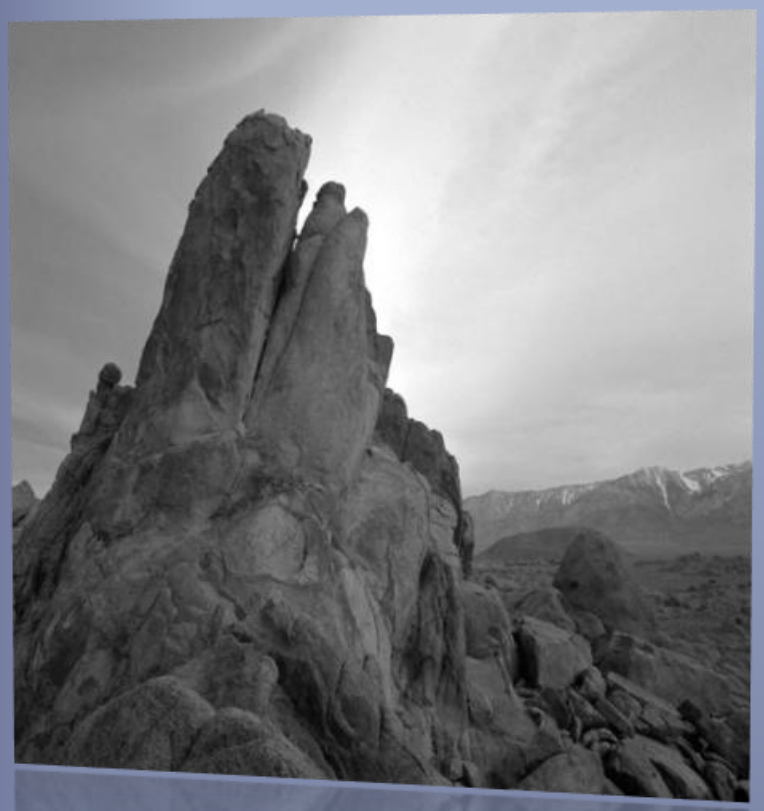


Congestive Heart Failure

CS-626 Data Mining And Warehouse

Members:
Mahnoor Ali (P19101031)
Nihan (P19101050)
Syeda Abeeha Hassan (P19101069)
Alija Faiz (P1710003)
MCS Final 2020
Department Of Computer Science
University Of Karachi.

Table of Contents

1. Abstract: 2

2. Introduction: 2

3. Methodologies: 4

I. 1. Dataset Description: 4

II. 2. Decision Trees: 8

III. 3. Random Forest: 10

IV. 4. K- Nearest Neighbor: 12

V. 5. Logistic Regression: 14

VI. 6. Fuzzy C means Clustering: 17

4. Conclusion: 20

5.References: 21

1. Abstract:

The best prediction of heart disorder can save you lifestyles threats, and incorrect prediction can show to be deadly on the same time, in this paper extraordinary gadget learning algorithms and deep studying are carried out to examine the outcomes and evaluation of the coronary heart failure prediction dataset. The dataset includes 13 major attributes used for appearing the evaluation. Various promising effects are completed and are confirmed the use of accuracy and confusion matrix.

Data mining transforms big quantities of uncooked facts generated with the aid of the fitness enterprise into beneficial facts which could assist in making knowledgeable choices, various research proved that huge features play a key role in improving overall performance of machine getting to know models. This study analyzes the heart failure survivors from the dataset of 299 sufferers admitted in clinic.

This article tries to use three of those AI-based methods namely Decision Tree, Random forest, K-Nearest Neighbor and Fuzzy c mean Clustering for forecasting cardiovascular or heart disorder.

All of these techniques can be evaluated based totally on exclusive particular & parameters with optimizations for higher accuracy. The accuracy of each method will then be compared relying on accuracy based on various parameters. The great & accurate technique is then implemented for predicting whether or not or no longer may a man or a woman have coronary heart sickness. This technique can be used by medical practitioners for early prediction of the disease in order that well timed care may be taken through the affected person.

