Introduction:

As we all know, liver disease is a major concern which, if not identified and dealt with immediately can lead to serious health issues. Luckily, there are a number of tests whose results, when analysed together can help doctors identify and prescribe the appropriate medication. Some such tests are the albumin test, bilirubin test and the alkaline phosphatase (ALP) test.

Objective:

Given a set of data, our main aim is to analyse it for each and every patient and determine whether or not He/She suffers from liver disease. In order to achieve this, we have used 4 models:

1. Naïve Bayes Model
2. The Logistic Regression Model
3. K-Nearest Neighbors (KNN) Model
4. Support Vector Machine(SVM)

Data info :

TOTAL BILIRUBIN (TBil) :

Normal values of total bilirubin range from 0.3–1.0 mg/dL.If bilirubin is not being attached to the glucose-derived acid (conjugated) in the liver or is not being adequately removed from the blood, it can mean that there is damage to your liver. Testing for bilirubin in the blood is therefore a good way of testing for liver damage.

DIRECT BILIRUBIN (DBil) :

The reference range of direct bilirubin is 0.1-0.4 mg/dL.Bilirubin is a substance made when your body breaks down old red blood cells. This is a normal process. Direct bilirubin travels freely through your bloodstream to your liver.PROTEIN LEVEL : The normal range for total protein is between 6 and 8.3 grams per deciliter (g/dL). This range may vary slightly among laboratories. These ranges are also due to other factors such as: age.

ALBUMIN :

It is the most abundant protein in human blood plasma; it constitutes about half of serum protein. It is produced in the liver. The reference range for albumin concentrations in serum is approximately 35 - 50 g/L (3.5 - 5.0 g/dL).

A/G RATIO :

The albumin to globulin (A/G) ratio has been used as an index of disease state, however, it is not a specific marker for disease because it does not indicate which specific proteins are altered. The normal A/G ratio is 0.8-2.0

SGPT :

An SGPT blood test is a test used to measure the amount of the enzyme glutamate pyruvate transaminase (GPT) in blood serum.This enzyme is found in much greater concentration in the liver. This test is also sometimes known as ALT or, where it is also combined with several other tests to find out how well the liver is functioning. The normal range of values SGPT is from 7 to 56 units per litre of serum.

SGOT :

The SGOT test measures one of two liver enzymes, called AST, which stands for aspartate aminotransferase. An SGOT test (or AST test) evaluates how much of the liver enzyme is in the blood.The normal range of values for AST (SGOT) is about 5 to 40 units per liter of serum (the liquid part of the blood).

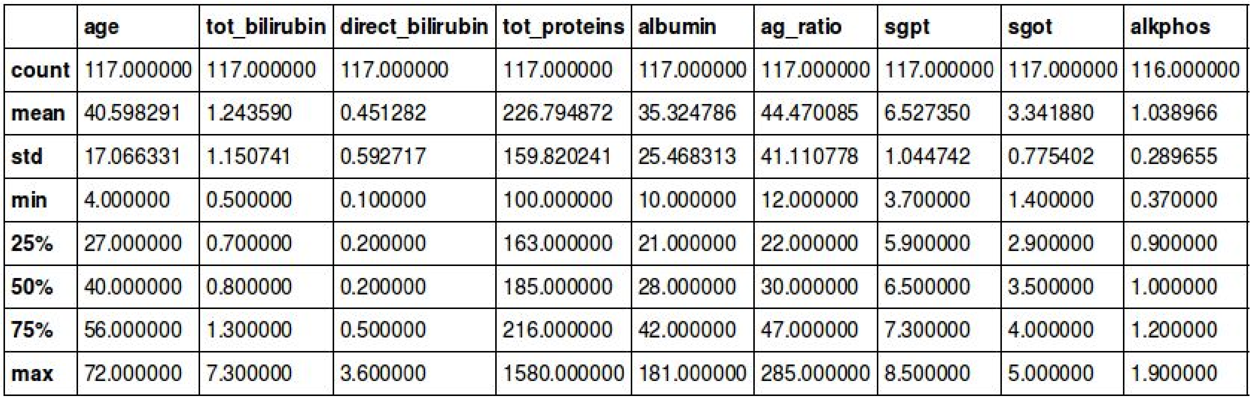
ALKPHOS :

An alkaline phosphatase (ALP) test is used measure the amount of the enzyme in your blood and help in diagnosing the problem. It checks how your liver is working. The normal range is 44 to 147 IU/L (international units per liter) or 0.73 to 2.45 microkat/L.

