STUDY ON LIVER DISEASE PREDICTION USING NAÏVE BAYES CLASSIFIER

Pre-synopsis for the submission of B.Tech 4th Year Project

Submitted by:

NAME	BRANCH	ROLL NO.	COURSE/BATCH
Shreya Kumari	CSE	1509033	B.tech/2015
Shubham Gupta	CSE	1509034	B.tech/2015
Swati Chandra	CSE	1509038	B.tech/2015
Nitish Kumar	CSE	1609004D	B.tech/2015

Under Supervision of:

Dr. S.C. Dutta

Head of Department

CSE & IT

B.I.T. Sindri

ABSTRACT

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 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 analyzed 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.

PROBLEM STATEMENT

There are many tools related to disease prediction. But particularly heart related diseases have been analyzed and risk level is generated. But generally there are no such tools that are used for prediction of general diseases. So, Liver Disease Predictor helps for the prediction of the general diseases.

Liver Disease Prediction system based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as dataset to the system. The system analyzes the symptoms provided by us as dataset and gives the probability of the disease as an output.

METHODOLOGY

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach we would come up with if we wanted to model a predictive modeling problem probabilistically. Naive Bayes simplifies the calculation of

probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

OUTPUT SUMMARY

Disease Prediction is done by implementing the Naive Bayes Classifier. Naive Bayes Classifier calculates the probability of the disease. Therefore, average prediction accuracy probability 54% is obtained.

ACKNOWLEDGEMENT

We wish to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude we give to our supervisor Dr. S.C. Dutta, Head of Department, B.I.T. Sindri whose contribution in stimulating suggestions, inspiring guidance and encouragement throughout the project work, helped us to coordinate my project especially in writing this report also gave the permission to use all required equipment and the necessary materials to complete the project. We are grateful to all the professors of the department of Computer Science and Engineering, B.I.T. Sindri, for their guidance and the support they provided us.

Last but not the least, our sincere thanks to all those individuals who motivated and provided us with valuable references contributing towards our work and well-wishers who have patiently extended all sorts of assistance for accomplishing this undertaking.

_____

Shreya Kumari
CSE 2015-19
1509033
B.I.T. SINDRI, DHANBAD

_____

Shubham Gupta
CSE 2015-19
1509034
B.I.T. SINDRI, DHANBAD

_____

Swati Chandra
CSE 2015-19
1509038
B.I.T. SINDRI, DHANBAD

_____

Nitish Kumar
CSE 2015-19
1609004D
B.I.T. SINDRI, DHANBAD

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

1.1 INTRODUCTION

At present, when one suffers from particular disease, then the person has to visit to doctor which is time consuming and costly too. Also if the user is out of reach of doctor and hospitals it may be difficult for the user as the disease cannot be identified. So, if the above process can be completed using an automated program which can save time as well as money, it could be easier to the patient which can make the process easier. There are other Liver related Disease Prediction System using data mining techniques that analyzes the risk level of the patient.

Liver is the largest gland in our body. The Liver's main job is to filter the blood coming from the digestive tract, before passing it to the rest of the body. The Liver is an important organ in our body, we can survive only one or two days if it shuts down.

The liver is an essential organ that has many functions in the body, including making proteins and blood clotting factors, manufacturing triglycerides and cholesterol, glycogen synthesis, and bile production.

The liver is a large organ that sits on the right hand side of the belly. The liver is the body's largest internal organ. Many different disease processes can occur in the liver, including infections such as hepatitis, cirrhosis (scarring), cancers, and damage by medications or toxins.

Symptoms of liver disease can include-

- jaundice
- abdominal pain and swelling
- confusion

- bleeding
- fatigue
- weight loss

Alcohol can be toxic to the liver (hepatotoxic), especially in high doses, and long-term alcohol abuse is a common cause of liver disease.

The liver is involved in metabolizing many toxins, including drugs and medications, chemicals, and natural substances.