2. Introduction:

The heart is the most crucial & critical organ of the human body. Life is completely dependent on the efficient working of our coronary heart. Its miles one of the major reasons of mortality in trendy world. Coronary heart disease remains one of the most serious fitness issues of our day. Its miles said to be the number one cause in demise globally. Frequently it's difficult for clinical specialists to count on a heart disorder on time. Coronary heart failure is a critical problem which has a massive effect on people’s life. With the increased tempo of existence, elevated element sizes and inactivity, most people usually forget about their fitness. Furthermore, because of the environmental deterioration, those elements can result in the issue of heart failure which could emerge as increasingly more common within the future. If humans did not be aware of the problem of heart failure, it might ultimately purpose the dying.

Medical diagnosis performs crucial function and yet complex project that desires to be executed efficiently and accurately. To reduce value forreaching scientific tests the right computer

primarily based statistics and choice support should be aided. Data mining is using software program strategies for locating styles and consistency in units of statistics.

Data mining inside the ultimate decades, there is a massive opportunity to permit computer systems to directly construct and classify the one of a kind attributes or training. Learning of the risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having Heart disease. Statistical analysis has diagnosed danger factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hyper tension, family history of coronary heart disease, obesity and lack of physical workout, fasting blood sugar and so on.

In literature, while feature engineering and feature selection are applied, the outcomes improve, both for class in addition to predictions. Machine learning and deep learning techniques for detecting the heart disease and additionally carried out hyper parameters tuning for growing the results accuracy.

Data mining is the process of coming across patterns in massive information units related to methods on the intersection of machine learning, records, and database structures. It is a vital system where useful techniques are applied to extract statistics patterns. The data mining may be carried out using category, clustering, prediction, association and time series analysis.

Data mining is the exploration of huge datasets to extract hidden and previously unknown patterns, relationships and knowledge that are tough to hit upon with traditional statistical methods. For that reason Data mining refers to mining or extracting understanding from huge quantities of statistics. Data mining applications can be used for better health coverage-making and prevention of hospital errors, early detection, prevention of diseases and preventable medical institution deaths. Heart disease prediction device can assist medical professionals in predicting coronary heart disorder based totally at the medical statistics of sufferers. Thereforeby means of implementing a coronary heart disease prediction device using information. Data mining strategies and performing some sort of information mining on numerous heart disorder attributes, it can able to predict extra probabilistically that the patients will be identified with coronary heart disease. This paper gives a new model that enhances the Decision Tree accuracy in identifying heart disease patients. It uses the different algorithm of Decision Trees. Hence Data mining refers to mining or extracting expertise from big quantities of information. Data mining programs will be used for higher fitness coverage-making and prevention of health center errors, early detection, prevention of diseases and preventable hospital deaths. Heart disorder prediction machine can help clinical professionals in predicting coronary heart disorder based totally on the scientific records of sufferers. for this reason by implementing a coronary heart disorder prediction machine the use of Data Mining strategies and doing some sort of Data mining on various heart disease attributes, it may able to are expecting more probabilistically that the

sufferers will be identified with heart sickness. This article offers a new model that enhances the Decision Tree accuracy in identifying coronary heart sickness patients.

3. Methodologies:

1. Dataset Description:

Study Procedures, Assessments: Primary and Secondary Variables:

Study involving 299 heart failure patient and 13 variables. Variables include Age, Anemia, Creatinine_Phosphokinase, Diabetes, Ejection_ Fraction, High Blood Pressure, Platelets, Serum_Creatinine, Serum_Sodium, Sex, Smoking, Time, and Death Event.

