



# Classification and Prediction Analysis of Diseases and Other Datasets Using Machine Learning

Junaid Nasir<sup>1</sup>, Alishba Ahsan<sup>1</sup>, Nadeem Sarwar<sup>1</sup>(✉), Wajid Rafique<sup>2</sup>, Sameer Malik<sup>3</sup>, Syed Zeeshan Hussain Shah<sup>4</sup>, Sarousha Nasir<sup>1</sup>, and Asma Irshad<sup>5</sup>

<sup>1</sup> Department of Computer Science, Bahria University, Lahore Campus, Lahore, Pakistan

Junaid.jans@gmail.com, alishbaahsan127@gmail.com, Nadeem\_srwr@yahoo.com, sapisces97@gmail.com

<sup>2</sup> Department of Computer Science and Technology, Nanjing University, Nanjing, People's Republic of China  
rafiqwa.jid@smail.nju.edu.cn

<sup>3</sup> Department of Computer Science, University of Sialkot, Sialkot, Pakistan  
Sameer.malik@uskt.edu.pk

<sup>4</sup> Department of Software Engineering, University of Sialkot, Sialkot, Pakistan  
zgellani@gmail.com

<sup>5</sup> Department of Life Science, University of Management & Technology, Lahore, Pakistan  
asmairshad76@yahoo.com

**Abstract.** Classification is one of the most used machine learning technique especially in the prediction of daily life things. Its first step is grouping, dividing, categorizing, and separation of datasets based on feature vectors. Classification procedure has many algorithms, some of them are Random Forest, Naïve Bayes, Decision Tree and Support Vector Machine. Before the implementation of every technique, the model is created and then training of dataset has been made on that model. Learning the algorithm-generated model must be fit for both the input dataset and forecast the records of class label. Many models are available for prediction of a class label from unknown records. In this paper, different classifiers such as Linear SVM, Ensemble, the Decision tree has been applied and their accuracy and time analyzed on different datasets. The Liver Patient, Wine Quality, Breast Cancer and Bupa Liver Disorder datasets are used for calculating the performance and accuracy by using 10 cross-fold validation technique. In the end, all the applied algorithm results have been calculated and compared in the terms of accuracy and execution time.

**Keywords:** Classification · Naïve Bayes · DT · SVM · CM · Scattered plot

## 1 Introduction

Data mining is a process of inferring knowledge from datasets [1]. It has three main techniques known as Classification, Clustering and Association Rule Mining. Classification of datasets is one of the most complicated tasks in the present era.

Classification of data is done by three phases which are training set, determine class attributes and goal. Taken data is always divided into two parts training set and testing data and for the sake of correct prediction 10 cross fold technique is used mostly [2].

In this paper, we are going to apply classification techniques on different datasets and will measure their accuracy rate along with execution time. For this paper, four different datasets Liver Disorder, Liver Patient Detection, Wine and Breast Cancer has been taken. All these datasets are taken from the UCI Repository. The link of UCI repository is <https://archive.ics.uci.edu/ml/datasets.html>.

Applied classification algorithms are Decision Tree and it's three sub-algorithms like Support Vector Machine, Fine tree, Coarse tree, Medium tree and Logistic Regression Classifier, and its sub-algorithms like Cubic SVM, Linear SVM, Gaussian SVM, Quadratic SVM, Fine, Coarse SVM and Ensemble Classifier with its sub-algorithms like busted trees and bagged trees (Fig. 1).

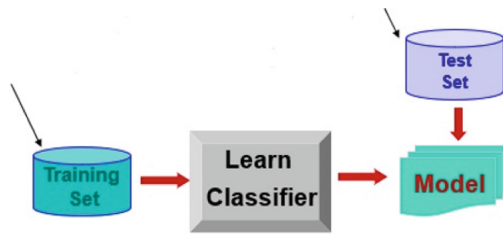


Fig. 1. Illustration of classification technique

After this the second section contains a literature review, third contains the methodology used for four used datasets, section four describes our experimental results and Sect. 5 offers the conclusion and future work.

