***SMARTBRIDGE SUMMER INTERNSHIP PROGRAM-2019***

***TEAM NAME: CRUSADERS***

***TEAM MEMBERS:***

*1. ATMAKURI MRUDURAVALI*

*2. BOLUKONDA NAVYA*

*3. VANCHA NAMITHA*

*4. GOVINDULA ANKITH*

Liver Patient Analysis

**INTRODUCTION:**

Liver disease is any trouble of liver function that causes sickness. The liver is responsible for many dangerous functions within the body and should it become diseased or damaged, the loss of those functions can cause significant injury to the body. Liver disease is also referred to as hepatic disease. Liver disease is a large term that covers all the potential problems that cause the liver to fail to perform its designated functions. Usually, more than 75% or three quarters of liver tissue needs to be affected before a decrease in function occurs.

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patient’s survival rate. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. Therefore, developing a machine that will enhance in the diagnosis of the disease will be of a great advantage in the medical field. These systems will help the physicians in making accurate decisions on patients and also with the help of Automatic classification tools for liver diseases (probably mobile enabled or web enabled), one can reduce the patient queue at the liver experts such as endocrinologists.

Classification techniques are much popular in medical diagnosis and predicting diseases. The main objective of this research is to use classification algorithms to identify the liver patients from healthy individuals. We used four classification algorithms namely: Logistic Regression, K Nearest Neighbors, Random Forest and Decision Tree have been considered for comparing a performance based on the liver patient data.

Further, the model with the highest accuracy is implemented as a user friendly Graphical User Interface (GUI) using IBM Watson Studio. The GUI can be readily utilized by doctors and medical practitioners as a screening tool for liver disease.

The dataset used is The Indian Liver Patient Dataset (ILPD) which was selected from UCI Machine learning repository for this study. It is a sample of the entire Indian population and comprises of 583 patient data.

**OBJECTIVES OF RESEARCH:**

The main objectives are:

* To estimate the number of patients suffering with liver disease.
* The poor performance in the training and testing of the liver disorder dataset as resulted from an insufficient in the dataset. Therefore, Sug suggested a method based on oversampling in minor classes in order to compensate for the insufficiency of data effectively. The author considered two algorithms of decision tree for the research work.
* These previously designed systems have been adequate but more works has to be done on their recognition rate for better accuracy in the diagnosis of the liver disease. In this case, this will make the diagnoses of the liver diseases to be more effective and efficient by preventing misdiagnosis of the liver disorder. Developing a system with better performance than the previous works will help in preventing misdiagnosis of the disease and help in providing the best and required medication for the patient.

**PROBLEM STATEMENT:**

Liver Patient Analysis is done to identify the patient is liver diseased or not based on liver fluids.

**REVIEW OF LITERATURE:**

Paul Mangiameli et al., proposed model selection affects the decision support systems accurately. In their model selection, how to affects the accuracy of decision support system hydrides by single model and ensembles. They proposed single model is not more accurate than ensembles. Ahmed M. Hashem et al., proposed to predict Liver Cirrhosis or fibrosis single stage classification model and multistage classification model. In their model based on Decision Tree, Neural Network, Nearest Neighborhood clustering and Logistic Regression

Ziol.M et al., proposed to evaluated liver fibrosis with chronic hepatitis C for patients using liver stiffness measurement (LSM).Z. Jiang.Z., proposed for discovering the corresponding degree of fibrosis by support vector machine (SVM)

Kemal Polat et al., proposed resource allocation mechanism of AIRS was changed with a new one decided by Fuzzy-Logic. This approach called as Fuzzy- AIRS was used as a classifier in the diagnosis of Liver Disorders. In this Classification accuracies were evaluated by comparing them with reported classifier’s accuracy, time and number of resources.

Piscaglia et al., proposed to predict Liver cirrhosis and other liver-related diseases used by Artificial neural network. Dong-Hoi Kim et al., proposed machine learning technique and decision tree(C4.5).In this method is used for to predict the susceptibility to two liver diseases such as chronic hepatitis and cirrhosis from single nucleotide polymorphism(SNP) data . They also used to identify a set of SNPs relevant to those diseases.

Anh Pham, developed optimizing the classification accuracy when analyzing some medical datasets. This proposed work done by new meta-heuristic approach, called the Homogeneity-Based Algorithm (or HBA).This approach used to predict error rates and associated penalty costs. These costs may be dramatically different in medical applications as the implications of having a false-positive and a false-negative case may be tremendously different.

