



## SCHOOL OF COMPUTER SCIENCE


### MID-EVALUATION (Weightage 10%) AUGUST 2023 SEMESTER

<b>MODULE NAME</b>	<b>: DATA MINING</b>
<b>MODULE CODE</b>	<b>: ITS61504</b>
<b>DUE DATE</b>	<b>: 01.11.2023- 10.11.2023, 8.00PM (MYT)</b>
<b>PLATFORM</b>	<b>: MyTIMES</b>

This paper consists of **THREE (3)** pages, inclusive of this page.

#### ***STUDENT DECLARATION***

- 1. I confirm that I am aware of the University's Regulation Governing Cheating in a University Test and Assignment and of the guidance issued by the School of Computing and IT concerning plagiarism and proper academic practice, and that the assessed work now submitted is in accordance with this regulation and guidance.*
- 2. I understand that, unless already agreed with the School of Computing and IT, assessed work may not be submitted that has previously been submitted, either in whole or in part, at this or any other institution.*
- 3. I recognise that should evidence emerge that my work fails to comply with either of the above declarations, then I may be liable to proceedings under Regulation*

No	Student Name	Student ID	Date	Signature	Score
	Felicia Dossou	0364975	10/11/2023		

#### Case Study - Hepatitis

Hepatitis diseases can be defined as inflammation of the liver which can lead to liver damage, cancer and liver failure. In this case study, the objective is to perform a prediction on diagnosis of hepatitis which involves several attributes and instances.

According to the dataset, there are 19 attributes that comprises demographic information and historic medical records of 155 instances have been collected from different patients.

## Instructions:

1. Perform Exploratory Data Analysis. Justify any feature selection and elimination that is necessary.
2. Prepare the dataset with various pre-processing techniques
3. Develop a predictive method or classification model to determine the factor and the condition of the patients.
4. Evaluate the performance of the model using a performance matrix and determine which attributes are mostly the reason for these diseases. Justify your answer.

## TABLE OF CONTENT

<b>1. Exploratory Data Analysis</b>	<b>3</b>
Data exploration	3
<b>2. Preprocessing</b>	<b>4</b>
Data cleaning	4
Data transformation	4
Distribution visualizations	5
Correlation analysis	6
<b>3. Method : Logistic Regression</b>	<b>7</b>
Summary	7
Predictions	8
<b>4. Performance evaluation</b>	<b>9</b>
Performance matrix	9
Conclusion	9

# 1. Exploratory Data Analysis

## Data exploration

Using some basic functions such as `dim(df)` or `colnames(df)`, we find the dimensions of our dataframe: 155 rows and 20 columns named "class", "age", "sex", "steroid", "antiviral", "fatigue", "malaise", "anorexia", "liver.big", "liver.firm", "spleen", "spiders", "ascites", "varices", "bilirubin", "phosphate", "sgot", "albumin", "protime" and "histology".

Moreover, `sum(is.na(df))` function gives us the confirmation of no missing value (NA)

In order to have an overall statistical view of our dataset, we use `summary(df)` to display the minimum, maximum, mean, median and quartiles of our numerical variables columns.

```
      class      age      sex      steroid      antiviral      fatigue
Min.   :1.000  Min.   : 7.0  Min.   :1.000  Length:155  Min.   :1.000  Length:155
1st Qu.:2.000  1st Qu.:32.0  1st Qu.:1.000  Class :character  1st Qu.:2.000  Class :character
Median :2.000  Median :39.0  Median :1.000  Mode  :character  Median :2.000  Mode  :character
Mean   :1.794  Mean   :41.2  Mean   :1.103  Mean   :1.845  Mean   :1.845
3rd Qu.:2.000  3rd Qu.:50.0  3rd Qu.:1.000  3rd Qu.:2.000  3rd Qu.:2.000
Max.   :2.000  Max.   :78.0  Max.   :2.000  Max.   :2.000

      malaise      anorexia      liver.big      liver.firm      spleen
Length:155  Length:155  Length:155  Length:155  Length:155
Class :character  Class :character  Class :character  Class :character  Class :character
Mode  :character  Mode  :character  Mode  :character  Mode  :character  Mode  :character

      spiders      ascites      varices      bilirubin      phosphate
Length:155  Length:155  Length:155  Length:155  Length:155
Class :character  Class :character  Class :character  Class :character  Class :character
Mode  :character  Mode  :character  Mode  :character  Mode  :character  Mode  :character

      sgot      albumin      protime      histology
Length:155  Length:155  Length:155  Min.   :1.000
Class :character  Class :character  Class :character  1st Qu.:1.000
Mode  :character  Mode  :character  Mode  :character  Median :1.000
                                          Mean   :1.452
                                          3rd Qu.:2.000
                                          Max.   :2.000
```

