Heart Failure Prediction Report

Mohammad S. Sammoudi

27 July, 2021

1.0 Introduction

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide. Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

1.1 Objective

The main objective of this project is to explore the Heart Failure Dataset, and to apply several models of machine learning on it. This aims to find the optimal model that gives best performance. The best model will give best predictions on heart failure.

1.2 Dataset overview

We used a dataset from kaggle website (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data). This dataset is tidy data and includes 299 observations with 13 variables as shown below:

```
## [1] 299  13
```

and the structure of the data is shown below:

```
## 'data.frame':    299 obs. of  13 variables:
##  $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
##  $ anaemia                 : int  0 0 0 1 1 1 1 1 0 1 ...
##  $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
##  $ diabetes                : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
##  $ high_blood_pressure     : int  1 0 0 0 0 1 0 0 0 1 ...
##  $ platelets               : num  265000 263358 162000 210000 327000 ...
##  $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
##  $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
##  $ sex                     : int  1 1 1 1 0 1 1 1 0 1 ...
##  $ smoking                 : int  0 0 1 0 0 1 0 1 0 1 ...
##  $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
##  $ DEATH_EVENT             : int  1 1 1 1 1 1 1 1 1 1 ...
```

Description of the variables

The dataset has 13 variables:- 1- age: the age of the patient and they are between 40 and 95 years old.(num) 2- anaemia: wheather the patient has anaemia or not(Decrease of red blood cells or hemoglobin) (int 0 or 1). 3- creatinine_phosphokinase: The level of creatinine phosphokinase in the blood.(int) 4- diabetes: wheatherthe patiens has diabetes or not(int 0 or 1). 5- ejection_fraction: how well your left ventricle (or right ventricle) pumps blood with each heart beat. 6- high_blood_pressure: wheather the patinet has hypertension or not (int 0 or 1). 7- platelets: number of platelets in the blood.(num) 8- serum_creatinine: The measure of creatinine in blood (num). 9- serum_sodium: The measure of Sodium in blood(int). 10- sex: The gender male(1) or female(0)(int). 11- smoking: wheather the patinet smoke or not (int 0 or 1). 12- time: Follow up period in days (int) (I will exclude this variable from analysis). 13- DEATH_EVENT: if the patient died during the follow-up period. 0 for no and 1 for yes.

Let's look at the first 6 results from the data set.

```
#show the first 6 rows in our dataset
head (heartfailure.dat)
```

```
##   age anaemia creatinine_phosphokinase diabetes ejection_fraction
## 1  75       0                      582        0                20
## 2  55       0                     7861        0                38
## 3  65       0                      146        0                20
## 4  50       1                      111        0                20
## 5  65       1                      160        1                20
## 6  90       1                       47        0                40
##   high_blood_pressure platelets serum_creatinine serum_sodium sex smoking time
## 1                   1    265000              1.9          130   1       0    4
## 2                   0    263358              1.1          136   1       0    6
## 3                   0    162000              1.3          129   1       1    7
## 4                   0    210000              1.9          137   1       0    7
## 5                   0    327000              2.7          116   0       0    8
## 6                   1    204000              2.1          132   1       1    8
##   DEATH_EVENT
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

2.0 Visualization and Exploratory Data Analysis EDA

In this section, I will start visualizing the variables to get insights about them, and to find the correlation between them.

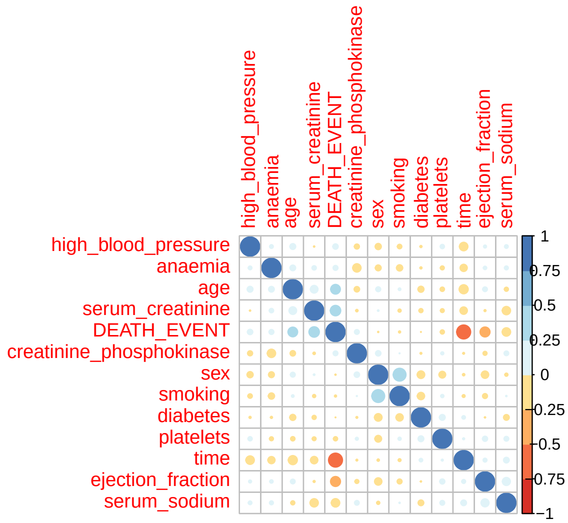
Let's check if there are any missing data in the dataset.

