

Received: 20.03.2017, Revised: 05.06.2017, Available online: 25 June 2017.
Volume 05, Issue 01, pp. 361-370
Original research paper, Article page: journalbinet.com/jstei-050117-38.html

Analysis of classification algorithms for liver disease diagnosis

Shapla Rani Ghosh and Sajjad Waheed

Dept. of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh.

Article info.

ABSTRACT

Key Words:

Classification Algorithms
(Bagging, K-star, NBC,
Logistic, Rep Tree), Liver
disease diagnosis



For any information:

ask.author@journalbinet.com

Now a days liver disease is extending markedly due to excessive alcohol consumption, smoking, drinking arsenic contaminated water, obesity, low immunity and by inheritance. Liver cancer symptoms may include jaundice, abdominal pain, fatigue, nausea, vomiting, back pain, abdominal swelling, weight loss, general itching. Selective algorithms may be used on medical instruments (e.g. CT scanner, MRI, Ultra sono, ECG etc.) to lessen time and cost on hepatic disease diagnosis. Here some of the algorithms such as, Naive Bayes classification (NBC), Bagging, KStar, Logistic and REP tree were used to evaluate the accuracy, precision, sensitivity and specificity. For these two data sets of UCLA and AP were considered to find out the best algorithm. The whole analysis was done using the software Weka 3.6.10. It was revealed that, KStar algorithm had the maximum accuracy, precision, sensitivity and specificity. On the other, minimum accuracy was obtained from NBC. Therefore K* algorithm can be used on diagnosis tools or instruments for rapid identification of specific liver disorder.

Citation: Ghosh, S. R. and Waheed, S. (2017). Analysis of classification algorithms for liver disease diagnosis. Journal of Science, Technology and Environment Informatics, 05(01), 361-370.
<https://doi.org/10.18801/jstei.050117.38>

© 2017, Ghosh and Waheed. This is an open access article distributed under terms of the Creative Common Attribution 4.0 International License.

I. Introduction

Liver cancer is also known as hepatic cancer. It is the growth and spread of unhealthy cells in the liver which is heritable. Cancer that starts in the liver is called primary liver cancer; which is a condition that happens when normal cells in the liver become abnormal in appearance and behavior. Cancer cells can then become destructive to adjacent normal tissues, and can spread both to other areas of the liver and to organs outside the liver which is called metastatic liver cancer or secondary liver cancer. Liver cancer consists of malignant hepatic tumors (growths) in the liver. The most common type of liver cancer is