To find the unique elements of each column, we use `unique(col)` and we observe that missing results are denoted '?'

## 2. Preprocessing

This second step is performed to clean, transform and organize the data for modeling.

### Data cleaning

#### Missing values

As I notice that my missing variables were denoted '?' in my dataset, I run a loop to `sum()` up their number for each attribute. Then I convert them into percentages.

```
[1] "There are 0 missing values in class meaning 0 %"
[1] "There are 0 missing values in age meaning 0 %"
[1] "There are 0 missing values in sex meaning 0 %"
[1] "There are 1 missing values in steroid meaning 0.65 %"
[1] "There are 0 missing values in antiviral meaning 0 %"
[1] "There are 1 missing values in fatigue meaning 0.65 %"
[1] "There are 1 missing values in malaise meaning 0.65 %"
[1] "There are 1 missing values in anorexia meaning 0.65 %"
[1] "There are 10 missing values in liver.big meaning 6.45 %"
[1] "There are 11 missing values in liver.firm meaning 7.1 %"
[1] "There are 5 missing values in spleen meaning 3.23 %"
[1] "There are 5 missing values in spiders meaning 3.23 %"
[1] "There are 5 missing values in ascites meaning 3.23 %"
[1] "There are 5 missing values in varices meaning 3.23 %"
[1] "There are 6 missing values in bilirubin meaning 3.87 %"
[1] "There are 29 missing values in phosphate meaning 18.71 %"
[1] "There are 4 missing values in sgot meaning 2.58 %"
[1] "There are 16 missing values in albumin meaning 10.32 %"
[1] "There are 67 missing values in protime meaning 43.23 %"
[1] "There are 0 missing values in histology meaning 0 %"
```

Knowing that I have a limited number of observations (155), I decide to take out attributes that have at least 10% of missing variables. This is why I get rid of "phosphate", "albumin" and "protime" columns.

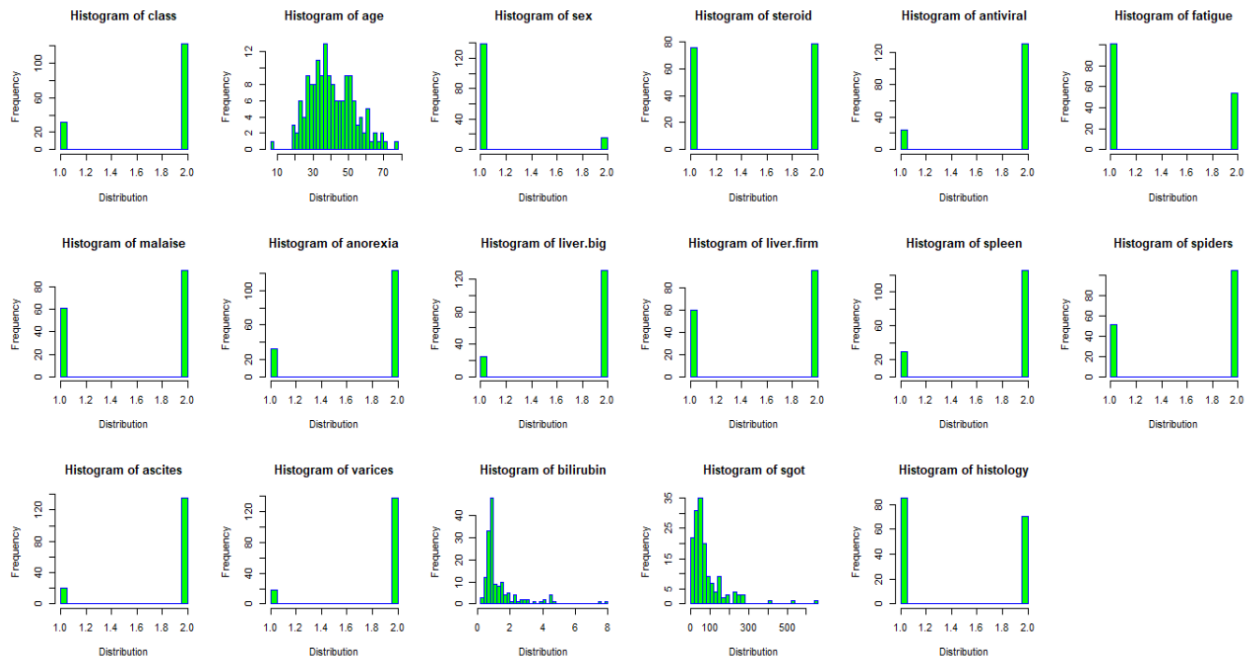
But there are still some rows with missing values. Therefore I proceed with imputation to replace missing data with substituted values. I use **mean imputation** for continuous variables and **mode imputation** for categorical variables.