## 2 Literature Review

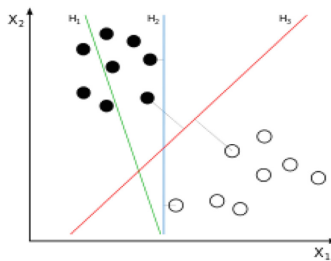
Before discussing the previous work on datasets, let's discuss the used algorithms first.

**Decision Tree:** Decision tree classification algorithms have major potential for covering and mapping problems and have not been confirmed in detail by the remote identifying public relative to more predictable pattern recognition techniques such as extreme likelihood classification [3].

**Logistic Regression Classifier:** It is Machine Learning technique used for binary classification problems. Logistic regression is entitled to the function used at the core of the method and the logistic function. It uses the equation as a representation, very much like linear regression [4]. It models the probability of the default class.

**Support Vector Machine:** It is one of the most talked and used classifier. It was extremely popular around 1990 and people are using it till date [5]. In Fig. 2 SVM is a

stimulating control and its ideas are outstandingly better. A Support Vector Machine (SVM) is a distinctive classifier officially categorized by an isolating hyper-plane.



**Fig. 2.** Support vector machine

This figure is describing the relationship between hyperplanes used in support vector machine. As all hyperplanes are distinguished as H1, H2, and H3 and separated by using lines.

**Ensemble Classifier:** Combination of at least two than two classifiers is known as outfit classifier. The gathering learning has to expand enthusiasm for the computational training society by producing better quality outcomes when contrasted with the single theory model [6]. Incremental learning calculation has a capacity to gain from the new approaching occurrences even after the classifier is as of now produced. It learns the novel data and pre-serves officially secure information without getting to the officially observed information so far [7]. One of the greatest advantages of a gradual model is it is existence efficient. Be that as it may while preparing it requires the little steady time per test, there ought to be just one example at the time in memory, it fabricates the model by examining the database and safeguard recently acquired information. The upside of utilizing steady ensemble is that it utilizes more than one classifier and different speculations which will be combined by utilizing casting a ballot rule.

For Liver Disease expectation [8] has inferred that these days, the liver is experiencing three noteworthy sicknesses Liver Cancer, Cirrhosis and Hepatitis with particular symptoms. They utilized Naïve Bayes and SVM for malady forecast. Correlation between these two calculations has been done dependent on their order precision measure. From the trial result, they have inferred that SVM is superior to Naïve Bayes as the exactness rate of SVM was 79.66% and for Naïve Bayes precision rate is 61.28%.

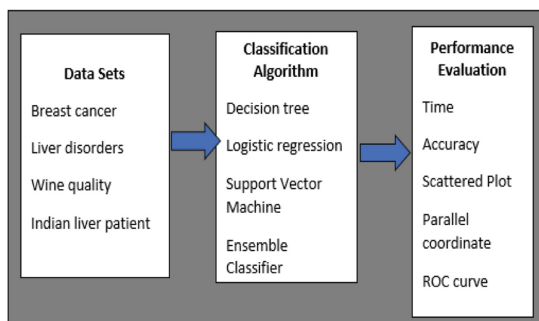
Bosom malignant growth is a standout amongst the most well-known diseases among ladies and the reason for ladies passing around the world [9]. Applied Naïve Bayes and J48 calculation on the bosom malignancy dataset for bosom disease forecast in ladies. He has presumed that Naïve Bayes is superior to J48 in light of the fact that the precision of Naïve Bayes was 97.80% and for J48 it was 96.05%.

The interest has been increased in wine from the last few years and people demands quality. To predict the wine quality data mining techniques has been applied to wine dataset by [10]. They applied the SVM and NN technique. From which he concluded that SVM is better than NN because the accuracy rate of SVM was 86.8% while for NN it was 64.3%.

The liver is the biggest internal organ of the human body with 4% of the body weight with a blood flow of 1.5 L per minute. The liver disorder is very common among people nowadays and it is very important to detect this disorder at an early stage. For this purpose [11] applied classification techniques cart and radial algorithm on dataset and concluded that radial algorithm is better than cart because in this specific scenario it was giving 70% accuracy while cart was giving almost 55% accuracy [12–15].