Summary Statistics:

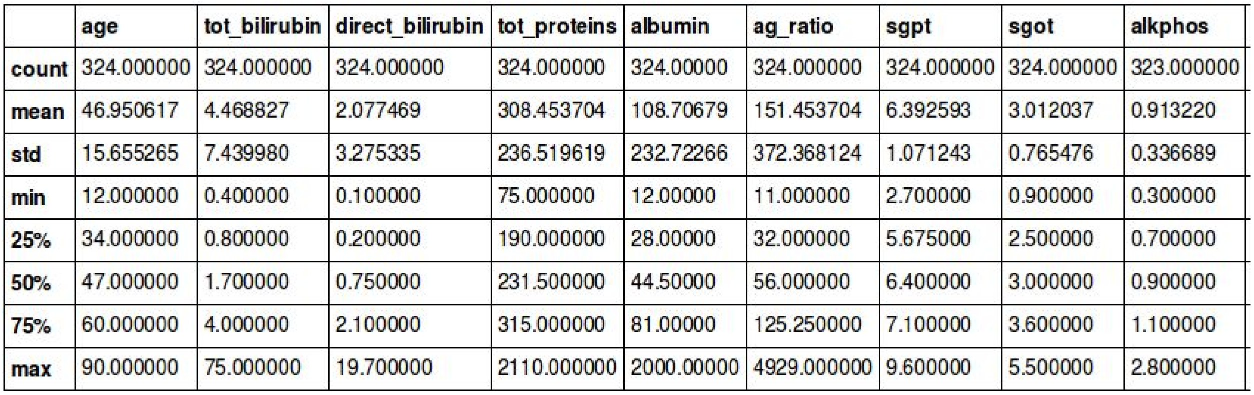
The various fundamental measures of the ILPD dataset is as shown below :

1. **Male who are not patients :**



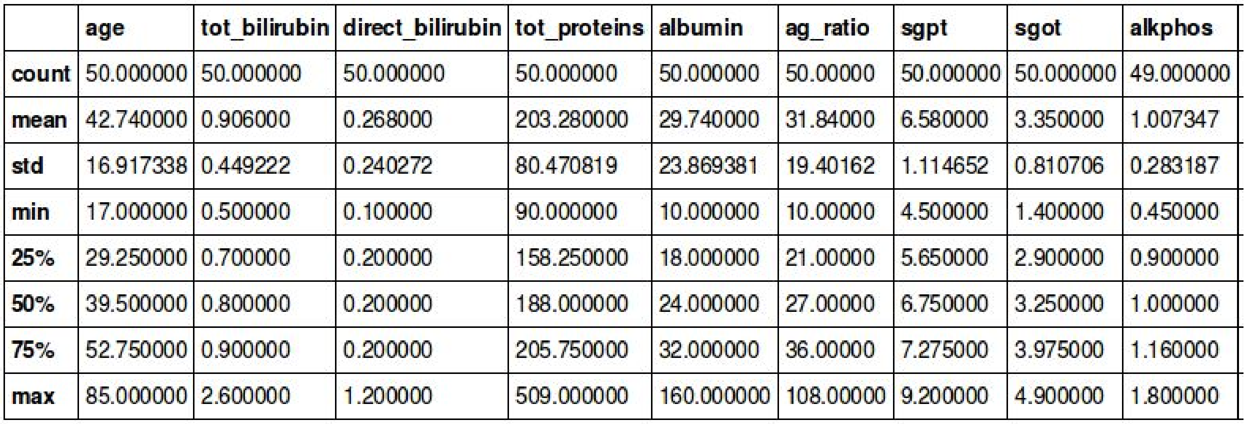
	age	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos
count	117.000000	117.000000	117.000000	117.000000	117.000000	117.000000	117.000000	117.000000	116.000000
mean	40.598291	1.243590	0.451282	226.794872	35.324786	44.470085	6.527350	3.341880	1.038966
std	17.066331	1.150741	0.592717	159.820241	25.468313	41.110778	1.044742	0.775402	0.289655
min	4.000000	0.500000	0.100000	100.000000	10.000000	12.000000	3.700000	1.400000	0.370000
25%	27.000000	0.700000	0.200000	163.000000	21.000000	22.000000	5.900000	2.900000	0.900000
50%	40.000000	0.800000	0.200000	185.000000	28.000000	30.000000	6.500000	3.500000	1.000000
75%	56.000000	1.300000	0.500000	216.000000	42.000000	47.000000	7.300000	4.000000	1.200000
max	72.000000	7.300000	3.600000	1580.000000	181.000000	285.000000	8.500000	5.000000	1.900000