Anemia: Decrease of red blood cells or hemoglobin (Boolean)
Creatinine_Phosphokinase: Level of the CPK enzyme in the blood (mcg/L)
Diabetes: If the patient has diabetes (Boolean)
Ejection_Fraction: Percentage of blood leaving the heart at each contraction (percentage)
High Blood Pressure: If the patient has hypertension (Boolean)
Platelets: Platelets in the blood (kilo platelets/ mL)
Serum_Creatinine: Level of serum creatinine in the blood (mg/dL)
Serum_Sodium: Level of serum sodium in the blood (mEq/L)
Sex: Woman or man (0 or 1)
Smoking: If the patient smokes or not (Boolean)
Time: Follow-up period (days)
Death Event: If the patient deceased during the follow-up period (Boolean).

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
1	75	0	582	0	20	1	265000	1.90	130	1	0	4	1
2	55	0	7861	0	38	0	263358	1.10	136	1	0	6	1
3	65	0	146	0	20	0	162000	1.30	129	1	1	7	1
4	50	1	111	0	20	0	210000	1.90	137	1	0	7	1
5	65	1	160	1	20	0	327000	2.70	116	0	0	8	1
6	90	1	47	0	40	1	204000	2.10	132	1	1	8	1
7	75	1	246	0	15	0	127000	1.20	137	1	0	10	1
8	60	1	315	1	60	0	454000	1.10	131	1	1	10	1

Figure of Heart failure prediction data

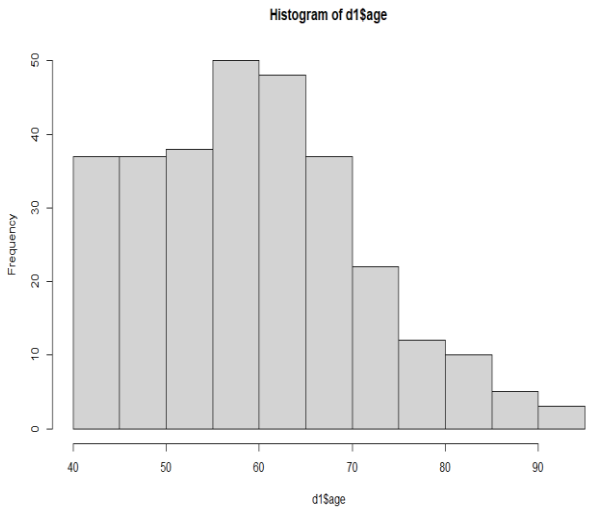
Primary Variables:

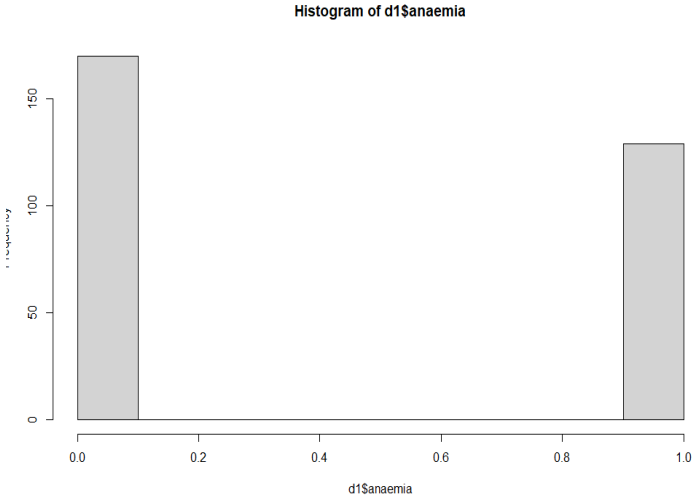
Age, Ejection_Fraction, Platelets, Serum_Creatinine, Serum_Sodium, Sex, Smoking, Time, Death Event.