```
# Checking if there are any missing values in the dataset
sum(is.na(heart_failure_data))
```

```
## [1] 0
```

We see that there is no missing data.

Now let's find the correlation between variables throug a correlation matrix.



From the plot above, we see that some variables have strong corrrelation with each other, but most of them have weak correlation. The average correlation in the dataset is 0.156153:

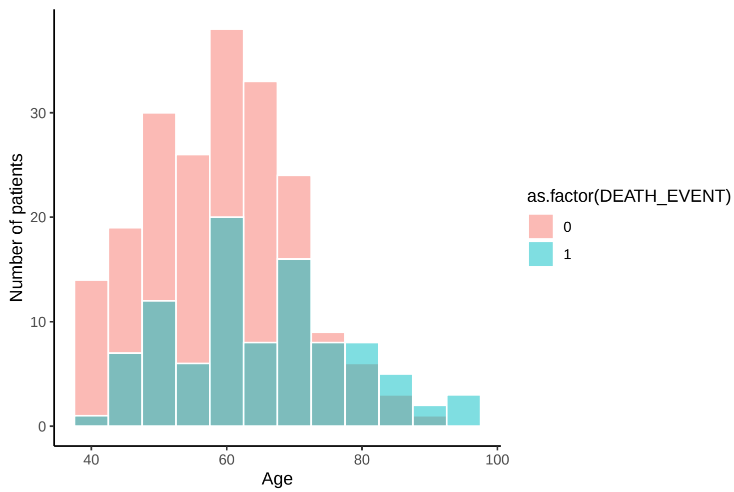
```r
mean(abs(df_cor))
```