Rong-Ho Lin proposed to predict accuracy of liver disease using case-based reasoning (CBR) and classification and regression tree (CART) approach. He also integrates CART and CBR for the diagnosis of liver diseases. In this model included two major steps. (1) CART To diagnose whether a patient suffers from liver disease using CART. (2)To predict which types of Liver disease affected for patients using CBR. He also , proposed to determine whether patients suffer from liver disease or not using case-based reasoning, artificial neural networks and analytic hierarchy methods . They also predict which types of liver disease suffered human body.

Chun-Ling Chuang et al., proposed to diagnosis early Liver disease and predict classification accuracy by integrated case- based reasoning into classification and regression tree, back-propagation neural network (BPN), discriminatory analysis and logistic regression of classification methods in data mining techniques. In their methods used a ten-fold cross-validation to select a best.

**DATA COLLECTION:**

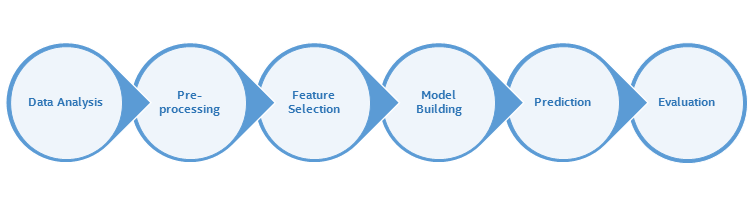
The Indian Liver Patient Dataset comprised of 11 different attributes of 583 patients. The patients were described as either 1 or 2 on the basis of liver disease. The detailed description of the dataset is shown in Table. The table provide details about the attribute and attribute type. As clearly visible from the table, all the features except Gender are real valued integers. The feature Gender is converted to numeric value (0 and 1) in the data pre-processing step.

|  |  |  |
| --- | --- | --- |
| **S.No** | **ATTRIBUTES** | **ATTRIBUTE TYPE** |
| 1. | Age | Numeric |
| 2. | Gender | Nominal |
| 3. | Total Bilirubin | Numeric |
| 4. | Direct Bilirubin | Numeric |
| 5. | Alkaline Phosphatase | Numeric |
| 6. | Alamine Aminotransferase | Numeric |
| 7. | Aspartate Aminotransferase | Numeric |
| 8. | Total Protiens | Numeric |
| 9. | Albumin | Numeric |
| 10. | Albumin and Globulin Ratio | Numeric |
| 11. | Dataset | Numeric(1,2) |

**METHODOLOGY:**

Highest accuracy is found in DECISION TREE algorithm among all the four classification algorithms.

Decision Trees are a type of supervised Machine Learning where the data is continuously split according to a certain parameter.The tree can be explained by two entities,namely decision nodes and and leaves.The leaves are the decisions or the final outcomes.And the decision nodes are where the data is split.



**DATA ANALYSIS:**

Before performing any processing on the available data, a data analysis is recommended. This process includes visualization of the data, identifying the outliers, and skewed predictors. These tasks help to inspect the data and thereby spot the missing values and irrelevant information in the dataset. A data cleanup process is performed to handle these issues and to ensure data quality. Gaining a better understanding of the dataset helps to identify useful information and supports decision making.

The Indian Liver Patient dataset consists of 583 records in which 416 are records of people with liver disease, and the remaining are records of people without any liver disease. The dataset has 10 features in which there is only one categorical data (Gender of the patient). The endmost column of the dataset represent the class in which each sample falls (liver patient or not). A value of 1 indicates the person has liver disease and a 2 indicates the person does not have the disease. There is no missing value in the dataset.

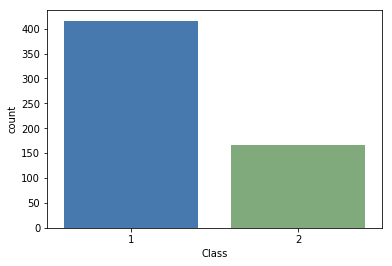


Figure1:Visualization: liver patient dataset class.

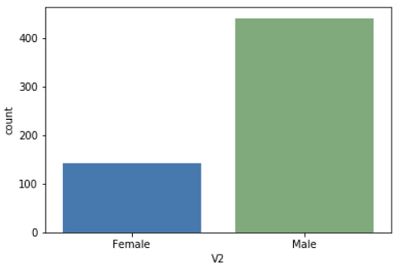
  
 Figure2: Visualization: male and female population.