2. Male who are patients :



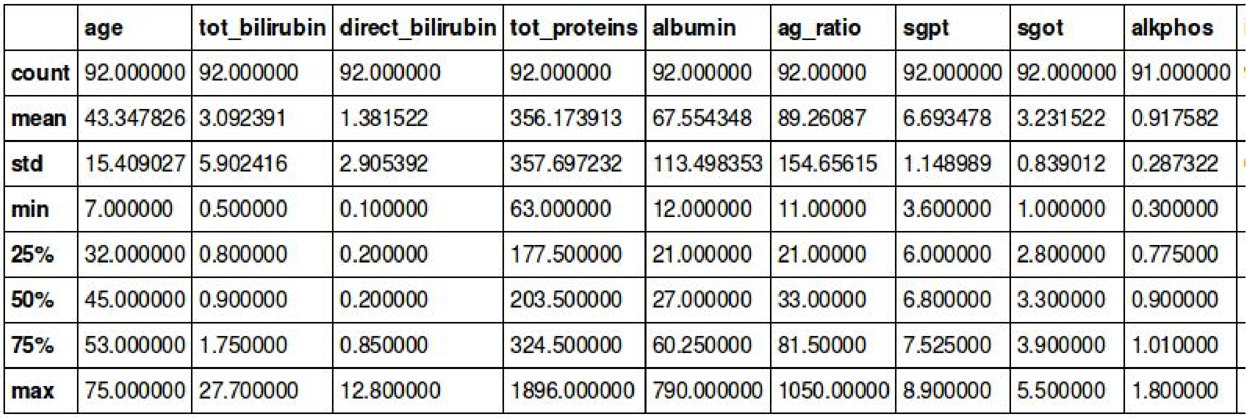
	age	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos
count	324.000000	324.000000	324.000000	324.000000	324.00000	324.000000	324.000000	324.000000	323.000000
mean	46.950617	4.468827	2.077469	308.453704	108.70679	151.453704	6.392593	3.012037	0.913220
std	15.655265	7.439980	3.275335	236.519619	232.72266	372.368124	1.071243	0.765476	0.336689
min	12.000000	0.400000	0.100000	75.000000	12.00000	11.000000	2.700000	0.900000	0.300000
25%	34.000000	0.800000	0.200000	190.000000	28.00000	32.000000	5.675000	2.500000	0.700000
50%	47.000000	1.700000	0.750000	231.500000	44.50000	56.000000	6.400000	3.000000	0.900000
75%	60.000000	4.000000	2.100000	315.000000	81.00000	125.250000	7.100000	3.600000	1.100000
max	90.000000	75.000000	19.700000	2110.000000	2000.00000	4929.000000	9.600000	5.500000	2.800000

3. Female who are not patient



	age	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos
count	50.000000	50.000000	50.000000	50.000000	50.000000	50.00000	50.000000	50.000000	49.000000
mean	42.740000	0.906000	0.268000	203.280000	29.740000	31.84000	6.580000	3.350000	1.007347
std	16.917338	0.449222	0.240272	80.470819	23.869381	19.40162	1.114652	0.810706	0.283187
min	17.000000	0.500000	0.100000	90.000000	10.000000	10.00000	4.500000	1.400000	0.450000
25%	29.250000	0.700000	0.200000	158.250000	18.000000	21.00000	5.650000	2.900000	0.900000
50%	39.500000	0.800000	0.200000	188.000000	24.000000	27.00000	6.750000	3.250000	1.000000
75%	52.750000	0.900000	0.200000	205.750000	32.000000	36.00000	7.275000	3.975000	1.160000
max	85.000000	2.600000	1.200000	509.000000	160.000000	108.00000	9.200000	4.900000	1.800000

