Juxtaposition on Classifiers in Modeling Hepatitis Diagnosis Data

Preetham Ganesh, Harsha Vardhini Vasu, Keerthanna Govindarajan Santhakumar, Raakheshsubhash Arumuga Rajan, and K. R. Bindu

1 Introduction

Machine learning, an advancing field of computer science, plays a crucial role in predicting unforeseeable parameters in different domains such as medical diagnosis, weather forecast, sports, and many more, which has always been very complicated for humans. A machine learning model is trained based on the inspection of data done by the algorithm with which mathematical equations can be developed to make better decisions in the future based on the observed trends. The preliminary objective is to make computers learn and make decisions without human intervention.

A recent trend observed in the medical field is the implementation of machine learning techniques to diagnose the presence of an infection/disease. Since medical datasets have loads of information, data mining also has a significant role in mining the necessary features for prediction. So it is fundamental to use both machine learning and data mining techniques to model and predict from hepatitis data.

A disease named Hepatitis C damages the liver by causing inflammation and infection in it. The condition aggravates after being infected with the Hepatitis C Virus (HCV). Identifying the presence of Hepatitis is one of the significant challenges faced by health organisations [1]. Worldwide around 130–170 million people have been infected by HCV [2]. Approximately 71 million among them have chronic hepatitis C, and 399,000 people die each year of Hepatitis C [3]. Accurate diagnosis and precise prediction at an early stage can help save the patient's life with minimum damage to the patient's health. This study intends to analyse Hepatitis Data and classify based on the observed patterns using different classifiers and check for the perfect classifier based on the performance measures.

P. Ganesh · H. V. Vasu · K. G. Santhakumar · R. A. Rajan · K. R. Bindu (✉)
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: j_bindu@cb.amrita.edu

S. Smys et al. (eds.), *New Trends in Computational Vision and Bio-inspired Computing*, https://doi.org/10.1007/978-3-030-41862-5_48

This study is segregated as sections and is as follows: Section 2 explores the literature survey in the areas related to data mining and machine learning. Section 3 discusses the details about the dataset used, machine learning models used for classification, and the performance measures used for evaluation. Section 4 presents the result obtained by the conducted study, and Sect. 5 concludes the paper based on the obtained result.

2 Related Works

The authors in [1] tested different decision tree algorithms on the hepatitis dataset from the UCI repository and evaluated the classification models using measures such as accuracy, precision, recall, and F1-Measure. Based on the results, it was concluded that the random forest classifier performed best with an accuracy of 87.5%.

Rosalina et al. [4] performed feature selection using the wrapper method on the same dataset mentioned above. The authors used Support Vector Machines (SVM) on both the feature selected data and the original data to compare its performance. An accuracy score was used to check the performance of the classifier model. It was concluded by the authors that SVM produced better results for the feature selected data than the original data.

Ekız et al. [5] used the Heart Diagnosis dataset from the UCI repository for analysis, where the classifiers used for analysing are Decision Tree, SVM, Ensemble Subspace on MATLAB and WEKA. Based on the values of accuracy, it was concluded that subspace discriminant performs better than the others, and among SVM, SVM with linear kernel surpasses the others.

This paper primarily anchors on finding the best classification model for the chosen dataset. The study is about the application of five classification algorithms—Random Forest Classifier, SVM, Logistic Regression, Naive Bayes Classifier and Decision Tree Classifier—on the hepatitis dataset and selecting the best by comparing its performance metrics such as accuracy, recall, specificity, precision, F1-Measure, Matthews Correlation Coefficient and many more.

3 Methodology

3.1 Dataset Description

The dataset was collected from the UCI Repository [6], which has 155 tuples, 19 self-dependent attributes, and a label named 'Class' for prediction. The columnwise details of the dataset are given in Table 1.

Table 1 Dataset description

S. No	Attribute	Type	Values
1	Age	Numerical	31, 34, 39, 32
2	Bilirubin	Numerical	0.7, 0.9, 1, 1.3
3	Alk. phosphate	Numerical	46, 95, 78, 59
4	SGOT	Numerical	52, 28, 30, 249
5	Albumin	Numerical	4, 4, 4.4, 3.7
6	Protime	Numerical	80, 75, 85, 54
7	Sex	Categorical	Male/female
8	Steroid	Categorical	Yes/no
9	Antivirals	Categorical	Yes/no
10	Fatigue	Categorical	Yes/no
11	Malaise	Categorical	Yes/no
12	Anorexia	Categorical	Yes/no
13	Liver Big	Categorical	Yes/no
14	Liver Firm	Categorical	Yes/no
15	Spleen Palpable	Categorical	Yes/no
16	Spiders	Categorical	Yes/no
17	Ascites	Categorical	Yes/no
18	Varices	Categorical	Yes/no
19	Histology	Categorical	Yes/no
20	Class	Categorical	Live/die

3.2 Process Flow

Figure 1 shows the process flow used in this study.