### 3 Methodology

In this paper, we selected different datasets and classified by using different algorithms as shown in Fig. 3:



**Fig. 3.** Evaluation of different datasets using various classification algorithms

The Table 1 is describing that we took four datasets named as breast cancer, liver disorder, wine quality and Indian liver patient on which the classification techniques like trees, logistic regression and SVM has been applied. The evaluation has been done and calculated in the form of accuracy and recalls precision etc. techniques. The datasets are Indian Liver Patient Dataset, Wine Quality, Breast Cancer and Liver Disorder to be classified. Datasets, number of instances, attributes and classes are given as follows:

**Table 1.** Datasets for classification used

Datasets	No of attributes	No of instances
Breast cancer	11	699
Liver disorders	7	345
Wine quality	12	4989
Indian liver patient	11	416

The sum total of what datasets has been taken from the UCI vault. The connection of UCI vault is <https://archive.ics.uci.edu/ml/datasets.html>. It is an open-source vault and any-one can download these datasets from here.

## 4 Results and Discussion

In this paper, we compute the investigational measures by using the performance factors such as the classification accuracy and the execution time. We draw the scattered plots, confusion matrix and ROC curve of every dataset. Let's see the experimental measure of datasets one by one:

### 4.1 Indian Liver Patient Dataset

The scattered plot (Fig. 4) is a line graph which is used to plot how much one variable is dependent on others. This scatter plot is showing the correlation between two variables known as male and female.

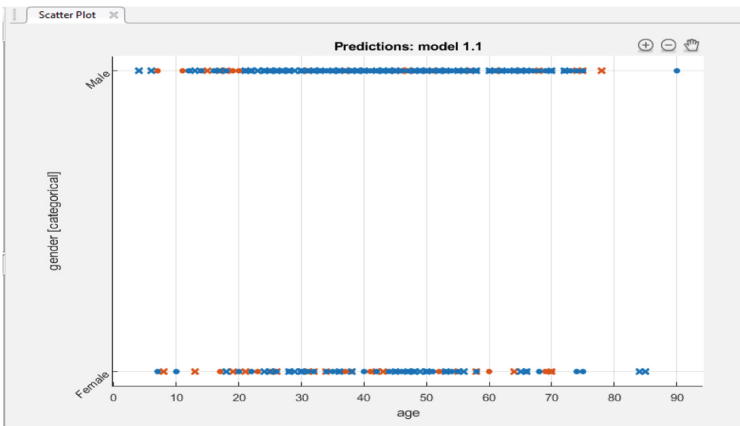


Fig. 4. Scattered plot Indian liver patient

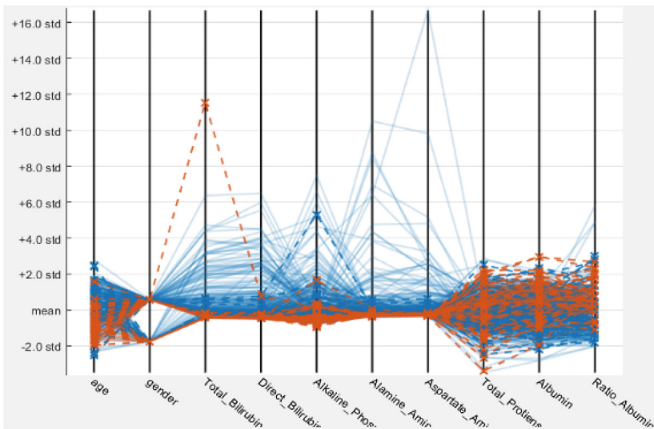
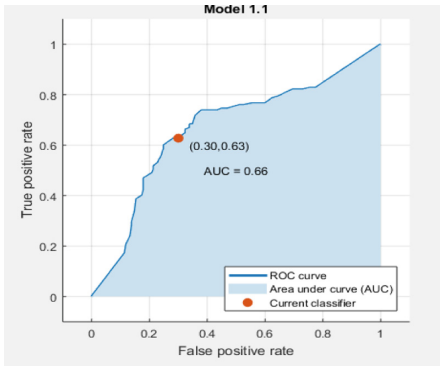
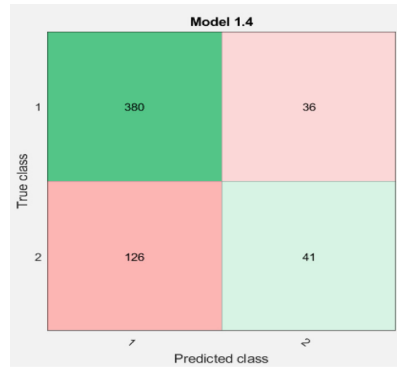


