

**Data Driven Decision Making**

**Acute Liver Failure**

*Rahul Sharma*

*Rishu Gandhi*

**School of Graduate Professional Studies**

MPS/MS in Data Analytics

DAAN 881 – Data Driven Decision Making

(Fall II, 2019)

# Document Control

## Work carried out by:

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Task description** |
| Rahul Sharma | rks5737@psu.edu |  |
| Rishu Gandhi | rmg5430@psu.edu |  |
|  |  |  |

## Revision Sheet

|  |  |  |
| --- | --- | --- |
| **Release No.** | **Date** | **Revision Description** |
| 1.0 | 10/27/2019 | Defined Research Goals and Queries |
| 2.0 | 11/03/2019 | Provided Data Description and Characteristics |
| 3.0 | 11/10/19 | Data Analysis and Issues Reporting |
| 4.0 | 11/17/2019 | Data Cleaning |
| 5.0 | 11/26/2019 | Modelling Steps |
| 6.0 | 12/07/2019 | Data Modelling |
| 7.0 | 12/13/2019 | Finishing Touch |

**Business Goal**

**The goal of our project is to analyze the factors causing acute liver failure.**

**Research Queries**

* What factors are causing acute liver failure?
* What are the major factors that lead to acute liver failure?
* What precautions and life-style habits one can take to reduce the risk of acute liver failure?
* What role the family hereditary diseases such as diabetes and cholesterol play in the risk of having an acute liver failure?

**Data Source**

This data is taken from the Kaggle organization. The data was collected by JPAC Center for Health Diagnosis and Control which includes information of 8,785 adults who are 20 years of age or older. This data is taken from the 2008–2009 and 2014–2015 surveys by collecting demographic and health information through direct interviews, examinations, and blood samples. The reference is given at the end of this document.

**Data Characteristics**

* The data consists of 8785 records and 30 columns (1 MB).
* The data captures multiple attributes that lead to acute liver failure

**Data Summary**

**Columns Names**- includes all the attributes present.

**Data Type**- explains the data type of each attribute.

**Scale**- explains the measurement scale of each attribute.

**Description**- explains each attribute.

**The number of Missing values**- includes the number of missing values in each attribute.

**Data Description:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Data Type** | **Type** | **Description** | **No. of Unique Values** | **No. of Missing Values** |
| **Age** | int64 | Continuous | Age of a patient | 66 | 0 |
| **Gender** | object | Categorical | Sex of the patient i.e. M or F | 2 | 0 |
| **Region** | object | Categorical | Geographical location | 4 | 0 |
| **Weight** | float64 | Continuous | Weight in Kgs | 1120 | 194 |
| **Height** | float64 | Continuous | Height of the patient in cm | 553 | 191 |
| **Body Mass Index** | float64 | Continuous | BMI i.e. weight divided by the square of height | 2439 | 290 |
| **Obesity** | float64 | Categorical | If the patient carries excess weight | 2 | 290 |
| **Waist** | float64 | Continuous | Waist size in cm | 800 | 314 |
| **Maximum Blood Pressure** | float64 | Continuous | Max blood pressure in mmHg | 142 | 304 |
| **Minimum Blood Pressure** | float64 | Continuous | Min blood pressure in mmHg | 107 | 376 |
| **Good Cholesterol** | float64 | Continuous | High-Density Lipoprotein (HDL) in mg/dL | 111 | 17 |
| **Bad Cholesterol** | float64 | Continuous | Low-Density Lipoprotein (LDL) in mg/dL | 279 | 18 |
| **Total Cholesterol** | float64 | Continuous | HDL + LDL | 284 | 16 |
| **Dyslipidemia** | int64 | Categorical | The abnormal amount of lipids in the blood | 2 | 0 |
| **PVD** | int64 | Categorical | Peripheral vascular disease | 2 | 0 |
| **Physical Activity** | float64 | Categorical | Physical Activity - Inactive, Low, Medium, High | 4 | 10 |
| **Education** | float64 | Categorical | If the patient is educated | 2 | 20 |
| **Unmarried** | float64 | Categorical | Marital Status of Patient | 2 | 452 |
| **Income** | float64 | Categorical | If the patient is earning | 2 | 1161 |
| **Source of Care** | object | Categorical | Hospital type | 5 | 0 |
| **Poor Vision** | float64 | Categorical | Information about the vision of the patient | 2 | 563 |
| **Alcohol Consumption** | int64 | Categorical | 0=No and 1=Yes | 2 | 0 |
| **HyperTension** | float64 | Categorical | abnormally high blood pressure - 0=No and 1=Yes | 2 | 80 |
| **Family HyperTension** | int64 | Categorical | Family History of Hypertension - 0=No and 1=Yes | 2 | 0 |
| **Diabetes** | float64 | Categorical | If the patient has diabetes - 0=No and 1=Yes | 2 | 2 |
| **Family Diabetes** | int64 | Categorical | Family History of Diabetes - 0=No and 1=Yes | 2 | 0 |
| **Hepatitis** | float64 | Categorical | inflammation of the liver - 0=No and 1=Yes | 2 | 22 |
| **Family Hepatitis** | float64 | Categorical | Family History of Hepatitis - 0=No and 1=Yes | 2 | 6 |
| **Chronic Fatigue** | float64 | Categorical | A disease characterized by profound fatigue | 2 | 35 |
| **ALF** | float64 | Categorical | Acute Liver Failure | 2 | 2785 |

