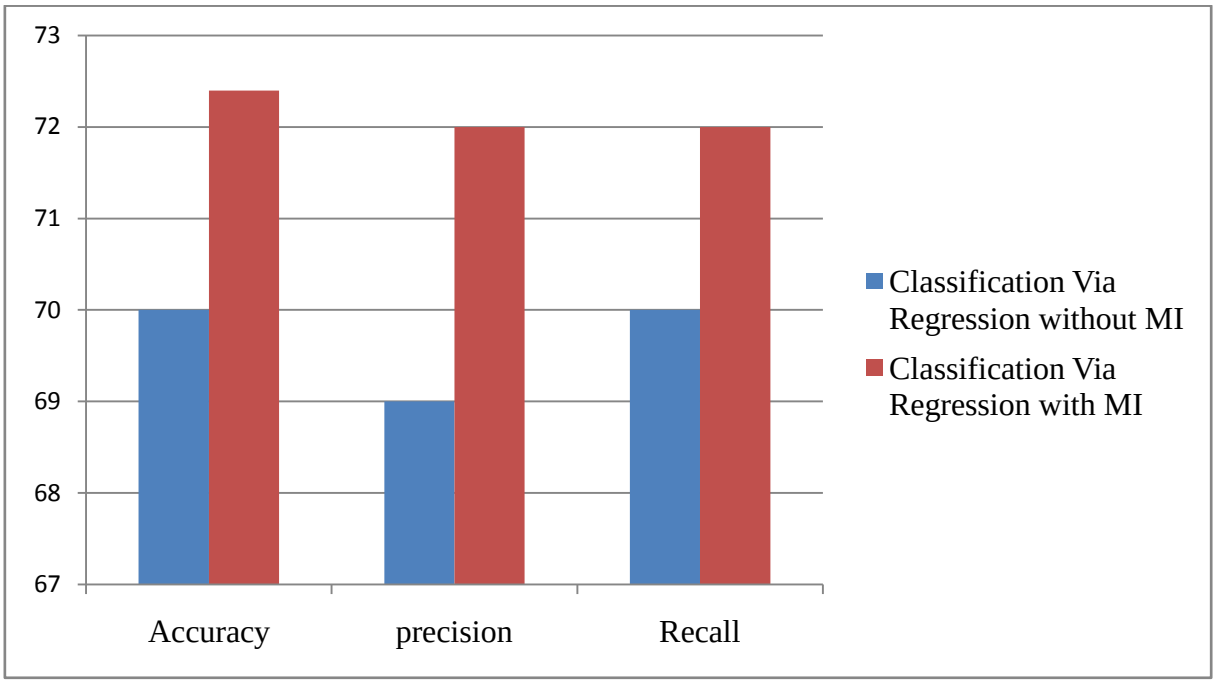

FIGURE 3: Performance of a classifier

From the figure-3, we observe the performance of Classification via Regression without MI based on accuracy has got 70%, whereas the performance of Classification via Regression with MI feature selection based on accuracy has achieved 72.4%. However, there is an improvement in the accuracy with feature selection. The accuracy rate is increased 2.4% with feature selection.

In our experimental result the Classification via Regression with MI feature selection algorithm shows the highest accuracy compared with Classification via Regression with MI. With the improvement the accuracy, the proposed model demonstrated that it performs well after selecting relevant features. This result provided new insight using a classification learning algorithm and reduction technique to selection relevant and important feature in order to improve the accuracy of the system and to identify possible features which may contribute to this improvement. Most of the proposed research system could effectively utilize feature selection process to improve detection rate of their system and minimize considerably the false alarm rate.

The experimental screen shots are shown in the figures-4 and figure-5.