hepatocellular carcinoma (or hepatoma, or HCC) and it tends to affect males more than females. Classification algorithms are very much suitable and used in different automated medical diagnosis tools. Stress, eating habits, drinking alcohol, exposure to toxic gas, and nature of drug intake etc. are the main causes of rising liver disease patients every year (Hoon et al., 2014). Symptoms of liver patient include jaundice, abdominal pain, fatigue, nausea, vomiting, back pain, abdominal swelling, weight loss, fluid in abnormal cavity, general itching, pale stool, enlarged spleen and gallbladder (Sindhuja and Priyadarsini, 2016). Certain abnormalities of liver patients are hard to find out in initial stage because it will function as usual even if it is partially damaged (Rong-Ho, 2009). Initial examination of liver disorders will decrease patients' mortality rate. Liver disease can be diagnosed by analyzing the levels of enzymes in the blood (Schiff et al., 2007). Besides, now a day's mobile devices are widely used to monitor human body conditions. Automatic classification algorithms can be applied in this case. It reduces the patient line at the liver experts. Otoom et al. (2015) proposed a system to detect and monitor coronary artery disease. Two tests with three algorithms- Bayes Net, Support vector machine, and Functional Trees FT were used. WEKA tool was used for detection. Test was done on 7 best selected features, Bayes Net attained 84.5% of correctness, SVM gave 85.1% accuracy and FT classified 84.5% correctly. Vembandasamy et al. (2015) used Naive Bayes algorithm for diagnosis heart disease. Bayes theorem is used in Naive Bayes. Therefore, it has a powerful independence assumption. Weka was used as a tool which executed 70% of percentage split. Naive Bayes had 86.419% of accuracy. Iyer et al. (2015) conducted an experiment to predict diabetes disease with the help of decision tree and Naive Bayes. Different tests were carried out using WEKA data mining tool. Naive Bayes showed 79.5652% correctness by using percentage split test. Maximum accuracy of algorithms was obtained by using percentage split test. Vijayarani and Dhayanand (2015) used Support vector machine and Naive bayes classification algorithms for liver disease prediction. Data analysis was done using MATLAB. Naive bayes gave 61.28% correctness in 1670.00 ms and SVM gave 79.66% accuracy in 3210.00 ms. Gulia et al. (2014) studied on intelligent techniques to classify liver patients using datasets from UCI. WEKA tool and five algorithms- J48, MLP, Random Forest, SVM and Bayesian Network were for experimentation. After FS, algorithms gave highest accuracy as- J48 70.669%, MLP 70.8405%, SVM 71.3551%, Random forest 71.8696% and Bayes Net gave 69.1252% accuracy. Tarmizi et al. (2013) investigated on Malaysia dengue outbreak detection with data mining models. Decision Tree (DT), Artificial Neural Network (ANN), and Rough Set theory (RS) classification algorithms were used. WEKA data mining tool with two tests (10 cross-fold validation and percentage split) were used. With 10-Cross fold validation DT gave 99.95%, ANN gave 99.98%, and RS showed 100% accuracy. With PS, Both DT and ANN exhibited 99.92% of exactness and RS showed 99.72% of accuracy. Ba-Alwi and Hintaya (2013) reported a comparative analysis on hepatitis disease diagnosis. Naive Bayes, Naive Bayes updatable, FT Tree, KStar, J48, LMT (Logistic Model Tree) and NN (Neural Network) algorithms were used. Classification results were measured in terms of accuracy and time. Analysis was done by neural connections and WEKA. Results obtained from neural connection were lower compared to algorithms used in WEKA. Rough set theory was applied as an alternative technique using WEKA. It performed better than NN especially for medical data analysis. Naive Bayes gave the accuracy of 96.52% in 0 sec, Naive Bayes Updateable algorithm showed 84% exactness in 0 sec, FT Tree presented 87.10% correctness in 0.2 sec, K star exhibited 83.47% accuracy in 0 sec, J48 achieved 83% accurateness in 0.03 sec, LMT provided 83.6% accuracy in 0.6 sec and NN showed 70.41% of accuracy. Naive Bayes was the best classification algorithm when rough set technique was used. It offered highest accuracy in lowest time. The main objective of this research work was to find the best classification algorithm in terms of precision, accuracy, specificity and sensitivity. Therefore the present investigation was done to determine the relative performance of five classification algorithms namely, Naive Bayes classification (NBC), Bagging algorithm, KStar algorithm, Logistic algorithm and REP Tree algorithm; based on collected liver patient data.

II. Materials and Methods

Two datasets of liver patients were used to estimate performance (i.e., of precision, accuracy, specificity and sensitivity) of classification algorithms. First one was collected from Andhra Pradesh state (AP) of India and the second one from the University of California at Irvine (UCI) (BUPA, 2010). First dataset contained 583 liver patient records with 10 attributes as shown in Table 01. Among them 441 were male and 142 were female patient. Second dataset contained 345 records with 6 attributes as shown in Table 02. The first 5 variables are obtained from blood tests which may be susceptible to liver disorders

that might arise from consumption of excessive alcohol. For the purpose of experimentation and analysis, Weka (version 3.6.10) data mining open source machine learning software (Weka, 2005). Accuracy, precession and specificity was calculated according to the formulae provided by Ramana et al. (2011); which can be defined as follows-

- Accuracy = $\frac{TP+TN}{TP+TN+FP+PN}$
- Precision = $\frac{TP}{TP+FP}$
- Specificity = $\frac{TN}{TN+FP}$

Four values, i.e., true positive, false positive, true negative and false negative are the basis to calculate all the measures. These values are described below.