The liver has multiple functions. It makes many of the chemicals required by the body to function normally, it breaks down and detoxifies substances in the body, and it also acts as a storage unit. Hepatocytes (hepar=liver + cyte=cell) are responsible for making many of the proteins (protein synthesis) in the body that are required for many functions, including blood clotting factors, and albumin, required to maintain fluid within the circulation system. The liver is also responsible for manufacturing cholesterol and triglycerides. Carbohydrates are also produced in the liver and the organ is responsible for turning glucose into glycogen that can be stored both in the liver and in the muscle cells. The liver also makes bile that helps with food digestion.

The liver plays an important role in detoxifying the body by converting ammonia, a byproduct of metabolism in the body, into urea that is excreted in the urine by the kidneys. The liver also breaks down medications and drugs, including alcohol, and is responsible for breaking down insulin and other hormones in the body. The liver is also stores vitamins and chemicals that the body requires as building blocks. These includes vitamin B12, folic acid, iron required to make red blood cells, vitamin A for vision, vitamin D for calcium absorption, and vitamin K to help blood to clot properly. The liver is a large organ and a significant amount of liver tissue needs to be damaged before a

person experiences symptoms of disease. Symptoms also may depend upon the type of liver disease.

The inflammation of hepatitis may be associated with pain in the right upper quadrant of the abdomen, nausea and vomiting. This may also be seen in people with gallstones.

People may have jaundice (have a yellow-orange hue to their skin) because the liver cannot metabolize bilirubin (the normal breakdown product of old red blood cells). There may be a tendency to bleed excessively or bruise easily because the liver is unable to manufacture blood clotting factors in adequate amounts.

Fatigue, weakness, weight loss, and shortness of breath because of muscle wasting; due to the inability of the liver to manufacture proteins. Because the liver is involved in the metabolism of sex hormones, gynecomastia (enlarged breast tissue in men) and impotence may occur.

In end-stage liver disease, ascites (fluid accumulation in the abdominal cavity), and leg swelling may occur because of inadequate production of albumin by the liver. There also may be difficulty in metabolizing ammonia causing its levels in the blood to rise, resulting in confusion due to encephalopathy (encephala=brain + pathy=dysfunction).

Liver Disease Predictor is a Machine Learning based application that predicts the disease of the user with respect to the given attribute values. Liver Disease Prediction system has data sets collected from different health related sites. With the help of Disease Predictor, the user will be able to know the probability of the disease with the given dataset’s attribute value.

As the use of internet is growing every day, people are always curious to know different new things. People always try to refer to the internet if any problem arises. People have access to internet than hospitals and doctors. People do not have immediate option when they suffer with particular disease. So, this system can be helpful to the people as they have access to internet 24 hours.

1.2 PROBLEM STATEMENT

There are many tools related to disease prediction. But particularly liver related diseases have been analyzed and risk level is generated. But generally there are no such tools that are used for prediction of general diseases. So, Liver Disease Predictor helps for the prediction of the liver diseases.

1.3 OBJECTIVE

1.3.1 GENERAL OBJECTIVE

To implement Naive Bayes Classifier that classifies the disease as per the dataset's attributes value.

1.3.2 SPECIFIC OBJECTIVE

To develop machine learning platform for the prediction of the disease.

CHAPTER 2 SYSTEM DESIGN

2.1 METHODOLOGY

Liver Disease Prediction has been already implemented using different techniques like Neural Network, decision tree and Naïve Byes algorithm. Particularly liver related disease is mostly analyzed. From the analysis it was found that Naive Bayes is more accurate than other techniques. So, Liver Disease Predictor also uses Naive Bayes for the prediction of different diseases.

For some types of probability models, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

2.1.1 DATA COLLECTION

Given a dataset containing various attributes of 583 Indian patients, use the features available in the dataset and define a supervised classification algorithm which can identify whether a person is suffering from liver disease or not.' The dataset for this problem is the ILPD (Indian Liver Patient Dataset) taken from the UCI Machine Learning Repository. Number of instances are 583. It is a multivariate data set, contain 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. All values are real integers. This data set contains 416 liver patient records and 167 non- liver patient records. The data set was collected from north east of Andhra Pradesh, India. This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90".

