Regression analysis for prediction heart disease

Outline:

- Introduction.
- The problem statement you addressed.
- How you addressed this problem statement
- Analysis.
- Implications.
- Limitations.
- Concluding Remarks

1. Introduction:

Heart disease is a leading cause of death for people of most races in the U.S. (African Americans, American Indians and Alaska Natives, and whites), according to the CDC. Heart Disease is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. It is the largest cause of mortality globally, resulting in the deaths of an estimated 17.9 million individuals each year. In the United States alone, heart disease claims roughly 647,000 lives each year — making it the leading cause of death. The buildup of plaques inside larger coronary arteries, molecular changes associated with aging, chronic inflammation, high blood pressure, and diabetes are all causes of and risk factors for heart disease.

About half of all Americans (47%) have at least 1 of 3 major risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicators include diabetes status, obesity (high BMI), not getting enough physical activity, or drinking too much alcohol. In addition, to the mentioned risk factors, Sex, smoking, age, family history, poor diet, physical inactivity and overweightness are some other key risk influences for heart disease.

The healthcare industry generates a lot of data regarding patient, diseases, and diagnoses, but it is not properly analyzed, so it does not have the same impact as it should on patient health. Identifying and preventing the factors that have the greatest impact on heart disease is very important in healthcare. The objective of this project is to analyze the importance of the 5 major predisposing factors, namely BMI- Body mass index, Hypertension, Triglycerides, Low-density lipoproteins, and diabetes. And perform regression analysis to detect "patterns" in the data that can predict heart disease using selected predisposing medical condition- underlying health conditions. In addition, predictive analysis also performed using a separate datatset from 299 patients with heart failures. Clinical laboratory test values collected from the patients, which can be used to perform machine learning analysis aimed at highlighting patterns and correlations to support the finding from the primary datatset with 396326 datapoints.

2. Problem Statement and Research Questions:

This project focuses on selected predictor variables and their relationship with the independent variables. As it can be mentioned above, the project focused on the major predictors for the heart failures such as BMI, Hypertension, Diabetes, history of COPD and Heart stroke.

- Investigate the relationship between BMI and the likelihood of developing heart disease. Does being overweight or obese increase the risk?
- Explore how elevated blood pressure contributes to cardiovascular risk. Does effective blood pressure management reduce the incidence of heart disease?
- Investigate the association between history of COPD and heart stoke, and heart disease.
- Investigate how diabetes impacts cardiovascular health. Having diabetes can be contributing factor for heart disease?

3. Approach/ Methodology:

We used classification technique machines algorithms to predict heart disease/ attack. Focuses on predicting cardiovascular disease with enhanced accuracy by employing three ML techniques such as logistic regression, Random Forest(RF) and K-Nearest Neighbors (KNN). ML based prediction performed in predicting the occurrence of heart disease and the survival status of patients. The dataset utilized in this study are the heart disease data with 40 different attributes and the Heart Failure Dataset, which comprises 13 attributes. The dataset was retrieved from Kaggle, and data preprocessing was performed for selecting the features and other data cleaning processes. Then, the dataset splitted into two parts: a training dataset and a test dataset; around 80% of the entire data is utilized for training, whereas the rest is used for testing. The test dataset is utilized to assess classifiers, while the training is to develop a model that predicts heart disease. To assess the models' performance, a confusion matrix was used to visualize metrics like accuracy, sensitivity, and specificity rate. R statistical programming package was utilized for the data analysis. The data analysis methodology steps can be seen below:

a. Data Preparation:

- The first step in data preprocessing was read in the dataset (e.g., heart_disease_data, heart_failure_data) using read.csv. Then the data cleaning step, cleaned column names to make them more readable , selected/filtered the desired variables, drop outliers, dropped NA values, etc.

b. **Exploratory Data Analysis (EDA):**
   - In this step, I explored the dataset using functions like summary(). Visualized distributions, correlations, and potential outliers. Identified missing values and handle them appropriately.

c. **Model Selection:**
   - In this step, I selected an appropriate model for the regression analysis. I applied the ML classifiers to the preprocessed dataset.

d. **Model Evaluation:**
   - Evaluate the model's performance using metrics include R-Squared, Root Mean Squared Error (RMSE) and Residual Standard Error (RSE). A confusion matrix was also utilized to visualize metrics like accuracy, sensitivity, and specificity rate.

e. **Interpretation:**
   - In this step, Interpreted the model coefficients. Identified the most important variables contributing to heart disease prediction. Then, I visualized the correlation matrix output to depict which attribute has high importance in impacting the model's prediction.

