REVIEW OF LIVER DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM

[1]Vijay Panwar,[2]Naved Choudhary, [3]Sonam Mittal, [4]Gaurav Sahu

[1]Student, B K Birla Institute of Engineering & Technology, Pilani

[2]Student, B K Birla Institute of Engineering & Technology, Pilani

[3]Associate Professor, B K Birla Institute of Engineering & Technology, Pilani

[4]Assistant Professor,B K Birla Institute of Engineering & Technology, Pilani

Abstract: Liver Disease is the leading cause of global death that impacts the massive quantity of humans around the world. This disease is caused by an assortment of elements that harm the liver. For example, obesity, an undiagnosed hepatitis infection, alcohol misuse which is responsible for abnormal nerve function, coughing up or vomiting blood, kidney failure, liver failure, jaundice, liver encephalopathy and there are many more. Diagnosis of liver infection at preliminary stage is important for better treatment. In today’s scenario devices like sensors are used for detection of infections. Accurate classification techniques are required for automatic identification of disease samples.This disease diagnosis is very costly and complicated. Therefore, the goal of this work is to evaluate the performance of different Machine Learning algorithms in order to reduce the high cost of chronic liver disease diagnosis by prediction. In this work, we used five algorithms Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest. The performance of different classification techniques was evaluated on different measurement techniques such as accuracy, precision, recall, and specificity. We found the accuracy 74%, 72%, 72%, 71%, and 57% for SVM,DT,RF,LR and NB. The analysis result shown the SVM achieved the highest accuracy. Moreover, our present study mainly focused on the use of clinical data for liver disease prediction and explores different ways of representing such data through our analysis.

Keywords: Classification, Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Random Forest

1. INTRODUCTION

Liver is the largest internal organ in the human body,it is essential for digesting food and releasing the toxic element of the body and plays a major role in metabolism and serving several vital functions. The liver is the largest glandular organ of the body. It weighs about 3 lb (1.36 kg) .The liver's main job is to strain the blood coming from the digestive tract, before passing it to the rest of the body. The liver also detoxifies chemicals and metabolizes drugs. As it does so, the liver hides bile that ends up back in the intestines. The liver also makes proteins important for blood clotting and other functions.The liver supports almost every organ in the body and is vital for our survival. Liver disease may not cause any symptoms at earlier stage or the symptoms may be vague, like weakness and loss of energy. Symptoms partly depend on the type and the extent of liver disease. Liver diseases are diagnosed based on the liver functional test[1].

Several diseases states can disturb the liver. Some of the diseases are Wilson's disease, hepatitis (an inflammation of the liver), liver cancer, and cirrhosis (a chronic inflammation that progresses ultimately to organ failure). Alcohol alters the metabolism of the liver, which can have on the whole detrimental effects if alcohol is taken over long periods oftime. Hemochromatosis can cause liver problems [2].

Common Liver Disorder

- **Fatty liver**is a revocable condition where large vacuoles of triglyceride fat acquire in liver cells via the process of limit. It can occur in people with a high level of alcohol consumption as well as in people who never had alcohol.
- **Hepatitis**(usually caused by a virus spread by excess contamination or direct contact with infected body fluids).
- **Cirrhosis** of the liver is one of the most serious liver diseases. It is an action used to indicate all forms of diseases of the liver characterized by the significant loss of cells. The liver gradually contracts in size and becomes leathery and hard. The regenerative action continues under liver cirrhosis but the progressive loss of liver cells exceeds cell replacement.
- **Liver cancer**. The risk of liver cancer is higher in those who have cirrhosis or who had valid types of viral hepatitis; but more often, the liver is the site of secondary (metastatic) cancers spread from other organs.

