Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 13 NO: 1 PAGE: 153-160 YEAR: 2024
DOI:10.17798/bitlisfen.1361016

Classification of Liver Disorders Diagnosis using Naïve Bayes Method

Özlem BEZEK GÜRE*

Batman University, Department of Medical Documentation, Secretariat Program Health Services Vocational School, Batman, Turkey
(ORCID: 0000-0002-5272-4639)

Keywords: Data mining, machine learning, Naive Bayes, liver disorder.

Abstract
Liver diseases pose a significant health challenge, necessitating robust predictive tools for early diagnosis. This study aims to determine the predictive performance of Naive Bayes classifier, one of the data mining algorithms, in the classification of liver patients. The study applied 2, 5, 10 and 20-fold cross-validation method. Trying to determine the effect of the cross-validation (CV) method used on the classification performance, this study used the "BUPA" dataset in the UCI Machine Learning Repository database for this purpose. The dataset consists of 6 variables and 345 examples. Orange program was used for data analysis. As a result of the analysis, the accuracy for the Naive Bayes method was determined to be 62.9%, 63.5%, 63.8%, and 64.3%, respectively. The AUC values were 0.68, 0.66, 0.66, and 0.67, respectively; the F1 scores were 0.56, 0.57, 0.58, and 0.58, respectively. On the other hand, the precision values were 0.60, 0.60, 0.60, and 0.62, respectively, while the recall values were determined to be 0.52, 0.53, 0.55, and 0.54. Additionally, the MCC values were determined to be 0.24, 0.26, 0.26, and 0.27, respectively. The analysis results indicate that the 20-fold CV method demonstrates marginally superior performance. The use of the free and easy-to-use program is recommended.

1. Introduction

Data mining can be characterized as a suite of techniques aimed at unveiling hidden patterns and trends within datasets. It encompasses the processes of model construction, data selection, and data discovery based on hitherto unknown patterns [1,2]. Briefly, this method serves to extract understandable and actionable information from datasets [3-5]. Data mining amalgamates methods from statistics, machine learning, and pattern recognition [6]. The accelerated advancements in computer technologies and the consequent proliferation of large datasets have fueled increased interest in data mining methodologies. These techniques have found extensive application across diverse fields such as engineering, economics, education, and healthcare. Due to their inherently complex and voluminous nature, traditional statistical methods often fall short in analyzing the extensive data generated within the healthcare sector. Data mining techniques are leveraged to transform these extensive data repositories into actionable insights for decision-makers. This, in turn, enhances healthcare operations and facilitates medical research [7]. Utilizing these methodologies enables capabilities such as predicting patient responses to medication dosages, identifying healthcare insurance fraud, and diagnosing or projecting specific diseases [2,7-8].

Being the body's second-largest organ, the liver plays a crucial role in maintaining human well-being.In addition to its role in nutrient storage, it performs a wide range of functions related to digestion, metabolism, and immunity, such as the breakdown of red blood cells, protein production, and the elimination of toxins from the body [9-11]. In recent years, an increase in liver-related diseases has been observed. Stress, inhalation of harmful gases, poor nutrition, excessive alcohol consumption, unnecessary medication intake, and viruses are

*Corresponding author: ozlem.bezekgure@batman.edu.tr
Received: 16.09.2023, Accepted: 17.01.2024

among the factors contributing to liver diseases [11-13]. Liver diseases are considered one of the most significant health issues globally [14]. If left untreated, liver diseases can lead to serious health problems. Therefore, early and accurate diagnosis of diseases related to this vitally important organ is crucial [15]. In recent years, an excessive increase in liver disorders has been observed in many countries. Consequently, liver diseases have begun to rank among the leading causes of death in these countries [8]. This study employed the BUPA Liver Disorder dataset to develop classification models using the Naïve Bayes method, aiming to predict liver diagnosis.

Given the vital functions of the liver and the rising prevalence of liver diseases globally, there is a clear need for accurate diagnosis of these conditions. Data mining methods can be leveraged to analyze healthcare data and gain insights into liver disease prediction and diagnosis.