- FN = false negatives: number of examples predicted negative that are actually positive
- TP = true positives: number of examples predicted positive that are actually positive
- FP = false positives: number of examples predicted positive that are actually negative
- TN = true negatives: number of examples predicted negative that are actually negative

Table 01. AP liver dataset and attributes

Attributes	Type
Gender	Categorical
Age	Real number
Total_bilirubin	Real number
Direct_bilirubin	Real number
Total_protiens	Real number
Albumin	Real number
A/G ratio	Real number
SGPT	Integer
SGOT	Integer
Alkphos	Integer

Table 02. UCLA liver dataset and attributes available

Attributes	Type
Mcv	Integer
Alkphos	Integer
SGPT	Integer
SGOT	Integer
Gammagt	Real number
Drinks	Real number

The flow diagram of the entire work is shown in Figure 01 below.

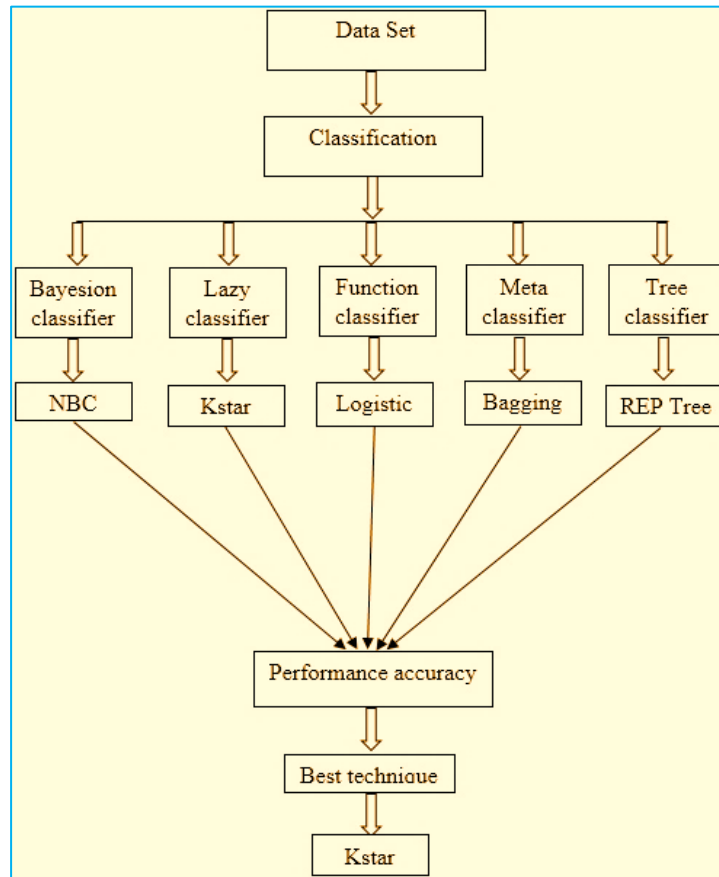


Figure 01. Flow diagram of classification algorithms.

Classification algorithms

Naive Bayes Algorithm: The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem. Bayes theorem provides a way of calculating the posterior probability, $P(c/x)$, from $P(c)$, $P(x)$, and $P(x/c)$. Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence (Ramana et al., 2011). The Naive Bayesian classification predicts that the tuple 'x' belongs to the class 'c' using the formula-

$$P(c/x) = \frac{P(x/c)P(c)}{P(x)}$$

- $P(c/x)$ is the posterior probability of class (target) given predictor (attribute).
- $P(c)$ is the prior probability of class.
- $P(x/c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor

KStar Algorithm: K^* is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function.

Bagging Algorithm: 'Bagging' stands for "bootstrap aggregating". Bagging is a method for improving results of machine learning classification algorithms. Bagging (Bootstrap aggregating) was proposed by

Leo Breiman to improve the classification by combining classifications of randomly generated training sets. More information about bagging was reported by [Friedman and springer \(2015\)](#), [Hastie \(2014\)](#), [Machova et. al. \(2006\)](#) and [Tibshirani \(2014\)](#). Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression.