2. LITERATURE REVIEW

Health care and medicine handles huge data on daily basis. This data comprises of information about the patients, diagnosis reports and medical images. It is important to utilize this information to decipher a decision support system. To achieve this it is important to discover and extract the knowledge domain from the raw data. It is accomplished by knowledge discovery and data mining (KDD) [3]. The implementation of data mining techniques is widespread in biological domain. In recent years, liver disorders have excessively increased and liver diseases are becoming one of the most fatal diseases in several countries. In this study, liver patient datasets are investigate for building classification models in order to predict liver disease. Several feature model construction and comparative analysis are implemented for improving prediction accuracy of Indian liver patients. Different studies have been conducted for classification of liver disorders, they are discussed briefly.

Classification algorithm is one of the greatest significant and applicable data mining techniques used to apply in disease prediction. Classification algorithm is the most common in several automatic medical health diagnoses. Many of them show good classification accuracy.

In another study the UCI liver dataset was used for selection of sub features based on random forest classifier with multi-layer perceptron induced [4].Different approaches for artificial intelligence for the liver patient dataset, precise predictions of liver failure were applied [5, 6, 7 and 8]. Identification of liver infection at preliminary stage is important to combat the frequency and severity deaths of patients in India. The patients must be screened based on initial symptoms for development of personalized therapy. In this study, an attempt is made for prediction of liver disease in patients using data mining techniques. Based on the review of literature, it was depicted that the past research studies have implemented different data mining techniques for classification of liver dataset. A hybrid model can be adapted to further increase the prediction accuracy of liver disease. It is followed by development of a graphical user interface would further aid the scientific community in early diagnosis of liver infection. It will provide a framework for end user application for generating promising treatment protocols.

3. METHODOLOGY

3.1. Data Collection

For this study, the Indian Liver Patient Dataset (ILPD) was selected from the UCI Machine Learning repository. It is a sample of the whole Indian population taken from the area of Andhra Pradesh. There were 583 instances based on ten different biological parameters in the dataset. Based on these criteria, the class value was stated as either yes (416 cases) or no (167 cases), reflecting the liver.

3.2. Pre-processing and Feature selection

To normalize the missing values, pre-processing techniques have been introduced. The missing values were replaced by null values along with their instances. Feature selection was followed to classify the appropriate attribute for classification. Using both filter and wrapper approaches, feature selection was carried out. The attributes with more than 70% correlation were initially excluded by correlation analysis from the dataset. The algorithm was implemented to estimate the value of different features in a dataset on the basis of random forests [9].

3.3. Randomization and splitting of dataset

To build classification models, the features selected in the preceding phase were accepted. The dataset was initially randomized to produce an arbitrary sample permutation. Splitting of the dataset into training (70 percent of the dataset) and test (30 percent) sets was followed. The training set consisted of 389 cases and the evaluation set consisted of the remaining 194 cases.

3.4. Classification algorithms

Classification algorithm is one of the greatest significant and applicable data mining techniques used to apply in disease prediction. Classification algorithm is the most common in several automatic medical health diagnoses. Many of them show good classification accuracy. Different data mining algorithms like Naïve Bayes, Decision Tree, Logistic Regression, Random forest and Support vectormachine (SVM) were implemented. The algorithms are briefly discussed below:

3.4.1. Naïve Bayes :

It is based on the Bayes theorem of conditional probability. The algorithm assumes that each attribute contributes to the total outcome independent of other attributes [10] .In machine learning we are often interested in selecting the best hypothesis (h) given data (d).In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d).One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Bayes' Theorem is stated as:

$$\mathbf{P(h|d) = (P(d|h) * P(h)) / P(d)}$$

Where,

- P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(d|h) is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(d) is the probability of the data (regardless of the hypothesis) [11].