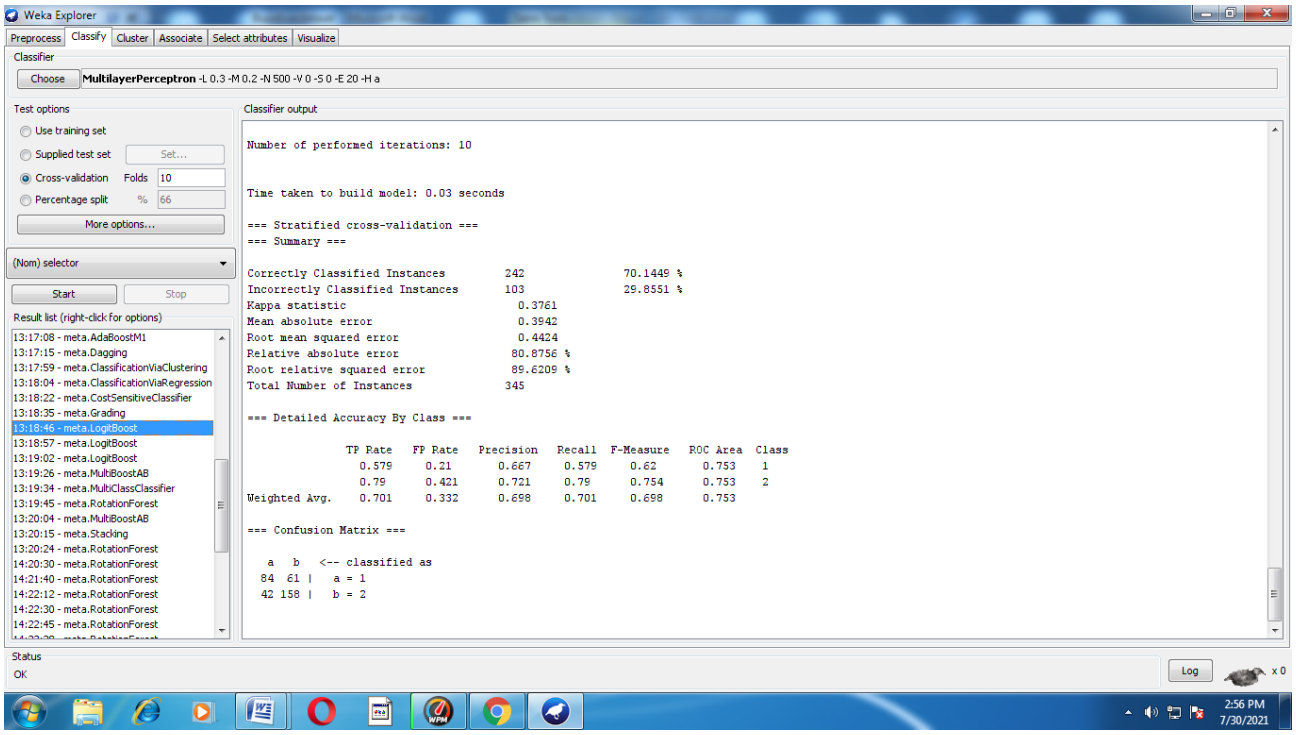

FIGURE 4: Experimental screen shot

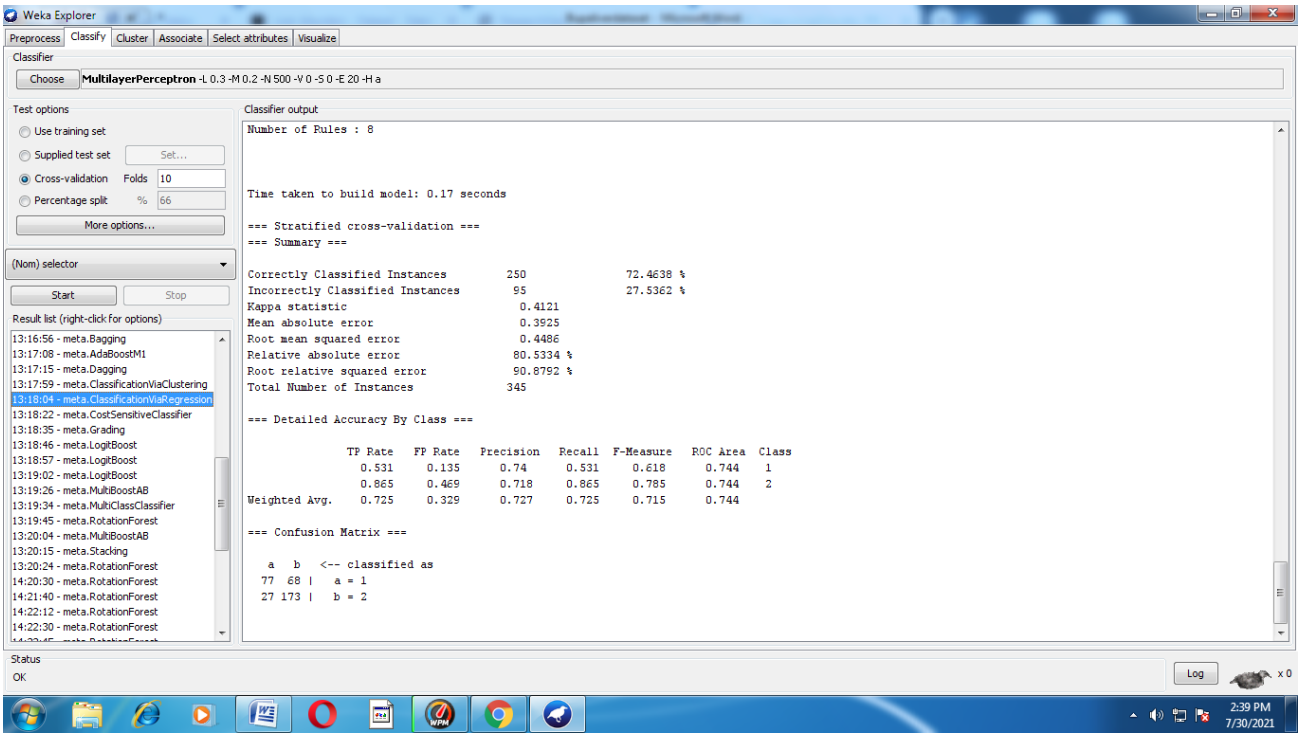

FIGURE-5: Experimental screen shot

V. CONCLUSION

This paper has investigated the approaches to solve the important classification problem of the feature selection. A presentation and proposition of a feature selection method which consist of a MI feature elimination using a Classification via Regression classifier to identify important features have been done. The feature selection, pre-processing, and classification techniques have produced a combination which provides promising results for classification. The evaluation the effectiveness of the method using different classification metric measurement has been made and it has been proved that by

reducing the number of features, the accuracy of the model was improved. In order to detect class from large dataset, detection algorithm, and feature selection method have too more efficient.

REFERENCES

[1] Abdar M, Yen NY, Hung JCS (2017) Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J Med Biol Eng 38(6):953–965

[2] Alfisahrin SNN, Mantoro T (2013) Data mining techniques for optimization of liver disease classification. In: 2013 International conference on advanced computer science applications and technologies. IEEE, pp 379–384

[3] Arora T and Dhir R 2017 Correlation-Based Feature Selection and Classification Via Regression of Segmented Chromosomes Using Geometric Features Medical & biological engineering & computing 55(5) 733-745