### Data transformation

As we can see in the summary, many variables are of type 'character' but are indeed numbers. Therefore we are going to perform hot encoding and convert them into numerical values using `as.numeric()` function

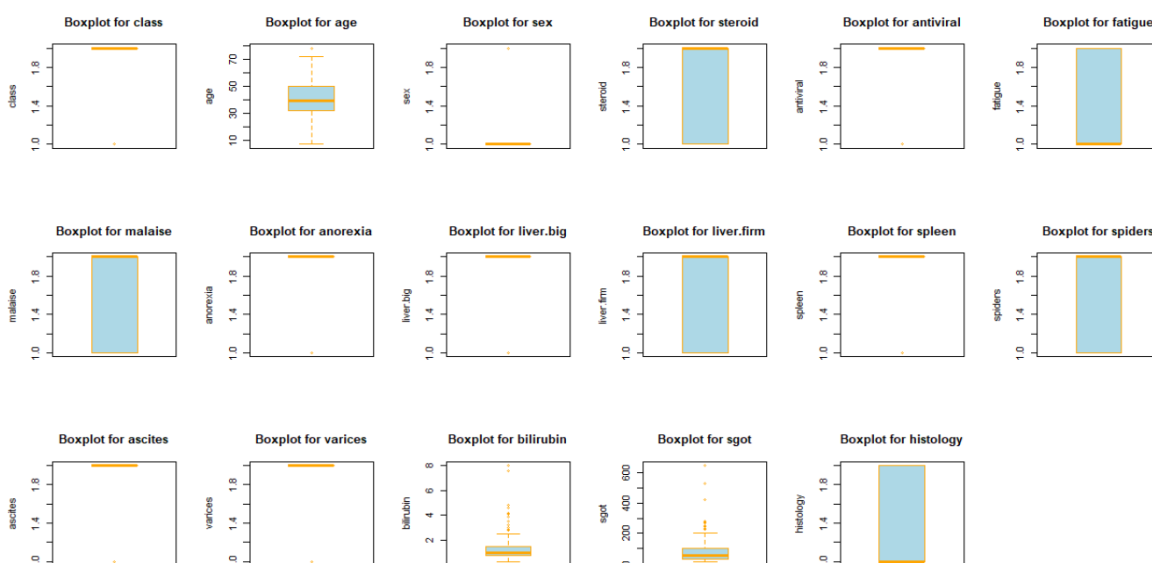
# Distribution visualizations

In order to have a better understanding on how the data is distributed, we created histograms with `hist()` for each of our attributes, that we placed into 6x3 subsets.