3.4.2. Logistic Regression:

Calculated Regression was for the most part utilized in naturalresearch and applications in the mid-20th century [12]. Logisticregression can deal with any number of numerical as well asabsolute factors. In addition, it introduces a discrete parallelitem somewhere in the range of 0 and 1. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Logistic regression can be divided into following types –

Binary or Binomial

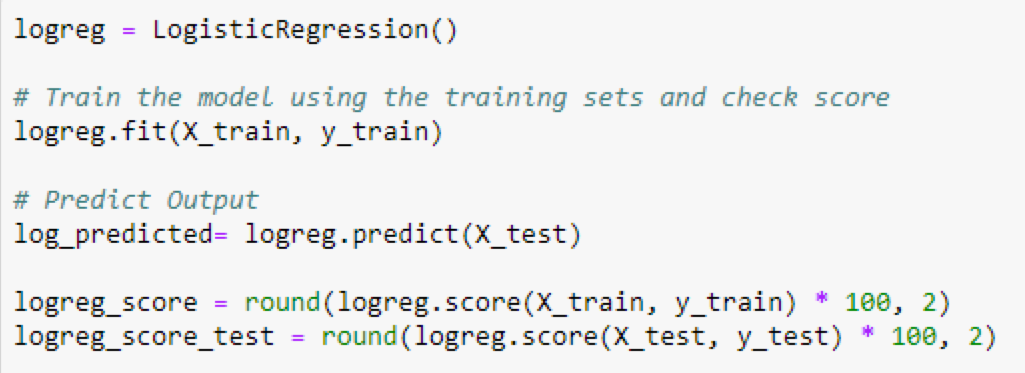
In such a kind of classification, a dependent variable will have only two possible types either 1 or 0. For example, these variables may represent success or failure, yes or no, win or loss etc.

Multinomial

In such a kind of classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.

Ordinal

In such a kind of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance. For example, these variables may represent “poor” or “good”, “very good”, “Excellent” and each category can have the scores like 0,1,2,3.

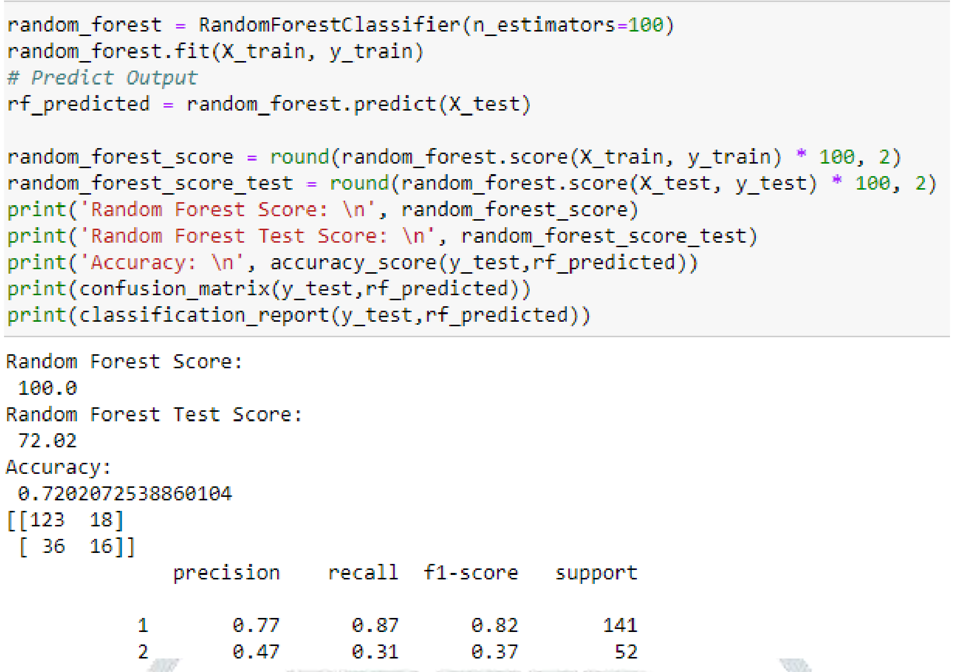


```
logreg = LogisticRegression()

# Train the model using the training sets and check score
logreg.fit(X_train, y_train)

# Predict Output
log_predicted= logreg.predict(X_test)

logreg_score = round(logreg.score(X_train, y_train) * 100, 2)
logreg_score_test = round(logreg.score(X_test, y_test) * 100, 2)
```

