SHUBHANGI SAILEJA RABICHANDRA GARNAIK

MAY 4, 2022, CSE 537

Final Project Report

**Liver Disease Prediction using Machine Learning**

**Introduction:**

Owing to the pandemic over the last three years, there has been a drastic shift of number of employees working from home and cutting back on travel to work. This in turn has led to an even more stagnant life style and sedentary work space giving rise to lifestyle disease like Liver (hepatic) disease. Due to the rise in the risk of liver disorder, it has become a need of the hour to have quick and efficient diagnostic systems in place. However, detecting trauma or abscess in the liver is a strenuous task, this is because liver is an internal organ and requires experienced physicians and expensive testing equipment. Hence to help the medical practitioners machine learning techniques have been used in healthcare sector to provide early diagnosis of diseases. Preventive measures applied during the early stage of liver disorder can help revert the condition of the damaged part. Hence, using machine learning algorithms with higher accuracy can help provide immediate diagnostic response.

**Problem Definition:**

A prediction system has been developed for detecting liver disorder based on a set of patients’ datasets (Indian Liver patient dataset) on machine learning algorithms.

**Dataset (Input):**

The dataset from UCI Machine Learning Repository shall be used for predicting the liver disorders. No new data shall be collected; since the data available in the Indian Liver Patient Dataset (ILPD), which exists in UCI Machine Learning Repository will be used.  
This dataset has 567 instances and 10 attributes.

**Accurately trained model (Output):**

The above-mentioned dataset is refined by cleaning and processing it, has been divided into training and testing set and has been used to train the machine learning models so as to detect accurately and precisely if user has liver disorder.

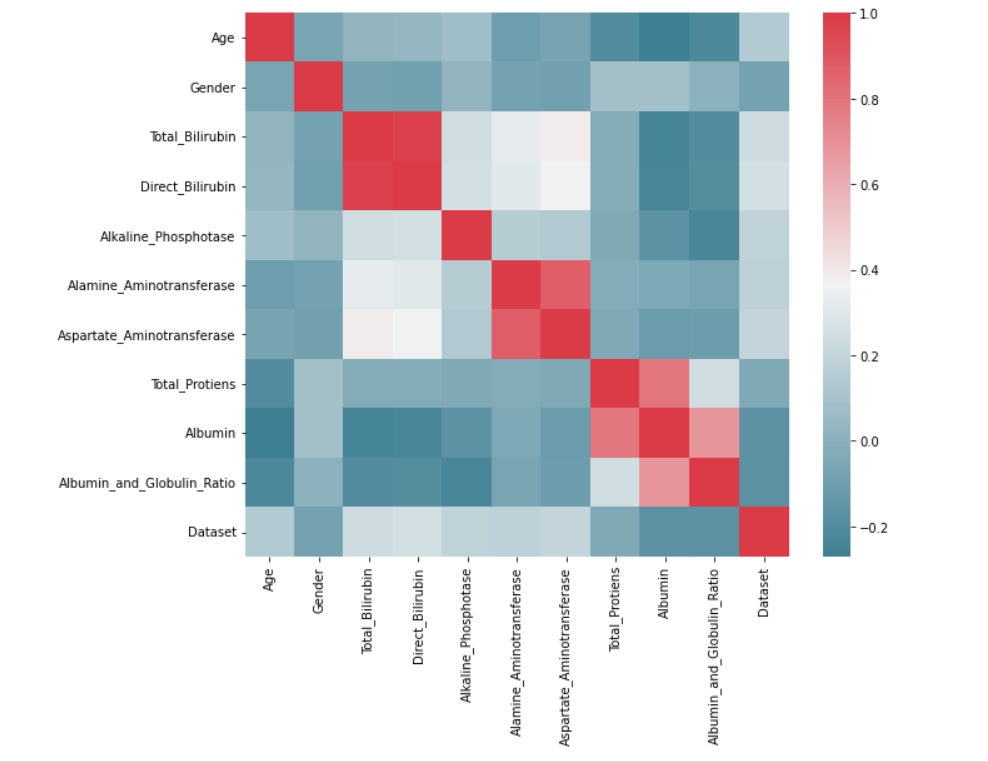
This is an interesting problem as it is required in today’s fast paced world and increasing lifestyle diseases. The dataset was small to train a model hence it was challenging to find ways to get more data and train the models well.

**Algorithm used:**

Set of machine learning algorithms have been used to build an accurate liver disease prediction system.

