International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
IJCAI 2021; 2(2): 12-16
Received: 07-05-2021
Accepted: 09-06-2021

Ahamed S
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

B Pavan Kumar Reddy
Assistant Professor, Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Corresponding Author:
Ahamed S
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

An efficient feature selection approach for detection of liver disease prediction

Ahamed S and B Pavan Kumar Reddy

DOI: https://doi.org/10.33545/27076571.2021.v2.i2a.32

Abstract
The human liver issue is a hereditary issue because of the habituality of liquor or impact by the infection. It can prompt liver disappointment or liver malignancy, if not been recognized in introductory stage. The point of the proposed strategy is to recognize the liver issue in beginning stage utilizing liver capacity test dataset. There are numerous problems of the liver that require clinical consideration by a doctor or other medical services proficient. The dataset contains a colossal volume of highlights estimations which are diminished using ReliefF based component assurance methodology. The dataset contains a gigantic rundown of abilities which is diminished using an improved segment decision procedure named as covering method. The current examination centers on different component determination methods, which is perhaps the most significant and every now and again utilized in information preprocessing for information mining. In this paper the impact of highlight determination on the exactness of Decision Tree and Naïve Bayes, classifiers is introduced utilizing Liver problem information. These two classifiers are contrasted and genuine dataset which are pre-prepared with highlight determination strategies. The proposed covering method depends on a choice tree - ReliffF and credulous bayes – ReliffF computation to pick the primary highlights from the given dataset. The picked subset of highlights then goes through a preprocessing step to introduce a consistency in the allotment of data. Since guileless bayes is seen to enjoy the benefit of giving a famous execution in portrayal stage.

Keywords: feature selection approach, detection, liver disease prediction

1. Introduction

The liver is thoroughly examined to be one of the focal organs in any living body with essential capacities like preparing extra items, creating compounds, and taking out depleted tissues or cells [3]. We can remain alive a few days if our liver closes down. The liver is the biggest glandular organ of the body. It weighs around 3 lb (1.36 kg). It is rosy brown in shading and is isolated into four projections of inconsistent size and shape [11]. The liver lies on the right half of the stomach depression underneath the stomach. Blood is conveyed to the liver through two enormous vessels called the hepatic supply route and the entrance vein. The heptic vein conveys oxygen-rich blood from the aorta (a significant vessel in the heart). The entry vein conveys blood containing processed food from the small digestive tract. These veins partition in the liver over and over, ending in minuscule vessels. Every slender prompts a lobule. Liver tissue is made out of thousands of lobules, and every lobule is comprised of hepatic cells, the fundamental metabolic cells of the liver [1, 2]. This paper portrays Excessive utilization of liquor can cause an intense or persistent irritation of the liver and may even damage different organs in the body, liquor prompted liver infection stays a significant issue.

At the point when the liver gets ailing, it might have numerous genuine outcomes. Liver illness (additionally called hepatic infection) is a wide term portraying any single number of sicknesses influencing the liver. Many are joined by jaundice brought about by expanded degrees of bilirubin in the framework. The bilirubin results from the separation of the hemoglobin of dead red platelets; ordinarily, the liver eliminates bilirubin from the blood and discharges it through bile.

1.1 Normal Liver Disorder: Fatty liver (otherwise called steatorrhoeic hepatosis or steatosis hepatitis) is a reversible condition where enormous vacuoles of fatty oil fat gather in liver cells by means of the cycle of steatosis.

It can happen in individuals with a significant degree of liquor utilization just as in individuals who never had liquor. Hepatitis (generally brought about by an infection spread by sewage tainting or direct contact with contaminated body liquids).

Cirrhosis of the liver is perhaps the most genuine liver sicknesses. It's anything but a condition used to signify all types of illnesses of the liver portrayed by the huge loss of cells. The liver steadily contracts in size and gets weathered and hard. The regenerative action proceeds under liver cirrhosis however the reformist loss of liver cells surpasses cell substitution.

Liver cancer The danger of liver malignant growth is higher in the individuals who have cirrhosis or who have had particular sorts of viral hepatitis; however more frequently, the liver is the site of optional (metastatic) tumors spread from different organs.

