The Impact of Feature Selection on Learning Accuracy in Liver Disease and Disorder Prediction

K Jyoshna

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *The effectiveness of AI models heavily relies on the selection of key features within the dataset. Feature selection is pivotal in model optimization, aiming to identify a concise set of features to build robust models with minimal redundancy. Utilizing advanced algorithms for feature selection enhances the predictive speed of the models, mitigating the impact of redundant and noisy data that can hinder data understanding and model efficacy. Consequently, practitioners seek to extract meaningful insights from vast datasets using feature selection techniques. Feature selection involves identifying the most relevant features while eliminating redundant and irrelevant ones. In this study, we conducted a comparison between SVM-RFE (Recursive Feature Elimination) based feature selection method using a prominent dataset (Liver disease and hepatitis dataset). Two classification algorithms, decision tree, and Naive Bayes, were employed to evaluate the performance of the algorithms. Results indicated that the decision tree classifier achieved higher accuracy rates on the dataset following the application of feature selection methods. The analysis highlights the efficacy of feature selection techniques in enhancing the performance of learning algorithms.*

I. INTRODUCTION

The liver stands as a vital organ within the human body, responsible for crucial functions such as metabolizing substances, synthesizing proteins, and eliminating waste products. Its significance is underscored by the fact that survival is limited to a few days if liver function ceases. This glandular organ, weighing approximately 3 lbs (1.36 kg), exhibits a reddish-brown hue and is segmented into four lobes of varying sizes and shapes. Positioned on the right side of the abdominal cavity beneath the diaphragm, blood is supplied to the liver through two major vessels: the hepatic artery and the portal vein. These vessels branch out within the liver into increasingly smaller vessels, with each capillary forming a lobule. Liver tissue comprises numerous lobules, each consisting of hepatic cells—the primary metabolic units of the liver.

The detrimental effects of excessive alcohol consumption on the liver remain a significant concern, leading to acute or chronic liver conditions and potential damage to other organs. Liver disease, encompassing various disorders affecting the liver, often manifests with jaundice due to elevated bilirubin levels in the bloodstream. Bilirubin, a byproduct of red blood cell breakdown, is normally eliminated by the liver through bile secretion.

Common liver disorders include fatty liver, hepatitis (often caused by viral infections), and cirrhosis—a severe condition characterized by extensive liver cell damage and scarring, resulting in liver shrinkage and hardening. Despite the liver's regenerative capacity, cirrhosis typically involves progressive cell loss outweighing regeneration efforts. Liver cancer risk escalates in individuals with cirrhosis or certain viral hepatitis infections, with the liver frequently serving as a site for secondary metastatic tumors originating from other organs. [1] [2][3][12].

II. FEATURE DETERMINATION

The issue of feature selection stands out as a primary concern in data representation. The goal of feature selection is to identify the most relevant features while minimizing their number to enhance accuracy and reduce data collection costs. Recently, with the emergence of high-dimensional datasets containing a limited number of samples, classification models have grappled with overfitting issues. Consequently, there is a growing need for feature selection techniques to eliminate redundancy and irrelevant features.

Effective prediction relies heavily on feature selection. Data mining algorithms employ feature selection methods to identify the best features from the dataset, which are then loaded into memory for preprocessing. Feature selection is a process whereby only a subset of the most relevant features is chosen, aiding in predicting the outcome and reducing dimensionality during

preprocessing. This method serves as a form of dimensionality reduction used for preprocessing. Notably, feature selection differs from dimensionality reduction in that it reduces features without altering the original dataset. As feature selection operates with fewer parameters, it helps reduce complexity.

Various feature selection algorithms are applied in data representation, including filter methods, wrapper techniques, and embedded procedures. Filter methods select features based on their scores in various statistical measures, while wrapper techniques employ a heuristic approach by evaluating all possible feature combinations. Embedded procedures, on the other hand, combine the advantages of both models, enhancing feature selection efficiency. [4][5][8][9][7][10][11]

2.1 Support Vector Machine-Recursive Feature Elimination (SVM-RFE)

The very much considered SVM-RFE calculation [6] is a covering highlight choice technique which creates the positioning of highlights utilizing in reverse element end. It was initially proposed to perform quality determination for disease order [13]. Its fundamental thought is to take out repetitive qualities and yields better and more smaller quality subsets. The highlights are wiped out as indicated by a basis identified with their help to the separation work, and the SVM [15] is re-prepared at each progression. SVM-RFE is a weight-based strategy; at each progression, the coefficients of the weight vector of a direct SVM are utilized as the element positioning model [16].