## Outliers

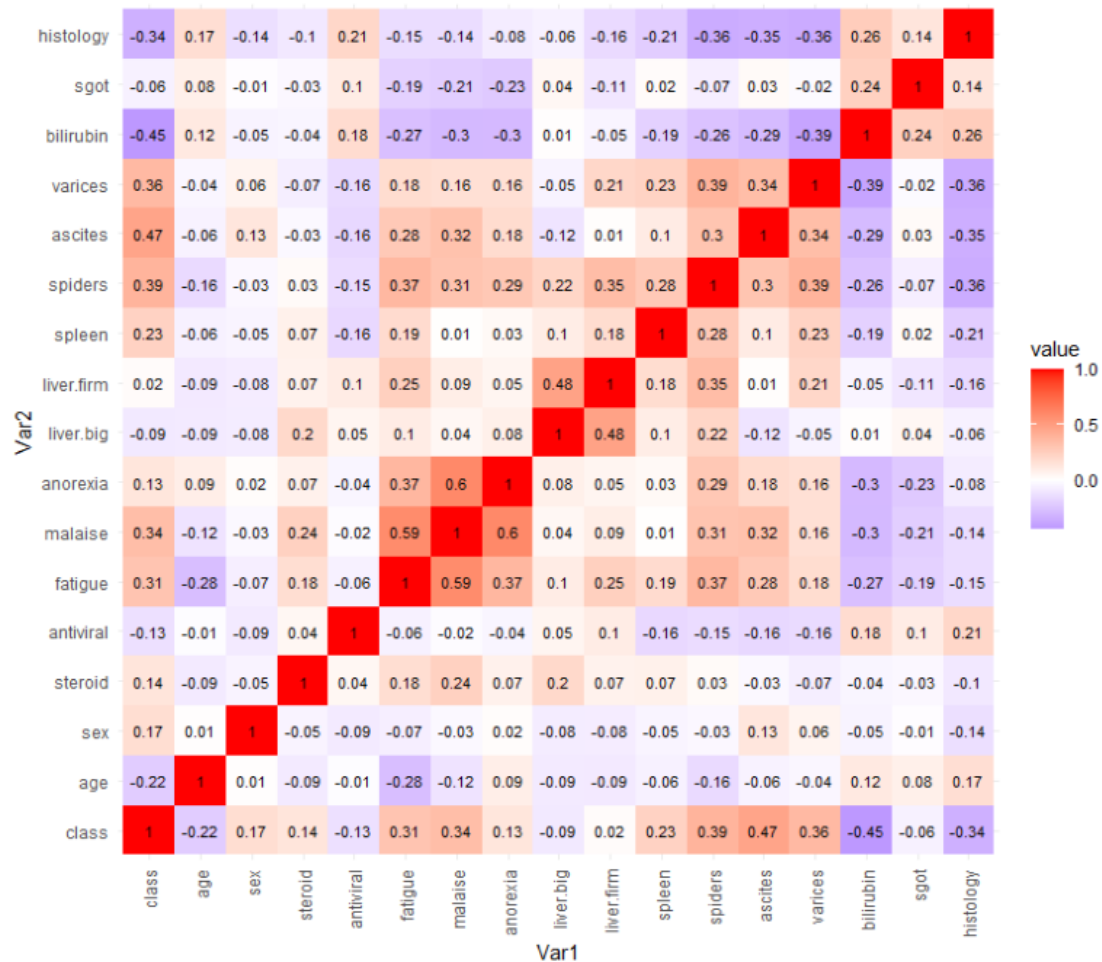
Aiming to identify and deal with outliers that would skew the analysis, we used `boxplot()` for a better visualization of the distribution according to median and quartiles.



Some variables have visible outliers but I think they are significant to diagnosis. Extreme values can be important for diagnosis therefore I can't exclude medical cases here. However, in order to decide which attributes have low impact on Hepatitis diagnosis and can be taken out, I proceed with correlation analysis.

## Correlation analysis

Using `ggplot()`, I can visualize and interpret the relationship between all my variables.



Making use of the following table, I decide to remove attributes that have a negligible correlation to "class" (.00 to -.30): "age", "sex", "steroid", "antiviral", "anorexia", "liver.big", "spleen", "sgot".

Size of Correlation	Interpretation
.90 to 1.00 (–.90 to –1.00)	Very high positive (negative) correlation
.70 to .90 (–.70 to –.90)	High positive (negative) correlation
.50 to .70 (–.50 to –.70)	Moderate positive (negative) correlation
.30 to .50 (–.30 to –.50)	Low positive (negative) correlation
.00 to .30 (.00 to –.30)	negligible correlation

### 3.Method : Logistic Regression

I chose the Logistic regression method because it is well suited to predict the outcome of a categorical dependent variable ("class" can be interpreted as binary : having Hepatitis or not) based on other predictor variables.