Data mining methods are frequently observed to be used for predicting diseases. The present study employs the Naive Bayes method from data mining techniques for the prediction of liver diseases. A review of the literature reveals numerous studies that have utilized this method for the same purpose [12,16-20]. Unlike other studies, this research also evaluates the classification performance of the Naive Bayes method when applied with different Cross-Validation techniques.

Other studies have also applied data mining techniques, including Naive Bayes, to liver disease prediction. However, this study takes the additional step of assessing the classification performance with different cross-validation methods. By evaluating multiple techniques, this allows for a more robust analysis of the predictive capabilities of data mining for liver diseases.

Numerous studies have applied various data mining methods to data related to liver diseases. Ram et al. [19] used Naive Bayes (NB), SMO, and Bayes Net methods in their study to predict liver diseases, noting that the SMO method displayed superior performance. Vijayarani and Dhayanand [21] implemented NB and Support Vector Machine (SVM) methods in their research, indicating that the SVM method performed better. Similarly, Kamruzzaman, Mahbub, and Hakim [22] used, in addition to these two methods, K-Nearest Neighbors (KNN) and Decision Tree (DT) methods, concluding that SVM showed better performance. Additionally, Abdar [16] used Linear Regression, KNN, C4.5, C5.0, CHAID, Neural Net, and Random Forest (RF) methods and found that the C4.5 algorithm demonstrated better classification performance. Sug [17] also utilized DT, C4.5, and CART algorithms.

The literature contains many examples of data mining methods being applied to predict liver diseases, with studies comparing the performance of different techniques. This highlights the utility of data mining for gaining insight into liver conditions, though it also suggests Naive Bayes may not always be the optimal method.

On the other hand, Nahar et al. [23] used decision tree methods like J48, LMT, Random Tree, RF, REPTree, Decision Stump, and Hoeffding Tree for predicting liver diseases. A study by Kuppan and Manoharan [11] employed J48 and NB methods. Baitharu and Pani [10] used J48 and NB methods along with DT, Multilayer Perceptron Neural Network (MLP), ZeroR, KNN, and VFI, determining that the MLP method outperformed others. Similarly, Al-Aidaroos, Bakar, and Othman [24] used NB, Logistic Regression, Kstar, DT, Neural Network, and Zero R methods and found that the NB method performed better compared to others. Bhardwaj, Mehta, and Ramani [25] used DT, RF, NB, KNN, SVM, Artificial Neural Networks (ANN), and Extreme Gradient Boost (XGBoost) methods for the same purpose. A study by Ramana, Babu, and Venkateswarlu [12] employed NB, C4.5, Back Propagation Neural Network, and SVM methods.

While some studies have found Naive Bayes to perform well for liver disease prediction, the literature also contains examples of other methods like decision trees, neural networks, and support vector machines outperforming Naive Bayes. More research is still needed to determine the optimal data mining techniques for this application.

2. Material and Method

The current study employs the "BUPA" data set. obtained from the UCI Machine Learning Repository. The data file was obtained from https://archive.ics.uci.edu/dataset/60/liver+disorders [26]. The dataset consists of 345 rows and 7 columns, with each row containing information about male individuals. The first 5 columns contain blood test results that can be used in the diagnosis of liver diseases related to alcohol. The 6th column includes the number of alcoholic drinks consumed daily as reported by the individuals. The last column contains a variable intended for use in dividing the data into training and testing sets [27]. In the study, the drinks (number of half-liter equivalents of alcoholic beverages consumed per day) variable was treated as the dependent variable; selector, mcv (mean corpuscular volume), alkphos (alkaline phosphatase),

sgpt (alanine aminotransferase), sgot (aspartate aminotransferase), and gammagt (gamma-glutamyl transpeptidase) were analyzed as independent variables. Subsequently, the dependent variable, drinks, was categorized into two groups: those with a value of 3 and below, and those above 3. Following this, the analyses using the Naïve Bayes method were conducted through Orange, a free Python-based software, to conduct the data analyses. [28]. Descriptive statistics of the independent variables are given in Table 1.

