Model Selection

Heart Attack Prediction

Done by:

2nd Year Bioinformatics Students…

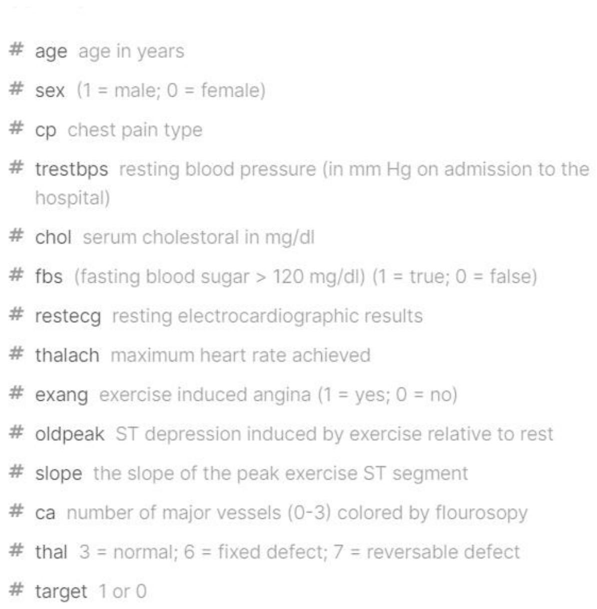
- Sahar Saber Ibrahim.
- Salma Mohamed Kamel.
- Salma Mohamed Saeid.
- Mohamed Mossad.
- Sara Ahmed Maher.
- Yomna Madgy.

Introduction:

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry.

Importing & Joining Data:

The dataset is taken from the UCI repository. It consists of 303 observation and 14 variables, which are described below.



```
# age  age in years
# sex  (1 = male; 0 = female)
# cp  chest pain type
# trestbps  resting blood pressure (in mm Hg on admission to the
  hospital)
# chol  serum cholestoral in mg/dl
# fbs  (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
# restecg  resting electrocardiographic results
# thalach  maximum heart rate achieved
# exang  exercise induced angina (1 = yes; 0 = no)
# oldpeak  ST depression induced by exercise relative to rest
# slope  the slope of the peak exercise ST segment
# ca  number of major vessels (0-3) colored by flourosopy
# thal  3 = normal; 6 = fixed defect; 7 = reversable defect
# target  1 or 0
```

It was separated into 2 data frames.

uci_data: training set

```
> dim(uci_data)
[1] 243  14
```

uci_test: test set

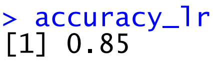
```
> dim(uci_test)
[1] 60 14
```

The Approach:

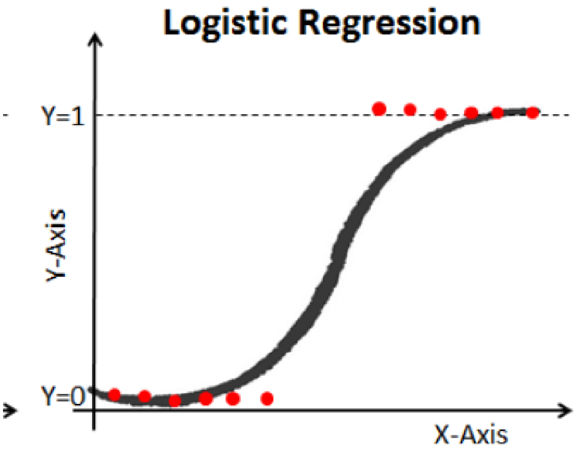
The code is implemented in R with different classification models like Logistic regression, random forest, kNN to determine which is the best model that fits the data.

Logistic regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. A binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1".