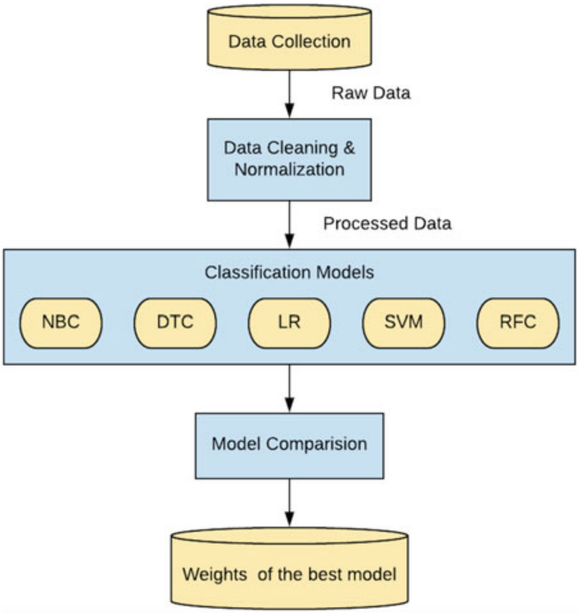
3.3 Classification Algorithms

Logistic Regression (LR) A logistic function is used to model the binary class variable, where the variable should be in the numerical form of 0 or 1. The class variable can be a combination of self-dependent binary variables/continuous variables. The respective probability of the value labeled '1' varies from 0.5 to 1, and '0' varies from 0 to 0.5 [7].

Naive Bayes Classifier (NBC) They are a family of uncomplicated probabilistic classifiers based on the implementation of Bayes Theorem, where the classifier works on the assumption that attributes are independent of each other [8]. There are six types of Naive Bayes classifiers out of which three are used in this study, namely Gaussian, Multinomial, and Bernoulli.

Support Vector Machine (SVM) It is used for finding the optimal dividing hyperplane between the classes using the statistical learning theory [9]. Overfitting

Fig. 1 Process flow



can be avoided by choosing the correct size of the margin separating the hyperplane from positive and negative classified instances [4].

Decision Tree Classifier (DTC) It resembles a structure similar to a flowchart, each interior node represents a try-out on a feature, and each limb represents the result of the try-out. Each leaf node represents any one of the class labels [10].

Random Forest Classifier (RFC) It is an ensemble learning method which mainly operates by building a swarm of decision trees during the training stage of the model and displaying the mode of the target class during the testing stage [11]. Usually, the model is overfitted to the training data.

3.4 Performance Measures

The performance of a classifier can be decided based on the instances the classifier has classified correctly in the test set after trained on the train set. A tool called Confusion Matrix plays a vital role in calculating the performance of the classifier [12]. The representation of the confusion matrix is given in Table 2. The performance measures used in this study are listed in Table 3, along with their definitions, are formulae [13].

$$MCC = \frac{TP * TN + FP * FN}{\sqrt{(TP + FP) * (FN + TP) * (TN + FP) * (TN + FN)}} \quad (1)$$

Table 2 Sample representation of confusion matrix

	Actual		
	Class	Class A	Class B
Predicted	Class A	True Positive (TP) Correctly classified as Positive	False Positive (FP) Incorrectly classified as Positive
	Class B	False Negative (FN) Incorrectly classified as Negative	True Negative (TN) correctly classified as Negative