4. Analysis and Implications:

The first analysis performed was the correlation matrix (COR), this is done to classify the optimal values or features importance in predicting the target variable. Correlation is a statistical feature that describes the strength and route of a linear relationship between two quantitative variables. A correlation matrix with heatmap is shown in Fig.1 and Fig 2. Most attributes have a moderate correlation with the "HEART ATTACK" and "DEATH_EVENT" variables. In addition to COR, variable selection analysis also completed to determine a set of variables that will provide the best fit for the model so that accurate predictions can be made. The next step, applying the ML classifiers to the pre-processed dataset, aiming for the highest achievable accuracy through feature reduction. And final step, the suggested classifiers are assessed for accuracy, sensitivity, and specificity.

The ML prediction analysis outcome shows, the accuracy of logistic regression (LR) model to predict Heart attack using underlying medical condition like stroke COPD and diabetes as predictor is 94.1%. However, the accuracy of LR dropped down to 85.7% when using other clinical laboratory test indicators. The accuracy Random Forest (RF) to predict Heart disease Fatality is 82.6%. With a Specificity of 77.2% and recall/sensitivity of 85.11%. This result indicates RF model was better at predicting individuals with underlying disease who survived from heart disease. The interesting finding on this analysis is the performance of logistic regression vs random forest. LG was more accurate and robust compared to Random Forest in predicting heart disease fatality. The accuracy of the last ML classifier -KNN, however, was low- 59.4%. See the outcome from both dataset using different predictor, Logistic Regression model performed better in predicting heart disease. The data analysis for the second dataset can be found on the last part of this article.

When we see the importance of the variable in the model, previous history of STROKE and COPD has the standardized coefficient with the largest absolute value. This measure suggests that Previous history of STROKE and COPD are the most important independent variable in the regression model followed by Diabetes and had the greatest effect on the dependent variable i.e Having Heart Attack. This result also supported on the odds ratio table and variable selection output tables- Stepwise summary (AIC and BIC outcome) and Parameter Estimates (Beta values). When we see the second dataset output however, diabetes has one of the lowest standardized coefficients. This is the other interesting finding in this analysis.