2. Feature Selection

Highlight decision has been a working and useful field of exploration area in plan affirmation, AI, experiences and data mining networks [6]. The essential goal of Feature assurance is to pick a subset of information factors by clearing out features, which are unnecessary or of no farsighted information. Highlight decision has shown in both speculation and practice to be reasonable in improving learning adequacy, growing judicious accuracy and diminishing multifaceted nature of learned results [7]. Highlight decision in regulated learning has an essential target of discovering a feature subset that produces higher portrayal precision. As the dimensionality of a space develops, the amount of highlights N increases. Finding an optimal component subset is obstinate and gives related part decisions have been wind up being NP-hard [9]. At this intersection, it is central to depict standard component decision measure, which involves four major advances, to be explicit, subset age, subset appraisal, ending standard, and endorsement [4, 5]. Subset age is a chase communication that produces contender incorporate subsets for evaluation dependent on a particular pursuit approach. Each up-and-comer subset is evaluated and differentiated and the previous best one as demonstrated by a particular appraisal. If the new subset goes to be better, it replaces best one. This cycle is reiterated until a given ending condition is satisfied.

2.1 Relief Feature Selection

Help was proposed by Kira and Rendell in 1994 [10]. Help is an element choice calculation for irregular determination of occasions for include weight computation. The Relief calculation embraces the arbitrary choice of occasions for weight assessment. An example is chosen from the information, and the closest adjoining test that has a place with a similar class (closest hit) and the closest adjoining test that has a place with the contrary class (closest miss) are distinguished. An adjustment of property estimation joined by an adjustment of class paves the way to weighting of the quality dependent on the instinct that the trait change could be answerable for the class change. Then again, an adjustment of quality worth joined by no adjustment of class prompts down weighting of the characteristic dependent on the perception that the trait change had no impact on the class. This strategy of refreshing the heaviness of the trait is performed for an arbitrary arrangement of tests in the information or for each example in the information. The weight refreshes are then arrived at the midpoint of so the last weight is in the reach [−1, 1]. The quality weight assessed by Relief has a probabilistic translation. It is relative to the contrast between two restrictive probabilities, in particular, the likelihood of the trait's worth being diverse adapted on the given closest miss and closest hit individually [12].

The accomplishment of the calculation is because of the way that it's quick, straightforward and execute and precise even with subordinate highlights and boisterous information. The calculation essentially comprises of three significant parts:

1. Compute the closest miss and closest hit;
2. Compute the heaviness of a component;
3. Return a positioned rundown of highlights or the top k highlights as indicated by a given edge.

3. Methodology

The information may contain excess and superfluous characteristics, there is a need to eliminate these traits without diminishing the precision utilizing a component determination method. Dimensionality decrease in Liver issue dataset prescient model comprises of the accompanying advances:

- To scale the information and to separate the highlights from the first dataset utilizing ReliefF.
- Create preparing and testing dataset.
- Apply choice tree and gullible bayes procedures to the preparation set.
- Generate the prescient model.
- Evaluate model utilizing testing dataset.
- Compare execution among the highlights and without include choice strategies.

3.1 Decision Tree

Decision tree learning is maybe the best techniques for controlled portrayal learning. Decision trees are a fundamental recursive plan for conveying a sequential course of action measure where a case, portrayed by a lot of qualities, is assigned to one of a disjoint course of action of classes [8]. A decision tree is a tree structure which bunches a data test into one of its likely classes. Decision trees are used to remove data by making decision rules from the colossal proportion of available information. A decision tree classifier has a fundamental construction which can be moderately taken care of and that adequately orchestrates new data.

Decision trees include center points and leaves. Each center in the tree incorporates testing a particular quality and each leaf of the tree demonstrates a class. For the most part, the test differentiates a property assessment and a steady. Leaf center points give a gathering that applies to all events that show up at the leaf, or a lot of requests, or a probability spread over each and every possible game plan. To describe a dark event, it is directed down the tree according to the potential gains of the characteristics attempted in reformist centers, and when a leaf is reached, the case is masterminded by the class given out to the leaf.

3.2 Naive Bayes

Naïve Bayes is maybe the awesome capable portrayal computations. Blameless Bayes Classifier that is the probabilistic classifier subject to the Bayes Theorem. Honest Bayes classifier expects that the effect of the qualities regard on a given class is self-sufficient on the value of various features [8]. The classifier essentially picks the imprint with the most raised probability, given the information features. The guileless piece of the classifier is that it's anything but's a strong opportunity between attributes, fundamentally it acknowledges the probabilities for all of the data features are self-governing of each other.

Let H be a hypothesis and X is a data residing in a certain C class. Then P (H / X) is called the posterior probability that expresses our confidence level on a hypothesis H after X data is given. P (H) represents the H prior probability for all sample data. P (H / X) is certainly more informative than P (H). Bayes's theorem describes the relationship between P (H / X), P (H), and P (X) is shown on equation 1 as follow:

$$P(H/X) = P(X/H) * P(H)/P(X) \quad (1)$$

4. Experimental Results

This section comprises the experimental analysis of Liver Disorder dataset was gathered from the UCI machine learning repository [13] as shown in Table 1.