4. **Female who are patients** :



	age	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos
count	92.000000	92.000000	92.000000	92.000000	92.000000	92.00000	92.000000	92.000000	91.000000
mean	43.347826	3.092391	1.381522	356.173913	67.554348	89.26087	6.693478	3.231522	0.917582
std	15.409027	5.902416	2.905392	357.697232	113.498353	154.65615	1.148989	0.839012	0.287322
min	7.000000	0.500000	0.100000	63.000000	12.000000	11.00000	3.600000	1.000000	0.300000
25%	32.000000	0.800000	0.200000	177.500000	21.000000	21.00000	6.000000	2.800000	0.775000
50%	45.000000	0.900000	0.200000	203.500000	27.000000	33.00000	6.800000	3.300000	0.900000
75%	53.000000	1.750000	0.850000	324.500000	60.250000	81.50000	7.525000	3.900000	1.010000
max	75.000000	27.700000	12.800000	1896.000000	790.000000	1050.00000	8.900000	5.500000	1.800000

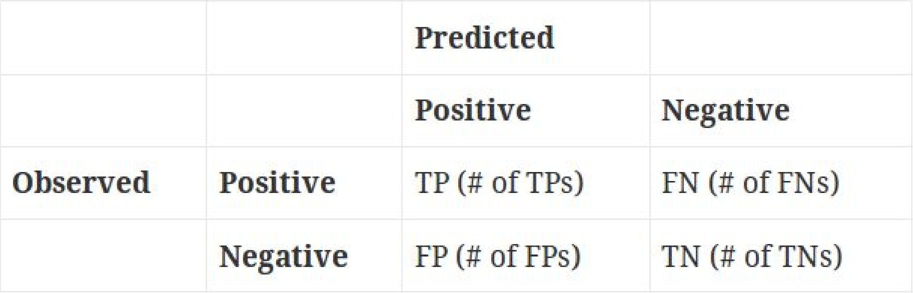
About Summary stats :

- Sgpt , Sgot and Alkphos does affect liver disease
- Alkphos is more in men compared to women
- Proteins contents in the body increases when patients suffering from liver disease

Methodology:

Confusion Matrix :

A confusion matrix of binary classification is a two by two table formed by counting of the number of the four outcomes of a binary classifier. We usually denote them as TP, FP, TN, and FN instead of “the number of true positives”, and so on.



		Predicted	
		Positive	**Negative**
Observed	**Positive**	TP (# of TPs)	FN (# of FNs)
	Negative	FP (# of FPs)	TN (# of TNs)

Measures from the confusion matrix :

1. **Accuracy** :

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}$$

2. **Sensitivity (**Recall or True positive rate**) :**

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$$

3. **Specificity (**True negative rate**) :**

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = \frac{\mathrm{TN}}{\mathrm{N}}$$

4. **Precision** (Positive predictive value) **:**

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

5. **F-score :**

F-score is a harmonic mean of precision and sensitivity.

Analysis and Prediction:

For predicting whether a person is liver patient or not, we are using **FOUR** different models and concept of Confusion matrix to predict and evaluate the result based on the data fields given in the dataset .

1. **Naive Bayes Algorithm :**

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). That is,

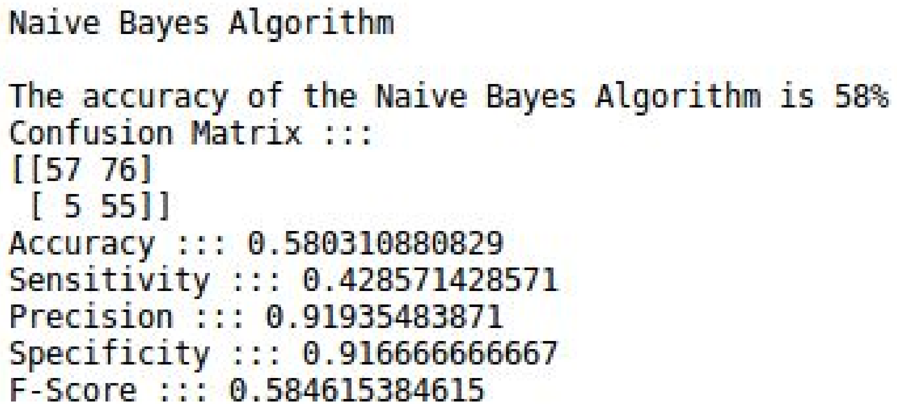
Likelihood Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