Table 1. Descriptive statistics of predictive variables

Variables	Mean	Standard deviation	Min	Max
Mcv	90,16	4,45	65	103
Alkphos	69,87	18,35	23	138
Sgpt	30,41	19,51	4	155
Sgot	24,64	10,06	5	82
Gammagt	38,28	39,25	5	297
Drinks	3,45	3,34	0	20

Naïve Bayes Method

The Naive Bayes method is a highly effective and robust probability-based classifier that employs Bayes' theorem in classification, under strong independence assumptions [29-30]. The method is an algorithm that is easy and quick to structure and interpret, without the need for complex iterative parameter estimations [31]. Because of these characteristics, the NB algorithm is commonly employed in extensive data analysis and various other domains [32]. This method analyzes the relationship between the independent and dependent variables by obtaining a conditional probability for each relationship [33]. NB computes probabilities by analyzing the frequency and combinations of values in a given dataset [34], and it is successful in high-dimensional datasets [35]. Compared to other data mining algorithms, NB is more resilient to overfitting and noisy data because it estimates fewer parameters [36].

The Naive Bayes classifier, one of the supervised learning algorithms, assumes that each attribute value is independent of the values of other attributes within the class. This assumption is termed class-conditional independence [37-38]. This assumption is referred to as "naive" because it is rarely valid in real-world applications. The algorithm tends to learn rapidly in various controlled classification problems [34]. Due to the independence assumption of the variables in this method, it is necessary to estimate each variable rather than the covariance matrix [39]. When this assumption holds, the learning process of Bayes classifiers becomes simpler, thereby achieving optimal assignment using the vector of observable factors. In addition to the independence assumption, it is assumed that the factors affecting the outcome of interest are not hidden [40]. Moreover, the method necessitates minimal training data to determine the parameters essential for classification [39,41].

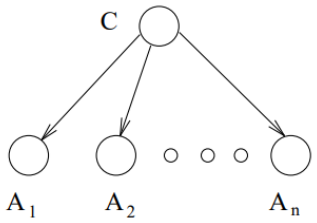

Figure 1. Structure of the Naive Bayes Classifier [29].

Different Naive Bayes (NB) classifiers vary chiefly due to the assumptions they hold about feature distribution. For discrete features, typically Multinomial or Bernoulli distributions are applied, whereas the Gaussian distribution is chosen for continuous features [42].

Bayes' Theorem

Developed by Thomas Bayes in the 1800s, this theorem pertains to probability and decision theory. Let Ω be a complete set and, $C_1, C_2, C_3, \dots \dots C_n \in \Omega$, C_i, represent ist category, where $P(C_i) > 0, i = 1,2, \dots \dots n$. Each category is distinct from the other, and $\bigcup_{i=1}^{n} C_i = \Omega$. For any X if $P(X) > 0$, then the Bayes equation is given in Equation 1.

$$P(C_i \backslash X) = \frac{P(X \backslash C_i)P(C_i)}{\sum_{i=1}^{n} P(X \backslash C_i)P(C_i)} \quad (1)$$

[39].

Naive Bayes Classifier

The Naive Bayes classifier uses the concept of maximum likelihood estimation to classify a sample according to the highest probable category.

$$P(C_i \backslash X) = Max\{P(C_1 \backslash X), P(C_2 \backslash X), P(C_3 \backslash X), \dots \dots \dots \dots P(C_n \backslash X)\} \quad (2)$$

Let $X = (A_1, A_2, A_3, \dots \dots \dots A_k)$ be a feature vector, where A_j represents the jth feature of x_j.