Fig. 1: Correlation Matrix heat map for the first dataset

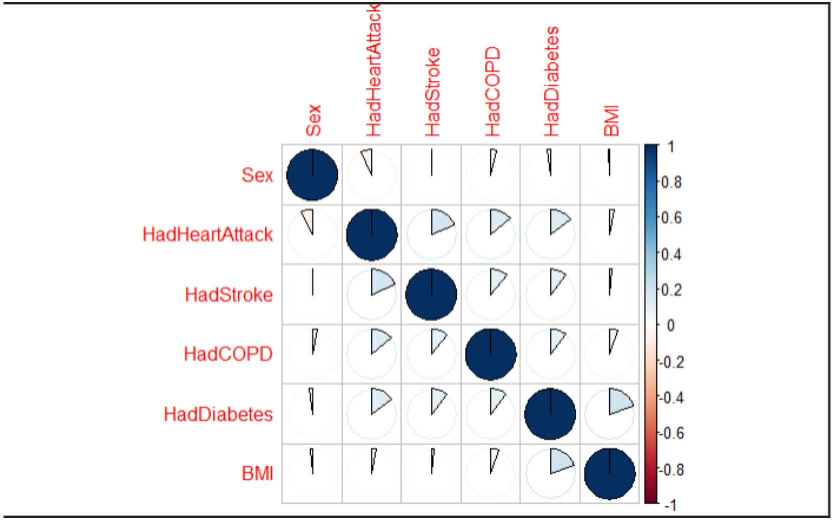


Fig. 2: Correlation Matrix heat map for the second dataset

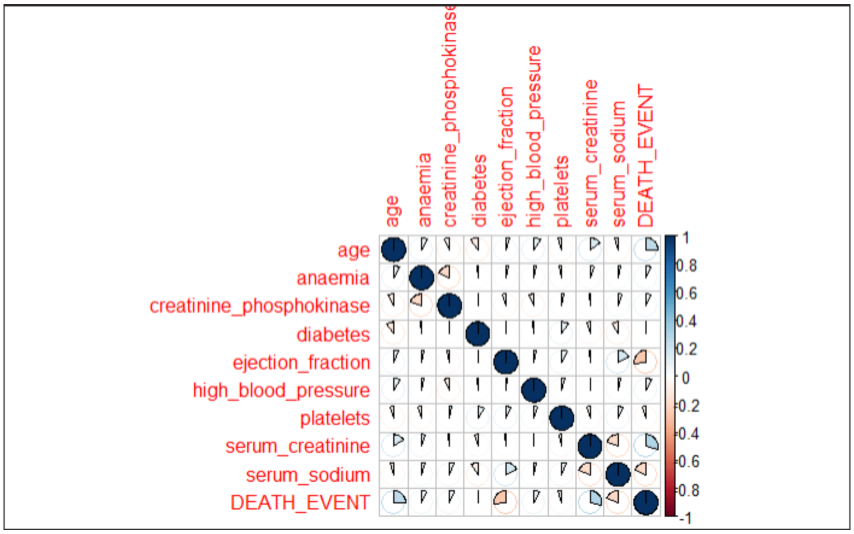
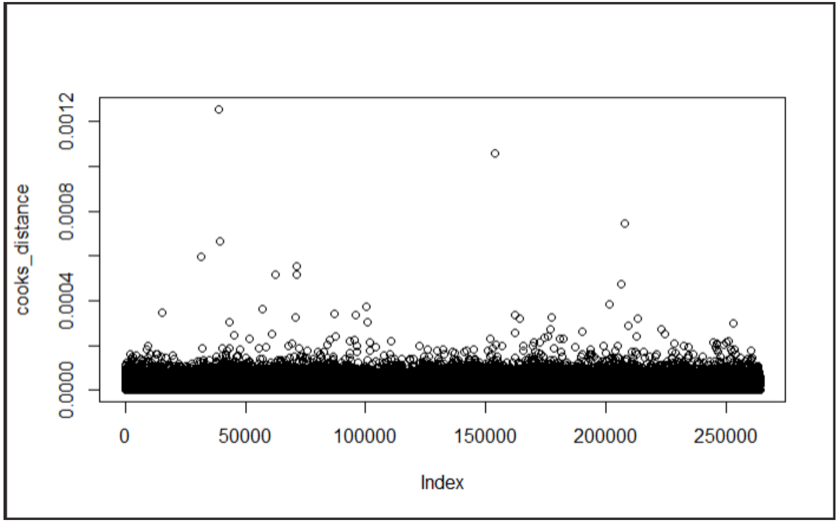


Fig 3: Output analysis of outliers



5. Limitations:

There are a few limitations encountered while analyzing the dataset with a potential impact on the efficiency and effectiveness of the models being developed. To mention a couple of them:

- Computational Resources: Analyzing large datasets like the heart_disease_data requires significant computational power. Training complex models on such bigger size data was time-consuming and resource intensive.
- Data Quality: The original dataset contained noisy, missing, or inconsistent data. Ensuring data quality is crucial for accurate model training. Cleaning and preprocessing large datasets were time-consuming and resource intensive.

6. Conclusion

In this analysis, machine learning classifiers are used to predict the presence of heart attack/ heart disease related death events. We used a couple of datasets from Kaggle. The original data is then preprocessed and cleaned. After that, machine learing models are applied for predicting. The analysis result showed us STROKE and COPD are the two most important independent variables in predicting heart attack. The resultant outcomes also reveal logistic regression achieve 94.1% accuracy in predicting the heart attack. This accuracy outcome can be further optimized using deep learning methods in future projects. However, that is not the scope of this project.

7. Dataset Links:

- heart_disease_data
- diabetes_prediction_dataset
- Hypertension-risk-data

8. Required Packages

Here is the some of the of R packages that will be employed to operate on the large dataset selected for this project:

- ✓ dplyr package
- ✓ magrittr package
- ✓ ggplot2 package
- ✓ coefplot package
- ✓ readxl package
- ✓ Metrics package

9. Reference:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9644238/

- CDC Heart Disease

finalProject_4

Abraham Abate

2024-05-28

Final Project section 3

```
# Load library
library(dplyr)
library(magrittr)
library(ggplot2)
library(tidyverse)
library(Hmisc)
library(olsrr)
library(corrplot)
# Data importing and cleaning steps
heart_fail_df <- read.csv(file='heart_failure_data.csv',
check.names=F,stringsAsFactors = F)

# check the number of rows and columns
nrow(heart_fail_df)
```