The result for the Naive Bayes Algorithm is as follows:



```
Naive Bayes Algorithm

The accuracy of the Naive Bayes Algorithm is 58%
Confusion Matrix :::
[[57 76]
 [ 5 55]]
Accuracy ::: 0.580310880829
Sensitivity ::: 0.428571428571
Precision ::: 0.91935483871
Specificity ::: 0.916666666667
F-Score ::: 0.584615384615
```

2. Logistic regression model :

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. Logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable.

The fundamental equation of generalized linear model is:

$g(E(y)) = \alpha + \beta x1 + \gamma x2$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ *is the linear predictor.*

Logit function :

$p = e^y/ 1 + e^y$ where p is the probability of success.

The result for the Logistic Regression is as follows:

(value corresponds to column in the data set)

Co-eff (beta values) :

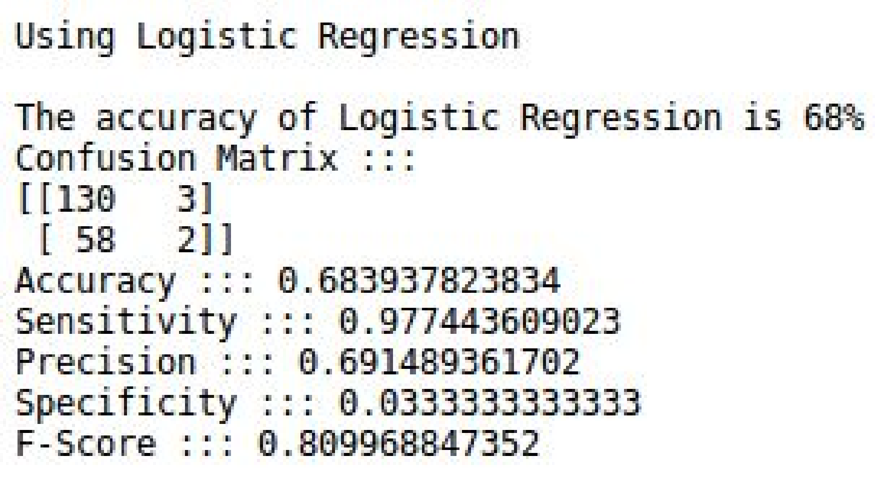
```
[ -1.37180847e-02   3.60746127e-01  -2.76733208e-02  -4.81686064e-01
  -6.98857934e-04  -1.81186879e-02  -1.35441653e-03  -3.89737253e-01
   7.08358791e-01 ]
```

Odds ratio :