The Naive Bayes classifier assumes that attributes function independently from one another. Therefore, the conditional probability, $P(X\backslash Ci)$ can be expressed as:

$$P(X\backslash C_i) = \prod_{j=1}^{k} P(A_j = x_j\backslash C_i) \quad (3)$$

When the third equation is substituted into the Bayes formula given in the first equation, we obtain:

$$P(C_i\backslash X) = \frac{\prod_{j=1}^{k} P(A_j=x_j\backslash C_i)P(C_i)}{P(X)} \quad (4)$$

When $\frac{1}{P(X)=\alpha(>0)}$, we have:

$$P(C_i\backslash X) = \alpha \prod_{j=1}^{k} P(A_j = x_j)C_i)P(C_i) \quad (5)$$

Let N(D) represent the total number of samples in the sample set $N(C_i)$, represent the number of samples in C_i, and $N(C = C_i, A_j = x_j)$, A_j C_i'deki x_j A_j represent the number of samples in C_i where the feature $A_j = x_j$

We can express $P(Ci)$ and $P(Aj=xj|Ci)$ as:

$$P(C_i) = \frac{N(C_i)}{N(D)} \quad (6)$$

$$P(A_j = x_j\backslash C = C_i) = \frac{N(C=C_i,A_j=x_j)}{N(C_i)} \quad (7)$$

Inserting equations 6 and 7 into equation 5 yields the subsequent result:

$$P(C_i\backslash X) = \alpha \prod_{j=1}^{k} \frac{N(C=C_i,A_j=x_j}{N(C_i)} . \frac{N(C_i)}{N(D)} \quad (8)$$

[39].

Performance Criteria

The confusion matrix indicates the degree to which the classifier used recognizes patterns in different classes.

Table 2. Confusion matrix

		Estimated		
		No	Yes	Total
Real	No	TN	FP	TN+FP
	Yes	FN	TP	FN+TP
	Total	TN+FN	FP+TP	TN+FN+FP+TP

This research intends to assess the effectiveness of the Naive Bayes classification algorithm concerning different Cross-Validation CV methodologies. CV serves as an instrumental technique for assessing the validity of the predictive model under consideration. In the instance of k-fold CV, the dataset is partitioned into k subsets. One of these subsets is employed as the test data, while the remaining subsets function as the training data. The mean error rate from these k subsets is calculated to estimate the overall error rate of the classification model [43]. In this research, performance metrics such as Area Under the Curve (AUC), accuracy, precision, recall, F1 score, and Matthews's correlation coefficient (MCC) have been utilized.

AUC: AUC, also known as the Area Under the Receiver Operating Characteristic (ROC) Curve, serves as a quantitative indicator of the model's accuracy. The magnitude of this area is a demonstrative measure of the classification model's performance efficacy [44].

MCC: Matthews correlation coefficient is a measure that shows the relationship between the predicted class and the actual class, ranging between [-1,1]. If the coefficient is +1, it indicates that the predictions made by the classifier are correct, whereas if it is -1, it indicates that the predictions are incorrect [45].

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad (9)$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$F1\ score = 2x\ \frac{(Precision\ x\ Recall)}{(Precision+ Recall)} \quad (13)$$

3. Results and Discussion

Table 3. Confusion Matrix Obtained Using the Naive Bayes Method

CV method		No	Yes
2	No	136	54
	Yes	74	81
5	No	137	53
	Yes	73	82
10	No	135	55
	Yes	70	85
20	No	138	52
	Yes	71	84

Table 4. Performance Criteria

CV method	AUC	CA	F1	Precision	Recall	MCC
2	0.678	0.629	0.559	0.600	0.523	0.242
5	0.660	0.635	0.565	0.607	0.529	0.255
10	0.657	0.638	0.576	0.607	0.548	0.262
20	0.669	0.643	0.577	0.618	0.542	0.273

Table 4 indicates that the 20-fold cross-validation method exhibits a marginally better performance compared to other methods.

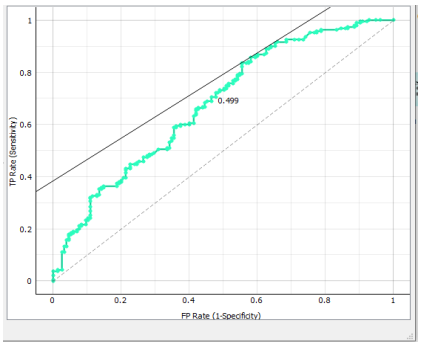

Figure 2. Area Under the ROC Curve (2-fold CV)