2.1.2 ATTRIBUTES INFORMATION

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.

The various fundamental measures of the Indian Liver Patient Dataset are 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

Table 2.1

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

Table 2.2

- As we can see from the tables above, the average age of males who are NOT patients is 40.6 and the average age of males who ARE patients is 46.9.

- We can also see that the mean of total bilirubin of NOT patients comes up to 1.24 whereas, in males who ARE patients, the mean of total bilirubin is 4.46.
- The mean albumin values for men who are Not patients is 35.32 whereas the men who ARE patients have a mean value of albumin 108.7.
- Therefore we can conclude that higher values of total bilirubin and albumin indicate that a patient is suffering from liver disease.

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

Table 2.3

4. Female who are patient:

	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

Table 2.4

- As we can see from the tables above, the average age of females who are NOT patients is 42.7 and the average age of females who ARE patients is also 43.3 .
- We can also see that the mean of total bilirubin of non patients comes up to 0.9 whereas, in females who ARE patients, the mean of total bilirubin is 3.0 .
- The mean albumin values for men who are Not patients is 29.7 whereas the men who ARE patients have a mean value of albumin 67.55 .
- Therefore we can conclude that higher values of total bilirubin and albumin indicate that a patient is suffering from liver disease.

2.2 ALGORITHM IMPLEMENTED

The algorithm implemented in this project is **Naive Bayes Classifier**.

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes has been studied extensively since the 1960s. It was introduced (though not under that name) into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the

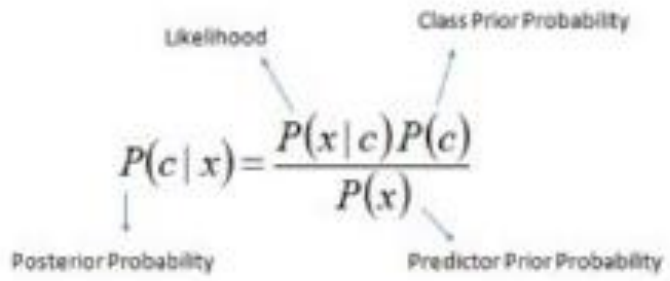
features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method.

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independent assumption. A more descriptive term for the underlying probability model would be the self-determining feature model. In basic terms, a Naive Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature . The Naive Bayes classifier performs reasonably well even if the underlying assumption is not true. The advantage of the Naive Bayes classifier is thatit only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. Because of independent variables are unspecified, only the variances of the variables for each label need to be determined and not the entire covariance matrix. In contrast to the Naive Bayes operator, the Naïve Bayes (Kernel) operator can be applied onnumerical attributes.