Figure1 shows a visualization of the number of patients with liver disease and patients with no liver disease, whereas figure2 represents a visualization of the male and female population in the dataset. Histograms of numerical variables are represented by figure 3.

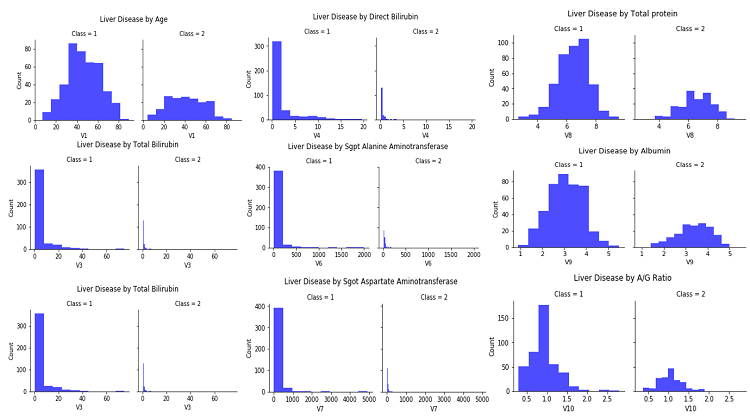


Figure 3. Visualization of numerical variables in the dataset.

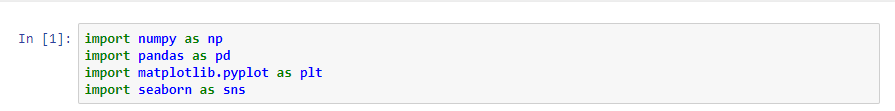
## DATA PRE-PROCESSING:

Some datasets contain irrelevant information, noise, missing values, and so on. These datasets should be handled properly to get a better result for the data mining process. Data preprocessing includes data cleaning, preparation, transformation, and dimensionality reduction, which convert the raw data into a form that is suitable for further processing.

The major objective of the experiment is to show the effect of various preprocessing methods on the dataset prior to classification. Different classification algorithms were applied to compare the results.

Some of the preprocessing includes:

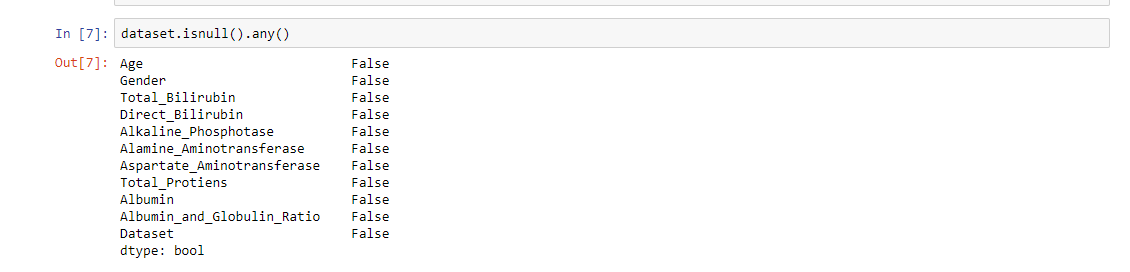
Step1:Importing Libraries.



Step2:Uploading Dataset(.csv)

C:\Users\DELL\OneDrive\Pictures\Screenshots\2019-06-21 (7).png

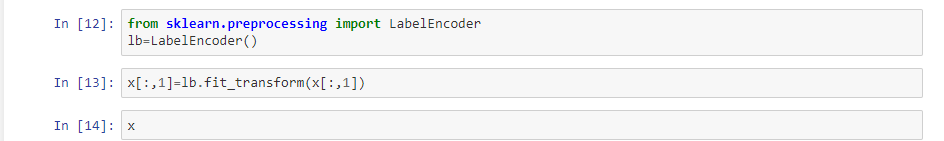
Step3:Checking missing values in the dataset



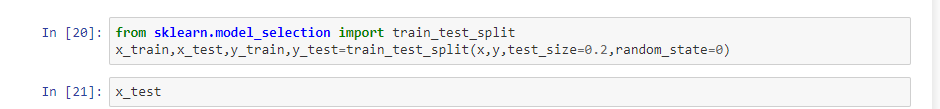
Step4:Seperating dependent variables and independent variables.



Step5:Perform Label Encoding.This method converts the categorical data into a numerical form. It is used when the feature column has a binary value. In the liver patient dataset, column(gender) has the values male/female, which is binary encoded into “0” and “1”.