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

```
ncol(heart_fail_df)
```

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

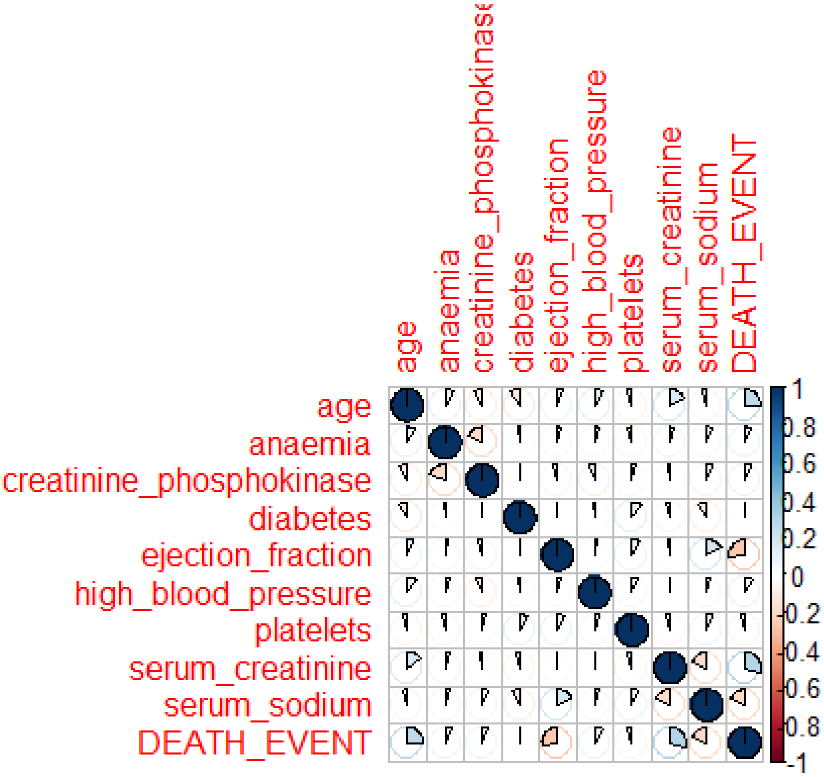
```
# check the classes of each of the columns
str(heart_fail_df)
```

```
## '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 ...
```

```
# count total missing values
colSums(is.na(heart_fail_df))
```

```
##                      age                  anaemia creatinine_phosphokinase
##                        0                        0                        0
##                 diabetes        ejection_fraction      high_blood_pressure
##                        0                        0                        0
##                platelets         serum_creatinine             serum_sodium
##                        0                        0                        0
##                      sex                  smoking                     time
##                        0                        0                        0
##              DEATH_EVENT
##                        0
```

```
# computing correlation matrix
cor_data <- cor(heart_fail_df %>% select(1:9,13))
corrplot(cor_data, method="pie")
```