The SVM-RFE calculation [6] can be broken into four stages:

1. Train a SVM on the preparation set;

2. Request highlights utilizing the loads of the subsequent classifier;

3. Dispose of highlights with the littlest weight;

4. Rehash the interaction with the preparation set limited to the leftover highlights

III. METHODOLOGY

This section gives the concise thought of chosen administered models of Decision Tree and Naive Bayes.

3.1 Machine Learning (ML) Techniques

ML is a piece of automated thinking that gets data from getting ready data reliant upon grounded real factors. ML is portrayed as an assessment that licenses PCs to learn data without being changed [9]. There are a couple of ML systems embraced to expect the attacks in the Test datasets which was used to set up the structure. These computations were used to arrange the attacks in other to find a viable procedure in expecting and organizing attacks. ML procedures are requested into three general classes, for instance, managed learning and independent learning [11]. Coordinated estimations learns for anticipating the article class from pre-named (portrayed) objects. Regardless, the independent estimation tracks down the trademark social event of things given as unlabeled data. In this work, the premium is with the going with managed learning estimations like Decision Tree and Naive Bayes procedures are surveyed.

3.1.1 Decision Tree

A choice tree is addressed as a tree. It addresses a great deal of the choice and these decisions are used to create rules for the classification of data plans. The major positive conditions of choice tree are that they are not difficult to understand and interpret. A center point of a choice tree identifies a quality by which the model is to be divided [9]. Every center point has a couple of edges, which are set apart by the conceivable assessment of the quality in the parent center point. An edge interfaces either two center points of a tree or a center point with a leaf. Leaf center points are named with class marks for classification of the case.

3.1.2 Naive Bayes

The Naive Bayes Classifier is a gathering strategy subject to the Bayes theory. It essentially improves learning by expecting that highlights are free given class. Despite the way that self-rule is generally a vulnerable assumption, before long guiltless

Bayes consistently battles well with more refined classifier [11]. Gullible Bayes Classifier is known to be better than some other portrayal procedures. Since first, the key nature of Naive Bayes is a very strong (gullible) speculation of self-sufficiency from each condition or event. Second, its model is clear and easy to make. Third, the model can be executed for enormous instructive lists.

Bayesian classifiers give out the most likely class to a given model depicted by its component vector. Learning such classifiers can be amazingly revamped by expecting that features are self-governing given class, that is, $P(X|C) = \prod_{i=1}^{n} P(Xi|C)$, where $X = (X_1, X_2, \ldots\ldots, X_n)$ is a component vector and C is a class.

IV. EXPERIMENTAL RESULTS

This section describes the experimental results obtained by applying the proposed algorithm to a two data sets namely Liver Disorder and Hepatitis are taken from the UCI machine learning repository [14] as shown in Table-1. In order to validate the prediction results of the comparison of the two classification (Decision tree and Nave Byes with SVM-RFE) techniques and the 10-fold crossover validation is used. The k-fold crossover validation is usually used to reduce the error resulted from random sampling in the comparison of the accuracies of a number of prediction models. We use 70% of records as the training data and the other 30% as the testing data.

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
2	Hepatitis	20	155	Die:32 Live:123

We have utilized the Weka tool compartment to try different things with these two information mining calculations [13]. The Weka is an outfit of apparatuses for information order, relapse, bunching, affiliation rules, and representation. WEKA was used as an information mining instrument to assess the exhibition and viability of the choice tree and guileless bayes and Proposed SVM-RFE method. This is on the grounds that the WEKA program offers a distinct structure for experimenters and designers to construct and assess their models. The grouping exactness is anticipated as far as accuracy and review. The assessment boundaries are the exactness, accuracy and review and in general precision of two UCI informational indexes are introduced in Table-2 and same are appeared in figure-1 with highlight and without include determination.

Table-2
Performance of Classification Algorithms

Dataset	Algorithm	Accuracy	Precision	Recall
Liver disorder	Naive Bayes	76.84	76.3	76.8
	Naive Bayes with SVM-RFE	75.78	75.2	75.8
	Decision Tree	82.81	83.1	82.8
	Decision Tree with SVM-RFE	85.72	85.4	85.7
Hepatitis	Naive Bayes	82.49	82.2	82.4
	Naive Bayes with SVM-RFE	84.6	84.2	84.6
	Decision Tree	87.76	87.32	87.54
	Decision Tree with SVM-RFE	89.56	89.2	89.3