Step6:Splitting train and testing data.



**FEATURE SELECTION:**

Feature extraction model for liver dataset and applied an improve probability in many medical application such as training artificial neural networks, linear constrained function optimization, wireless network optimization, data classification, and many other areas where GA can be applied. Computation in PSO is based on a swarm of processing elements called particles in which each particle represent a candidate solution.

**STATISTICAL TECHNIQUES AND DATA VISUALIZATION:**

In this, we used four classification algorithms:

1. Decision Tree
2. Random Forest
3. Logistic Regression
4. K Nearest Neighbors

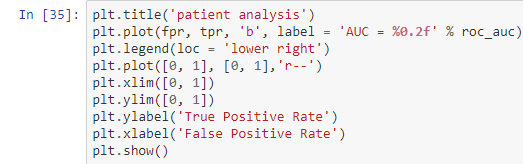
**DECISION TREE:**

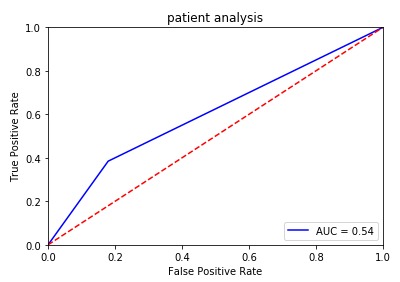
A decision tree is a Supervised machine learning algorithm. As the name goes it is a tree-like structure with its root node on the top. This algorithm is one of the most popular techniques used for both classification and regression tree(CART) tasks. Mostly it is used for classification as it doesn’t work that well for regression problems.

If you can interpret the relationship between the target variables and the input variables, seeing the plots then you can straight away go for Linear regression. But in case there exists a non-linear or some other complex relationship which you are not able to visualize seeing the data or the plot then decision tree should be preferred. However, cross-validation can only confirm it.

### ****Terminologies related to Decision Tree:****

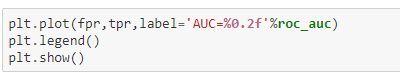
* **Root node:** the topmost tree node which divides into two homogeneous sets.
* **Decision node:** a sub-node which further splits into other two sub-nodes.
* **Terminal/Leaf node:** the lowermost nodes or the nodes with no children that represents a class label (decision taken after computing all attributes)
* **Splitting**: dividing a node into two or more nodes. The splitting technique results in fully grown trees until the criteria of a class attribute are met. But a fully grown tree is likely to over-fit the data which leads to poor accuracy on unseen observations. This is when Pruning comes into the picture.
* **Pruning**: Process of reducing the size of the tree by removing the nodes which play a minimal role in classifying an instance without reducing the predictive accuracy as measured by a cross-validation set.
* **Branch:** a sub-section of a decision tree is called a branch.

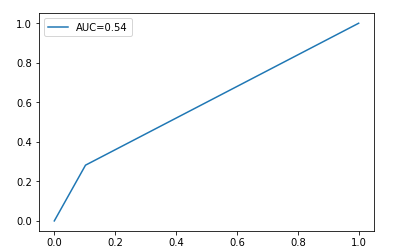




**RANDOM FOREST:**

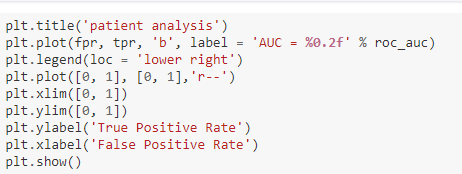
Random forests is a machine learning regression method for classification that drive by constructing liver data into a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. It is unexcelled in accuracy among current algorithms. It output classification efficiently on large liver dataset. It can handle thousands of input attributers without variable deletion. It gives estimates of what variables are important in the classification. Random Forests grows many classification trees. To classify a new liver object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and says the tree "votes" for that class. The forest chooses the classification having the most votes.

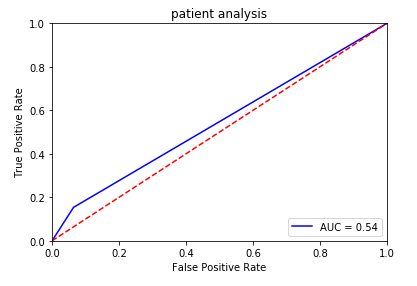




**LOGISTIC REGRESSION:**

Logistic regression is one of the simpler classification models. Because of its parametric nature it can to some extent be interpreted by looking at the parameters making it useful when experimenters want to look at relationships between variables. A parametric model can be described entirely by a vector of parameters = (0, 1... p). An example of a parametric model would be a straight-line y = kx + m where the parameters are k and m. With known parameters the entire model can be recreated. Logistic regression is a parametric model where the parameters are coefficients to the predictor variables written as 0 +1 +X1 + ...PXp Where 0 is called the intercept. For convenience we instead write the above sum of the parameterized predictor variables in vector form as X. The name logistic regression is a bit unfortunate since a regression model is usually used to find a continuous response variable, whereas in classification the response variable is discrete. The term can be motivated by the fact that we in logistic regression found the probability of the response variable belonging to a certain class, and this probability is continuous.