```
## [1] 0.156153
```

And now we'll start exploring varibles one by one and plot them to conclude a results about them on how they can affect our classification purpose.

2.1 Age

The first variable in our dataset is the Age. The follwing table shows the density plot of patients ages.



Age Distribution with Death Event

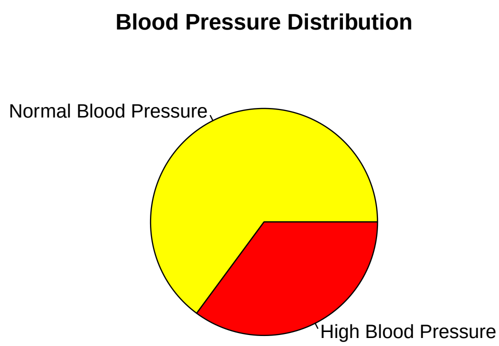
From above figure, we can condlude the follwing:-

1- The average age of the patients seems to be between 55 to 75 years, With the maximum age being 95 and the minimum being 40 years.

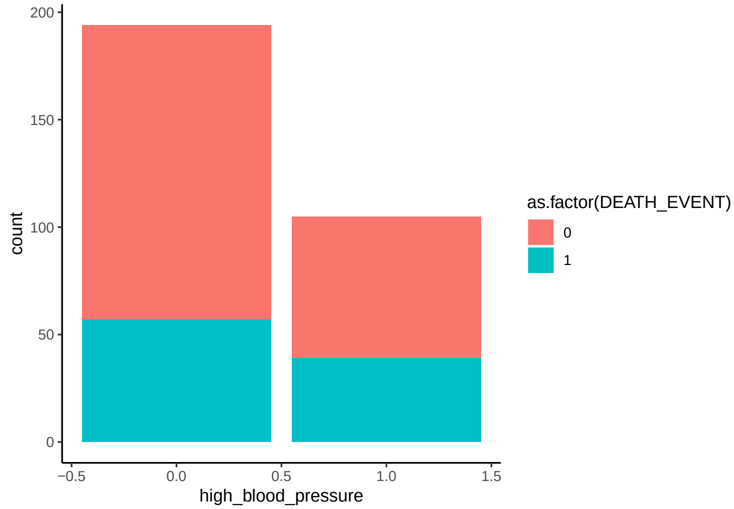
2- As the age increses, the probability of death increases.

2.2 Blood Pressure

The figure below shows the distribution of patiens with presence of high blood pressure or not.



Now, let's see how blood pressure affect the heart failure. This is shown in the next figure.

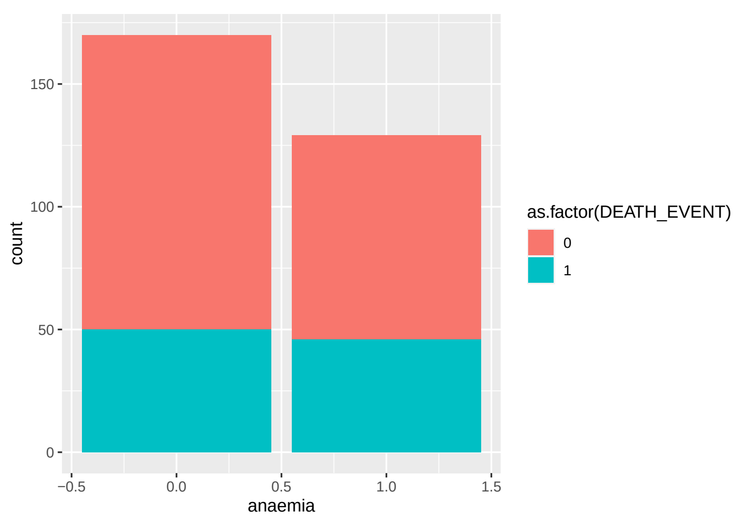


The figure above shows the realtion of blood pressure with heart failure. We can see from this figure that presence of high blood pressure not increase the probability of heart failure.

2.3 Anaemia

Anaemia is the Decrease of red blood cells or hemoglobin, so does thre is a relation between anaemia and heart failure.

The follwing figure shows the anaemia distribution and there are no effect on DEATH_EVENT.

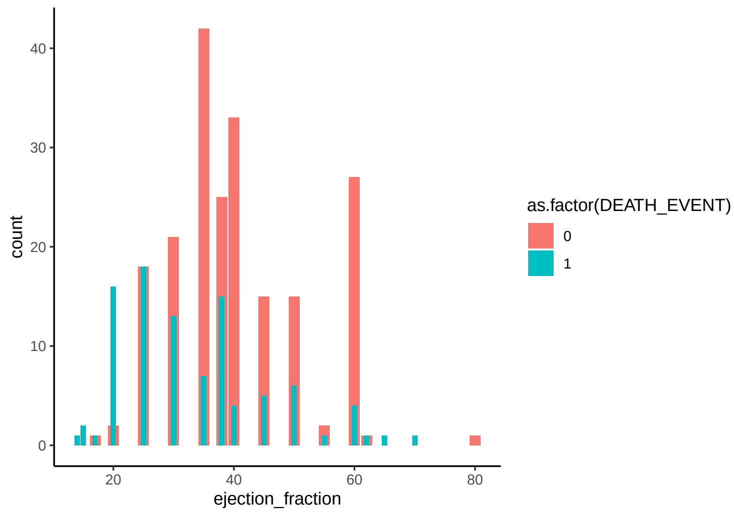


From above figure, we conclude that approximately heart failure does not has a realtion with anaemia.

2.4 Ejectino Fraction

Ejection fraction (EF) is a measurement, expressed as a percentage, of how much blood the left ventricle pumps out with each contraction. An ejection fraction of 60 percent means that 60 percent of the total amount of blood in the left ventricle is pushed out with each heartbeat.

This indication of how well the heart is pumping out blood can help to diagnose and track heart failure. A normal heart's ejection fraction may be between 50 and 70 percent.

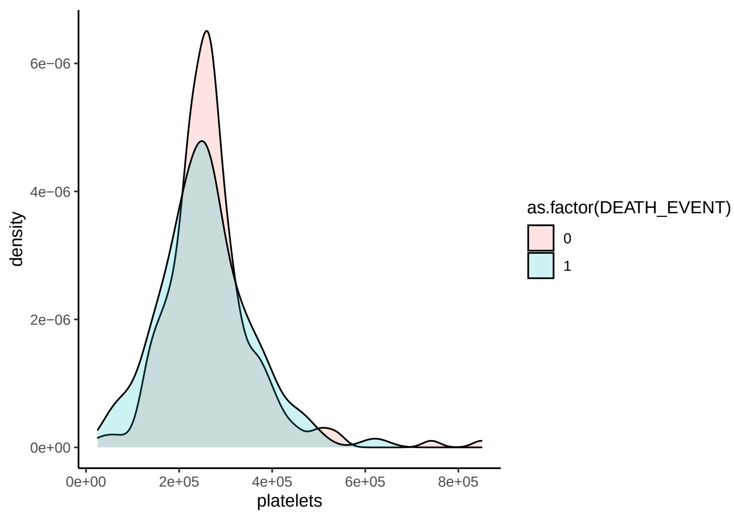


According to the plot above when ejection fraction is low, then the heart failure becomes more likely to happen.

2.5 PlateLets

This variable is the number of platelets in the blood.

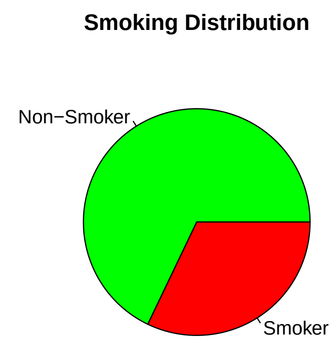