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



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

Above,

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

2.2.1 LEARNING PROCESS

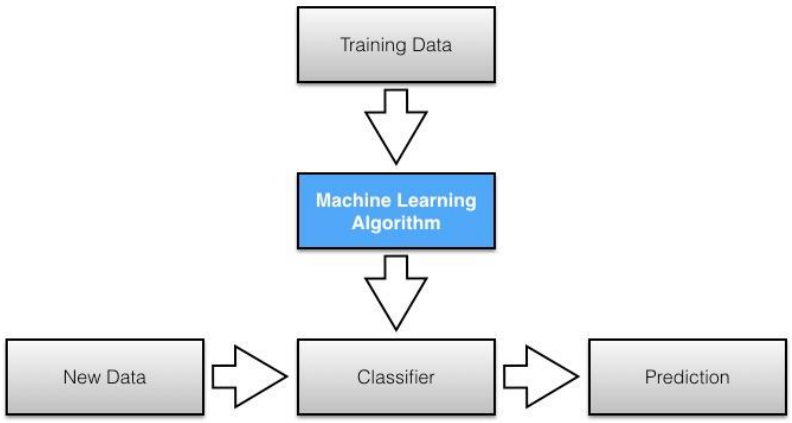


Fig 2.1

2.2.2 MODEL ARCHITECTURE

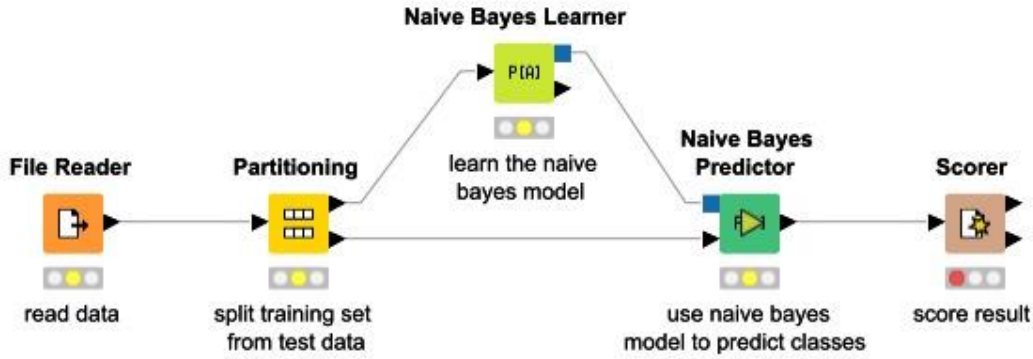


Fig 2.2

2.2.3 EXAMPLE MODEL

Sex classification

Problem: classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size.

Training

Example training set below:

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Table 2.5

The classifier created from the training set using a Gaussian distribution assumption would be (given variances are *unbiased* sample variances):

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	$3.5033*10^{-2}$	176.25	$1.2292*10^{2}$	11.25	$9.1667*10^{-1}$
female	5.4175	$9.7225*10^{-2}$	132.5	$5.5833*10^{2}$	7.5	1.6667

Table 2.5

Let's say we have equiprobable classes so P(male)= P(female) = 0.5. This prior probability distribution might be based on our knowledge of frequencies in the larger population, or on frequency in the training set.

Testing

Below is a sample to be classified as male or female.

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

Table 2.6

We wish to determine which posterior is greater, male or female. For the classification as male the posterior is given by:

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

For the classification as female the posterior is given by

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

The evidence (also termed normalizing constant) may be calculated:

$$\begin{aligned}\text{evidence} = P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male}) \\ + P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})\end{aligned}$$

However, given the sample, the evidence is a constant and thus scales both posteriors equally. It therefore does not affect classification and can be ignored. We now determine the probability distribution for the sex of the sample.

$$\begin{aligned}P(\text{male}) &= 0.5 \\ p(\text{height} \mid \text{male}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789\,,\end{aligned}$$

where μ=5.855 and σ^2=3.5033x10^{-2} are the parameters of normal distribution which have been previously determined from the training set. Note that a value greater than 1 is OK here – it is a probability density rather than a probability, because height is a continuous variable.

$$\begin{aligned}p(\text{weight} \mid \text{male}) &= 5.9881 \cdot 10^{-6} \\ p(\text{foot size} \mid \text{male}) &= 1.3112 \cdot 10^{-3} \\ \text{posterior numerator (male)} &= \text{their product} = 6.1984 \cdot 10^{-9} \\ P(\text{female}) &= 0.5 \\ p(\text{height} \mid \text{female}) &= 2.2346 \cdot 10^{-1} \\ p(\text{weight} \mid \text{female}) &= 1.6789 \cdot 10^{-2} \\ p(\text{foot size} \mid \text{female}) &= 2.8669 \cdot 10^{-1} \\ \text{posterior numerator (female)} &= \text{their product} = 5.3778 \cdot 10^{-4}\end{aligned}$$

Since posterior numerator is greater in the female case, we predict the sample is female.

CHAPTER 3 RESULT

3.1 EXPERIMENTAL RESULT

In this section, the results are analyzed which are given by the classification algorithms such as naïve Bayes. This work is implemented in PYTHON using KERAS API. Below figure represents the accuracy measure for the Naïve Bayes algorithms which is used in our model.