Fig. 5. Parallel coordinate plot Indian liver patient



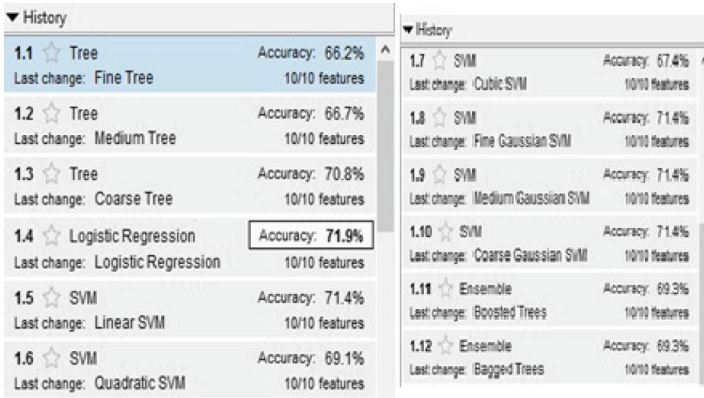
**Fig. 6.** ROC curve Indian liver patient dataset



**Fig. 7.** Confusion matrix Indian liver patient dataset

ROC curve is a measure of usefulness of a test (Fig. 5, 6 and 7). This curve is showing that the range of data is between 0 and 1 and every other thing lies between this [16–18].

As confusion matrix has four components known as True Positive, False Positive, True Negative and False Negative. This confusion matrix shows that the Liver Patient Dataset has 380 TP, 36 TF, 126 FP, and 41 FN values. Out of all applied algorithms, logistic regression gives the best accuracy of 71% (Fig. 8).



**Fig. 8.** Algorithm results Indian liver patient dataset

This plot (Fig. 9 and 10) shows that all dimensions of the dataset are dependent upon each other.

All data dimensions lie between the range of 0 to 1 where the orange dot represents the details of current applied classifier. True positive values of this dataset are 442, True negative is 16, false positive is 4 and false negative is 237, respectively (Fig. 11 and 12).

### 4.2 Breast Cancer Dataset

Out of all applied algorithms, KNN gives the best accuracy of around 96.9% (Fig. 13).