Fig 1. Implementation of Logistic Regression

3.4.3. Random Forest:

Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach. They are classifiers that construct decision trees for training input. Arandom value is assigned as range to feature space for splitting the tree. Based on the training ensemble class value is predicted as the modal value of distinct tree. The algorithm was implemented by Breiman [13]. The algorithm is used for ranking the features by estimating out-of-bag error. It is followed by computing of important score for each feature.



```
random_forest = RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, y_train)
# Predict Output
rf_predicted = random_forest.predict(X_test)

random_forest_score = round(random_forest.score(X_train, y_train) * 100, 2)
random_forest_score_test = round(random_forest.score(X_test, y_test) * 100, 2)
print('Random Forest Score: \n', random_forest_score)
print('Random Forest Test Score: \n', random_forest_score_test)
print('Accuracy: \n', accuracy_score(y_test,rf_predicted))
print(confusion_matrix(y_test,rf_predicted))
print(classification_report(y_test,rf_predicted))
```

```
Random Forest Score:
 100.0
Random Forest Test Score:
 72.02
Accuracy:
 0.7202072538860104
[[123  18]
 [ 36  16]]
              precision    recall  f1-score   support

           1       0.77      0.87      0.82       141
           2       0.47      0.31      0.37        52
```

Fig 2. Implementation of Random Forest

3.4.4. Support Vector Machine:

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems.They are kernel based supervised classifier developed by Vapnik [14].The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. In the SVM algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

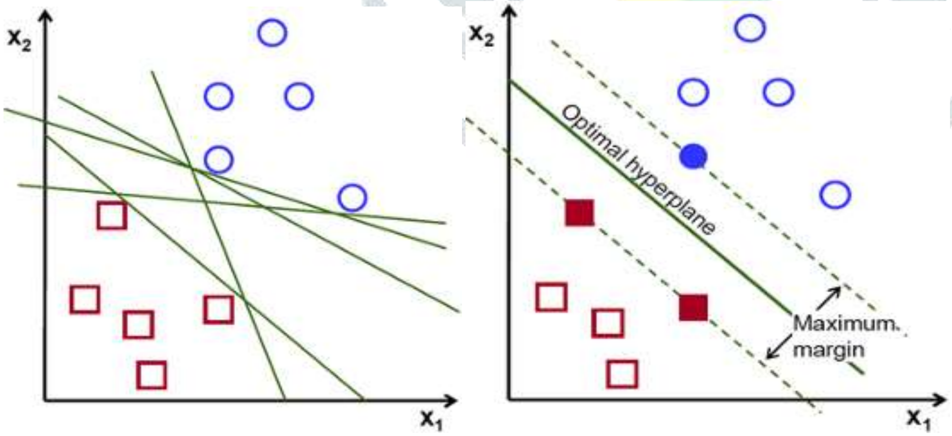
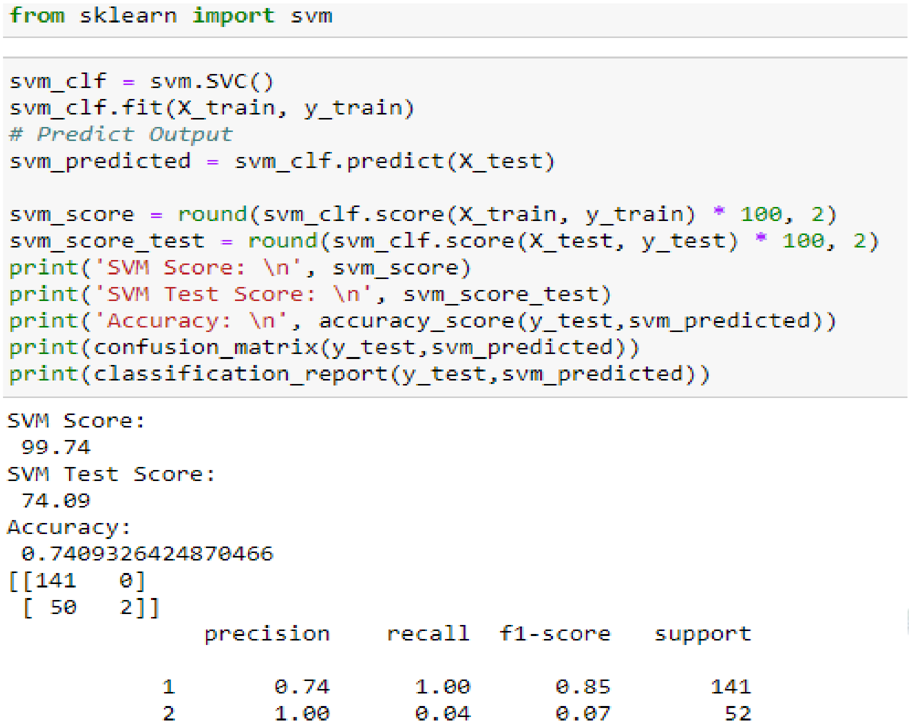