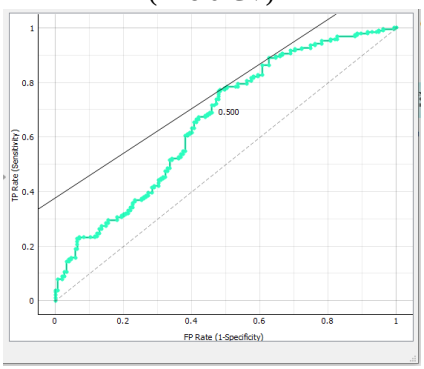

Figure 3. Area Under the ROC Curve (5-fold CV)

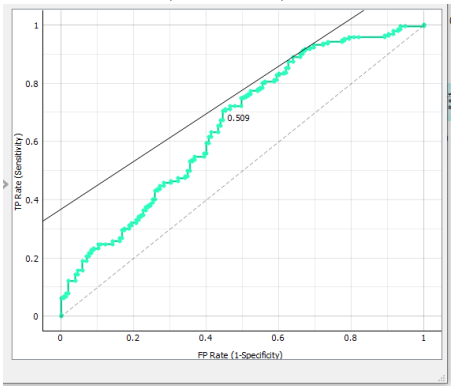

Figure 4. Area Under the ROC Curve (10-fold CV)

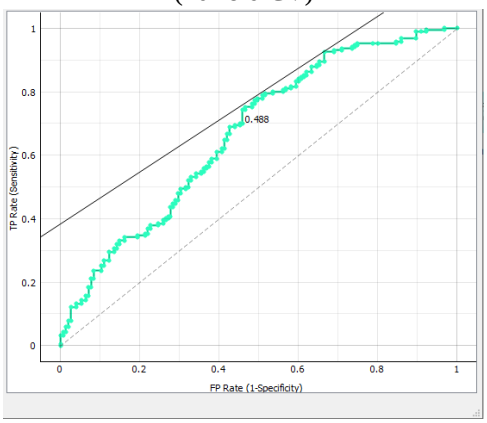

Figure 5. Area Under the ROC Curve (20-fold CV)

In the academic literature, there are numerous studies that utilize the BUPA dataset for the classification of liver diseases. It has been observed that the Naive Bayes method is frequently employed in these studies. Similar to our research, a study conducted by Sujana, Rao, and Reddy [46] used the Weka software and applied 2, 5, and 10-fold CV methods; according to their analysis, the accuracy rates were determined to be 60.58%, 60.89%, and 62.32%, respectively. Another study by Ramana, Babu, and Venkateswarlu [12] also used the Weka software and implemented the 10-fold CV method, determining the accuracy, precision, sensitivity, and specificity of the Naive Bayes method to be 51.59%, 45.17%, 71.03%, and 37.5%, respectively. Ruengdetkharn and Lohpetch [47] applied a 5-fold cross-validation method in their study and determined the accuracy to be 62.90%. Pradhan et al. [48] employed a two-fold CV method in their study and used performance indicators such as Type I error, Type II error, true negative rate, true positive rate, accuracy, and F1 score; these metrics were determined to be 0.51, 0.4, 0.49, 0.6, 0.53, and 0.52, respectively. Compared to similar studies, the analysis results obtained using the Orange software appear to yield slightly better outcomes than those from other programs.

In the existing literature, datasets focusing on liver diseases have been employed in various studies. Abdar [16]) utilized the Indian Liver Patient Dataset (ILPD) from the UCI database for this purpose. The study employed both RapidMiner and SPSS Modeler for analysis. According to the results, the accuracy and precision rates obtained through the RapidMiner program using the Naive Bayes method were 66.92% and 45.13%, respectively. In contrast, the SPSS Modeler yielded an accuracy of 74.26% and a precision of 40.24%. Similarly, Ram et al. [19]

conducted analyses using Python to predict liver diseases, employing methods such as Naive Bayes (NB), SMO, and Bayes Net. On the other hand, Alam, Rahman, and Rahman [49] used the Weka software in their study, applying the Bayes Net method. They reported an accuracy rate of 0.68 using ten-fold CV.