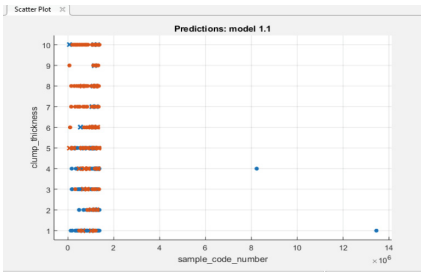


Fig. 9. Scattered plot breast cancer

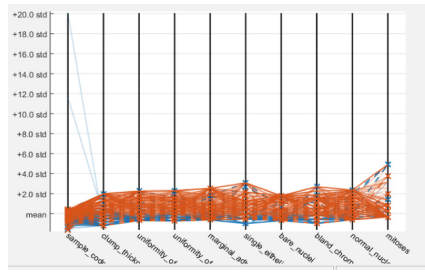


Fig. 10. Parallel coordinate plot breast cancer

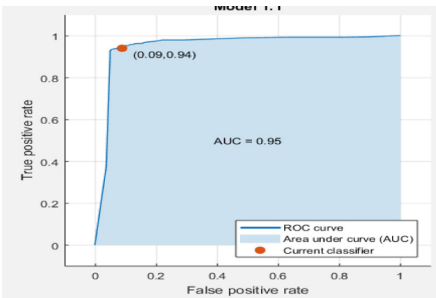


Fig. 11. ROC curve breast cancer dataset

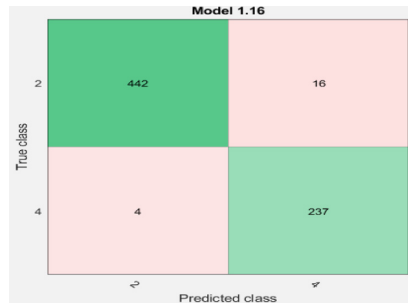


Fig. 12. Confusion matrix breast cancer dataset

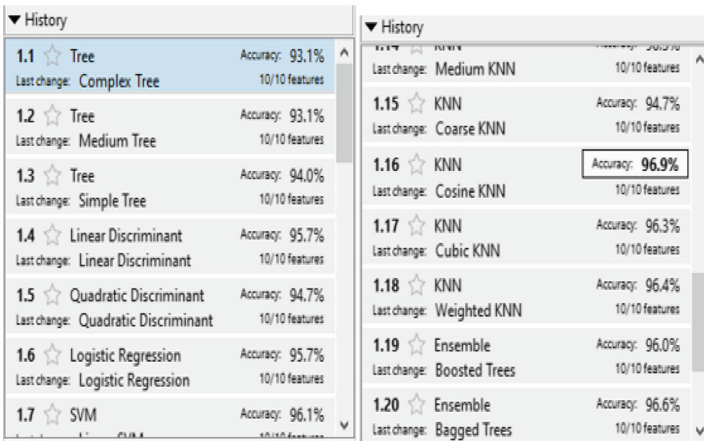


Fig. 13. Result of applied algorithm breast cancer dataset

### 4.3 Wine Quality

This plot (Fig. 14 and 15) shows that all dimensions are dependent on each other but their correlation is very high so that they are scattered in each other.

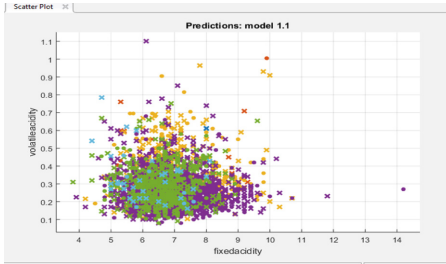


Fig. 14. Scattered plot wine quality dataset

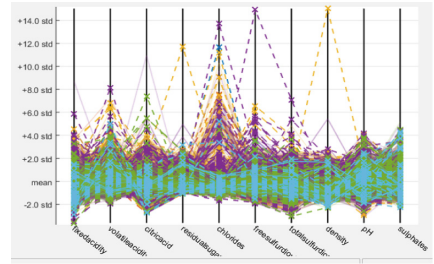


Fig. 15. Parallel coordinate plot wine quality

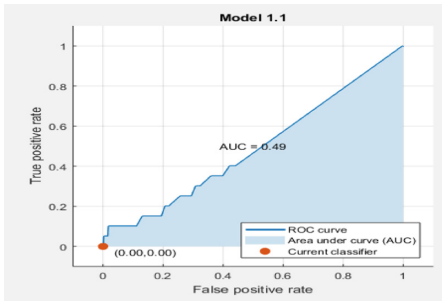


Fig. 16. ROC curve wine quality

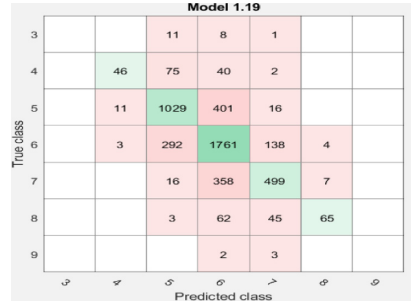


Fig. 17. Confusion matrix wine quality

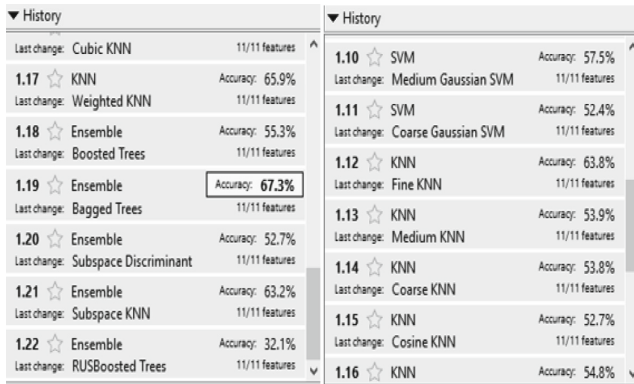


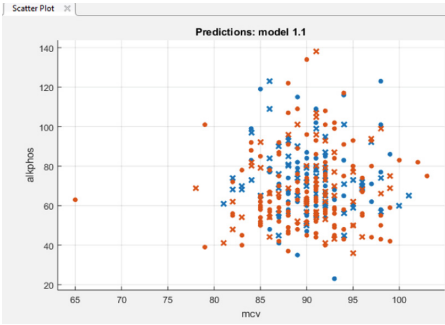
Fig. 18. Result of applied algorithm wine quality dataset

Dimensions of this set also lie between 0 to 1.