Fig 3. Possible hyperplanes in SVM

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features.



```
from sklearn import svm
```

```
svm_clf = svm.SVC()
svm_clf.fit(X_train, y_train)
# Predict Output
svm_predicted = svm_clf.predict(X_test)

svm_score = round(svm_clf.score(X_train, y_train) * 100, 2)
svm_score_test = round(svm_clf.score(X_test, y_test) * 100, 2)
print('SVM Score: \n', svm_score)
print('SVM Test Score: \n', svm_score_test)
print('Accuracy: \n', accuracy_score(y_test,svm_predicted))
print(confusion_matrix(y_test,svm_predicted))
print(classification_report(y_test,svm_predicted))
```

```
SVM Score:
 99.74
SVM Test Score:
 74.09
Accuracy:
 0.7409326424870466
[[141   0]
 [ 50   2]]
              precision    recall  f1-score   support

           1       0.74      1.00      0.85       141
           2       1.00      0.04      0.07        52
```

Fig 4. Implementation of SVM

3.4.5. Decision Tree:

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter.Decision Tree calculation has a place with the supervised learning algorithms [15]. It is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels.The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. The paths from root to leaf represent classification rules. Below diagram illustrate the basic flow of decision tree for decision making with labels (Rain(Yes), No Rain(No)).