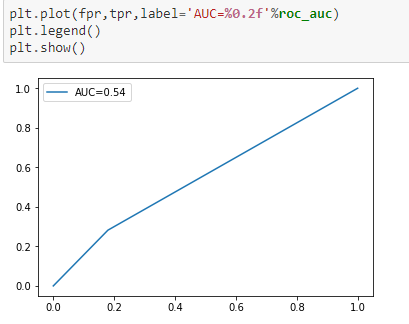




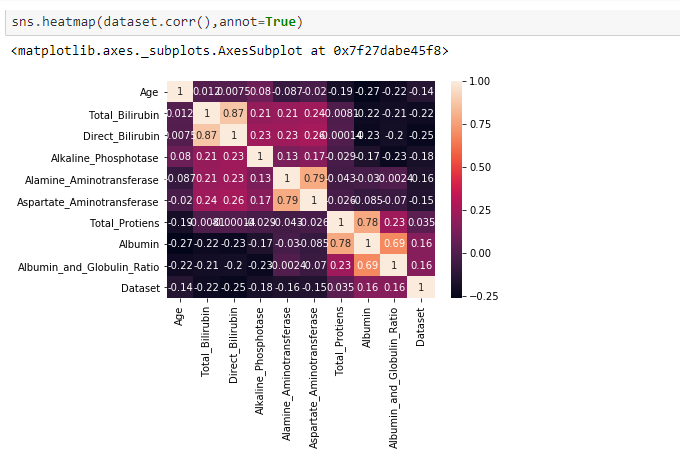
**K NEAREST NEIGHBORS:**

This section describes the implementation details of KNN algorithm. The model for KNN is the entire training dataset. When a prediction is required for a unseen data instance, the KNN algorithm will search through the training dataset for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance.

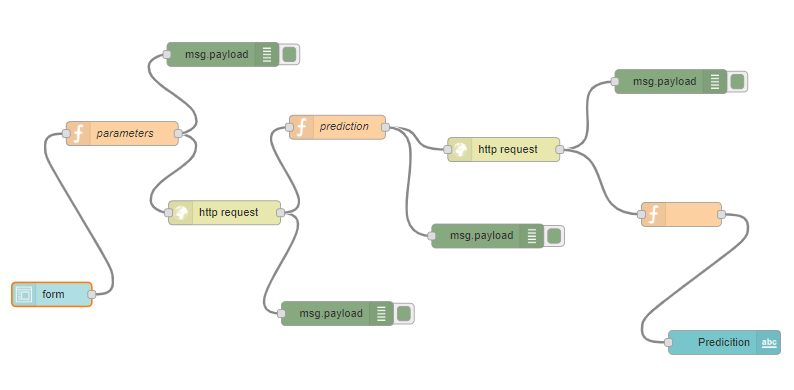
The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. Other types of data such as categorical or binary data,Hamming distance can be used. The KNN algorithm is belongs to the family of instance-based, competitive learning and lazy learning algorithms. Instance-based algorithms are those algorithms that model the problem using data in-stances (or rows) in order to make predictive decisions. The KNN algorithm is an extreme form of instance-based methods because all training observations are retained as part of the model. It is a competitive learning algorithm, because it internally uses competition between model elements (data instances) in order to make a predictive decision. The objective similarity measure between data instances causes each data instance to compete to win or be most similar to a given unseen data instance and contribute to a prediction.