Logistic Model tree regression: Logistic model tree (LMT) is a classification model with an associated training algorithm that combines logistic regression (LR) and decision tree learning ([Wikipedia, 2017](#)). It is also called a logit model, which is used to model dichotomous outcome of variables.

REP Tree Algorithm: REP means reduced-error pruning. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning.

III. Results and Discussion

The performance all algorithms were analyzed with AP dataset. Contributions of all the classification algorithms for the given attributes are shown in [Table 03](#) and [Table 04](#). Among the classification algorithms K Star gave better accuracies with the considered attributes.

Table 03. Accuracy, Precision, Sensitivity and Specificity percent of selected algorithms with AP dataset

Algorithm name	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	TTBM (s)
Bagging	88	88.5	76.6	54.6	0.11
KStar	100	100	100	100	0.00
NaiveBayes	39	70.6	55.2	86.5	0.02
Logistic	75.4	63.5	50	75.4	0.19
REPTree	79	77.4	60.4	24.1	0.02

TTBM =Total time to built model

Table 04. Error rate of the algorithmsfor all featuresof AP dataset

Algorithm name	MAE(%)	RMSE(%)	RAE(%)
Bagging	23.83	30.3	64.837
KStar	0	0.02	0.0105
NaiveBayes	51.36	58.13	139.716
Logistic	35.26	41.99	95.9099
REPTree	31.62	39.76	86.0198

MAE= Mean Absolute Error, RMSE= Root Mean Square Error, RAE= Relative Absolute Error

[Figure 02](#) indicates that K* algorithm had highest accuracy, precision, sensitivity and specificity. Whereas, NBC gave the lowest accuracy with AP dataset. In [Figure 03](#), error rate of the selected algorithms with AP data set is shown. We carried out similar experiment with UCLA dataset also ([Table 05](#) and [Table 06](#)); there K* algorithm was found to be giving maximum value in all parameters.

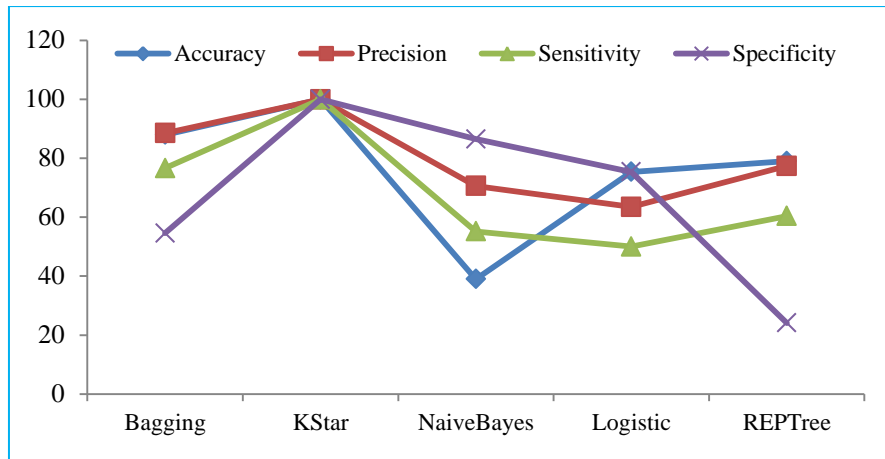


Figure 02. Accuracy, Precision, Sensitivity and Specificity percent of selected algorithms with all attributes of AP dataset.