Before training the models, the dataset was analyzed and preprocessed to improve the accuracy of the model by refining the input data:

1. Dataframe.isnull().sum() - Checked for sum of null values in each row of the dataset.
2. Dataframe.fillna() – Filled rows with null value with the mean of rest of the data from that row.
3. Dataframe.duplicated().sum() – Checked for duplicate values in the dataset and removed all except its first occurrence.
4. Removed Outliers: some of the features had abrupt large values which would impact the accuracy of the trained model; hence, those values were removed.
5. Converted categorical data to binary data – for Gender column.
6. Dataframe.corr() – It depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.
7. Sns.heatmap() – use to plot the correlation matrix

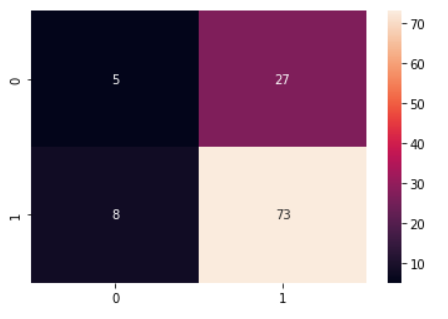
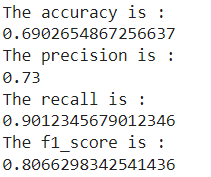


Correlation matrix of the dataset used in the project

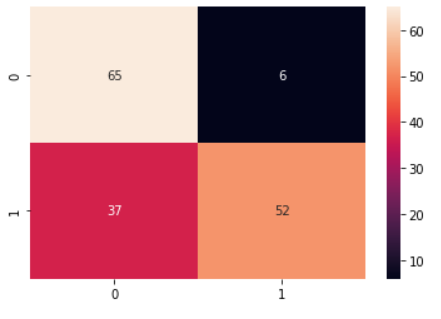
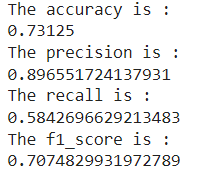
1. Assigned object for input features and target variable in X and y respectively.
2. Over sampling method – The dataset used for the project is imbalanced, hence to use it effectively it should be over sampled. It is sum of the number of records of two different categories and doubles it. It generates other feasible value in all the features so that both the category has equal distribution.
3. Feature scaling: Performed feature scaling on X\_train and X\_test data using Standard scalar. It is used to resize the distribution of values ​​so that the mean of the observed values ​​is 0 and the standard deviation is 1.

Depending on the type of dataset and the associated attributes, a set of machine learning algorithms have been selected. They are as follows:

1. Logistic Regression
   1. Confusion matrix and output on using the input dataset without over sampling

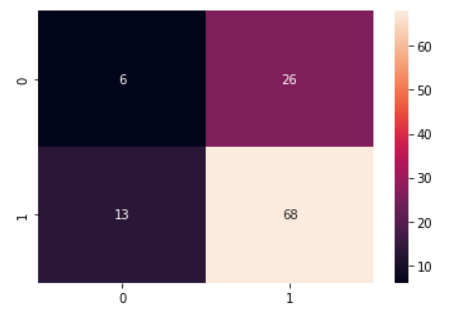
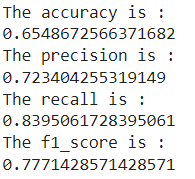
 

* 1. Output on oversampling the input dataset

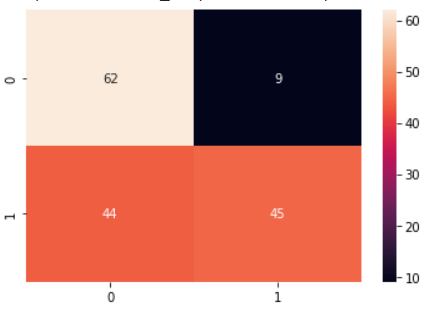
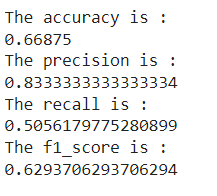
 

It is observed that on over sampling the input dataset, the accuracy and precision of the model improves.