Our model result are as follows:

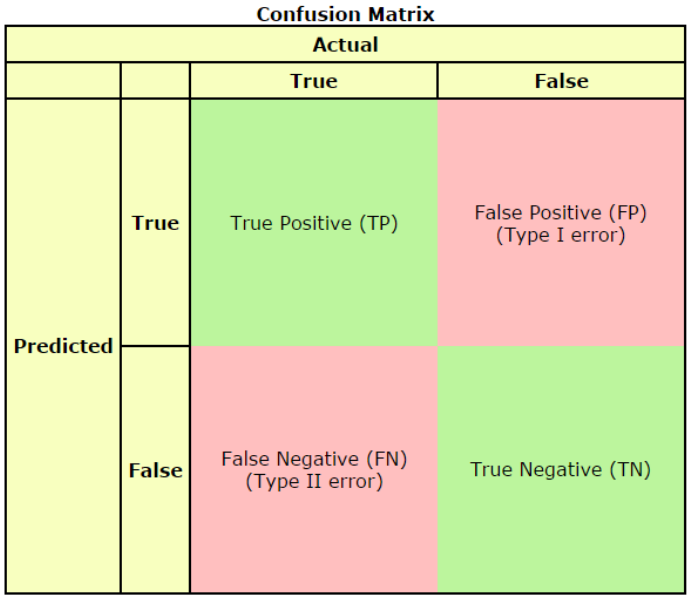
	precision	recall	f1-score	support
1	0.95	0.34	0.50	121
2	0.39	0.96	0.56	54
avg / total	0.78	0.53	0.52	175

3.2 RESULT UNDERSTANDING

3.2.1 CONFUSION MATRIX

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix,[4] is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).[2] The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).



3.2.2 CONFUSION MATRIX TERMINOLOGY

- ACCURACY

  Overall, how often is our model correct?

$$Accuracy = \frac{truepositives + truenegatives}{totalexamples}$$

As a heuristic, or rule of thumb, accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally. However, it does not give detailed information regarding its application to the problem.

The problem with using accuracy as your main performance metric is that it does not do well when you have a severe class imbalance. Let's use the dataset in the confusion matrix above. Let's say the negatives are normal transactions and the positives are fraudulent transactions. Accuracy will tell you that you're right 99% of the time across all classes.

But we can see that for the fraud class (positive), you're only right 50% of the time, which means you're going to be losing money. Hell, if you created a hard rule predicting that all transactions were normal, you'd be right 98% of the time. But that wouldn't be a very smart model, or a very smart evaluation metric. That's why, when your boss asks you to tell them "how accurate is that model?", your answer might be: "It's complicated."

To give a better answer, we need to know about *precision*, *recall* and *f1 scores*.

- PRECISION

  When the model predicts positive, how often is it correct?

$$Precision = \frac{truepositives}{truepositives + falsepositives}$$

Precision helps when the costs of false positives are high. So let's assume the problem involves the detection of skin cancer. If we have a model that has very low precision, then many patients will be told that they have melanoma, and that will include some misdiagnoses. Lots of extra tests and stress are at stake. When false positives are too high, those who monitor the results will learn to ignore them after being bombarded with false alarms.

- RECALL

$$Recall = \frac{truepositives}{truepositives + falsenegatives}$$

Recall helps when the cost of false negatives is high. What if we need to detect incoming nuclear missiles? A false negative has devastating consequences. Get it wrong and we all die. When false negatives are frequent, you get hit by the thing you want to avoid. A false negative is when you decide to ignore the sound of a twig breaking in a dark forest, and you get eaten by a bear. (A false positive is staying up all night sleepless in your tent in a cold sweat listening to every shuffle in the forest, only to realize the next morning that those sounds were made by a chipmunk. Not fun.) If you had a model that let in nuclear missiles by mistake, you would want to throw it out. If you had a model that kept you awake all night because *chipmunks*, you would want to throw it out, too. If, like most people, you prefer to not get eaten by the bear, and also not stay up all night worried about chipmunk alarms, then you need to optimize for an evaluation metric that's a combined measure of precision and recall.