Table 1: Dataset Information

S. No	Name of the Dataset	No. of Attributes	No. of Instances	No. of Classes
1	Liver Disorder	7	345	Presence:145 Absence:200

The two ML classifiers are assessed on the dataset. To approve the expectation consequences of the correlation of the two characterizations (choice tree and gullible bayes) with highlight choice procedures and the 10-overlap hybrid approval is utilized. The k-crease hybrid approval is generally used to lessen the mistake came about because of irregular examining in the examination of the exactnesses of various forecast models. The fundamental motivation behind this examination is to foresee and assess the liver illnesses effectively utilizing highlight choice procedure and grouping calculations productively. Additionally, analyze the aftereffects of both component determination and without choice procedure on two classifiers in particular choice tree and gullible bayes to gauge which technique gives the more accurately ordered outcome for conclusion of liver illnesses. Highlight choice method was carried out to decrease the properties from liver infections dataset to discover better outcomes. The itemized Statistical synopsis of the dataset displayed in the figure-1 and figure-2.

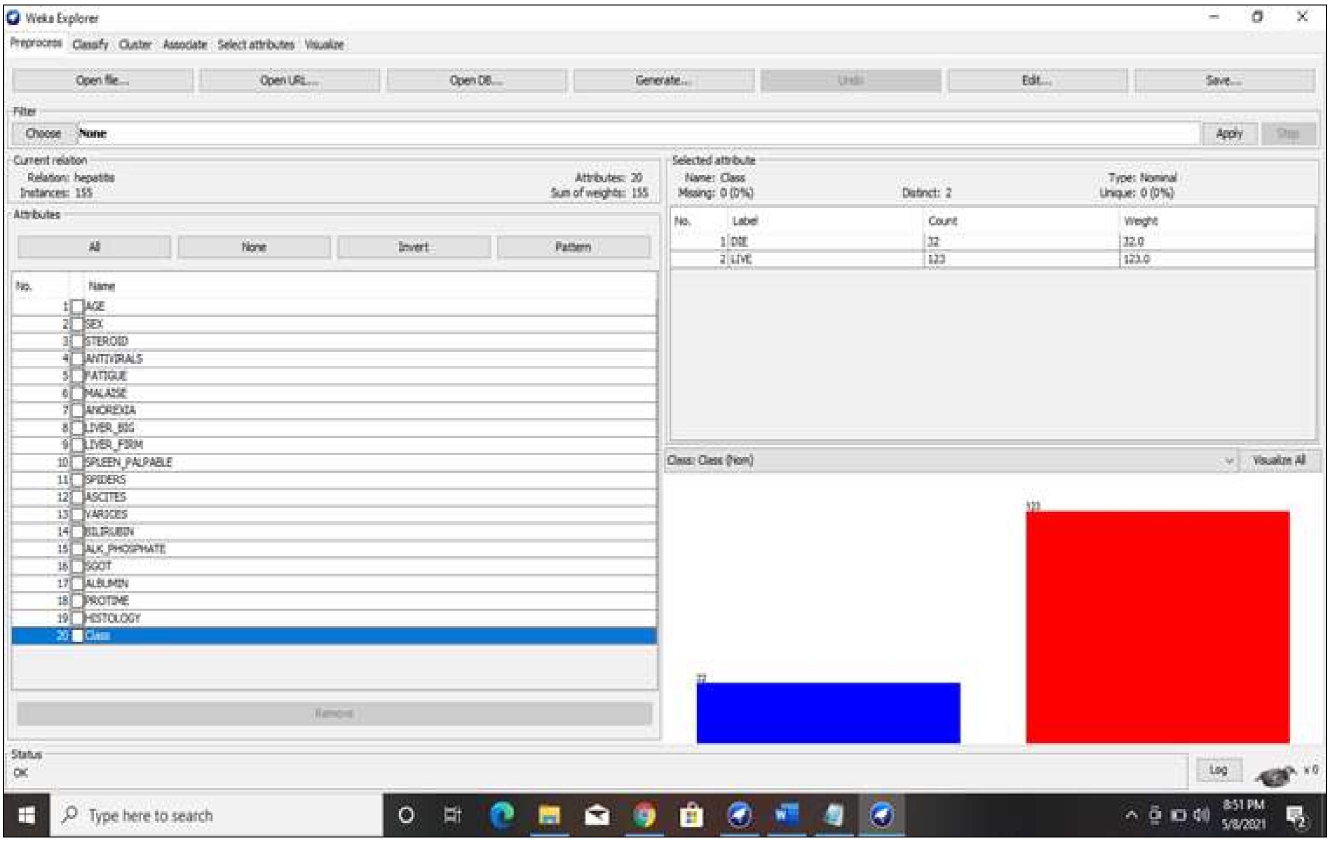

Fig 1: Summary of the Liver Dataset

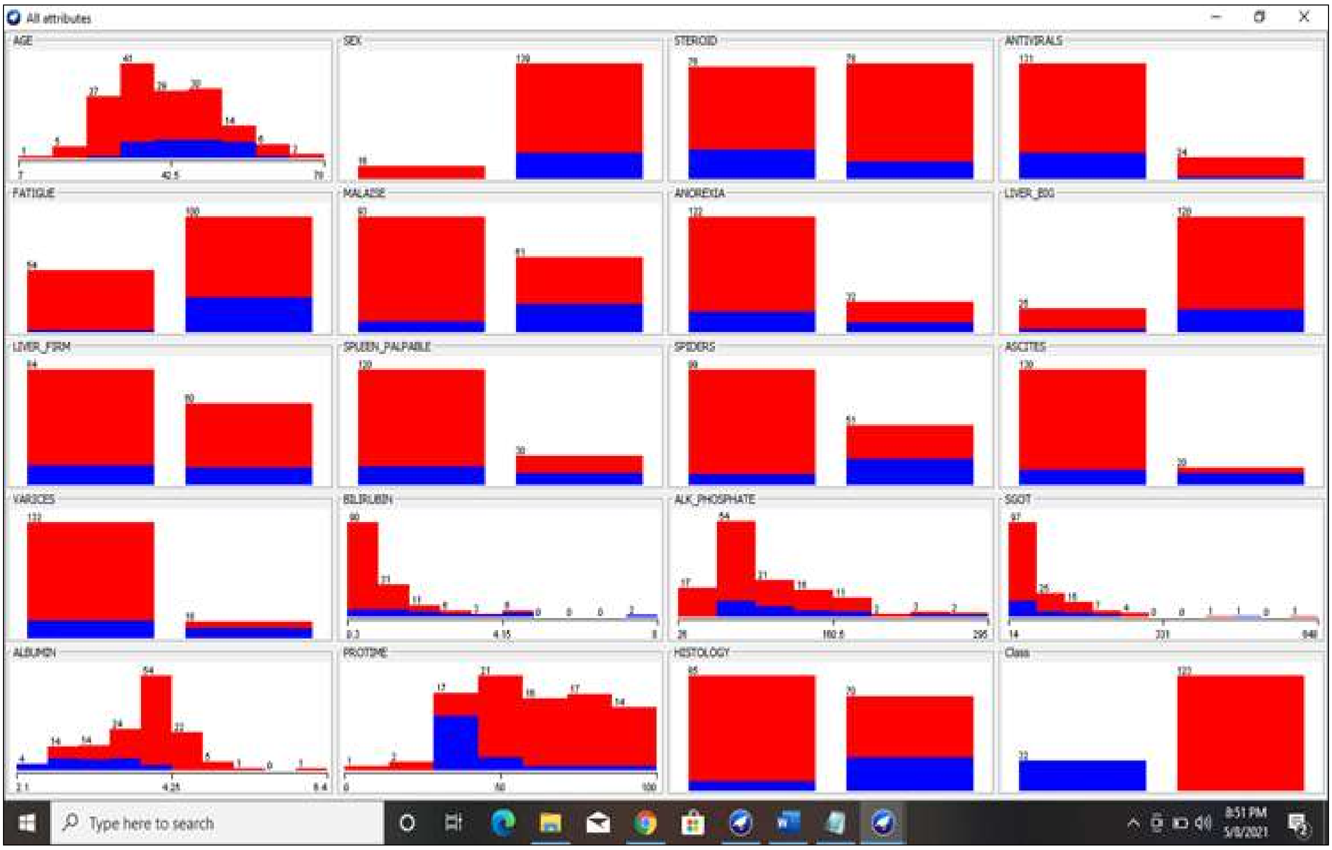

Fig 2: Statistical Summary of Dataset

The results of two classifiers are compared the on basis of correctly classified instances with feature selection techniques and without using feature selection techniques shown in table-2 and same shown in the figure-1.