[4] A. N. Arbain and B. Y. P. Balakrishnan, "A comparison of data mining algorithms for liver disease prediction on imbalanced data," International Journal of Data Science and Analytics, vol. 1, 2019.

[5] G. Ravi Kumar, K. Nagamani and G. Anjan Babu, "A Framework of Dimensionality Reduction Utilizing PCA for Neural Network Prediction", Lecture Notes on Data Engineering and Communications Technologies, ISBN 978-981-15-0977-3, Volume 37, PP:173-180, Springer Nature Singapore Pte Ltd. 2020

[6] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE Trans. Knowl. Data Eng, vol. 17, no. 4, (2005), pp. 491–502.

[7] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," Pattern Recognition, vol. 42, no. 7, pp. 1330–1339, 2009.

[8] H. Witten and E. Frank, "Data mining – Practical Machine Learning tools and Techniques (2nd Ed.)," San Francisco, CA: Morgan Kaufmann Publisher, An imprint of Elsevier, 2005.

[9] Kira, K., and Rendell, L. A. The feature selection problem: Traditional methods and a new algorithm. In Proceedings of the AAAI-92 (1992), AAAI Press, pp. 129–134.

[10] N. Nahar and F. Ara, "Liver disease prediction by using different decision tree techniques," International Journal of Data Mining & Knowledge Management Process, vol. 8, no. 2, pp. 01–09, 2018.

[11] Ruan Y, Lin H and Tsai M 2014 Improving Ranking Performance With Cost-Sensitive Ordinal Classification Via Regression Information retrieval 17(1) 1-20

[12] UCI machine learning repository. http://archive.ics.uci.edu/ml/.