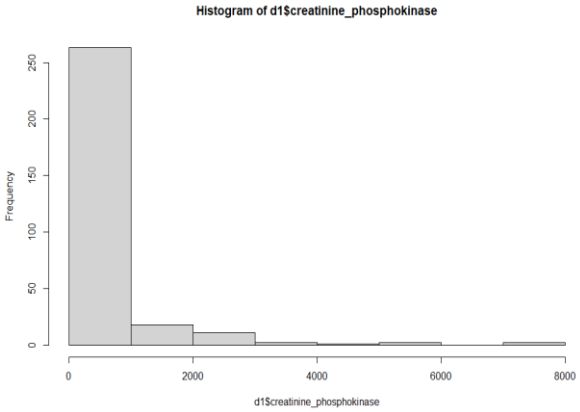
Secondary Variables:

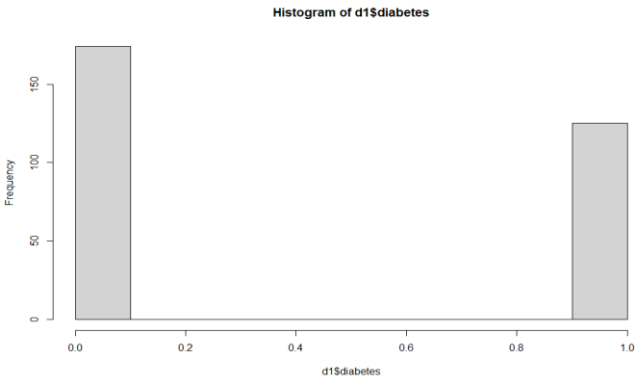
Anemia, Creatinine_Phosphokinase, Diabetes, High Blood Pressure.

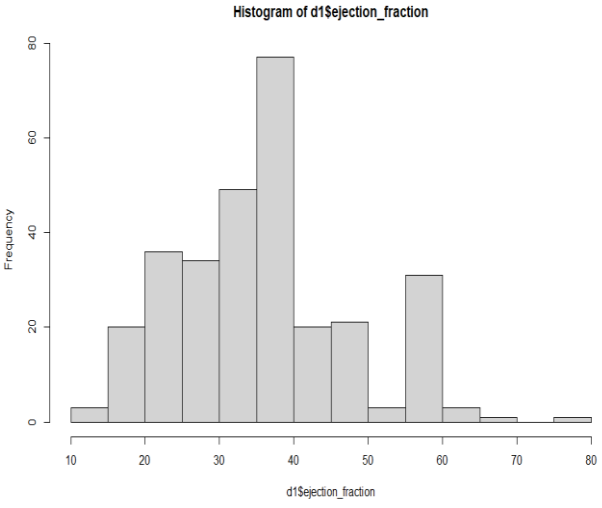
Histogram of all attributes of Heart failure prediction data