Table 2: Performance of classifiers

Algorithm	Accuracy	Precision	Recall	F-Measure
Decision Tree with all features	89.87	89.5	89.9	89.5
Decision Tree with reduced features	92.16	92	92.2	92.9
Naïve Bayes with all features	90.84	90.4	90.8	90
Naïve Bayes with reduced features	94.72	95.4	95.7	95.5

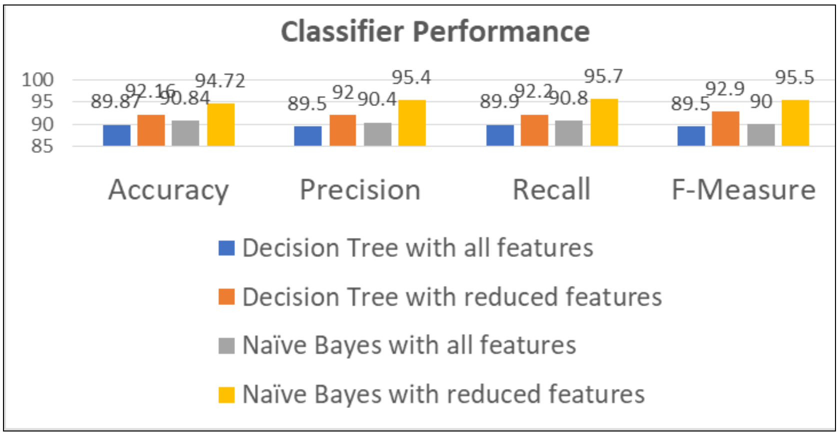

Fig 3: Performance of Classification with and without feature selection

From the figure-3, we notice the exhibition of choice tree without include determination, the exactness has 89.87%, while with highlight choice dependent on precision has accomplished 92.16%. Thus, there is improvement in the exactness with include choice. The exactness rate is expanded 2.29% with highlight determination.

We notice the exhibition of guileless bayes calculation without highlight determination, the exactness has 90.84%, though with include choice dependent on precision has accomplished 94.72%. Be that as it may, there is an improvement in the precision with include determination. The exactness rate is expanded 3.88% with include determination. In this way, in both datasets, there is an improvement with include determination.

5. Conclusion

In the proposed work, two classifiers were carried out on liver patient illness dataset to anticipate liver sicknesses. The consequences of the proposed work were thought about utilizing highlight choice and without utilizing highlight choice procedures after the execution of choice tree and innocent bayes classifiers in wording and exactness, accuracy and review. The best outcome was accomplished utilizing innocent bayes classifier with highlight choice strategies on liver patient infections dataset.

6. References

1. Abdar M, Yen NY, Hung JCS. Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J Med Biol Eng 2017;38(6):953-965.
2. Alfisahrin SNN, Mantoro T. Data mining techniques for optimization of liver disease classification. In: 2013 International conference on advanced computer science applications and technologies. IEEE 2013, 379-384
3. Arbain N, Balakrishnan BYP. A comparison of data mining algorithms for liver disease prediction on imbalanced data, International Journal of Data Science and Analytics 2019, 1.
4. Ravi Kumar G, Nagamani K, Anjan Babu G. 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, Springer Nature Singapore Pte Ltd 2020;37:173-180.
5. Ravi Kumar G, Venkata Sheshanna Kongara, Dr Ramachandra GA. An Efficient Ensemble Based Classification Techniques for Medical Diagnosis, International Journal of Latest Technology in Engineering, Management and Applied Sciences 2013;2(8):5-9. ISSN-2278-2540.
6. Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering, IEEE Trans. Knowl. Data Eng 2005;17(4):491-502.
7. Liu H, Sun J, Liu L, Zhang H. Feature selection with dynamic mutual information, Pattern Recognition 2009;42(7):1330-1339.
8. Witten H, Frank E. Data mining – Practical Machine Learning tools and Techniques (2nd Ed.), San Francisco, CA: Morgan Kaufmann Publisher, An imprint of Elsevier 2005.
9. Guyon, Elisseeff A. An introduction to variable and feature selection, J Mach.
10. Kira K, Rendell LA. The feature selection problem: Traditional methods and a new algorithm. In Proceedings of the AAAI-92, AAAI Press 1992, 129-134.
11. Nahar N, Ara F. Liver disease prediction by using different decision tree techniques, International Journal of Data Mining & Knowledge Management Process 2018;8(2):01-09.
12. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 2003;53:23-69.
13. UCI machine learning repository. http://archive.ics.uci.edu/ml/.