**Data Description and Analysis**

1. **Continuous Variables**
   * The continuous variables data description is provided in the table below:

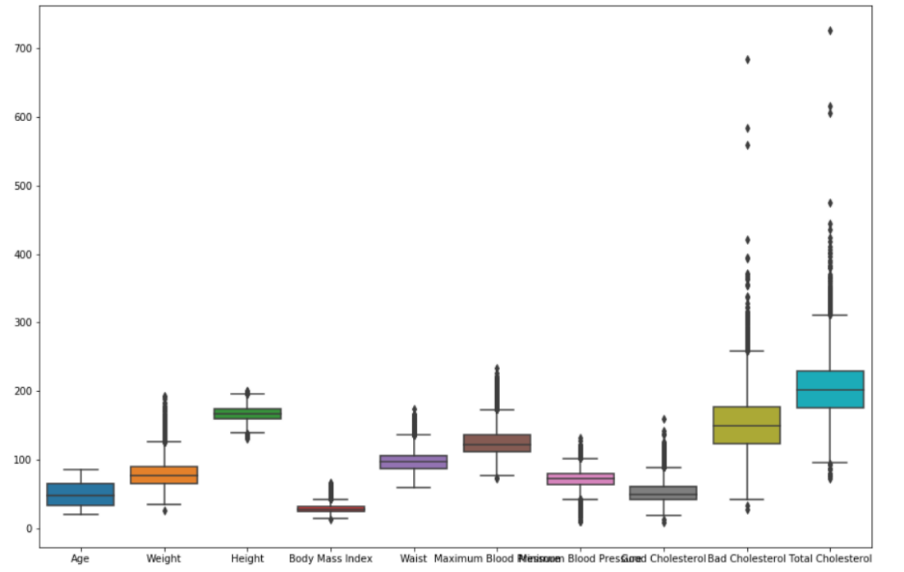
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Continuous Variables | count | mean | std | min | max |
| Age | 8785 | 49.34991 | 18.83131 | 20 | 85 |
| Weight | 8591 | 79.1002 | 19.40697 | 25.6 | 193.3 |
| Height | 8594 | 167.0281 | 10.10345 | 130.4 | 200.1 |
| Body Mass Index | 8495 | 28.29602 | 6.184846 | 12.04 | 66.44 |
| Waist | 8471 | 96.84518 | 15.10117 | 58.5 | 173.4 |
| Maximum Blood Pressure | 8481 | 125.8322 | 21.04686 | 72 | 233 |
| Minimum Blood Pressure | 8409 | 71.51683 | 12.67235 | 10 | 132 |
| Good Cholesterol | 8768 | 51.82254 | 15.78913 | 8 | 160 |
| Bad Cholesterol | 8767 | 152.5875 | 42.9761 | 27 | 684 |
| Total Cholesterol | 8769 | 204.4138 | 42.78593 | 72 | 727 |

* + The Skewness and Kurtosis values were obtained from the following code snippet:

|  |  |  |
| --- | --- | --- |
| **Continuous Variables** | **Skewness Values** | **Kurtosis Values** |
| **Age** | 0.218269 | -1.097473 |
| **Weight** | 0.97215 | 1.857392 |
| **Height** | 0.107659 | -0.29649 |
| **Body Mass Index** | 1.14343 | 2.416513 |
| **Waist** | 0.490752 | 0.500805 |
| **Maximum Blood Pressure** | 1.033431 | 1.318187 |
| **Minimum Blood Pressure** | -0.272185 | 1.399803 |
| **Good Cholesterol** | 1.019316 | 1.584573 |
| **Bad Cholesterol** | 1.028096 | 5.678164 |
| **Total Cholesterol** | 0.954568 | 5.320727 |

The above table shows the skewness values for each continuous variable in our data. All the continuous variables other than Minimum Blood Pressure are positively skewed. The skewness values show that the data is not highly skewed because the variables are close to being normal hence skewness values being close to 0.