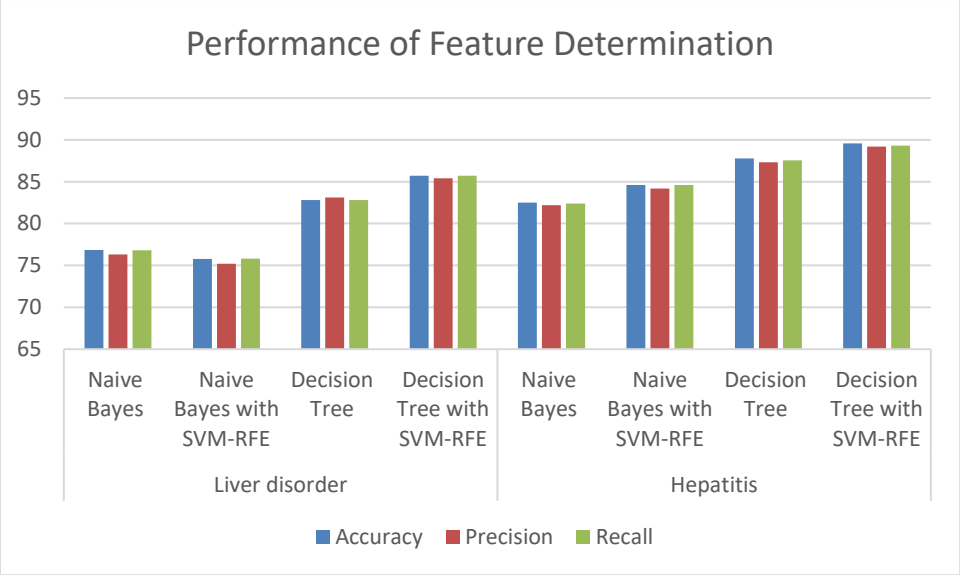


Figure-1: Performance of Classification with and without feature selection

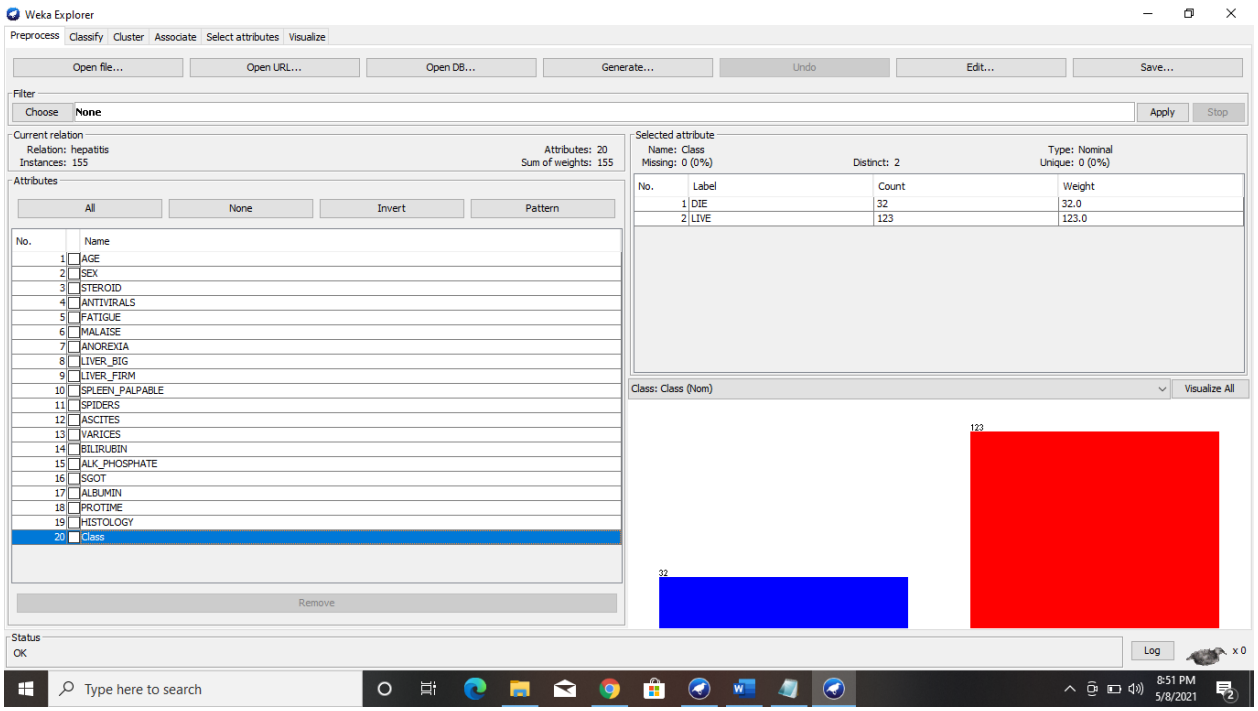
We see in the figure-1, the presentation of the two order calculations with SVM-RFE based component determination and without highlight choice on the two datasets. The accuracy of decision tree calculation on Hepatitis dataset utilizing decision tree has accomplished 87.76% while decision tree with SVM-RFE 89.56%. The accuracy of naive bayes calculation on Hepatitis dataset without SVM-RFE has 82.49%, while utilizing naïve bayes with SVM-RFE has 84.6%.