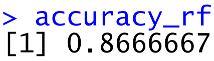


```
> accuracy_lr
[1] 0.85
```

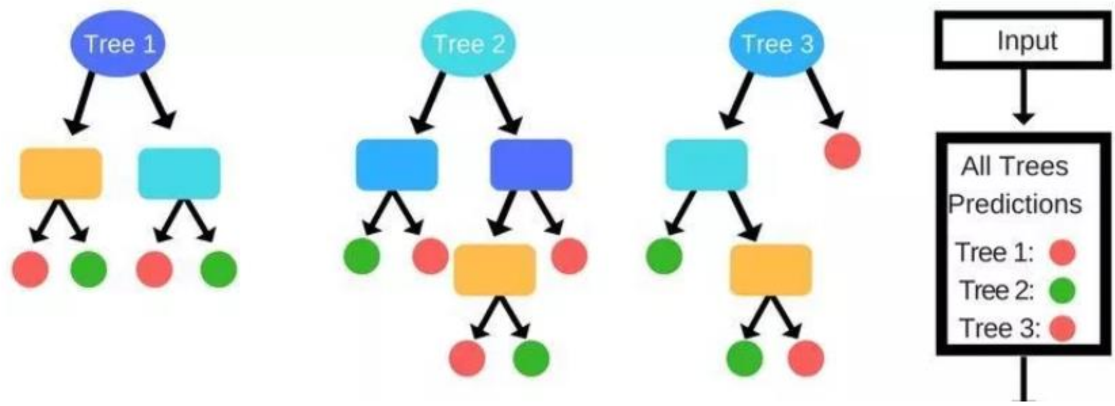


Random Forest:

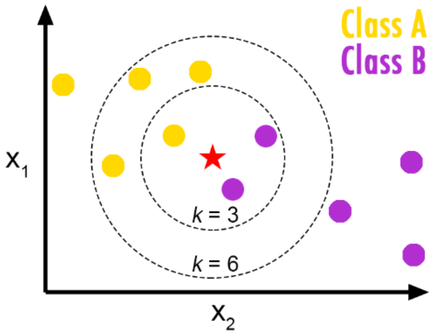
A random forest consists of multiple random decision trees. It is a general solution that can be applied to classification and regression problems and doesn't require a certain form of data, it was worth trying.



```
> accuracy_rf
[1] 0.8666667
```



K-Nearest Neighbors, kNN:



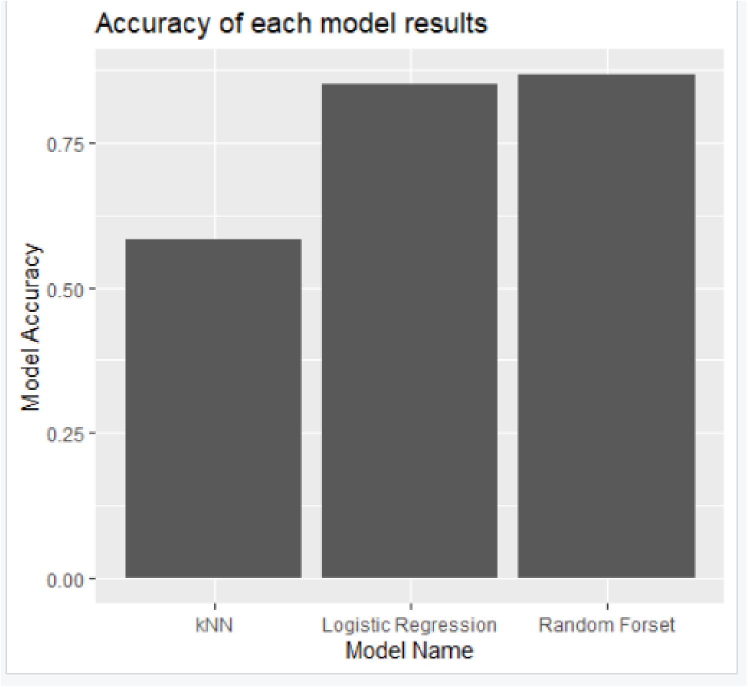
KNN algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

The suitable value for K is 10 which is determined through measuring accuracies corresponding to several values of k.

```
> accuracy_knsn
[1] 0.5833333
```

Conclusion:

We see that the highest accuracy for the test set is achieved by Random Forest is equal to 86.67%.


Accuracy of each model results
Model Accuracy
kNN
Logistic Regression
Random Forset
Model Name