The kurtosis values foreshadow the outliers present in our data. The variables with small kurtosis values represent light-tailed or lack of outliers. Bad Cholesterol and Total Cholesterol have slightly higher kurtosis values which represent moderate tailed outliers.

* + The Box Plot distribution is provided as below:

1. **Categorical Variables** 
   * The categorical variable data description is provided in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Count | Missing | Description | Missing Values Handling |
| Obesity | 8495 | 290 | 0  - 5811  1  - 2684 | Replaced by -1 |
| Dyslipidemia | 8785 | 0 | 0 -  7859  1  - 926 | Replaced by -1 |
| PVD | 8785 | 0 | 0  - 8440  1  - 345 | Replaced by -1 |
| Physical Activity | 8775 | 10 | 2  - 4632  1  - 2229  3  - 1349  4  - 565 | Replaced by -1 |
| Education | 8765 | 20 | 0  - 4982  1  - 3783 | Replaced by -1 |
| Unmarried | 8333 | 452 | 0  - 5261  1  - 3072 | Replaced by -1 |
| Income | 7624 | 1161 | 0  - 4442  1  - 3182 | Replaced by -1 |
| Poor Vision | 8222 | 563 | 0  - 7699  1  - 523 | Replaced by -1 |
| Alcohol Consumption | 8785 | 0 | 0  - 6114  1  - 2671 | Replaced by -1 |
| HyperTension | 8705 | 80 | 0  - 5202  1  - 3503 | Replaced by -1 |
| Family  HyperTension | 8785 | 0 | 0  - 6736  1  - 2046 | Replaced by -1 |
| Diabetes | 8783 | 2 | 0  - 7805  1  - 978 | Replaced by -1 |
| Family Diabetes | 8785 | 0 | 0  - 6046  1  - 2739 | Replaced by -1 |
| Hepatitis | 8763 | 22 | 0  - 8180  1  - 583 | Replaced by -1 |
| Family Hepatitis | 8779 | 6 | 0  - 8599  1  - 180 | Replaced by -1 |
| Chronic Fatigue | 8750 | 35 | 0  - 8496  1  - 254 | Replaced by -1 |
| ALF | 6000 | 2785 | 0  - 5536  1  - 2785 | Replaced by -1 |

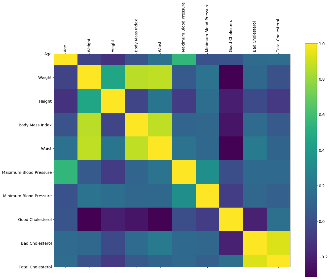
The missing values have been replaced by -1 which implies that we have created one more category among these variables

1. **Duplicate Values**

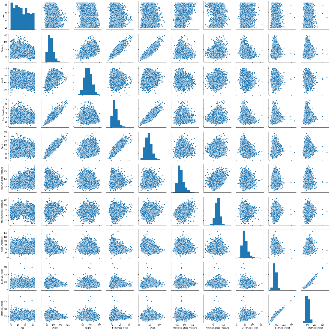
****

* + As shown above, we performed a duplicate check using Python; it is found that there are no duplicate rows present in our data.

1. **Outliers**
   * The variables namely *Total Cholesterol* and *Bad Cholesterol* has high kurtosis values and it is evident from box plot as well that there are outliers exists in these two variables. We will take the relevant measures to remove these outliers in Week 4.
2. **Multicollinearity**
   * Interaction between Continuous variables
     + The HeatMap and Pair plot were created for continuous variables