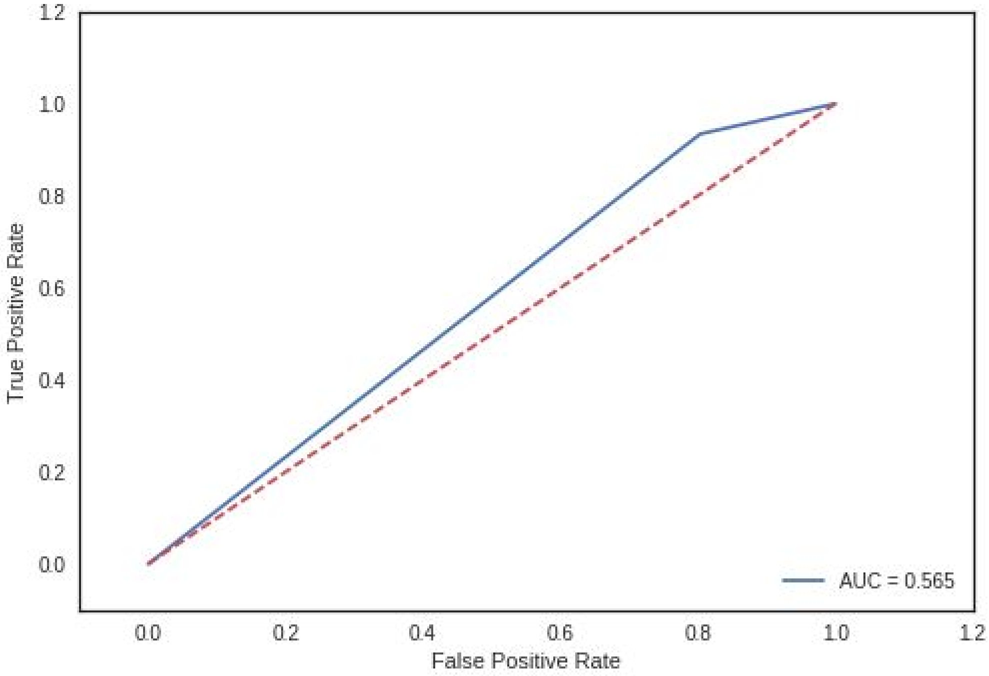
```
[ 0.98637558  1.43439926  0.97270608  0.61774096  0.99930139  0.98204447
  0.9986465   0.67723479  2.03065579 ]
```

p_values :

```
[  7.53184160e-11   8.40956918e-01   1.55016543e-50   8.79158793e-28
   0.00000000e+00   0.00000000e+00   0.00000000e+00   8.05830852e-01
   1.39095787e-01 ]
```



```
Using Logistic Regression

The accuracy of Logistic Regression is 68%
Confusion Matrix :::
[[130   3]
 [ 58   2]]
Accuracy ::: 0.683937823834
Sensitivity ::: 0.977443609023
Precision ::: 0.691489361702
Specificity ::: 0.0333333333333
F-Score ::: 0.809968847352
```