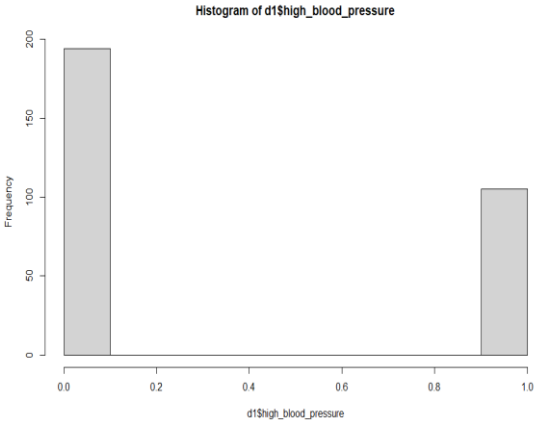


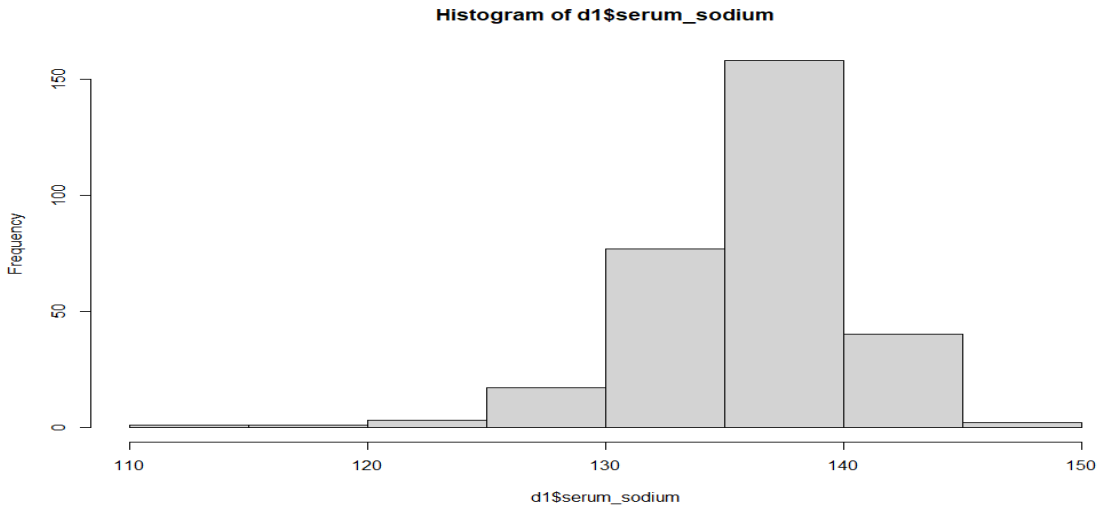


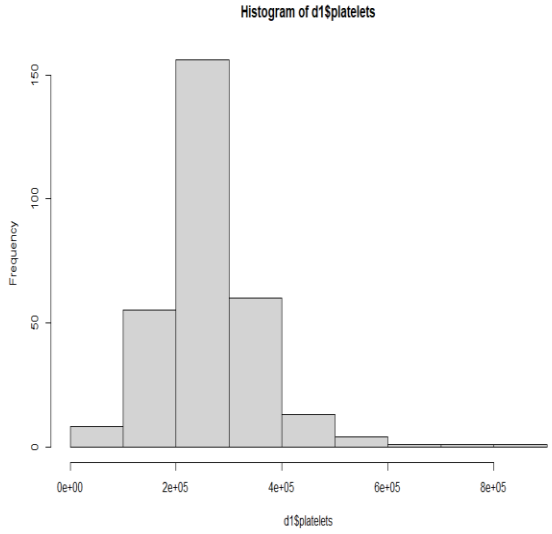


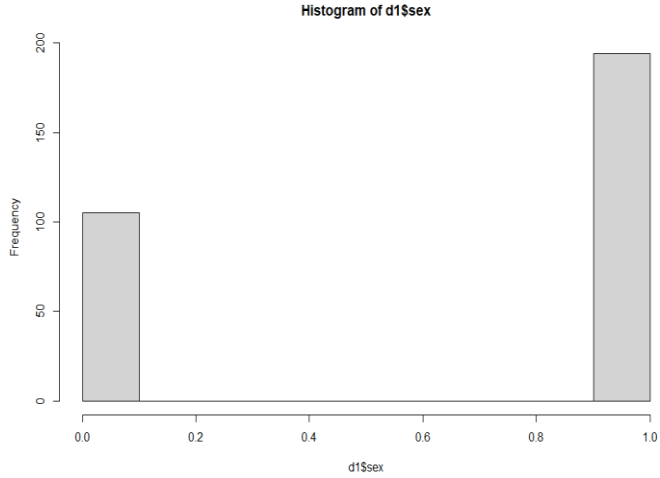


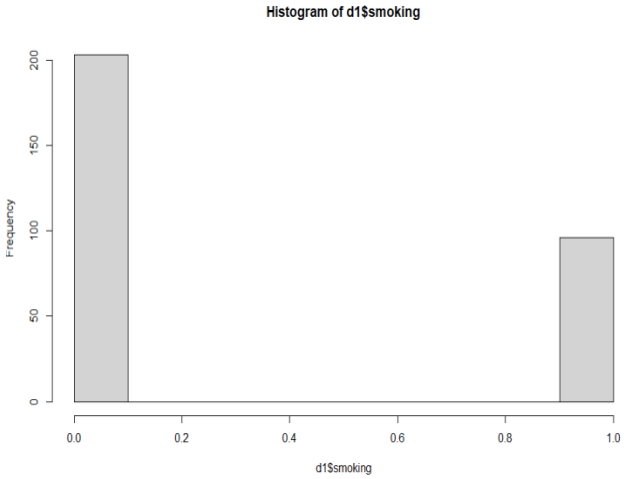