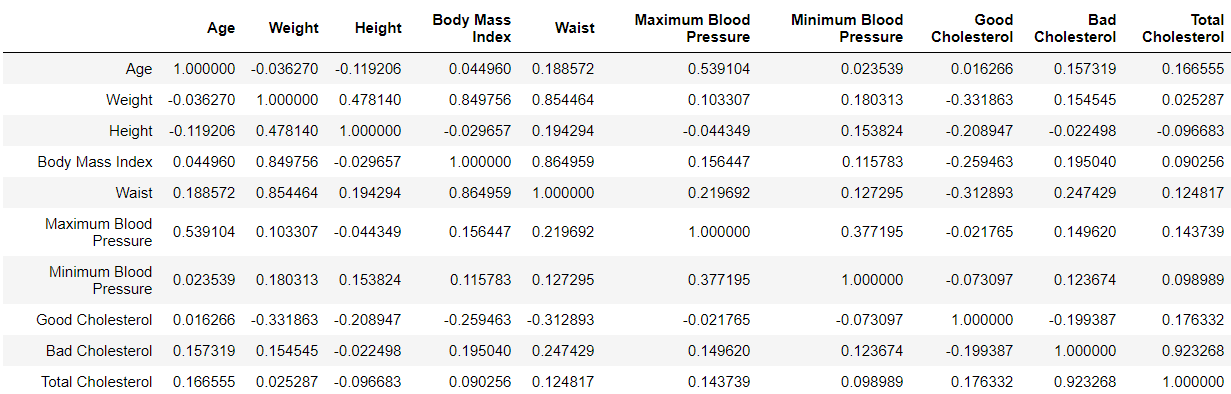
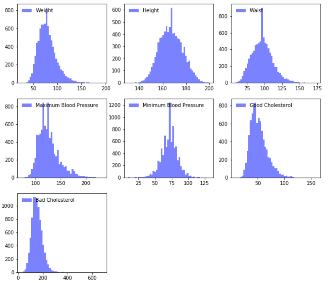
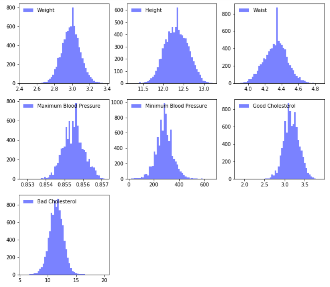
****

* + - The visual representation of correlation matrix in presented in the form of pair plot:

****

As it can be seen from the HeatMap and pairplot, Weight is highly correleted with Body Mass Index and Waist. Also, total cholestrol and Bad cholestrol are highly correlated.

**Data Cleaning**

1. **Catering Missing Values**
   1. **Continuous Variables-** There were many missing values in our data for continuous variables, which have been replaced by their respective mean values.
   2. **Categorical Variables-** There were many missing values in our data for categorical variables, which have been replaced by -1.
2. **Catering Correlation Check**
   1. The above table shows the correlation matrix values for continuous variables; it is shown that Total Cholesterol is highly correlated with Bad Cholesterol with a value of 0.923. It was also observed that Total Cholesterol is the sum of Good Cholesterol and Bad Cholesterol. Therefore, we have removed Total Cholesterol variable and kept Bad Cholesterol variable, which will help us analyze accurately.
   2. It is observed that Weight and Height are used in determining Body Mass Index; this explains that we do not need the column of Body mass index as Weight and height are sufficient to provide similar information. Therefore, we have removed the column of Body Mass Index and kept all the other columns.
3. **Catering Skewness**
   1. Even though all the continuous variables contain skewness value close to 0, however, they cannot be completely neglected therefore we performed Box-Cox Transformation to normalize the variables.
      1. Below is the histogram output BEFORE applying Box-Cox Transformation
      2. Below is the histogram output AFTER applying Box-Cox Transformation
      3. The BEFORE and AFTER Box-Cox Transformation Histogram Shape explains the normality of each continuous variable.
         1. The table below contains skewness values after performing Box-Cox Transformation; they are very close to 0.

|  |  |
| --- | --- |
| **Continuous Variables** | **Skewness Values** |
| **Weight** | -0.0030 |
| **Height** | -0.0013 |
| **Waist** | -0.0001 |
| **Maximum Blood Pressure** | 0.0017 |
| **Minimum Blood Pressure** | 0.1040 |
| **Good Cholesterol** | -0.0020 |
| **Bad Cholesterol** | 0.0209 |

1. **Outlier Detection**
   1. In order to reduce outliers, we removed unnecessary columns and obtained new Kurtosis values as show in the table below.