4. Conclusion and Suggestions

This research seeks to ascertain the predictive performance of the Naive Bayes classifier of the Naive Bayes classifier when categorizing liver disorder diagnosis through data mining methods. The study has implemented 2, 5, 10, and 20-fold CV methods. An attempt has been made to ascertain the impact of the employed CV method on classification performance. Analyses were conducted in the Orange program, which is a Python-based free software. Performance measures such as AUC, accuracy, precision, recall, F1 score and MCC were used. As a result of the analysis, the accuracy for the Naive Bayes method was determined to be 62.9%, 63.5%, 63.8%, and 64.3%, respectively. The AUC values were 0.68, 0.66, 0.66, and 0.67, respectively; the F1 scores were 0.56, 0.57, 0.58, and 0.58, respectively. On the other hand, the precision values were 0.60, 0.60, 0.60, and 0.62, respectively, while the recall values were determined to be 0.52, 0.53, 0.55, and 0.54. Furthermore, the MCC values were identified as 0.24, 0.26, 0.26, and 0.27, respectively. Based on the analysis results, it can be claimed that the 20-fold CV method exhibited better performance Based on these results, it is observed that the 20-fold CV method showed slightly better performance.

In the current study, the Orange program was utilized to examine the performance of cross-validation methods. A review of the literature indicates that programs such as Weka and SPSS Modeler are preferred for implementing machine learning methods, while the use of the Orange program is limited. It is recommended to use this free and user-friendly program since it appears based on the findings that the Orange software may offer advantageous outcomes. The dataset used in this study is small; therefore, larger datasets can be utilized to examine cross-validation methods. On the other hand, investigations can be conducted using different data mining methods.

Acknowledgment

If necessary, the people, institutions and organizations that helped in the study should be thanked for their help and support.

Statement of Research and Publication Ethics

The dataset used in the present study was sourced from the publicly accessible UCI database; therefore, the research did not necessitate ethical committee approval."

References

[1] M.Kayri, İ.Kayri and M.T. Gencoglu, "The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data", *IEEE 14th International Conference on Engineering of Modern Electric Systems (EMES),* pp.1-4, June 2017.

[2] H. C. Koh and G. Tan, "Data mining applications in healthcare", *Journal of Healthcare Information Management,* vol.19, no.2, pp.65-72, 2011.

[3] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works", *Expert systems with applications,* vol.41, no.4, pp.1432-1462, 2014.

[4] M., Kayri and, İ. Kayri, "The comparison of Gini and Twoing algorithms in terms of predictive ability and misclassification cost in data mining: an empirical study", *International Journal of Computer Trends and Technology (IJCTT),* vol. 27, no. 1, pp.21-30, 2015.

[5] Ö. B. Güre, M. Kayri and F.Erdoğan, "Analysis of Factors Effecting PISA 2015 Mathematics Literacy via Educational Data Mining", *Education & Science/Egitim ve Bilim,* vol.*45,* no.202, pp.393-415, 2020.

[6] M. Sharma, "Data mining: A literature survey", *International Journal of Emerging Research in Management & Technology*, vol.3, no.2, pp.1-4, 2014.

[7] R. H. Khokhar, R. Chen, B.C. Fung and S.M. Lui, "Quantifying the costs and benefits of privacy-preserving health data publishing", *Journal of biomedical informatics,* vol.50, pp.107-121, 2014.

[8] S. Bahramirad, A. Mustapha and M. Eshraghi," Classification of liver disease diagnosis: A comparative study", *IEEE 2013 Second International Conference on Informatics & Applications (ICIA),* pp.42-46, September 2013.

[9] P. Kumar and R.S. Thakur, "Liver disorder detection using variable-neighbor weighted fuzzy K nearest neighbor approach", *Multimedia Tools and Applications*, vol.80, pp.16515-16535, 2021.

[10] T. R. Baitharu and S.K. Pani, "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset", *Procedia Computer Science*, vol.85, pp.862-870, 2016.

[11] P. Kuppan and N. Manoharan, "A Tentative analysis of Liver Disorder using Data Mining Algorithms J48, Decision Table and Naive Bayes", *International Journal of Computing Algorithm*, vol.6, no.1, pp.2278-239, 2017.