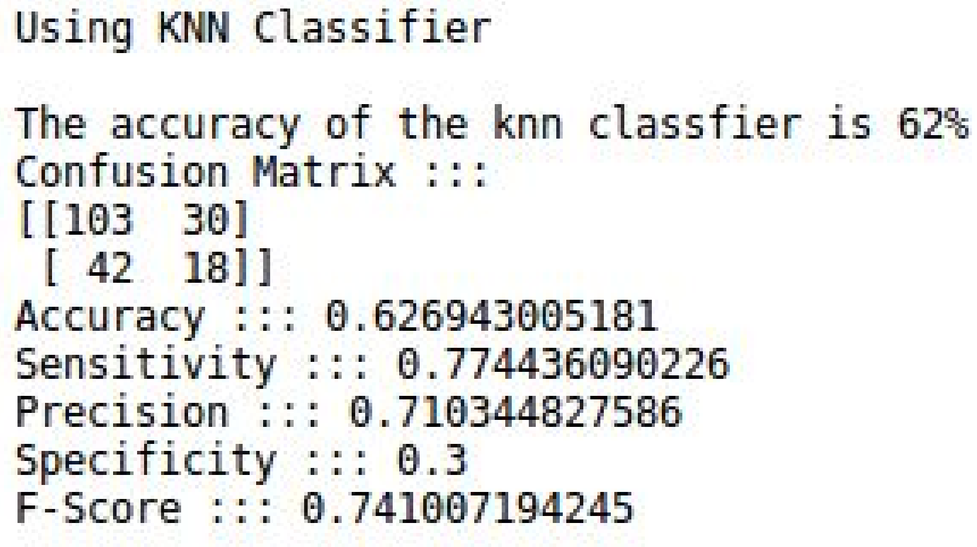
ROC curve

3. K nearest neighbors (KNN)

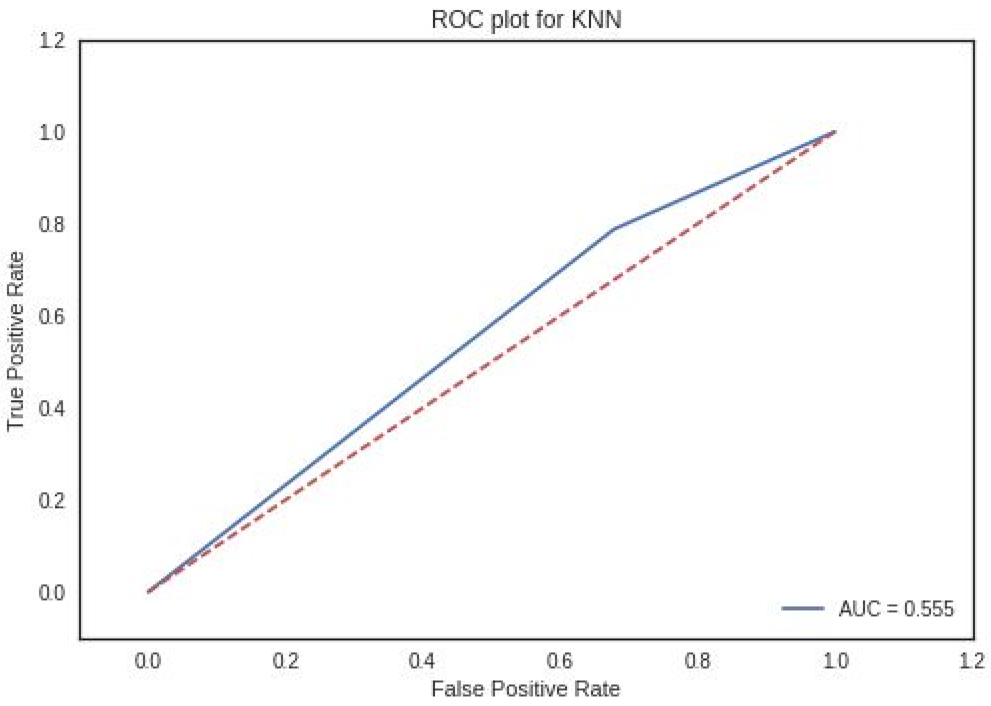
K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure . KNN searches the memorized training observations for the K instances that most closely resemble the new instance and assigns to it the their most common class.Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by,

$$d(x, x') = \sqrt{\left(x_1 - x_1'\right)^2 + \left(x_2 - x_2'\right)^2 + \ldots + \left(x_n - x_n'\right)^2}$$

The result for K nearest neighbors is as follows:



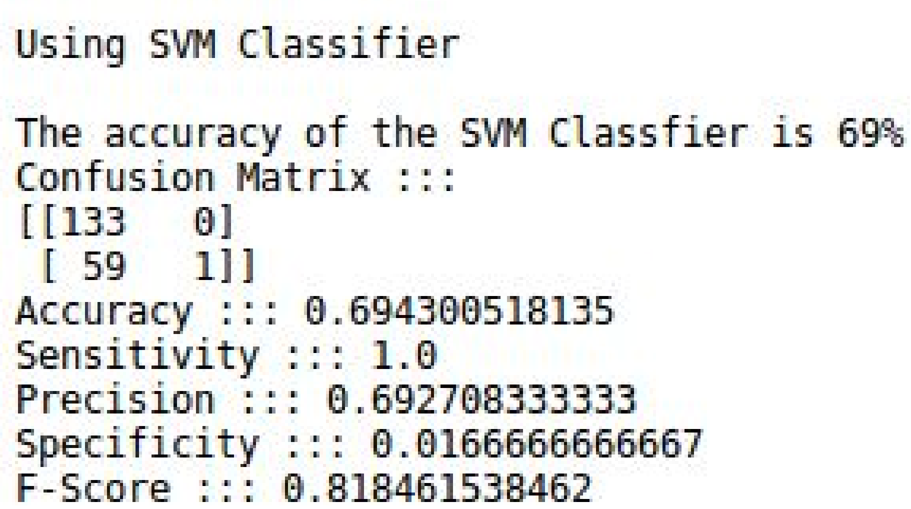
```
Using KNN Classifier

The accuracy of the knn classfier is 62%
Confusion Matrix :::
[[103  30]
 [ 42  18]]
Accuracy ::: 0.626943005181
Sensitivity ::: 0.774436090226
Precision ::: 0.710344827586
Specificity ::: 0.3
F-Score ::: 0.741007194245
```