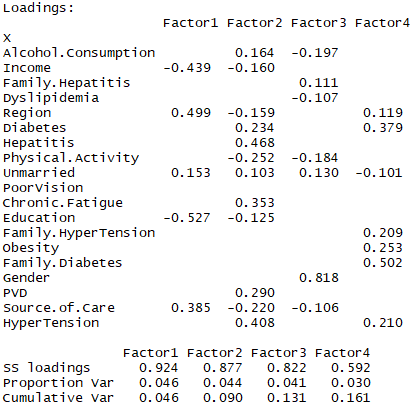
|  |  |
| --- | --- |
| **Continuous Variables** | **Kurtosis Values** |
| **Weight** | 0.2109 |
| **Height** | -0.2231 |
| **Waist** | 0.0464 |
| **Maximum Blood Pressure** | 0.0580 |
| **Minimum Blood Pressure** | 1.2735 |
| **Good Cholesterol** | 0.2158 |
| **Bad Cholesterol** | 0.8349 |

* + - 1. It can be seen that all the kurtosis values are less than 3, which represent moderate very less outliers or no outliers present in our data.

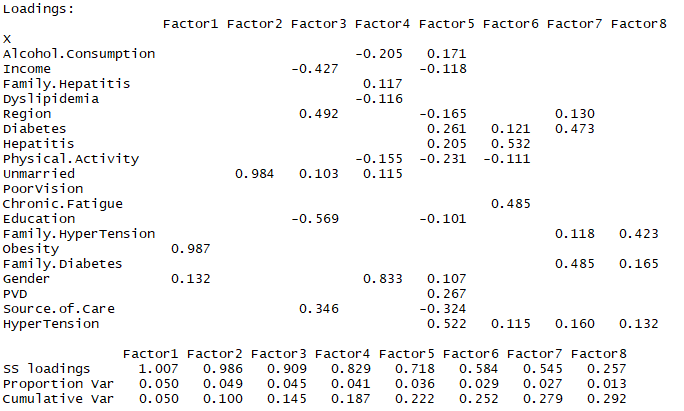
1. **Model Validity**
   1. The data consists of 8785 records in which there are no ALF (response variable) values for 2785 records. Therefore, we decided to divide the data into Training (6000 records) and Test (2785 records) Sets; The Model will be performed based on 6000 records and obtain an equation, which then will be verified by using Test Set of 2785 records. This will provide us accurate validity of our Model.

**Pre-processing Steps:**

* As performed in week 4, we have created a correlation matrix for continuous variables and found out that Total Cholesterol and Body Mass Index are highly correlated with Bad Cholesterol and Weight respectively, so we removed these two variables.
* As our dataset doesn’t have many records, so in order to get the good result we are keeping the remaining 28 variables post removing 2 attributes during cleaning process
* We normalized the continuous variables in order take them in a common scale. And, we performed Box Cox transformation to remove the skewness present in the dataset.
* **Reduce the Number of Categorical Attributes:**
  + Exploratory Factor Analysis is being performed on the dataset to reduce the categorical attributes using 4 and 8 factors. The results are shown below:
  + **4-Factor:**

****

* + **8-Factor:**

****

As we can see that the variation explained by the above factors is not significant, hence we will not use dimensionality reduction on the categorical variables and consider all of them for developing the model.

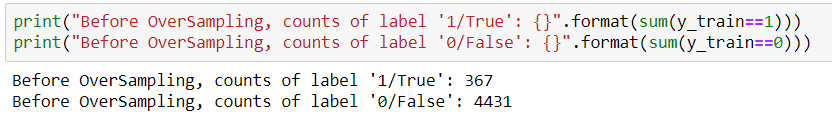
**Data Modelling:**

As we have a lot of continuous and categorical predictor variables, and our response variable is binary, so we will be performing Logistic Regression to evaluate the effect of each variable on to the possibility of having an Acute Liver Failure. We will create the model based on 6000 training records and then apply it to the remaining 2785 test records to identify if any of these patients have Acute Liver Failure.

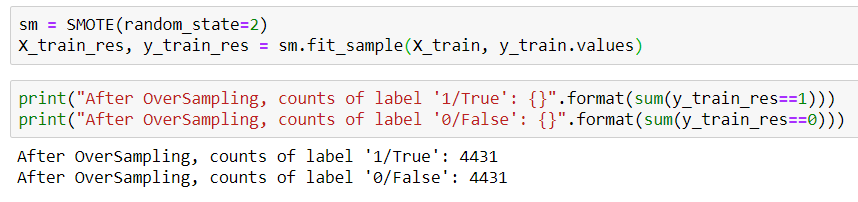
**Data Modeling**

* **SMOTE Oversampling**

We have divided our dataset into two parts 80% Training and 20% Test samples. It has been seen that the training dataset is quite unbalanced i.e. it contains 4431 records for ‘False’ class and 367 records for ‘True’ class.