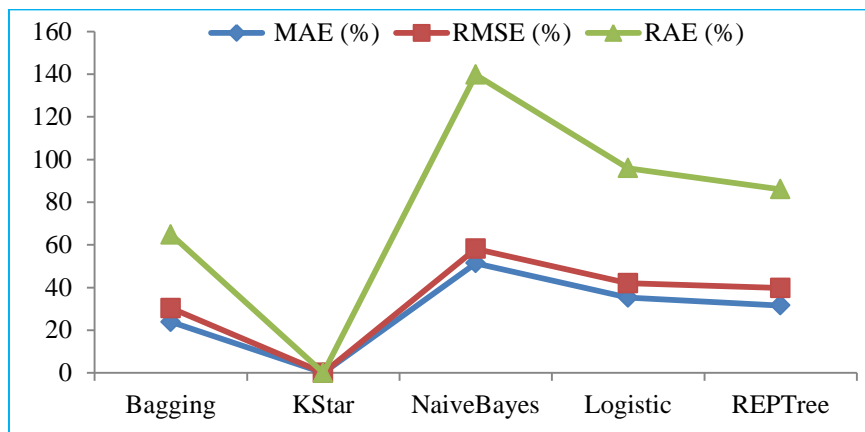


Figure 03. Error analysis of algorithms for all features of AP dataset.

Table 05. Accuracy, Precision, Sensitivity and Specificity percent of selected algorithms with UCLA liver dataset

Algorithm name	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	TTBM (s)
Bagging	87.5	88.4	85.9	96.5	0.06
KStar	100	100	100	100	0.00
NaiveBayes	56.8	56.8	59.9	40.5	0.01
Logistic	70.4	70.2	68.2	82.5	0.03
REPTree	74.2	76.6	70.6	93.5	0.01

TTBM =Total time to built model

Table 06. Error rate of the algorithms for all features of UCLA dataset

Algorithm name	MAE (%)	RMSE (%)	RAE (%)
Bagging	30.82	34.48	63.2404
KStar	0.11	0.53	0.2258
NaiveBayes	45.23	50	92.804
Logistic	41.02	45.17	84.1645
REPTree	35.42	42.09	72.6844

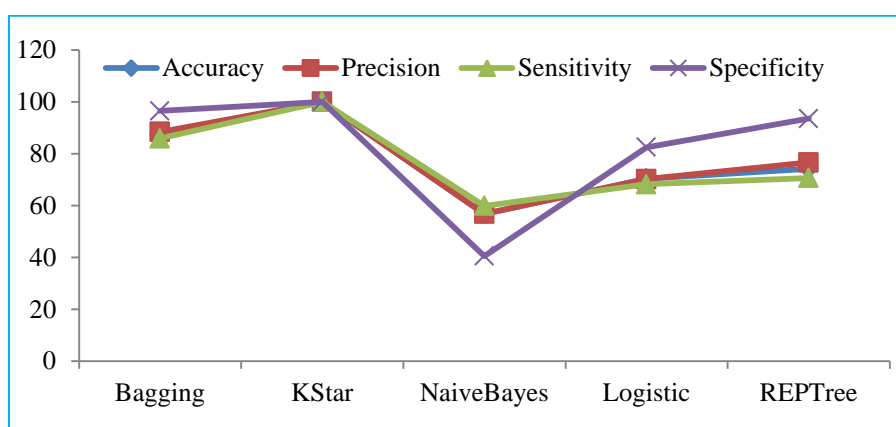


Figure 04. Accuracy, Precision, Sensitivity and Specificity percent of selected algorithms with all attributes of UCLA dataset.

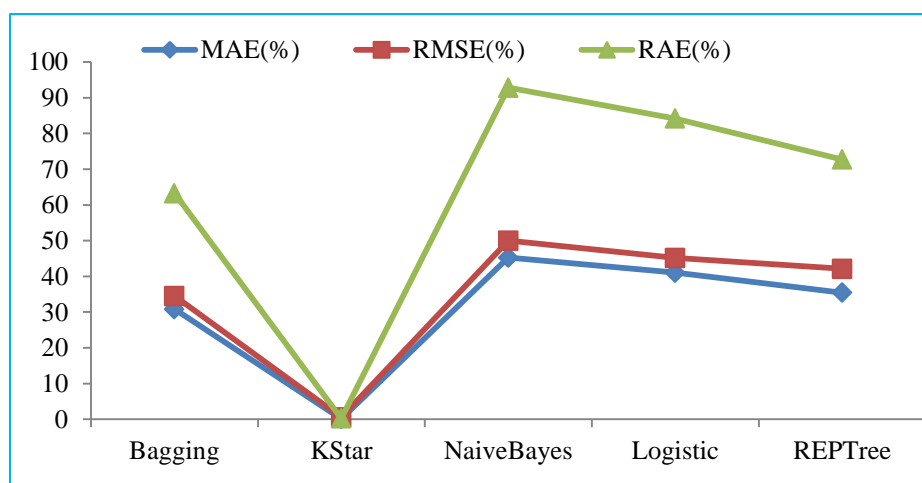


Figure 05. Error analysis of algorithms for all features of UCLA dataset.