[12] B. V. Ramana, M. S. P. Babu and N.B. Venkateswarlu, "A critical study of selected classification algorithms for liver disease diagnosis*", International Journal of Database Management Systems*, vol.3, no.2, pp.101-114, 2011.

[13] R. Kalaviselvi and G. Santhoshni, "A Comparative Study on Predicting the Probability of Liver Disease", *International Journal of Engineering Research & Technology (IJERT)*, vol.8, no. 10, pp.560-564, 2019.

[14] R. H. Lin, "An intelligent model for liver disease diagnosis", *Artificial Intelligence in Medicine*, vol.47 no.1, pp.53-62, 2009.

[15] S. N. N. Alfisahrin and T. Mantoro, "Data mining techniques for optimization of liver disease classification" *IEEE 2013 International Conference on Advanced Computer Science Applications and Technologies*, pp.379-384, *December 2013*.

[16] M. Abdar, "A survey and compare the performance of IBM SPSS modeler and rapid miner software for predicting liver disease by using various data mining algorithms", *Cumhuriyet University Faculty of Science Science Journal (CSJ)*, vol.36, no.3, pp.3230-3241, 2015.

[17] H. Sug, "Improving the prediction accuracy of liver disorder disease with oversampling", *Proc. of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics, Wisconsin United States, January, 25-27, 2012*.

[18] H. Subhani and S. Badugu, "A study of liver disease classification using data mining and machine learning algorithms", in Learning and Analytics in Intelligent Systems : Proc. of the the Advances in Decision Sciences, Image Processing, Security and Computer Vision: International Conference on Emerging Trends in Engineering (ICETE), Hyderabad, India, March 22–23, 2019, George A. Tsihrintzis, Maria Virvou, Lakhmi C. Jain, Eds. Berlin: Springer,2019. Vol. 2, pp. 630-640.

[19] M. K. Ram, C. Sujana, R. Srinivas and G. S. N. Murthy, "A fact-based liver disease prediction by enforcing machine learning algorithms". in *Advances in Intelligent Systems and Computing Proc. Of the Computational Vision and Bio-Inspired Computing: ICCVBIC*, Coimbatore, India, November 19-20, 2020. Janusz Kacprzyk Eds. Berlin: Springer, 2020. pp.567-586

[20] S.Wang, J. Ren and R. Bai, "A semi-supervised adaptive discriminative discretization method improving discrimination power of regularized Naive Bayes", *Expert Systems with Applications*, vol. 225, no. 120094, pp. 1-7, 2023.

[21] S. Vijayarani and S. Dhayanand, "Liver disease prediction using SVM and Naïve Bayes algorithms", *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol.4, no.4, pp.816-820, 2012.

[22] T. M., Kamruzzaman, M. S., Mahbub and M. A. Hakim, "A Structured Method For Predicting Liver Disease Using Machine Learning Techniques & Improvements In Correctness", *IEEE 12th International Conference on Computing Communication and Networking Technologies (ICCCNT*), pp.01-07, July 2021.

[23] 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 .

[24] K. Al-Aidaroos, A. A., Bakar and Z. Othman, "Medical data classification with Naive Bayes approach", *Information Technology Journal*, vol.11, no.9, pp.1166-1174, 2012.

[25] R. Bhardwaj, R. Mehta and P. Ramani, "A comparative study of classification algorithms for predicting liver disorders", *Intelligent Computing Techniques for Smart Energy Systems. Lecture Notes in Electrical Engineering*, vol 607. Springer, Singapore.

[26] UCI Machine Learning Repository: BUPA data Set. Available: https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation+Dataset#

[27] J. McDermott and R.S. Forsyth, "Diagnosing a disorder in a classification benchmark", *Pattern Recognition Letters*, vol.73, pp. 41-43, 2016.