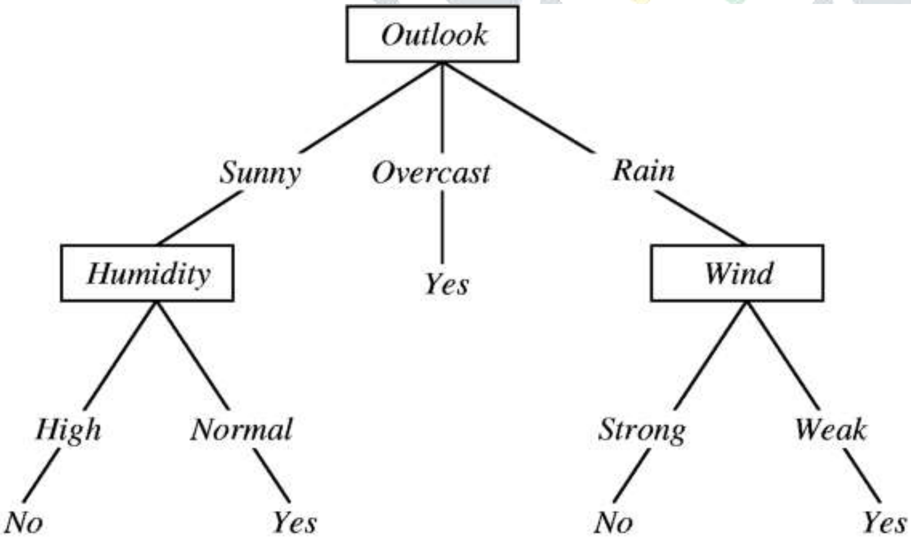
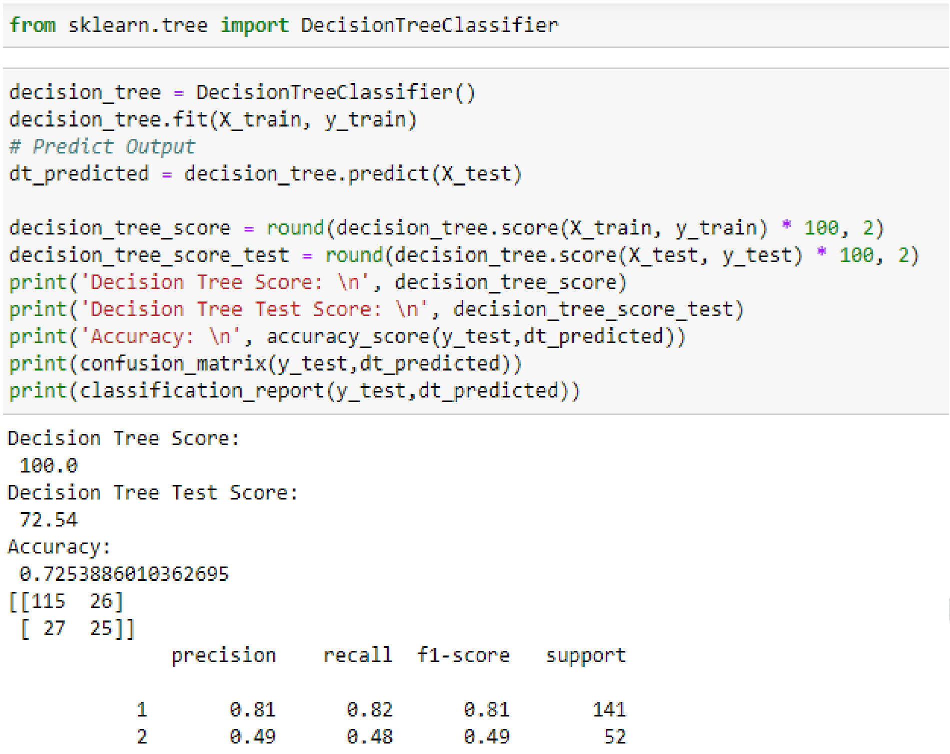


Fig 5. Example of Decision Tree

Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.



```
from sklearn.tree import DecisionTreeClassifier
```

```
decision_tree = DecisionTreeClassifier()
decision_tree.fit(X_train, y_train)
# Predict Output
dt_predicted = decision_tree.predict(X_test)

decision_tree_score = round(decision_tree.score(X_train, y_train) * 100, 2)
decision_tree_score_test = round(decision_tree.score(X_test, y_test) * 100, 2)
print('Decision Tree Score: \n', decision_tree_score)
print('Decision Tree Test Score: \n', decision_tree_score_test)
print('Accuracy: \n', accuracy_score(y_test,dt_predicted))
print(confusion_matrix(y_test,dt_predicted))
print(classification_report(y_test,dt_predicted))
```

```
Decision Tree Score:
 100.0
Decision Tree Test Score:
 72.54
Accuracy:
 0.7253886010362695
[[115  26]
 [ 27  25]]
              precision    recall  f1-score   support

           1       0.81      0.82      0.81       141
           2       0.49      0.48      0.49        52
```