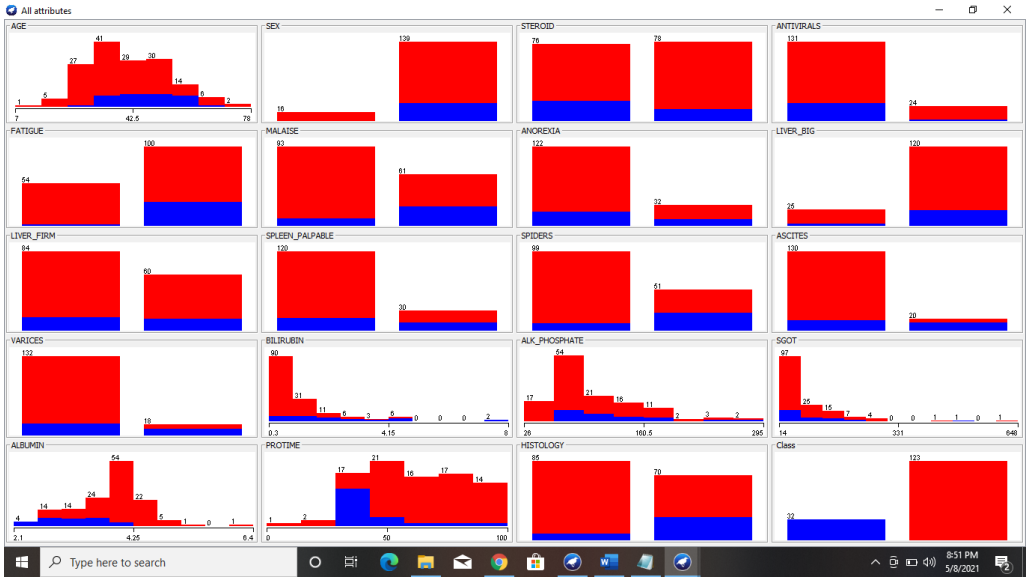
The Accuracy of decision tree calculation on Liver problem dataset utilizing decision tree has accomplished 82.81% while decision tree with SVM-RFE 85.72%. The accuracy of naive bayes calculation on Liver issue dataset without SVM-RFE has 76.84%, while utilizing credulous bayes with SVM-RFE has 75.78%.

So, in these two datasets, decision tree and naive bayes calculations with SVM-RFE highlight determination has hot most noteworthy correct nesses when contrasted with just choice tree and innocent bayes order.