The following figure shows that the distributions of Platlets count in the absence or presence of death events are similar.



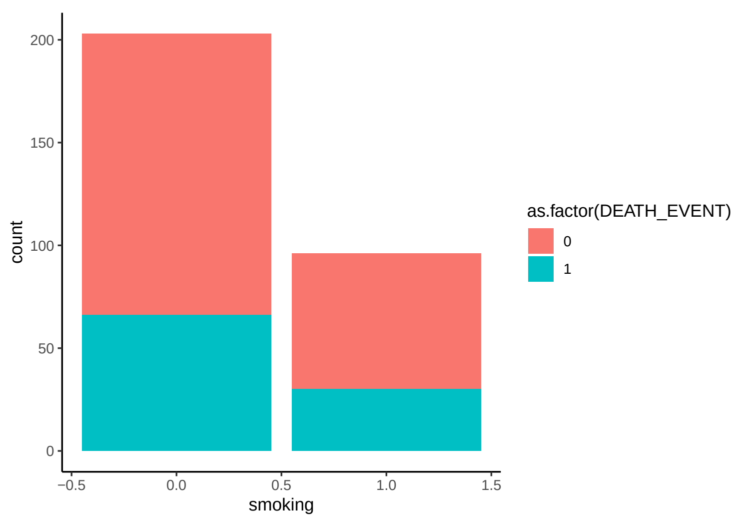
2.6 Smoking

This variable is factor and shows the patients wheather they are smoking or not.

The distribution of smokers in the data set is shown below:



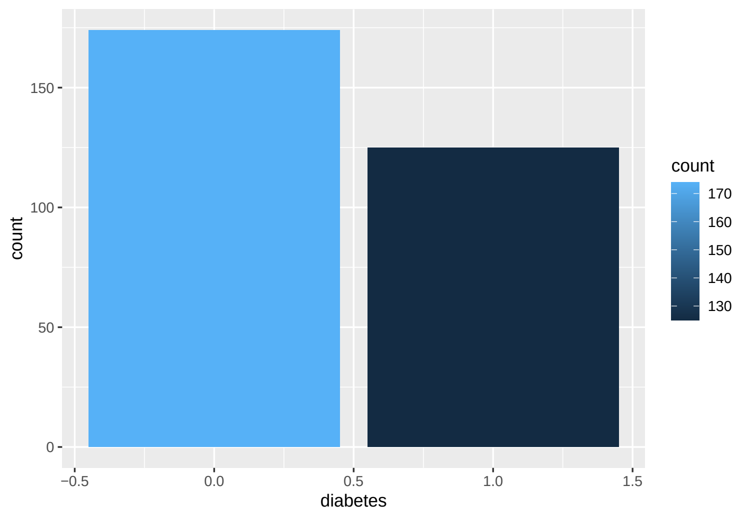
And the Smoking distribution with death event is shown below:



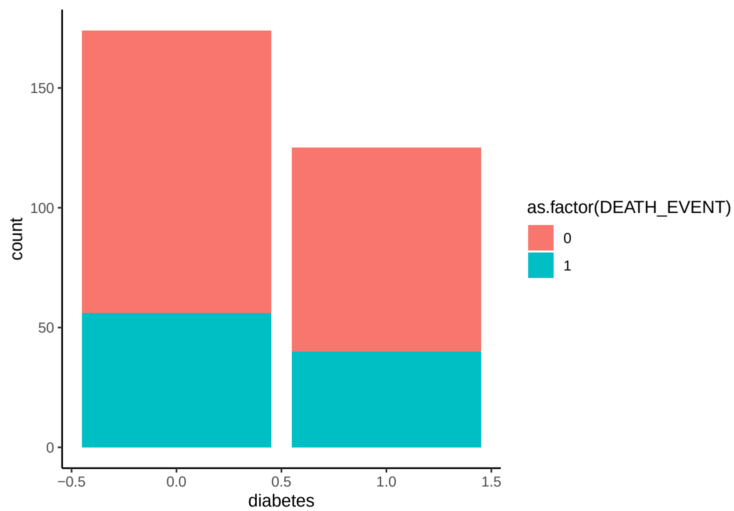
We can conclude from the figure above that Smokiers are more likely to have heart failure than non smokers.

2.7 diabets

Some of patients on the dataset have diabetes, and the distribution is shown below:



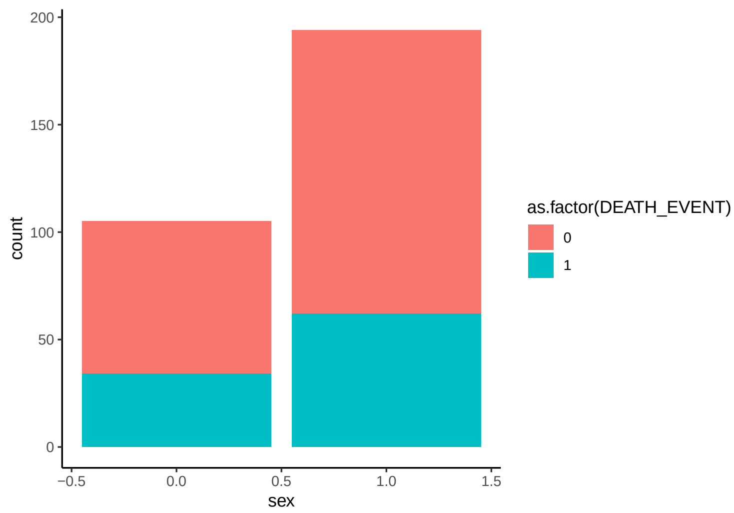
So, does if the patient has diabetes, will have more probability to have a heart failure? The relationship between death_event and diabetes is shown in the next figure:



The above figure shows that diabetes has no effect on heart failure.

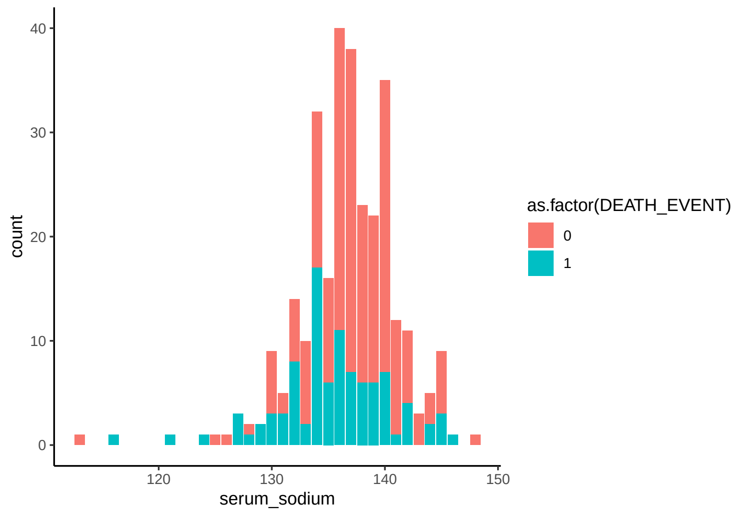
2.8 Sex

According to the figure below, it appears that females are more likely to have a heart failure.



2.9 Serum Sodium

According the next figure the serum sodium not affect the death event a lot.



3.0 Results

This section will have methodology that we work with and the resutls obtained from applying five different machine learning models.

The methods used for prediction of heart failure are:

1- Logistic Regression.
2- Random Forest. 3- Decision Tree. 4- Quadratic Discriminant Analysis. 5- Linear Discriminant Analysis

3.1 Project Methodology

We will follow the following steps to analyze the data and reach our goal of a masximum accuracy:-

- Firstly we need to download data and explore its observations and variables, then we'll make some visualizations to better understanding the data and this will help us later in choosing the appropriate model, and this is done in section 2.
- Then We'll start building models with the ideas gaind from the first step using machine learning models.
- Before start building models, we will split the data to trainig set and testing set, the training set will be used to train the models and evaluation will be done using the testing set.