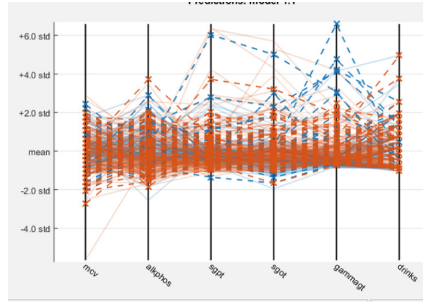


As the label classes of this dataset are not binary that's why this confusion matrix is not in its general form. It has more than one TP, TN, FP, and FN classes (Fig. 16 and 17).

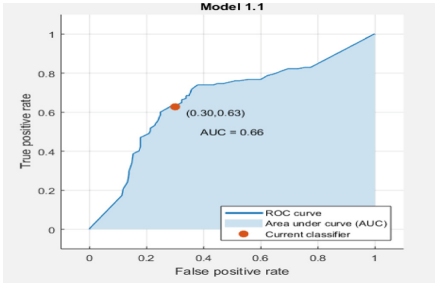
Out of all applied algorithms, the ensemble classifier gave its best accuracy at 67.3% (Fig. 18).



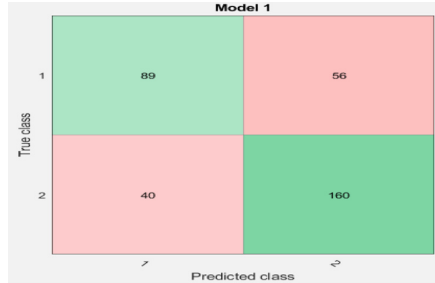
**Fig. 19.** Scattered plot Bupa liver disorder



**Fig. 20.** Parallel coordinate plot Bupa liver disorder



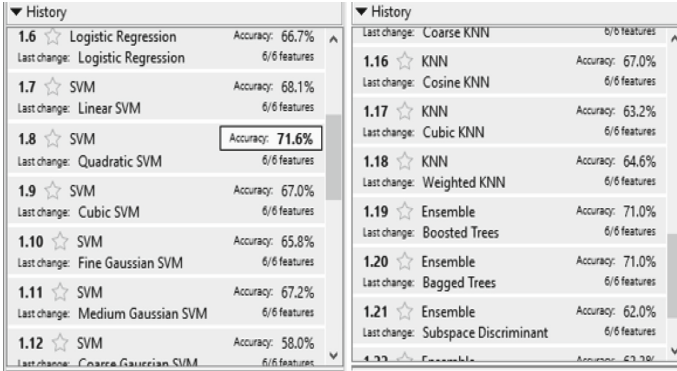
**Fig. 21.** ROC curve Bupa liver disorder



**Fig. 22.** Confusion matrix Bupa liver disorder

**4.4 Bupa Liver Disorder**

TP values of this dataset are 89, TN is 56, FP is 40 and FN are 160 respectively.



**Fig. 23.** Result of applied algorithm Bupa liver disorder

Out of all applied algorithms SVM gave its best accuracy at 71.6% (Fig. 19, 20, 21, 22 and 23 and Table 2).