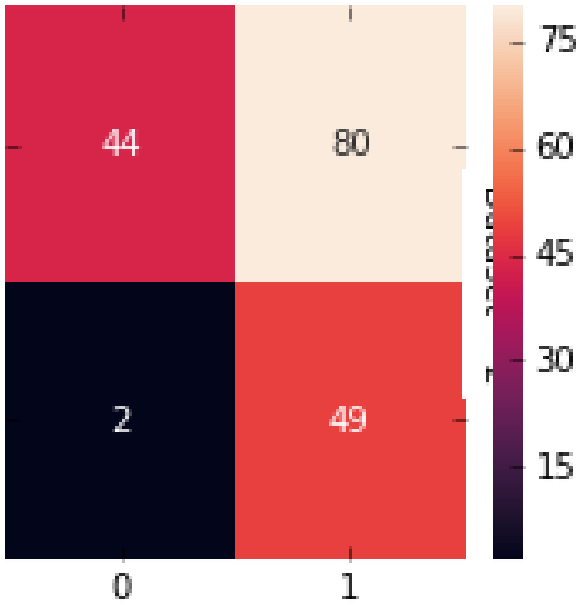
- F1 SCORE

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

F1 is an overall measure of a model's accuracy that combines precision and recall, in that weird way that addition and multiplication just mix two ingredients to make a separate dish altogether. That is, a good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

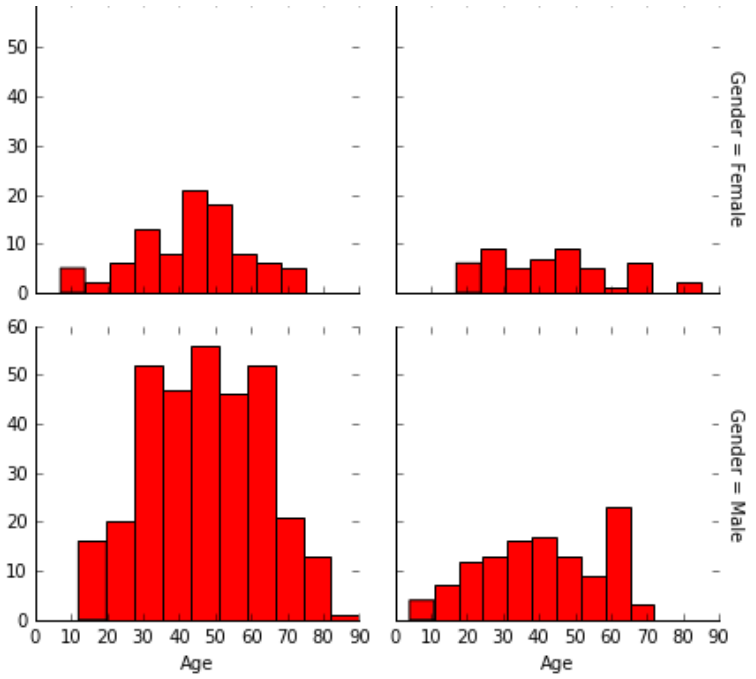
- ✓ true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
- ✓ true negatives (TN): We predicted no, and they don't have the disease.
- ✓ false positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- ✓ false negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