Fig 6. Implementation of Decision Tree

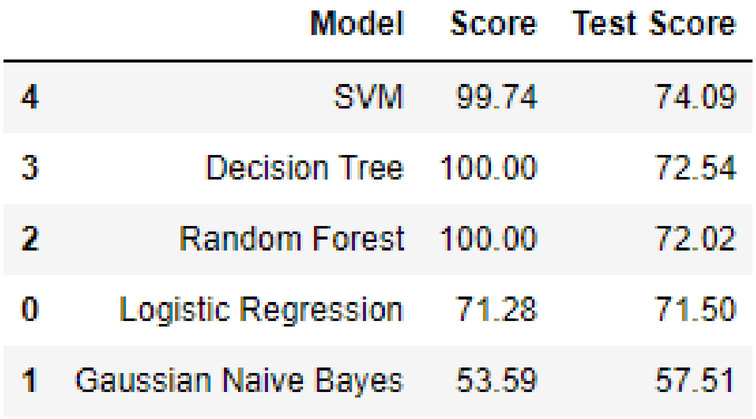
4. RESULTS AND DISCUSSION

4.1. Dataset

It is a sample of the entire Indian population collected from Andhra Pradesh region. The dataset comprised of 583 instances based on ten different biological parameters. The class value was reported based on these parameters as either yes (416 cases) or no (167 cases) that represent the liver infection. The patients were described as either '1' or '2' on the basis of liver disease.

4.2. Application of classification algorithms

Five algorithms namely Naïve Bayes, Logistic Regression, Decision Tree, Random forest and SVM were implemented for classification of the Indian Liver Patient Dataset. The models were generated for the training set and evaluated on the test set. Based on the accuracy of prediction it was observed that SVM achieved an accuracy of 74.09%.



	Model	Score	Test Score
4	SVM	99.74	74.09
3	Decision Tree	100.00	72.54
2	Random Forest	100.00	72.02
0	Logistic Regression	71.28	71.50
1	Gaussian Naive Bayes	53.59	57.51

5. REFERENCES

[1] S. Karthik, A. Priyadarishini and J. Anuradha and B. K. Tripathy,"Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types", Pelagia Research Library,Advances in Applied Science Research, 2011.

[2] https://www.medicinenet.com/liver_disease/article.htm

[3] Hastie T, Robert, T, Jerome F (2009). The Elements of Statistical Learning: Data mining, Inference and Prediction. Springer. 485–586.

[4] Jankishran Pahariyavohra, Jagdeesh makhijani and sanjay patsariya. Liver patient classification using intelligence techniques. International journal of advanced research in computer science and software engineering. 4(2): 295-299.

[5] Anju Gulia, Dr. Rajan Vohra, Praveen Rani (2014). Liver patient classification using intelligent techniques. International Journal of Computer Science and Information Technologies. 5(4): 5110-5115.

[6] Prasad Babu et. al. (2014). An implementation of hierarchical clustering on Indian Liver Patient Dataset. International Journal of Emerging Technologies in Computational and Applied Sciences. 8(6): 543-547.

[7] Hoon Jin, Seoungcheon Kim, Jinhong Kim (2014). Decision factors on effective liver patient data prediction. International Journal of Bio-Science and Bio-technology. 6(4): 167-178.

[8] Esraa M Hashem, Mai S Mabrouk (2014). A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis. American Journal of Intelligent Systems. 4(1): 9-14.

[9] Miron Kursa, Witold Rudnicki (2010). Feature selection with Boruta package. Journal of Statistical Software. 36(11): 1-13.

[10] Kim Larsen (2005). Generalized Naïve Bayes Classifier. SIGKDD Explorations. 7(1): 76-81.

[11] https://machinelearningmastery.com/naive-bayes-for-machine-learning/

[12] Logistic Regression, Retrieve from:HTTPS://WWW.SAEDSAYAD.COM/LOGISTIC _REGRESSION.HTM, LAST Accessed: 5 Octobor,2019

[13] Breiman Leo (2001). Random Forests. Machine Learning 45 (1): 5–32.

[14] Cortes C, Vapnik V (1995). Support-vector networks. Machine Learning 20 (3): 273.

[15] Decision Trees, Retrieve from:https://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/, Last Accessed: 5 Octobor,2019.