1. K Nearest Neighbor
   1. Output on using the input dataset without over sampling

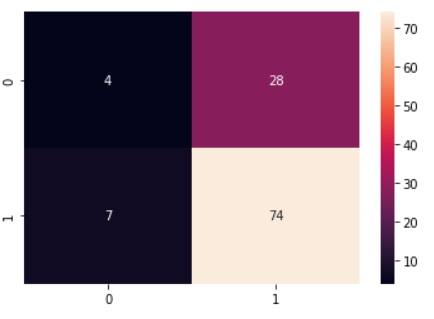
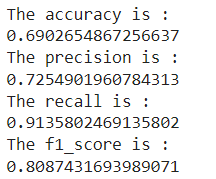
 

* 1. Output on oversampling the input dataset

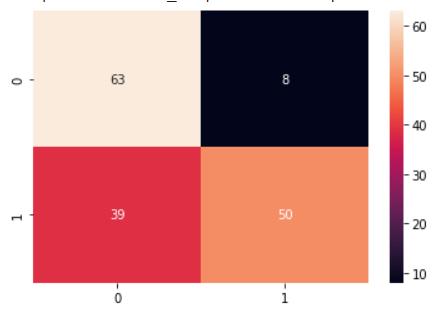
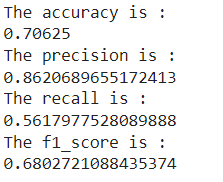
 

It is observed that on over sampling the input dataset, the accuracy and precision of the model improves.

1. Support Vector Machine
   1. Output on using the input dataset without over sampling

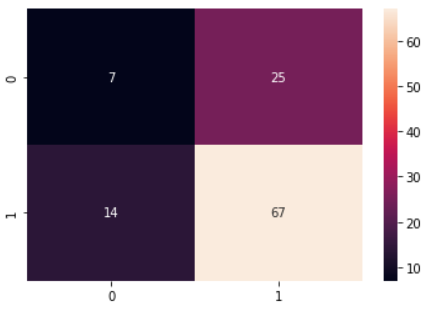
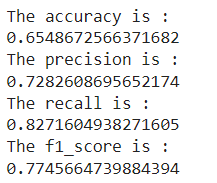
 

* 1. Output on oversampling the input dataset

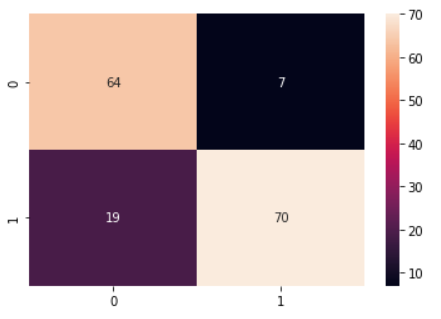
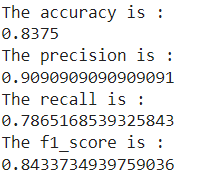
 

It is observed that on over sampling the input dataset, the accuracy and precision of the model improves.

1. Random Forest classifier
   1. Output on using the input dataset without over sampling

* 1. Output on oversampling the input dataset

It is observed that on over sampling the input dataset, the accuracy and precision of the model improves.

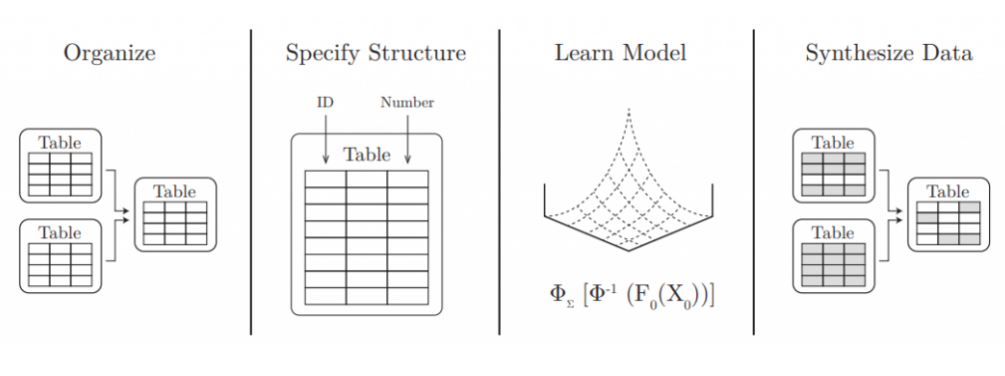
The random forest algorithm performed well with the highest accuracy when over sampled input data was used to train the model.

It was observed that oversampling the input data does not result in higher performance in general. Hence, as a next step data augmentation of the actual data was considered as an alternative to yeild superior performance.

The SDV (Synthetic Data Vault) was used to perform data augmentation using existing ILPD dataset.