- We will use 5 classification and machine learning models which are (Logistic regression, Random Forest, Disision trees model, Quadratic Discriminant Anallysis QDA and Linear Discriminant Analysis LDA)
- The different used models needs some tuning, so we will use cross validation technique to have the best tuning and get the best accuarcy.
- We will evaluate all models using the accuracy, sensitivity and specificty.

3.2 Model Evaluation.

We will choose the best machine learning model by the follwing criteria.

1- Maximum accuracy. (The proportion of cases that were correctly predicted in the test set) 2- Maximum sensitivity.(Also known as the true positive rate (TPR) or recall, is the proportion of actual positive outcomes correctly identified as such.) 3- Maximum specificity. (Also known as the true negative rate (TNR), is the proportion of actual negative outcomes that are correctly identified as such.)

and all these results can be got from the confusion matrix for each model. The confusion matrix tabulates each combination of prediction and actual value, it determines the results by combining the referenced and predicted outputs.

3.3 Splitting the dataset into training and testing sets.

Before we start building models, it is necessary to split our data into two parts, the first set is called training set and will be used to train models.The other set is called testing set and will be used to test the model.

The train set will be called train and has 80% of the data. The testing set will have 20% of data and called test.x set.

test.y will be a vectory that has the DEATH_EVENT variable, this variable will be the classification outcome.

3.4 Building models

We will start building different models and after each model built, we will check the accuracy, sensitivity and specificty values, so at the end we will have our final model.

3.4 Model 1: Logistic Regression Model (GLM)

The general form of a logistic regression model is:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \mathbf{x}_i^T\beta \tag{1}$$

where $\hat{\pi}_i$ is the estimated probability that observation i is positive, $\mathbf{x}_i$ is the i^{th} vector in the design matrix and β is the vector of coefficients.

Let's fit the model using the base general linear modeling function in R, glm.

The output of the glm model is shown below:

```
##
## Call:  glm(formula = DEATH_EVENT ~ ., family = "binomial", data = train)
##
## Coefficients:
##              (Intercept)                       age                  anaemia1
##                1.883e+00                 5.049e-02                 4.218e-01
## creatinine_phosphokinase                 diabetes1         ejection_fraction
##                3.395e-04                 4.335e-01                -7.381e-02
##      high_blood_pressure1                 platelets          serum_creatinine
##                4.859e-01                -1.483e-06                 6.900e-01
##             serum_sodium                      sex1                  smoking1
##               -3.091e-02                -2.607e-01                -1.062e-01
##
## Degrees of Freedom: 238 Total (i.e. Null);  227 Residual
## Null Deviance:      304.7
## Residual Deviance: 240.9    AIC: 264.9
```

Now, lets define the predictions for this glm_model using the predict function.

And for this model we will use a cutoff of 0.5 to make our decision.

```
y_hat_glm <- ifelse(preds_glm > 0.5, 1, 0)
```

and finally, the results are showin in the follwing confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 40 10
```

```
##          1  4  6
##
##                Accuracy : 0.7667
##                  95% CI : (0.6396, 0.8662)
##     No Information Rate : 0.7333
##     P-Value [Acc > NIR] : 0.3377
##
##                   Kappa : 0.3226
##
##  Mcnemar's Test P-Value : 0.1814
##
##             Sensitivity : 0.9091
##             Specificity : 0.3750
##          Pos Pred Value : 0.8000
##          Neg Pred Value : 0.6000
##              Prevalence : 0.7333
##          Detection Rate : 0.6667
##    Detection Prevalence : 0.8333
##       Balanced Accuracy : 0.6420
##
##        'Positive' Class : 0
##
```

We can see that the accuracy is 0.7666667, sensitivity is 0.9091 and specificity is 0.3750.

We will make a dataframe that handle all results for all models for the purpose on comparison.