Table 3 Performance measures description along with their formulae

S. No	Performance measure	Definition	Formula
1	Accuracy	The fraction of tuples the model has classified correctly	$\frac{TP+TN}{TP+FP+TN+FN}$
2	Balanced Accuracy	Average of correctly classified tuples for each class	$\frac{\frac{TP}{P}+\frac{TN}{N}}{2}$
3	Recall (R)/Sensitivity (SN)	The fraction of tuples correctly classified as positive	$\frac{TP}{TP+FN}$
4	Specificity (SP)	The fraction of tuples correctly classified as negative	$\frac{TN}{FP+TN}$
5	Precision (Pr)	The fraction of tuples correctly classified as positive among predicted positives	$\frac{TP}{TP+FP}$
6	Negative Predictive Value	The fraction of tuples correctly classified as negative among predicted negative	$\frac{TN}{TN+FN}$
7	Fall-out	The fraction of tuples incorrectly classified as positive	$\frac{FP}{FP+TN}$
8	False Discovery Rate	The fraction of tuples incorrectly classified as negative among predicted negatives	$\frac{FP}{TP+FP}$
9	False Negative Rate	The fraction of tuples incorrectly classified as negative among actual negatives	$\frac{FN}{TP+FN}$
10	F1-Measure	Harmonic mean of precision and recall	$\frac{2*\mathrm{Pr}*R}{\mathrm{Pr}+R}$
11	Matthews Correlation Coefficient (MCC) [14]	Correlation coefficient between observed and predicted tuples	Eq. (1)
12	Informedness [15]	Evaluates how informed a model is for the specified condition	$SP + SN - 1$
13	Markedness [15]	Evaluates how marked a condition is for the model	$Pr + NPV - 1$

4 Results and Discussion

This chapter discusses in detail the outcomes of the five classifier models that have been used for the study based on different measures mentioned in Table 3. The programming was done with the help of R Programming language in R Studio.

Table 4 Performance measure values based on the formulas in Table 3

S. No	Performance measures	NBC	DTC	LR	SVM	RFC
1	Accuracy	0.823	0.86	0.867	0.873	**0.907**
2	Balanced Accuracy	0.749	0.843	0.836	0.842	**0.885**
3	Recall	0.536	0.808	0.702	0.788	**0.845**
4	Specificity	**0.962**	0.878	0.904	0.896	0.926
5	Precision	**0.867**	0.45	0.717	0.55	0.683
6	Negative Predictive Value	0.813	**0.963**	0.904	0.954	**0.963**
7	Fall-Out	**0.038**	0.122	0.096	0.104	0.074
8	False Discovery Rate	**0.133**	0.55	0.283	0.45	0.317
9	False Negative Rate	0.464	0.192	0.298	0.212	**0.155**
10	F1-Measure	0.66	0.535	0.68	0.622	**0.734**
11	Matthews Correlation Coefficient	0.116	**0.206**	0.158	0.189	0.189
12	Informedness	0.499	0.687	0.606	0.684	**0.771**
13	Markedness	**0.679**	0.413	0.621	0.504	0.646

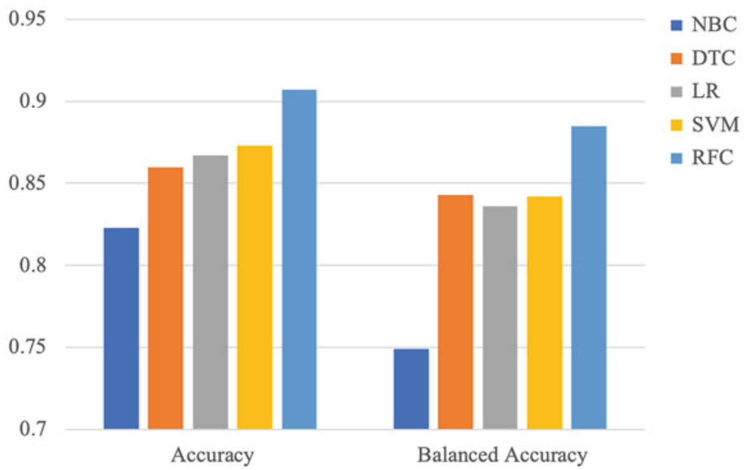


Fig. 2 Accuracy and balanced accuracy for all the classifiers

The dataset collected had missing values, which was imputed using Predictive Mean matching [16], and the numerical attributes were normalized using Z-Score normalization [17]. The processed dataset was split into a train and test set using the Holdout method. Table 4 discusses in detail the various performance measures for each of the classifiers in the test set. The graphical representation of the same is given in Figs. 2, 3, and 4.