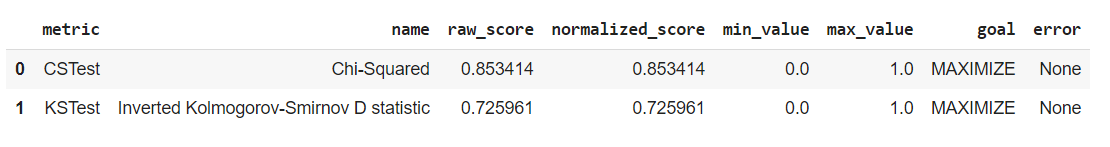
**SDV (Synthetic Data Vault):**

SDV is a set of synthetic data generating open source library. It can handle multiple datatypes and missing data with minimum user input. For the project, synthetic data generators for single table was used. It uses Deep learning models , mathematical techniques and Coupulas to analyze data. Then, it generates new synthetic data with similar format and statistical properties to that of the original data. Gaussian Copulas model and recursive methods for data sampling are used by SDV to synthesize new data from multi-tabular data.

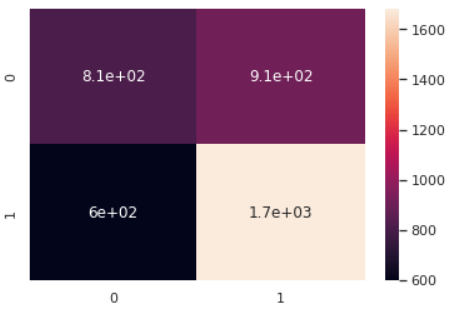


Workflow of SDV

SDV was applied to the original data set to generate 8000 synthetic data. This data was then evaluated using KSTest and CSTest(details below).



1. This Synthetic data was then used on two prevoiusly well performing machine learning models namely:
2. Random forest Algorithm
3. Output on using synthetically generated data from actual dataset



The accuracy is:

0.623

The precision is:

0.6489197530864198

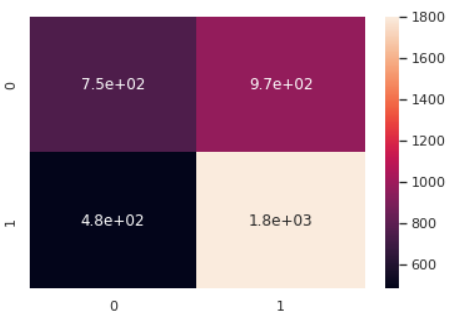
The recall is:

0.737719298245614

The f1\_score is:

0.6904761904761905

1. Support Vector Machine
2. Output on using synthetically generated data from actual dataset



The accuracy is:

0.63825

The precision is:

0.6505240332490061

The recall is:

0.7894736842105263

The f1\_score is:

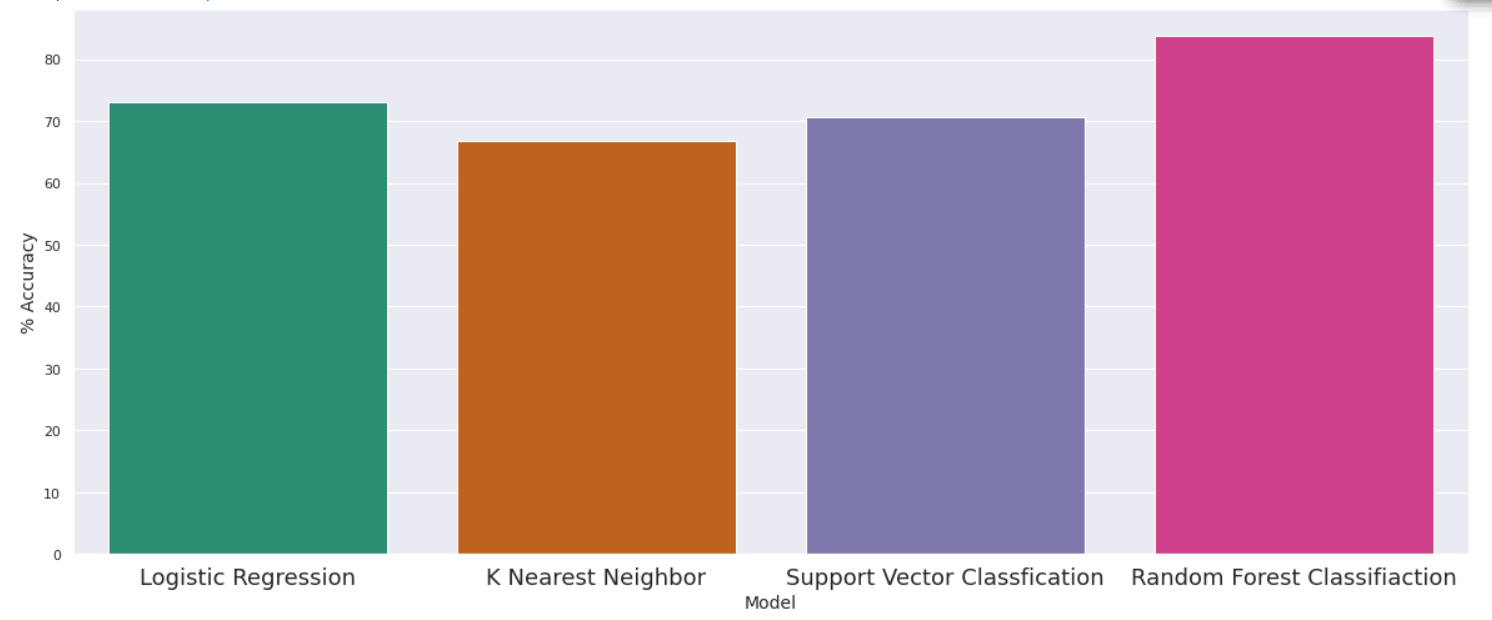
0.7132950267485635

**Evaluation:**

The performance of the models has been evaluated using Confusion matrix, F1 score, Precision, Recall and accuracy. As the data in the dataset is unbalanced, random over sampling has been applied on the training data to improve the accuracy and precision of the models.

Key Model evaluators are as follows:

1. Confusion matrix: Matrix for summarizing the performance of a classification algorithm using true positives, true negatives, false positive and false negative.
2. Accuracy: It states the number of times the Machine Learning model was correct overall.
3. Precision: Measure of how good the model is at predicting a specific category.
4. Recall: It states the number of times the model was able to detect a specific category.
5. F1 score: It compare the performance of two classifiers.
6. Two statistical tests were used for evaluating the synthetically generated data. They are KSTest and CSTest.
7. CSTest (Chi-Squared): This metric uses the Chi-Squared test to compare the distributions of two discrete columns. The output for each column is the CSTest p-value, which indicates the probability of the two columns having been sampled from the same distribution. It compares the discrete columns.
8. KSTest: This metric uses the two-sample Kolmogorov–Smirnov test to compare the distributions of continuous columns using the empirical CDF. The output for each column is 1 minus the KS Test D statistic, which indicates the maximum distance between the expected CDF and the observed CDF values. Basically, it compares the continuous columns in the dataset.



Comparing Model Performance (Using over sampled input data)

**Conclusion:**

Diseases related to liver are becoming more and more common with time. Keeping track of the liver heath can be beneficial for working individuals who work for long hours and miss out some daily exercise.

So, in such a scenario, this project will be extremely helpful to the society. Until now using the (ILPD) dataset for this project, 83.750 % accuracy, 90.9 % precision, 78.65% recall and 84.33% f1 score was achieved using Random Forest model on oversampled data. It could be difficult to get higher accuracies with very small datasets, however one can achieve an improved accuracy by resampling the existing data, collecting more data of the similar standard or by developing a model to augment specifically for the existing type of data.

**Following are the links that have been referred;**

<https://www.researchgate.net/publication/318410773_Diagnosis_of_liver_diseases_using_machine_learning>

<https://www.analyticsvidhya.com/blog/2021/06/indian-patients-liver-dataset-analysis-and-classification/>

https://github.com/devprabal/indian-liver-patients-prediction

<https://drive.google.com/file/d/1JZ9LBsKqfc4wxcJGq64aU_1fiTX5dNbn/view?usp=sharing>

<https://drive.google.com/file/d/1UzmBJe0jsgMltaL-McfOEIWhnkxDV_Z-/view?usp=sharing>

<https://drive.google.com/file/d/1ImlK_PfGfgQBU3spt5ZbzgP4o-gnTphU/view?usp=sharing>

https://dai.lids.mit.edu/wp-content/uploads/2018/03/SDV.pdf

<https://dai.lids.mit.edu/wp-content/uploads/2018/12/Andrew_MEng.pdf>