**Table 2.** Best results given by algorithms on datasets:

Dataset	Best algorithm	Accuracy	Prediction speed	Training time	True positive rate for class 1 & class 2	False negative rate for class 1 & class 2	Positive prediction value for class 1 & class 2	False discovery rate for class 1 & class 2
Indian liver patient dataset	Logistic regression	72.2%	~ 4500 obs/sec	28.756 s	91%, 25%	9%, 75%	75%, 53%	25%, 47%
Breast cancer dataset	K Nearest Neighbor (KNN)	97.1%	~ 5100 obs/sec	17.239 s	97%, 98%	3%, 2%	99%, 94%	1%, 6%
Liver disorder	Quadratic SVM	72.2%	~ 1100 obs/sec	2.179 s	61%, 80%	39%, 20%	69%, 74%	31%, 26%
Wine quality	Bagged tree	69.3%	~ 7200 obs/sec	294.1 s	Class is not binary for this dataset			

## 5 Conclusion

In this paper, we analyzed the performance of 12 classifiers, all of them are described above. Performance factors such as classification accuracy and execution time are compared. From the experimental results, it is found that the different algorithms are best for different types of datasets. Like, from all algorithms logistic regression proved best for Indian liver patient dataset, KNN for Breast Cancer Dataset, Quadratic SVM for Liver Disorder and Bagged Tree for Wine Quality. In future various tests like KS, MAE, and RAE, etc. can be applied to these datasets for the sake of more accurate results.

## References

1. Ahuja, S., Angra, S.: Machine learning and its applications: a review. In: 2017 International Conference on Big data and Computational Intelligence, pp. 57–60 (2017)
2. Brownlee, J.: Machine Learning Mastery (2018). <https://machinelearningmastery.com/k-fold-cross-validation/>
3. Brodley, C., Friedl, M.A.: Decision tree classification of land cover from remotely sensed data. *Remote Sens. Environ.* **61**, 399–409 (1997)
4. Brownlee, J.: Logistic regression for machine learning, 1 April 2016. <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>
5. Brownlee, J.: Support Vector machine for machine learning, 20 April 2016. <https://machinelearningmastery.com/support-vector-machines-for-machine-learning/>
6. Woods, K., Kegelmeyer, W.P., Bowyer, K.: Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 405–410 (1997)
7. Lange, S., Zilles, S.: Formal models of incremental learning and their analysis. In: Proceedings of the International Joint Conference on Neural Networks, pp. 2691–2699 (2003)
8. Dhayanand, S., Vijayarani, S.: Liver disease prediction using SVM and Naive Bayes. *Int. J. Sci. Eng. Technol. Res. (IJSETR)* **4**(4), 816–820 (2015)
9. Borges, L.R.: Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection. In: Proceedings of XI Workshop de Visão Computacional 2010, 05th–07th October 2015, pp. 15–19 (2015)
10. Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Using data mining for wine quality assessment. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds.) *DS 2009. LNCS (LNAI)*, vol. 5808, pp. 66–79. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-04747-3\\_8](https://doi.org/10.1007/978-3-642-04747-3_8)
11. Olaniyi, E.O., Adnan, K.: Liver disease diagnosis based on neural networks, pp. 48–53 (2015)
12. Ahmed, F., et al.: Wireless mesh network IEEE 802.11 s. *Int. J. Comput. Sci. Inf. Secur.* **14** (12), 803–809 (2016)
13. Aslam, N., Sarwar, N., Batool, A.: Designing a model for improving CPU scheduling by using machine learning. *Int. J. Comput. Sci. Inf. Secur.* **14**(10), 201 (2016)
14. Bilal, M., Sarwar, N., Saeed, M.S.: A hybrid test case model for medium scale web based applications. In: 2016 Sixth International Conference on Innovative Computing Technology (INTECH), pp. 632–637 (2016)
15. Bajwa, I.S., Sarwar, N.: Automated generation of express-G models using NLP. *Sindh Univ. Res. J.-SURJ (Sci. Ser.)* **48**(1), 5–12 (2016)
16. Cheema, S.M., Sarwar, N., Yousaf, F.: Contrastive analysis of bubble & merge sort proposing hybrid approach. In: 2016 Sixth International Conference on Innovative Computing Technology (INTECH), pp. 371–375 (2016)
17. Sarwar, N., Latif, M.S., Aslam, N., Batool, A.: Automated object role model generation. *Int. J. Comput. Sci. Inf. Secur.* **14**(9), 301–308 (2016)
18. Ibrahim, M., Sarwar, N.: NoSQL database generation using SAT solver. In: 6th International Conference on Innovative Computing Technology, INTECH 2016, no. August 2016, pp. 627–631 (2016)