Regression Analysis

```
# Load Library
library('foreign')
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.2.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
library(ggplot2)
library(dplyr)


# Split the data
data_split <- sample.split(heart_fail_df, SplitRatio = 0.8)

# Training data
train_df <- subset(heart_fail_df, data_split=="TRUE")

#Testing data
test_df <- subset(heart_fail_df, data_split=="FALSE")

#logistic regression model
# Note: Target Variable is Death_Event- the survived patients (death event =
0)
# and the dead patients (death event = 1)
logReg_model <- glm(DEATH_EVENT ~ anaemia + ., data=train_df, family =
binomial)
summary(logReg_model)
```

```
##
## Call:
## glm(formula = DEATH_EVENT ~ anaemia + ., family = binomial, data =
train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2826  -0.5964  -0.2331   0.4819   2.7181
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               6.804e+00  6.147e+00   1.107 0.268373
## anaemia                  -3.832e-01  4.146e-01  -0.924 0.355318
## age                       4.027e-02  1.798e-02   2.239 0.025143 *
## creatinine_phosphokinase  2.510e-04  1.986e-04   1.264 0.206356
## diabetes                  2.840e-01  3.918e-01   0.725 0.468552
## ejection_fraction        -7.047e-02  1.836e-02  -3.838 0.000124 ***
## high_blood_pressure       4.893e-02  4.195e-01   0.117 0.907156
## platelets                 5.131e-07  1.887e-06   0.272 0.785633
## serum_creatinine          4.278e-01  2.272e-01   1.883 0.059671 .
## serum_sodium             -4.338e-02  4.305e-02  -1.008 0.313672
## sex                      -9.288e-02  4.697e-01  -0.198 0.843263
## smoking                  -3.644e-01  4.809e-01  -0.758 0.448645
## time                     -2.112e-02  3.431e-03  -6.154 7.57e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 282.67  on 229  degrees of freedom
## Residual deviance: 174.33  on 217  degrees of freedom
## AIC: 200.33
##
## Number of Fisher Scoring iterations: 6
```

```
# Variable Selection
ols_step_forward_p(logReg_model)
```

```
##
##
##                                        Stepwise Summary
## --------------------------------------------------------------------------
------------------
## Step    Variable                      AIC        SBC          SBIC
R2       Adj. R2
## --------------------------------------------------------------------------
------------------
##  0      Base Model                  299.639    306.515    -2095201.205
0.00000    0.00000
##  1      time                        228.286    238.601    -3965087.142
0.27307    0.26988
##  2      ejection_fraction           209.667    223.419    -4741758.399
0.33540    0.32955
##  3      serum_creatinine            203.159    220.350    -5103231.254
0.35954    0.35104
##  4      age                         200.792    221.421    -5297853.634
0.37158    0.36041
##  5      creatinine_phosphokinase    199.800    223.866    -5434499.770
0.37971    0.36586
## --------------------------------------------------------------------------
------------------
##
## Final Model Output
## ------------------
##
##                           Model Summary
## ---------------------------------------------------------------
## R                       0.616       RMSE                 0.362
## R-Squared               0.380       MSE                  0.135
## Adj. R-Squared          0.366       Coef. Var          120.656
## Pred R-Squared          0.343       AIC                199.800
## MAE                     0.295       SBC                223.866
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##  AIC: Akaike Information Criteria
##  SBC: Schwarz Bayesian Criteria
##
##                                 ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares         DF    Mean Square      F          Sig.
## --------------------------------------------------------------------
## Regression     18.490          5          3.698    27.424    0.0000
## Residual       30.206        224          0.135
## Total          48.696        229
## --------------------------------------------------------------------
##
##                                           Parameter Estimates
## --------------------------------------------------------------------------
---------------------------
##                    model      Beta    Std. Error    Std. Beta      t
Sig      lower     upper
## --------------------------------------------------------------------------
---------------------------
##              (Intercept)     0.609         0.163                  3.738
0.000     0.288     0.931
##                     time    -0.003         0.000       -0.465    -8.593
0.000    -0.003    -0.002
##        ejection_fraction    -0.010         0.002       -0.246    -4.641
0.000    -0.014    -0.006
##         serum_creatinine     0.080         0.032        0.134     2.485
0.014     0.017     0.143
##                      age     0.005         0.002        0.122     2.228
0.027     0.001     0.009
## creatinine_phosphokinase     0.000         0.000        0.091     1.713
0.088     0.000     0.000
## --------------------------------------------------------------------------
---------------------------
```

```
# Run the test data through the model
pred_val <- predict(logReg_model, type="response")

# Validate the model using- confusion matrix
confMatrix <- table(Actual_Value = train_df$DEATH_EVENT,
                    Predicted_Value = pred_val > 0.5)
confMatrix
```