A good classifier model should have high accuracy, recall, precision, sensitivity, specificity, and F1-Measure [18] and low false-negative rate, false-discovery rate, and false-positive rate. The dataset used for analysis is biased, i.e., class 'live' has 123 tuples, and class 'die' has 32 tuples. Therefore, accuracy, balanced accuracy, precision, recall, and F1-measure will not be sufficient to judge as to whether a

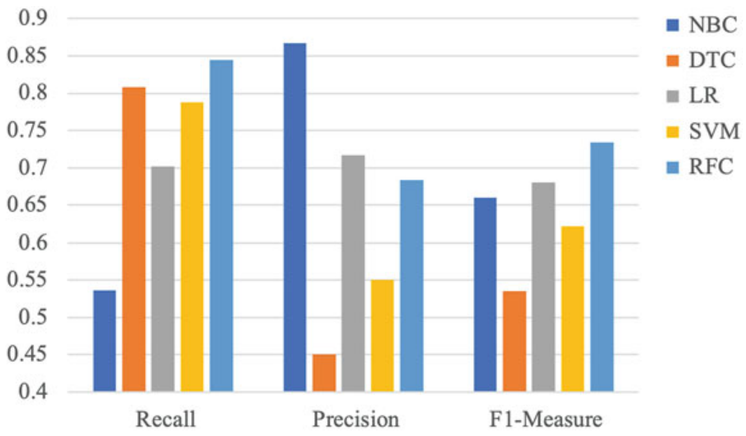


Fig. 3 Recall, precision, and F1-measure for all the classifiers

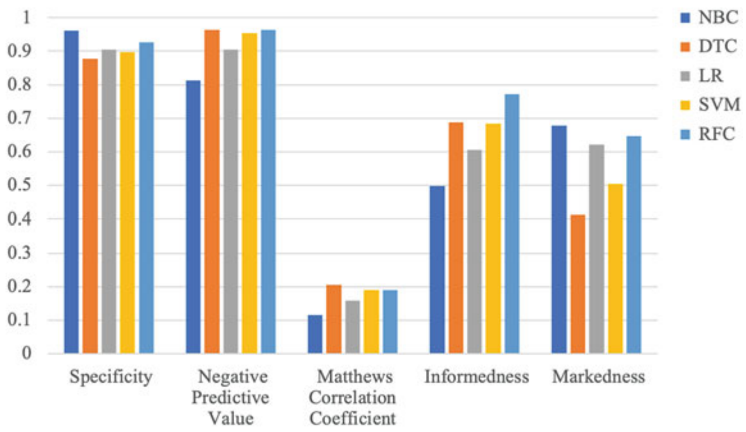


Fig. 4 Other performance measures for all the classifiers

classifier performed well or not. From the observation in Table 4, it can be inferred that Random Forest Classifier outperformed the other models. Even though the other models had better values in a few performances measures better than Random Forest Classifier, but the difference was very minimal. Hence, it can be concluded that Random Forest Classifier performed best for the chosen dataset.

5 Conclusion

In this study, the performance of the different classifiers modeled on the hepatitis data from the UCI Repository was inspected. The classifiers used in this study are Logistic Regression, Naive Bayes Classifier, Support Vector Machine, Decision

Tree Classifier, and Random Forest Classifier. Various performance measures were used for evaluating and comparing the performance of the classifier models. Based on the obtained results, it was inferred that Random Forest Classifier outperformed the other classifiers and provided an accuracy of 90.7%. The model produced good accuracy for a sparse dataset, so there is a higher probability that the model would work even better in a denser dataset, which would help diagnose Hepatitis C at an earlier stage.

References

1. M. Ramasamy, S. Selvaraj, and M. Mayilvaganan. An empirical analysis of decision tree algorithms: Modeling hepatitis data. In *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 1–4, 2015.
2. Huda Yasin, Tahseen A. Jilani, and Madiha Danish. Hepatitis-c classification using data mining techniques. *International Journal of Computer Applications*, 24(3):1–6, 2011.
3. World Health Organization et al. *Global hepatitis report 2017*. World Health Organization, 2017.
4. A.H. Roslina and A. Noraziah. Prediction of hepatitis prognosis using support vector machines and wrapper method. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, pages 2209–2211, 2010.
5. S. Ekız and P. Erdogmus. Comparative study of heart disease classification. In *2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–4, 2017.
6. Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
7. Strother H. Walker and David B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1/2):167–179, 1967.
8. David J. Hand and Keming Yu. Idiot's bayes: Not so stupid after all? *International Statistical Review/Revue Internationale de Statistique*, 69(3):385–398, 2001.
9. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
10. J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
11. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
12. Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
13. Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
14. B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451, 1975.
15. David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.
16. Donald B. Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business Economic Statistics*, 4(1):87–94, 1986.
17. D. Freedman, R. Pisani, and R. Purves. *Statistics: Fourth International Student Edition*. International student edition. W.W. Norton & Company, 2007.
18. Vikas K Vijayan, KR Bindu, and Latha Parameswaran. A comprehensive study of text classification algorithms. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1109–1113. IEEE, 2017.