3.2.3 MODEL OUTPUT CONFUSION MATRIX

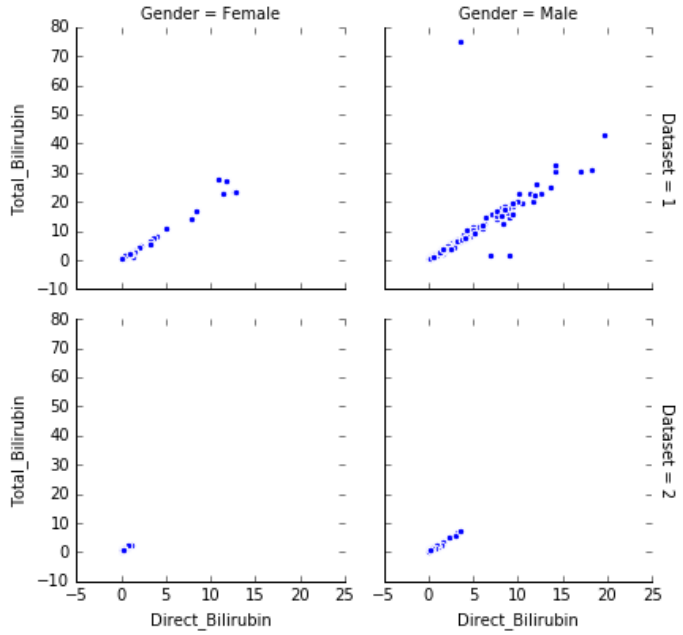


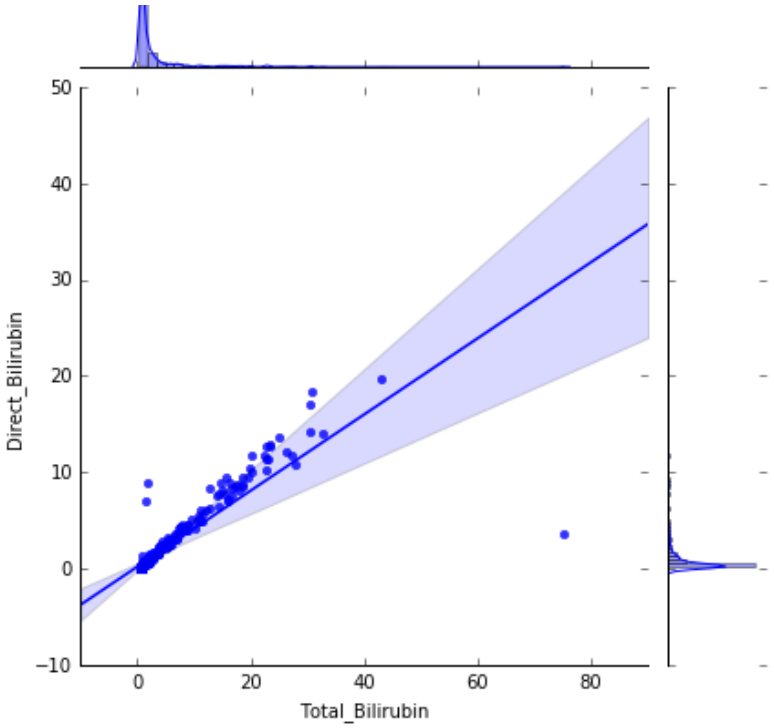
3.3 OUTPUT AS GRAPH

3.3.1 DISEASE BY GENDER AND AGE

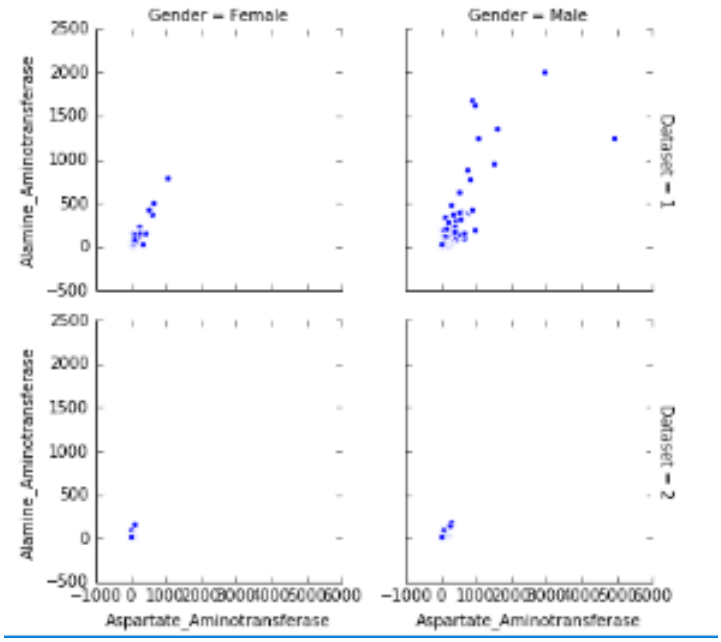


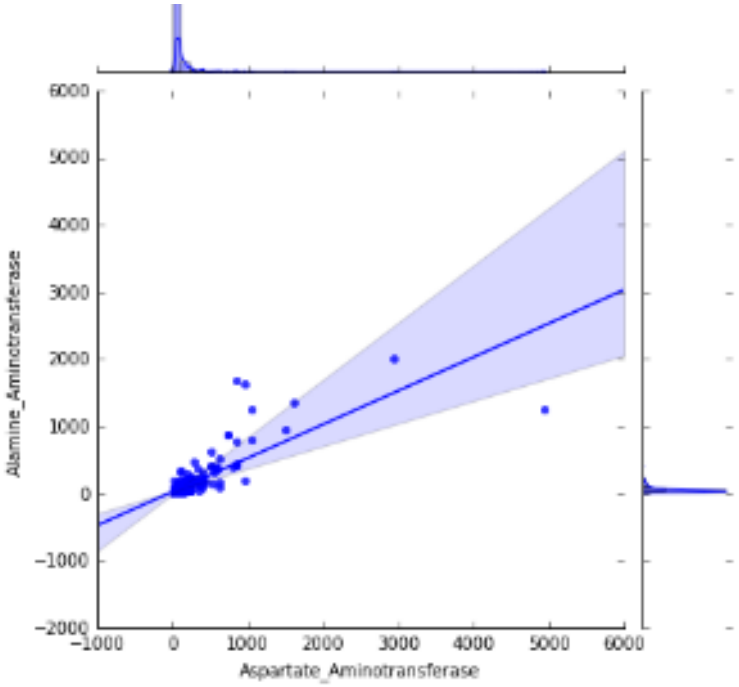
3.3.2 TOTAL-BILURUBIN AND DIRECT-BILURUBIN





3.3.3 ALAMINE AMINOTRANSFERASE AND ASPARTATE AMINOTRANSFERASE





3.4 CONCLUSION

This project aims to predict the liver disease on the basis of the values of certain attributes. The project is designed in such a way that the system takes dataset as input and produces output i.e. predict disease. Average prediction accuracy probability of 54% is obtained. Liver Disease Predictor was successfully implemented using Naïve Bayes Classifier.

3.5 CORRECTIVE MAINTENANCE

In case of any bugs left in the system, the bugs and issues will be fixed for smooth running of the application. The accuracy of the system can be further improved with other algorithms if needed.

3.6 ADAPTIVE MAINTENANCE

The features in the application can be added such as history of the disease can be kept in the log. The available list of symptoms can also be added for covering more number of diseases.

3.7 SCOPE

This project aims to provide a machine learning application to predict the liver disease on the basis of various dataset's attribute. The user can select various dataset's attribute and can find the diseases with their probabilistic figures.

3.8 LIMITATIONS

The limitations of this project are:

a. Liver Disease Predictor does not recommend medications of the disease.

b. Past history of the disease has not been considered.

REFERENCES

[1] Sa’diyah Noor Novita Alfisahrin, Teddy Mantoro, Data Mining Techniques For Optimatization of Liver Disease Clasification 978-1-4799-2758-6/13 $31.00 © 2013 IEEE
DOI10.1109/ACSAT.2013.81

[2] Sina Bahramirad, Aida Mustapha, Maryam Eshraghi, Classification of Liver Disease Diagnosis: A Comparative Study, ISBN: 978-1-4673-5256-7/13/$31.00 ©2013 IEEE

[3] https://en.wikipedia.org/wiki/Confusion_matrix

[4] Adam, S., & Parveen, A. (2012). Prediction System for Liver Disease Using Naive Bayes.

[5] K.M. Al-Aidaroos, A. B. (n.d.). 2012. Medical Data Classification With Naive Bayes Approach

[6] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification With Naive Bayes Approach. Information Technology Journal.

[7] https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[8]https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

[9] https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

[10] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[11] https://www.coursera.org/learn/machine-learning/home/welcome