Table 07 shows accuracy, precision, sensitivity and specificity of common attributes (SGOT, SGPT, ALP) of both AP and UCLA datasets. Table 08 shows error rates of common attributes (SGOT, SGPT, ALP) of both AP and UCLA datasets. Figure 06, Figure 07, Figure 08 and Figure 09 shows accuracy, precision, sensitivity, specificity performance of the selected algorithms with both datasets with common attributes.

Table 07. Accuracy, Precision, Sensitivity and Specificity performance of classification algorithms

Classification Algorithms	Accuracy (%)		Precision (%)		Sensitivity (%)		Specificity (%)	
	AP liver dataset	UCLA liver dataset	AP liver dataset	UCLA liver dataset	AP liver dataset	UCLA liver dataset	AP liver dataset	UCLA liver dataset
Bagging	84	82	84.3	82.3	84	80.4	97.7	90.5
K-Star	98.5	99.7	98.5	99.7	98.5	99.7	99.5	100
Navie Bayes	35	52.8	71.5	57.3	54.2	55.3	16.8	39
Logistic	75.6	66.1	57.2	65.6	50	63.6	100	79.5
REPTree	80.4	75.7	78.8	77.3	67	72.5	93.6	92.5

Table 08. Error performance of classification algorithms

Classification Algorithms	MAE		RMSE		RAE (%)	
	AP liver dataset	UCLA liver dataset	AP liver dataset	UCLA liver dataset	AP liver dataset	UCLA liver dataset
Bagging	0.2621	0.3436	0.336	0.379	71.038	70.504
K-Star	0.1622	0.1594	0.2179	0.1934	43.953	32.713
Navie Bayes	0.5068	0.4789	0.5365	0.5056	137.38	98.272
Logistic	0.3639	0.4366	0.4264	0.467	98.628	89.575
REPTree	0.2768	0.3369	0.372	0.4104	75.03	69.126

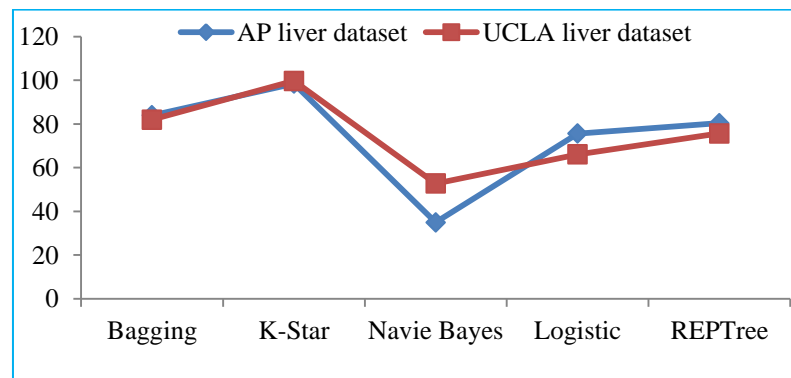


Figure 06. Accuracy of classification algorithms with AP and UCLA dataset.

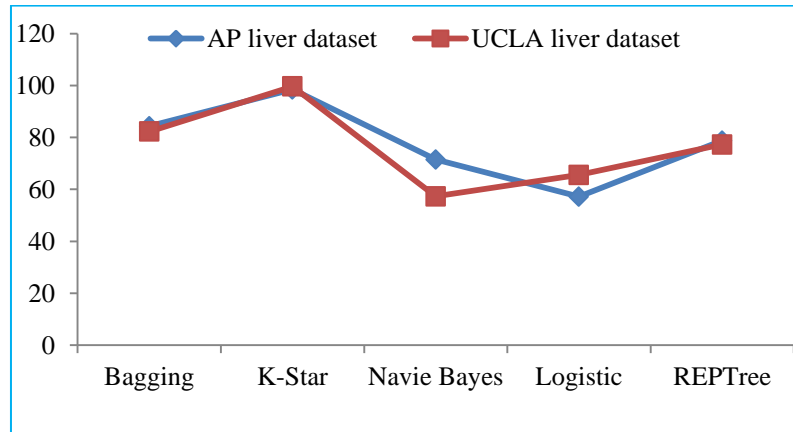


Figure 07. Precision level of classification algorithms with both dataset.