[28] Orange programming. Available: https://orangedatamining.com/
[29] N. Friedman, D. Geiger and M. Goldszmidt, M.,"Bayesian network classifiers", *Machine learning*, vol.29, pp.131-163, 1997.
[30] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation", *Soft Computing*, vol.25, no.3, pp.2277-2293, 2021.
[31] X. Wu and V. Kumar, *The top ten algorithms in data mining*, CRC press, 2009.
[32] A. Choi, N. Tavabi, and A. Darwiche, "Structured features in Naive Bayes classification" in the *AAAI Conference on Artificial Intelligence*, vol.3 no.1, pp.3233-3240, *February, 2016*.
[33] Z. Muda, W. Yassin, M.N. Sulaiman and N.I. Udzir, "A K-Means and Naive Bayes learning approach for better intrusion detection", *Information technology journal*, vol.10 no.3, pp.648-655, 2011.
[34] M. M. Saritas and A. Yasar, "Performance analysis of ANN and Naive Bayes classification algorithm for data classification", *International journal of intelligent systems and applications in engineering*, vol.7 no.2, pp.88-91, 2019.
[35] S. S. Nikam, "A comparative study of classification techniques in data mining algorithms", *Oriental Journal of Computer Science and Technology*, vol.8 no.1, pp.13-19, 2015.
[36] R. Blanquero, E. Carrizosa, E., P.Ramírez-Cobo and M.R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification", *Computers & Operations Research*, vol.135, no.105456, pp.1-11, 2021.
[37] J. Han and M. Kamber, M. *Data mining: concepts and techniques*, Second Edi. TM KSIDMA Systems, ed., Morgan Kaufmann Publisher, 2006
[38] S. Mukherjee and N. Sharma, "Intrusion detection using naive Bayes classifier with feature reduction", *Procedia Technology*, vol.4, pp.119-128, 2012.
[39] H. Chen, S. Hu, R. Hua and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management", *EURASIP Journal on Advances in Signal Processing*, vol. 2021 no.1, pp.1-12, 2021.
[40] S. K. Depren, Ö. E. Aşkın and E. Öz, "Identifying the classification performances of educational data mining methods: A case study for TIMSS", *Educational Sciences: Theory & Practice*, vol.17, no.5, pp.1605-1623, 2017.
[41] G. Kaur and E.N. Oberai, "A review article on Naive Bayes classifier with various smoothing techniques", *International Journal of Computer Science and Mobile Computing*, vol.3, no.10, pp.864-868, 2014.
[42] S. Xu, "Bayesian Naïve Bayes classifiers to text classification", Journal of Information Science, vol.44, no.1, pp.48-59, 2018.
[43] D. Berrar, Cross-validation, Encyclopedia of Bioinformatics and Computational Biology, Vol. 1, Elsevier, pp. 542–545, 2018.
[44] H. Şevgin and E. Önen, "Comparison of Classification Performances of MARS and BRT Data Mining Methods: ABİDE- 2016 Case", Education & Science/Egitim ve Bilim, vol.47, no.211, pp.195-222, 2022.
[45] G. Akgül, A.A. Çelik, Z.E. Aydın and Z.K. Öztürk, "Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı", *Bilişim Teknolojileri Dergisi*, vol.13, no.3, pp.255-268, 2020.
[46] T. S. Sujana, N. M. S. Rao and R. S. Reddy, "An efficient feature selection using parallel cuckoo search and naïve Bayes classifier", *IEEE 2017 International Conference on Networks & Advances in Computational Technologies (NetACT)*, pp.167-172, July 2017.
[47] C. Ruengdetkhachorn and D. Lohpetch, "Feature Selection using Parallel Cuckoo Algorithm with Naïve Bayes Classifier based on Two Different Strategies", *IEEE 22nd International Computer Science and Engineering Conference (ICSEC)*, pp.1-4, November 2018,
[48] D. Pradhan, B.B. Misra, B. Sahoo and D.K. Jena, "Evolutionary Teaching-Learning Based Modified Polynomial Classifier", *IEEE 19th OITS International Conference on Information Technology (OCIT)*, pp.313-318, December 2021.
[49] M. Z. Alam, M. S. Rahman and M. S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking", *Informatics in Medicine Unlocked*, vol.15, no.100180, pp.1-11, 2019.