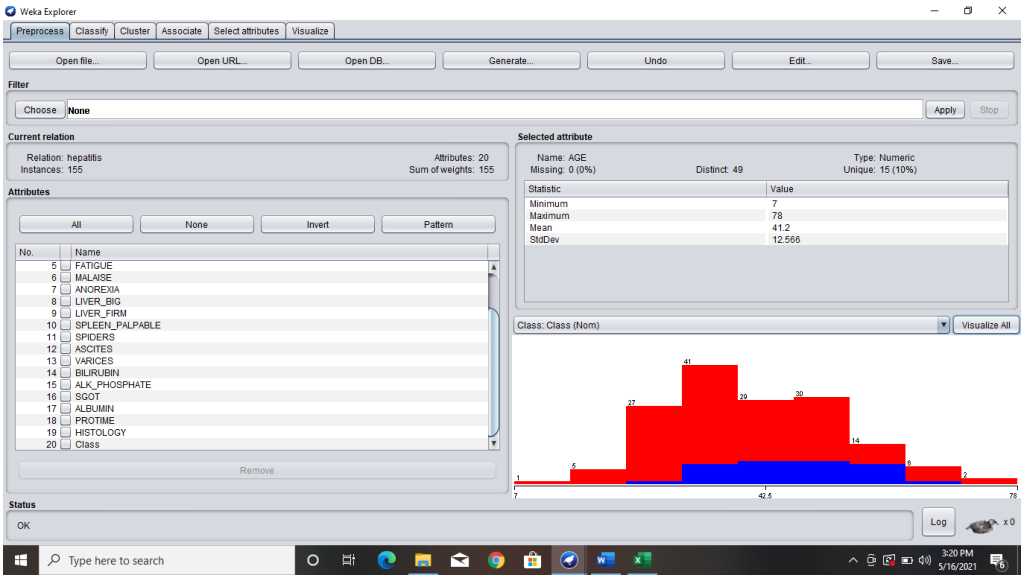
4.1 Screen Shots

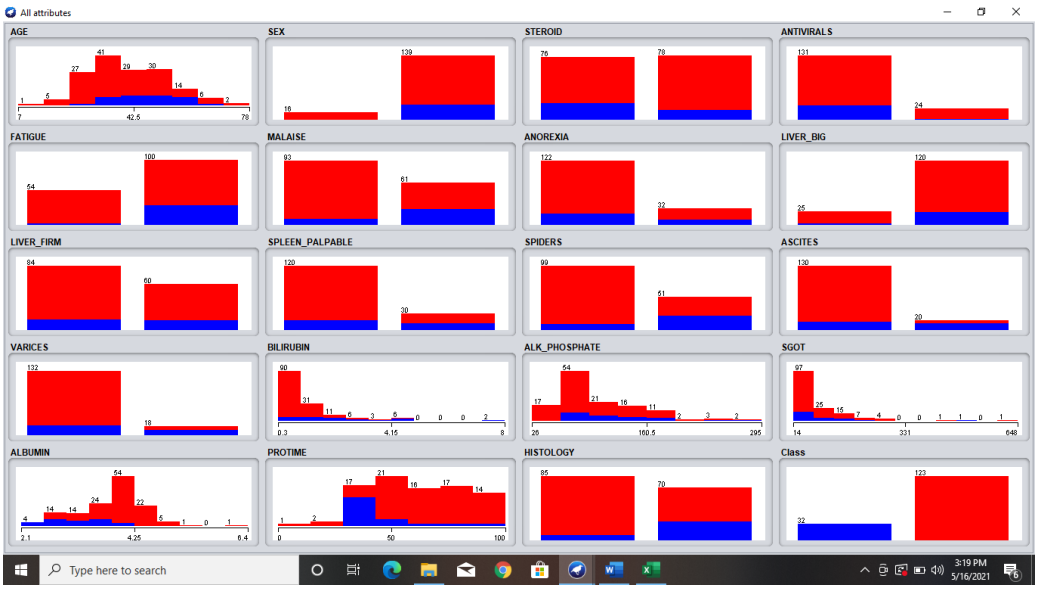
4.1.1 Data Visualization of Liver disorder Data





4.1.2 Data Visualization of Hepatitis dataset





V. CONCLUSION

Feature selection plays a critical role in data mining studies, as many AI algorithms struggle to cope with an abundance of irrelevant features. Therefore, feature selection methods have become essential in numerous analyses. In this study, we conducted a comparative assessment using SVM-RFE-based feature selection algorithms to predict the risks of liver disease and hepatitis infection. We proposed an SVM-RFE-based feature selection approach for disease classification, aiming to enhance classifier accuracy by integrating SVM-RFE with decision tree and naive Bayes algorithms. Our experimental findings suggest that reusing features eliminated during the SVM-RFE process can improve the SVM-RFE classifier. Moving forward, we plan to evaluate our approach on datasets with a large number of features.

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] 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.

[3] Chuang CL (2011) Case-based reasoning support for liver disease diagnosis. Artif Intell Med 53(1):15–23

[4] 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

[5] S.Rahamat Basha and Surya Bhupal Rao G.Ravi Kumar, "A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, PP: 324-332, 2020, Institute of Mechanics of Continua and Mathematical Sciences

[6] Guyon, Weston, Barnhill, and Vapnik, "Gene selection for cancer classification using support vector machines," MACHLEARN: Machine Learning, vol. 46, (2002).

[7] 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.

[8] 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.

[9] H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations", San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., (2000)

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach.
Learn. Res., vol. 3, (2003) March, pp. 1157–1182

[11] J. Han and M. Kamber," Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[12] 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.

[13] Pember A. Mundra and J. C. Rajapakse, "SVM-RFE with relevancy and redundancy criteria for gene selection," in PRIB, J. C. Rajapakse, B. Schmidt, and L. G. Volkert, Eds., vol. 4774, Springer, (2007), pp. 242–252.

[14] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/.

[15] V. N. Vapnik, "The nature of statistical learning theory", New York, NY, USA: Springer-Verlag New York, Inc., (1995).

[16] Y. Tang, Y.-Q. Zhang and Z. Huang, "Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis", IEEE/ACM Trans. Comput. Biology Bioinform, vol. 4, no. 3, (2007), pp. 365–381.