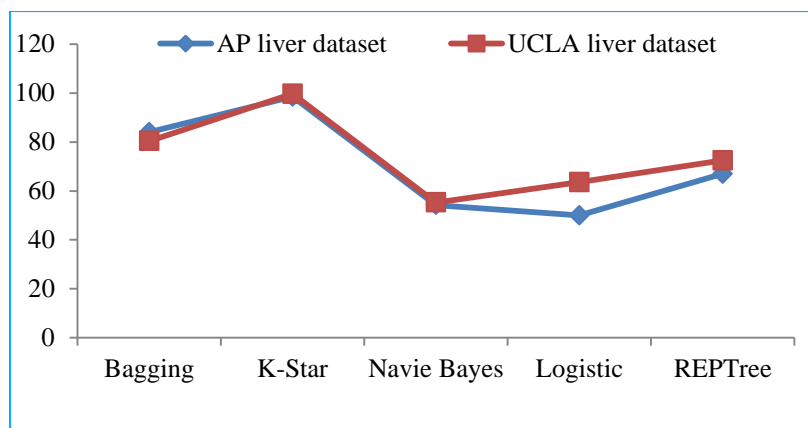


Figure 08. Sensitivity status of classification algorithms with AP and UCLA liver dataset.

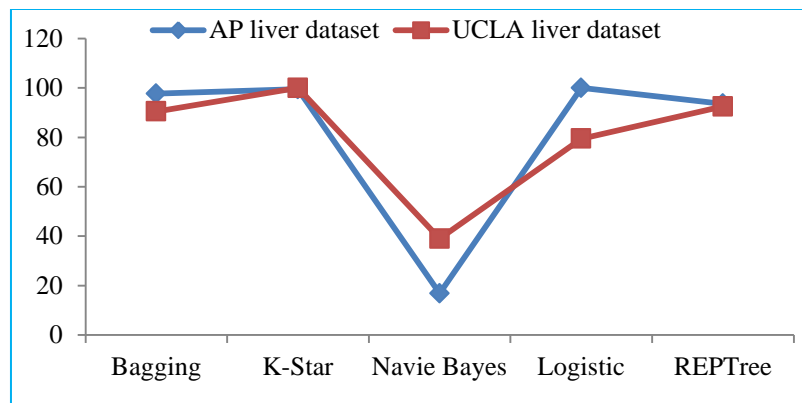


Figure 09. Specificity for selected classification algorithm

IV. Conclusion

In terms of accuracy, precision, sensitivity and specificity K* algorithm was found to be superior because it had the lowest error rate with highest accuracy compared to NBC, Bagging, Logistic and Rep Tree with both AP and UCLA data sets. Therefore this algorithm (K star) is most suitable for liver disease diagnosis.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Source of Funding

The entire research work was funded by author herself.

Author Contribution

This work was carried out in collaboration between the authors. Author S. R. Ghosh managed the experimental process, performed Weka analyses, collected the related review and wrote the whole manuscript. Author S. Waheed supervised designed and made linguistic corrections of the manuscript. All authors read and approved the final manuscript.