#### Summary

After executing a binary transformation on values in "class", I build a logistic regression model using `glm()` to obtain statistical information such as **coefficients of estimate** (change in the log odds of the outcome), **standard error** (variability of the coefficient estimate), **z-value** (number of sd the coefficient is from zero) and **pr(>|z|)** (statistical significance of the coeff) . Here is the outcome:

Logistic Regression for fatigue against Hepatitis

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.5351     0.8420  -1.823  0.06826 .
fatigue      2.3966     0.7528   3.184  0.00145 **
---
```

Logistic Regression for malaise against Hepatitis

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.2412     0.6341  -1.958  0.0503 .
malaise      1.7433     0.4389   3.972 7.13e-05 ***
---
```

Logistic Regression for liver.firm against Hepatitis

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.1841     0.6772   1.749  0.0804 .
liver.firm   0.1011     0.4050   0.250  0.8029
---
```

Logistic Regression for spiders against Hepatitis

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.6882     0.6560  -2.573  0.0101 *
spiders      1.9645     0.4365   4.500 6.79e-06 ***
---
```

Logistic Regression for ascites against Hepatitis

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.5664     1.0082  -3.537 0.000404 ***
ascites      2.7191     0.5497   4.946 7.56e-07 ***
---
```

Logistic Regression for varices against Hepatitis

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.6130     0.9956  -2.624  0.00868 **
varices      2.1611     0.5385   4.013 6e-05 ***
---
```

Logistic Regression for bilirubin against Hepatitis

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.7434     0.3958   6.932 4.16e-12 ***
bilirubin    -0.8875     0.2069  -4.290 1.78e-05 ***
---
```

Logistic Regression for histology against Hepatitis

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	4.2338	0.8276	5.116	3.13e-07 ***
histology	-1.8230	0.4668	-3.905	9.41e-05 ***

---

As we can see "fatigue", "malaise", "spiders", "ascites", "varices", "bilirubin", "histology" show significant coefficients, suggesting that they are important predictor variables in determining the likelihood of hepatitis.

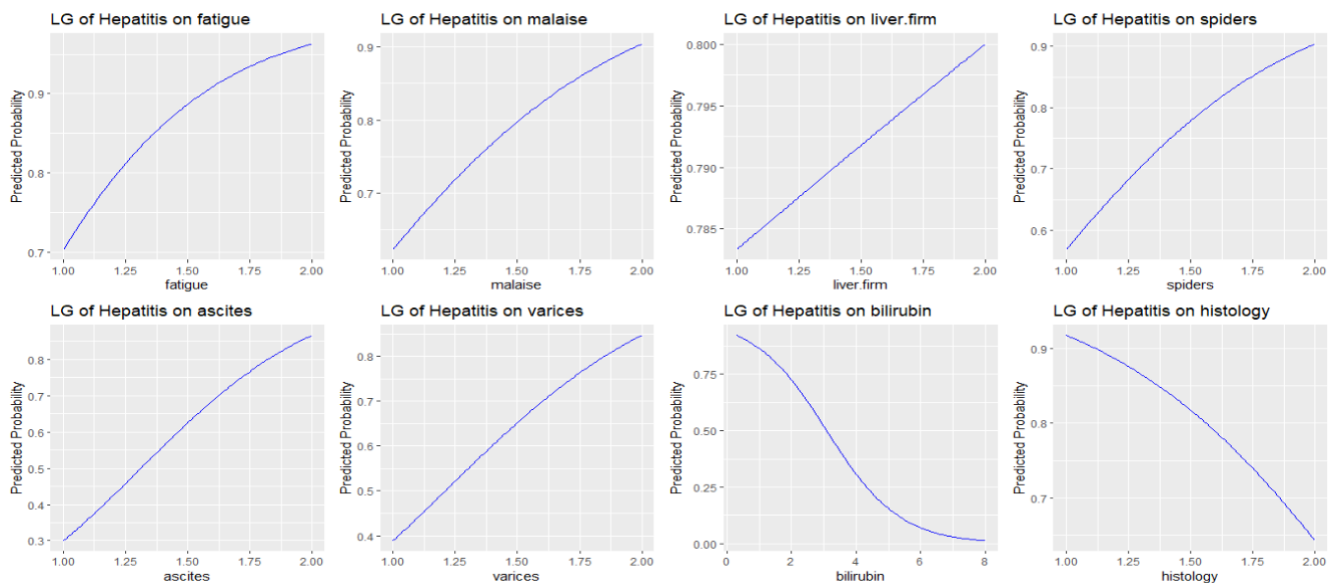
## Predictions

Next, we are trying to generate predictions building the same logistic regression model but instead for new data points created within the range of each attribute. For this purpose, we create **x\_values** spanning from the minimum to the maximum value.

The outcome shows predicted probabilities for our new data points, as we can see the outcome for instance for fatigue :