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##            0   147   13
##            1    22   48
```

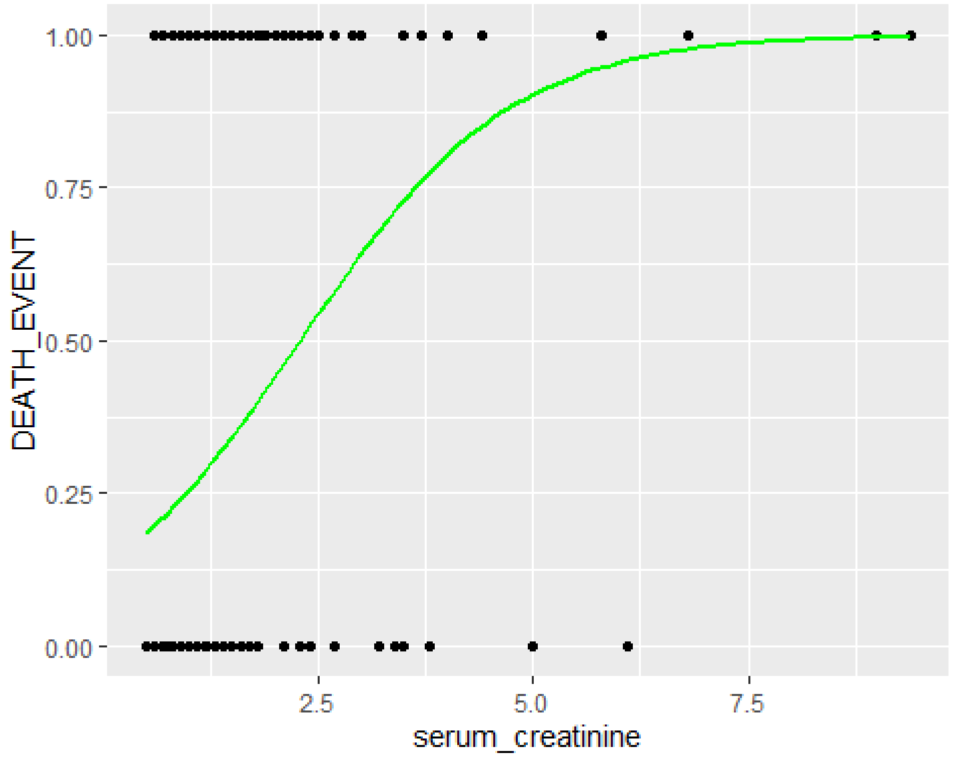
```
# compute the accuracy = (TP + TN) / (TP + TN + FP + FN)
model_accuracy <- (confMatrix[[1,1]] + confMatrix[[2,2]]) / sum(confMatrix)
model_accuracy
```

```
## [1] 0.8478261

# computing correlation matrix
cor_data <- cor(heart_fail_df)

# Plot Predicted data and original data points
ggplot(data = heart_fail_df,
       mapping = aes(x=serum_creatinine, y=DEATH_EVENT)) +
  geom_point() +
  geom_smooth(method="glm", color="green", se=FALSE,
                method.args = list(family=binomial))

## `geom_smooth()` using formula = 'y ~ x'
```



```
# multivariate linear regression model
library(randomForest)

## Warning: package 'randomForest' was built under R version 4.2.3

rf_model <- randomForest(as.factor(DEATH_EVENT) ~ anaemia + .,
                         data=train_df, ntree=500,
                         proximity=T, importance=T)
# Make a prediction on the fitted model with the test data
predict_cat <- predict(rf_model, newdata = test_df, type = "response")

# create a df and convert the output to factors
pred_df <- data.frame(Predicted = predict_cat, Actual = test_df$DEATH_EVENT)
```

```
pred_df$Predicted <- as.factor(pred_df$Predicted)
pred_df$Actual = as.factor(pred_df$Actual)

# Confusion matrix
confusionMatrix(pred_df$Actual, pred_df$Predicted)

## Confusion Matrix and Statistics
## 
##           Reference
## Prediction  0  1
##          0 41  2
##          1  9 17
## 
##                Accuracy : 0.8406
##                  95% CI : (0.7326, 0.9176)
##     No Information Rate : 0.7246
##     P-Value [Acc > NIR] : 0.01774
## 
##                   Kappa : 0.6415
## 
##  Mcnemar's Test P-Value : 0.07044
## 
##             Sensitivity : 0.8200
##             Specificity : 0.8947
##          Pos Pred Value : 0.9535
##          Neg Pred Value : 0.6538
##              Prevalence : 0.7246
##          Detection Rate : 0.5942
##    Detection Prevalence : 0.6232
##       Balanced Accuracy : 0.8574
## 
##        'Positive' Class : 0
## 

# Training KNN Classifier and Predicting
library(class)
prediction <- knn(
                 train = train_df,
                 test = test_df,
                 cl = train_df$DEATH_EVENT,
                 k=10
                 )

# Model Evaluation
actual <- test_df$DEATH_EVENT

cm <- table(actual,prediction)
cm
```

```
##       prediction
## actual  0  1
##      0 43  0
##      1 24  2
```

```
# Calculate accuracy
accuracy <- sum(diag(cm))/length(actual)
sprintf("Accuracy: %.2f%%", accuracy*100)
```

```
## [1] "Accuracy: 65.22%"
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.