V. References

- [1]. Ba-Alwi, F. M. and Hintaya, H. M. (2013). Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach. *International Journal of Scientific & Engineering Research*, 4 (8), 680-685.
- [2]. BUPA Liver Disorders Dataset (2010). UCI repository of machine learning databases. Available from <ftp://ftp.ics.uci.edu/pub/machinelearningdatabases/liverdisorders/bupa.data>.
- [3]. Friedman, J. and Springer, L. A. (2015). Comparison of the bagging and the boosting methods using the decision trees classifiers. Available from <http://www.stat.stanford.edu/people/faculty/friedman.html>.
- [4]. Gulia, A., Vohra, R. and Rani, P. (2014). Liver Patient Classification Using Intelligent Techniques. *International Journal of Computer Science and Information Technologies*, 5 (4), 5110-5115.
- [5]. Hastie, T. (2014). Retrieved from web browsing <http://www.stat.stanford.edu/%7Ehastie/>Robert
- [6]. Hoon, J., Seoungchon, K. and Jinhong, K. (2014). Decision factors on effective liver patient data prediction. *International journal of Bio-Science and Bio-Technology*, 6(4), 167-168. <https://doi.org/10.14257/ijbsbt.2014.6.4.16>
- [7]. Iyer, A., Jeyalatha, S. and Sumbaly, R. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process*, 5 (1), 1-14. <https://doi.org/10.5121/ijdkp.2015.5101>
- [8]. Machova, K., Barcak, F. and Bednar, P. (2006). A bagging method using decision trees in the role of base classifiers. *Acta Ploytechnica Hungarica*, 3(2), 121-132.
- [9]. Otoom, A. F., Abdallah, E. E., Kilani, Y., Kefaye, A. and Ashour, M. (2015). Effective Diagnosis and Monitoring of Heart Disease. *International Journal of Software Engineering and Its Applications*, 9 (1), 143-156.
- [10]. Ramana, B. V., Babu, M. S. P. and Venkateswarlu, N. B. (2011). A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *International Journal of Database Management Systems (IJDMS)*, 3(2), 1-14.
- [11]. Rong-Ho, L. (2009). An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine*, 47(1), 53-62. <https://doi.org/10.1016/j.artmed.2009.05.005>
PMid:19540738
- [12]. Schiff, E. R., Sorell, M. F. and Maddrey, W. C. (2007). *Diseases of the liver*, 10th ed., Philadelphia, PA19106, USA: Lippincott Williams and Wilkins.
- [13]. Sindhuja, D. and Jemina Priyadarsini, R. (2016). A survey on classification techniques in data mining for analyzing liver disease disorder, *International Journal of Computer Science and Mobile Computing*, 5(5), 483-488.
- [14]. Tarmizi, N. D. A., Jamaluddin, F., Abu Bakar, A., Othman, Z. A., Zainudin, S. and Hamdan, A. R. (2013). Malaysia Dengue Outbreak Detection Using Data Mining Models. *Journal of Next Generation Information Technology*, 4(6), 96-107.
- [15]. Tibshirani, R. (2014). Available from <http://www.stat.stanford.edu/%7Etibs/>Jerome.

- [16]. Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology*, 2 (9), 441-444.
- [17]. Vijayarani, S. and Dhayanand, S. (2015). Liver Disease Prediction using SVM and Naive Bayes Algorithms. *International Journal of Science, Engineering and Technology Research*, 4(4), 816-820.
- [18]. Weka (2005). v3.6.10jre: data mining with open source machine learning software © 2002-2005 David Scuse and University of Waikato.
- [19]. Wikipedia (2017). Logistic model tree. Available from https://en.wikipedia.org/wiki/Logistic_model_tree.

HOW TO CITE THIS ARTICLE?

Crossref: <https://doi.org/10.18801/jstei.050117.38>

APA (American Psychological Association)

Ghosh, S. R. and Waheed, S. (2017). Analysis of classification algorithms for liver disease diagnosis. *Journal of Science, Technology and Environment Informatics*, 05(01), 361-370.

MLA (Modern Language Association)

Ghosh, S. R. and Waheed, S. "Analysis of classification algorithms for liver disease diagnosis". *Journal of Science, Technology and Environment Informatics*, 05.01 (2017): 361-370.

Chicago and or Turabian

Ghosh, S. R. and Waheed, S. "Analysis of classification algorithms for liver disease diagnosis". *Journal of Science, Technology and Environment Informatics*, 05, no. 01 (2017): 361-370.