```
Logistic Regression for fatigue against Hepatitis
1 2 3 4 5 6 7 8 9 10 11
0.7029703 0.7080001 0.7129795 0.7179078 0.7227844 0.7276087 0.7323801 0.7370981 0.7417623 0.7463722 0.7509274
12 13 14 15 16 17 18 19 20 21 22
0.7554277 0.7598726 0.7642620 0.7685955 0.7728731 0.7770945 0.7812597 0.7853685 0.7894210 0.7934170 0.7973568
23 24 25 26 27 28 29 30 31 32 33
0.8012401 0.8050673 0.8088384 0.8125535 0.8162128 0.8198165 0.8233648 0.8268580 0.8302964 0.8336802 0.8370098
34 35 36 37 38 39 40 41 42 43 44
0.8402855 0.8435077 0.8466767 0.8497930 0.8528570 0.8558690 0.8588296 0.8617393 0.8645984 0.8674074 0.8701670
45 46 47 48 49 50 51 52 53 54 55
0.8728775 0.8755396 0.8781536 0.8807203 0.8832401 0.8857135 0.8881412 0.8905237 0.8928616 0.8951554 0.8974058
56 57 58 59 60 61 62 63 64 65 66
0.8996133 0.9017784 0.9039019 0.9059842 0.9080260 0.9100279 0.9119904 0.9139142 0.9157998 0.9176478 0.9194588
67 68 69 70 71 72 73 74 75 76 77
0.9212335 0.9229722 0.9246758 0.9263447 0.9279794 0.9295807 0.9311490 0.9326849 0.9341890 0.9356617 0.9371037
78 79 80 81 82 83 84 85 86 87 88
0.9385156 0.9398978 0.9412508 0.9425752 0.9438716 0.9451404 0.9463822 0.9475974 0.9487865 0.9499501 0.9510887
89 90 91 92 93 94 95 96 97 98 99
0.9522026 0.9532924 0.9543585 0.9554015 0.9564217 0.9574196 0.9583956 0.9593503 0.9602839 0.9611969 0.9620898
100
0.9629630
```

To have a better global understanding, I create a **plot()** for each of my attributes.





## 4. Performance evaluation

### Performance matrix

After splitting the data into **training set** and **testing set**, we make prediction on both and print the `table()` of both predicted and actual classes :

```
Distribution of Predicted Classes:
```

```
0 1  
6 25
```

```
Distribution of Actual Classes:
```

```
0 1  
8 23
```

We observe that the distribution of predicted classes closely matches the distribution of actual classes, which is a first good indicator of model accuracy.

Moreover once I check the predictions, I use `confusionMatrix()` to evaluate the model performance.

	Reference	
Prediction	0	1
0	4	2
1	4	21

		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative

According to my data

**TP** = 4 cases were correctly predicted as class Hepatitis

**TN** = 21 cases correctly predicted as non Hepatitis

**FP** = 4 cases wrongly predicted as Hepatitis

**FN** = 2 cases wrongly predicted as non Hepatitis

The formula of accuracy is  $A = (TP+TN)/(TP+TN+FP+FN)$

$A = (4+21)/(4+21+2+4) = 0.8065$  approximately

The accuracy shows 80% on average which is a fair score of prediction.

### Conclusion

The prediction model, with an accuracy of 80%, confirms that the attributes of "fatigue", "malaise", "spiders", "ascites", "varices", "bilirubin" and "histology" are predominantly responsible for Hepatitis.