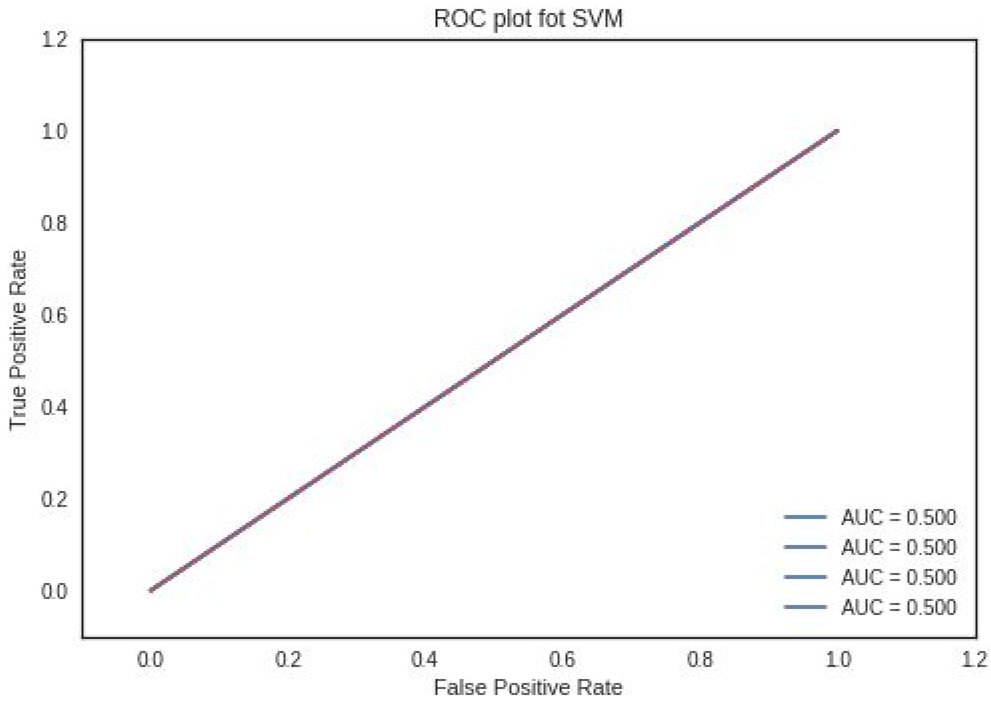
4. **Support Vector Machine**

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for classification problems.In this 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 differentiate the two classes very well.

The result for Support Vector Machine is as follows:



```
Using SVM Classifier

The accuracy of the SVM Classfier is 69%
Confusion Matrix :::
[[133   0]
 [ 59   1]]
Accuracy ::: 0.694300518135
Sensitivity ::: 1.0
Precision ::: 0.692708333333
Specificity ::: 0.0166666666667
F-Score ::: 0.818461538462
```



ROC Curve

Analysis and Comparison of the results:

Reasons for choice of algorithms:

- All are Supervised Classification Methods and are classified based on features.
- Naive Bayes : It is very simple to implement. When assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression and you need less training data.
- KNN : The K-nearest neighbor algorithm is that is simple to understand and easy to implement. With zero to little training time
- Logistic Regression : Based on probabilistic results.
- SVM : It works really well with clear margin of separation.

As we can see from the results :

- Naive Bayes has max Specificity out of all followed by KNN which implies that it has the best ability to classify people who don't have the disease and are not patients and we can conclude it from these models.

- The sensitivity for the SVM classifier is very high (close to 1), thus we can conclude that it correctly classifies the people who have the disease and are patients. This measure is relatively lower in KNN and Logistic Regression. (Logistic being higher than KNN) and Naive Bayes has low sensitivity among all.

- The Accuracy of Logistic Regression and SVM goes hand in hand with with very little difference where as KNN is slightly less and Naive Bayes Accuracy is very less compared to other models.

- All models have similar F-Square values but Naive Bayes has very less value relatively to other models

- Other measures are roughly the same and hence you can't make any conclusions based on them.

- Also, SVM works extremely well for binary classification.