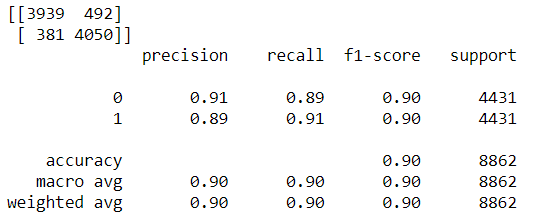
To resolve this issue, we will perform the SMOTE (*Synthetic Minority Oversampling TEchnique*) oversampling. The representation is shown below:



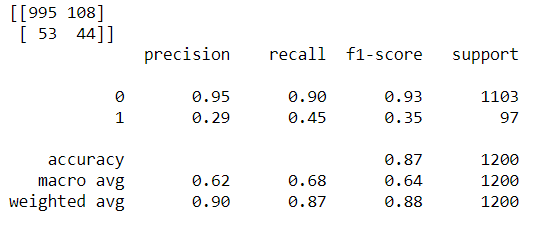
Now, our training dataset is balanced with 4431 records in each class. SMOTE just replicated the samples multiple times to make the training dataset balanced.

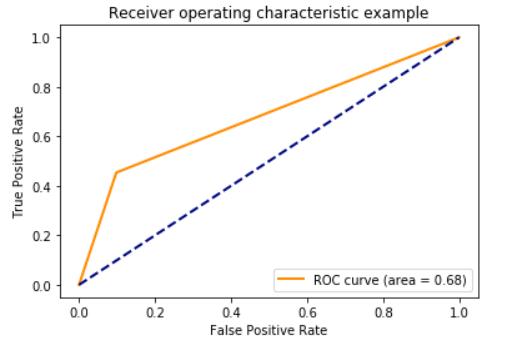
* **Logistic Regression** 
  + Training Data:

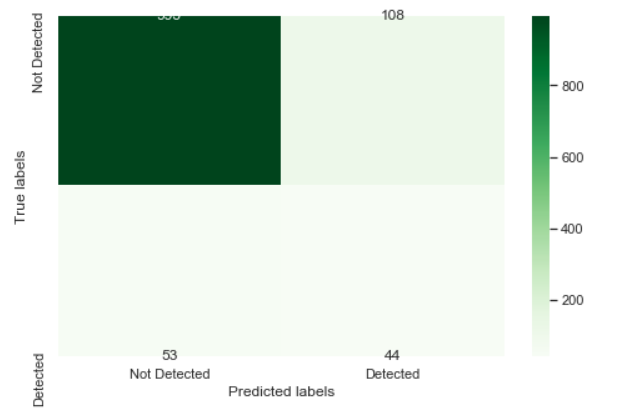
After performing the logistics model, we got the following confusion and classification report for the training dataset:



* + We performed logistic regression on the test data and below is the output we received:

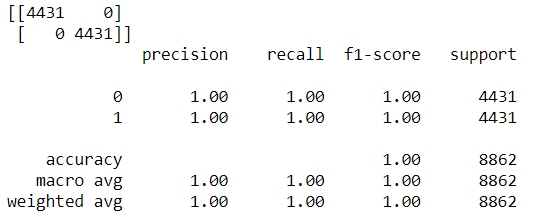




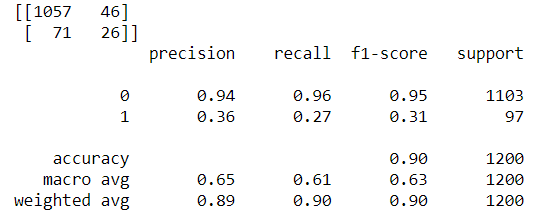


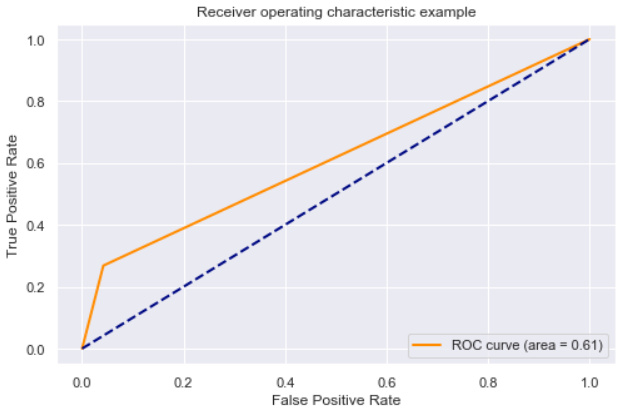
* + Based on the output we received, the accuracy is 87% which shows satisfactory result for logistic regression.
  + The ROC curve produces a higher value for area under the curve (0.68), which validated the logits regression model more.
  + The accuracy is quite comparable for the training and test datasets; hence we can say that the model doesn’t overfit.
* **Random Forest**
  + Training Data:

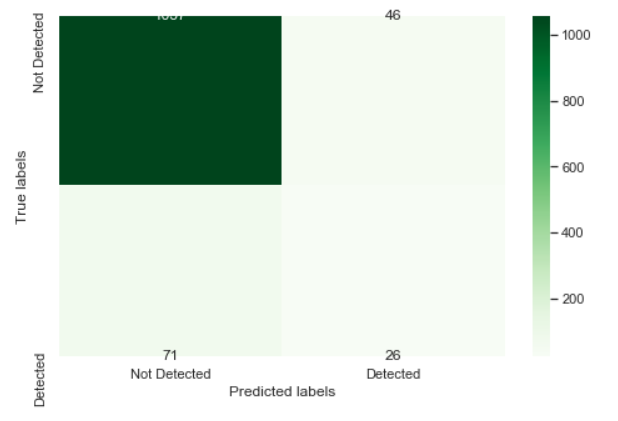
We got the following confusion matrix and classification report:



* + In order to obtain more accuracy and validate our logistic regression result, we performed random forest on the SMOTE data. Below is the output we received for the test data:

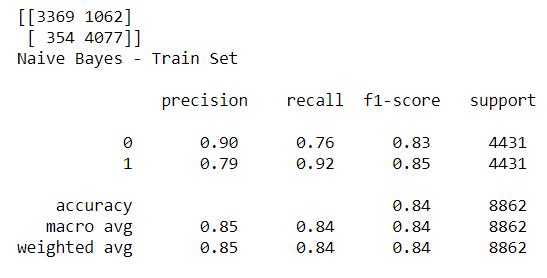




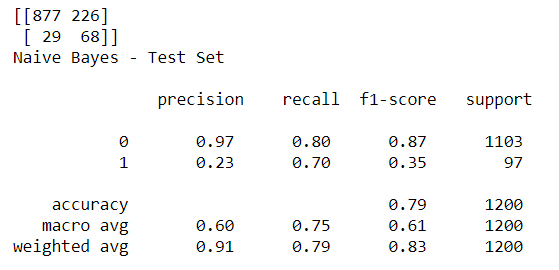


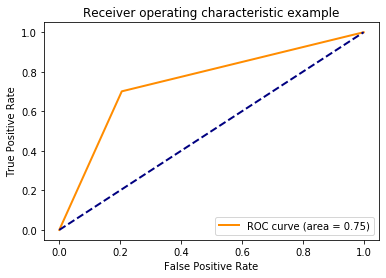
* + We obtained the accuracy of 90% which further validates our Model.
  + It is clearly indicated that the values for precision, recall and f1-score differs highly for both training and test samples.
* **Naïve Bayes**
  + Training Data:

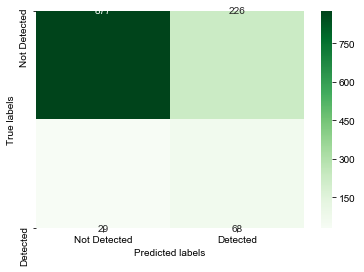
We got the following confusion matrix and classification report:



* + Test Data:



****

****

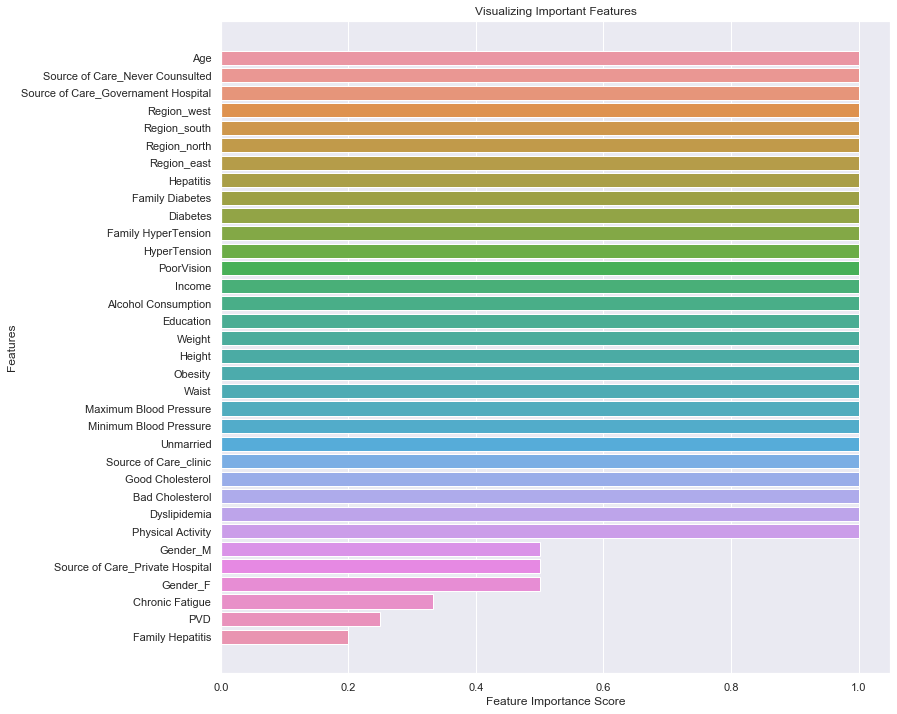
* + Though the accuracy is not much higher than the other models, our recall value has increased significantly in test data as compared to other models.
  + The ROC curve area is largest among all three model.
  + The recall value is not quite similar in test and train data, so we can say that the model exhibits overfitting.
* **Conclusion:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic | Random Forest | Naïve Bayes |
| Precision | *0.29* | *0.36* | *0.23* |
| Recall | *0.45* | *0.27* | *0.70* |
| Accuracy | *0.87* | *0.90* | *0.79* |
| F1 - Score | *0.35* | *0.31* | *0.35* |

The F1 score for the Random Forest and Naïve Bayes is same. The best model among all three of the above is Naïve Bayes, as the recall value is higher in Naïve Bayes.

**Important Feature**

* + Appropriate Python coding was used to produce a chart of important features/attributes contributing to Acute Liver Failure Risk. Below:

****

* + The major attributes/factors causing Acute Liver Failure Risk are: Age, Source of Care Never Consulted, Source of Care Government Hospital, Region West, Region South, Region North, Region East, Hepatitis, Family Diabetes, Diabetes, Family Hypertension, Hypertension, Poor Vision, Income, Alcohol Consumption, Education, Weight, Height, Obesity, Waist, Maximum Blood Pressure, Minimum Blood Pressure, Unmarried, Source of Care Clinic, Good Cholesterol, Bad Cholesterol, Dyslipidemia, Physical Activity. All these features have a score of 1.
* **Addressing Business Queries/Research** 
  + **What factors are causing acute liver failure?/ What are the major factors that lead to acute liver failure?**
    - As the chart shows, the major factors leading to acute liver failure are Age, Source of Care Never Consulted, Source of Care Government Hospital, Region West, Region South, Region North, Region East, Hepatitis, Family Diabetes, Diabetes, Family Hypertension, Hypertension, Poor Vision, Income, Alcohol Consumption, Education, Weight, Height, Obesity, Waist, Maximum Blood Pressure, Minimum Blood Pressure, Unmarried, Source of Care Clinic, Good Cholesterol, Bad Cholesterol, Dyslipidemia, Physical Activity. They all contain Feature Importance Score of 1.
  + **What precautions and life-style habits one can take to reduce the risk of acute liver failure?**
    - As the chart shows, the major factors such as Diabetes, Hypertension, Maximum and Minimum Blood Pressure, and Weight lead to acute liver failure. Therefore, it is advisable to maintain a healthy food diet helps to avoid diabetes and hypertension. The sugar and salt intake should be moderate in daily food intake. One should also exercise on a regular basis to avoid such issues.
  + **What role the family hereditary diseases such as diabetes and cholesterol play in the risk of having an acute liver failure?**
    - Diabetes and Cholesterol both play a huge role in the risk of having an acute liver failure as they are both part of the major features contributing to the risk. Both have a score of 1 and therefore they have significant impact on developing acute liver failure risk if they are not maintained via healthy diet.

**REFERENCES:**

Kumar, R. (2018, September 15). Acute Liver Failure. Retrieved October 31, 2019, from <https://www.kaggle.com/rahul121/acute-liver-failure>