**HEAT MAP:**

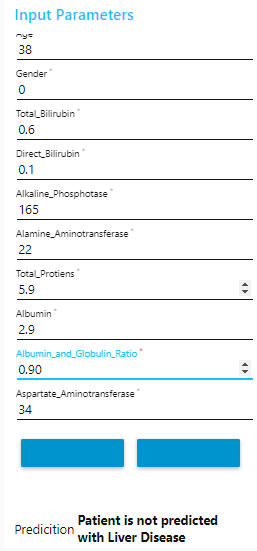
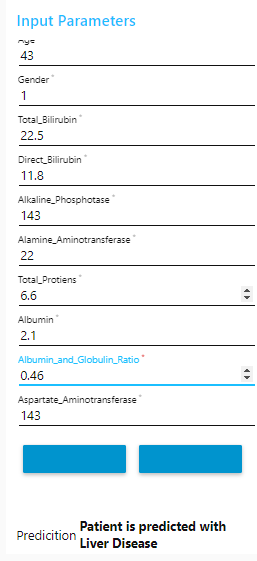
****

**MODEL DEPLOYMENT:**

****

**DEVELOPMENT OF GUI:**

The model that gave the maximum accuracy for the test data was the Decision tree. So,Decision tree is used for creating the GUI. The GUI is created using Node-Red editor.The GUI contains input fields for all attributes in the dataset. The system will predict whether the patient has liver disease or not based on the trained model. The GUI will be a useful tool for medical staff in the early diagnosis of liver disease in patients. A picture of the developed GUI is shown below.

**DATA MODELING USING SUPERVISED ML TECHNIQUES:**

**MODEL BUILT USING LOGISTIC REGRESSION:**

STEP1:Importing Logistic Regression package.

STEP2:Predicting values and fitting model by using x\_train and y\_train parameters.

STEP3:Finding accuracy score using y\_test and y\_predict.

**MODEL BUILT USING DECISION TREE:**

STEP1:Importing DecisionTreeClassifier package.

STEP2:Predicting values and fitting model by using x\_train and y\_train parameters.

STEP3:Finding accuracy score using y\_test and y\_predict.

**MODEL BUILT USING RANDOM FOREST:**

STEP1:Importing RandomForestClassifier package.

STEP2:Predicting values and fitting model by using x\_train and y\_train parameters.

STEP3:Finding accuracy score using y\_test and y\_predict.

**MODEL BUILT USING K NEAREST NEIGHBORS:**

STEP1:Importing KNeighborsClassifier package.

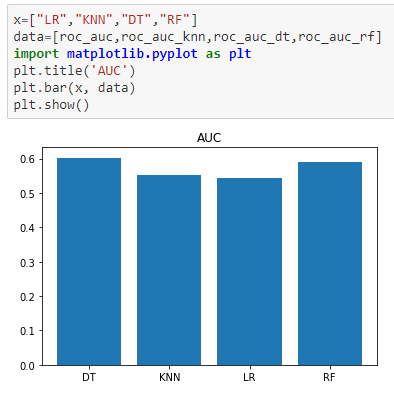
STEP2:Predicting values and fitting model by using x\_train and y\_train parameters.

STEP3:Finding accuracy score using y\_test and y\_predict.

**COMPARITIVE ANALYSIS OF CLASSIFICATION ALGORITHMS:**

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm Name | Accuracy score | ROC | CM |
| Decision Tree | 0.6752136752136 | 0.6025641025641 | [[64, 14],  [24,15]] |
| Logistic Regression | 0.6837606837606 | 0.5512820512820512 | [[74, 4],  [33, 6]] |
| Random Forest | 0.6523076923076 | 0.5897435897435 | [[70, 8],  [28,11]] |
| KNN | 0.6410256410256 | 0.5512820512820 | [[64, 14],  [28, 11]] |

**ROC\_AUC curves of four classification algorithms using bar plot:**



**FINDINGS AND SUGGESTIONS:**

From the above classification algorithms,we have calculated the ROC\_AUC curves by using logistic regression ,decision tree ,random forest, KNN. Decision tree is slightly better than the other algorithms used.

To improve the accuracy of the model,

1. use better classification algorithms.
2. High quality dataset.

**To avoid liver diseases:**

Maintain a healthy weight.

Eat a balanced diet

Exercise regularly.

Avoid toxins.

Use alcohol responsibly.

Avoid the use of illicit drugs.

Avoid contaminated needles.

Get medical care if you’re exposed to blood.

Don’t share personal hygiene items.

Wash your hands.

Get vaccinated.

**CONCLUSION:**

In this project, we have proposed methods for diagnosing liver disease in patients using machine learning techniques. The four machine learning techniques that were used include Logistic Regression, KNN, Random Forest and Decision tree. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. Decision tree was the model that resulted in the highest accuracy. A GUI, which can be used as a medical tool by hospitals and medical staff was implemented using IBM Watson Studio.