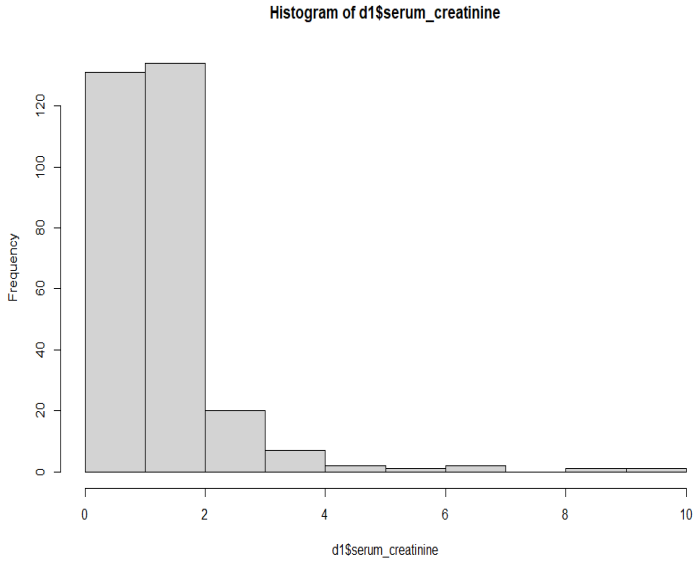


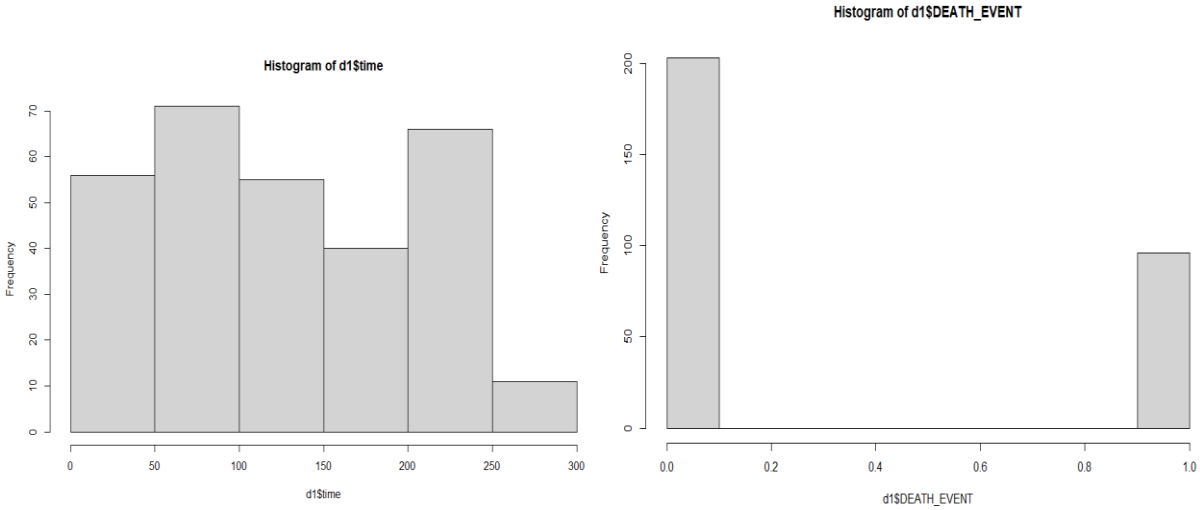

Histogram of d1$serum_sodium
Frequency
0
50
100
150
110
120
130
140
150
d1$serum_sodium

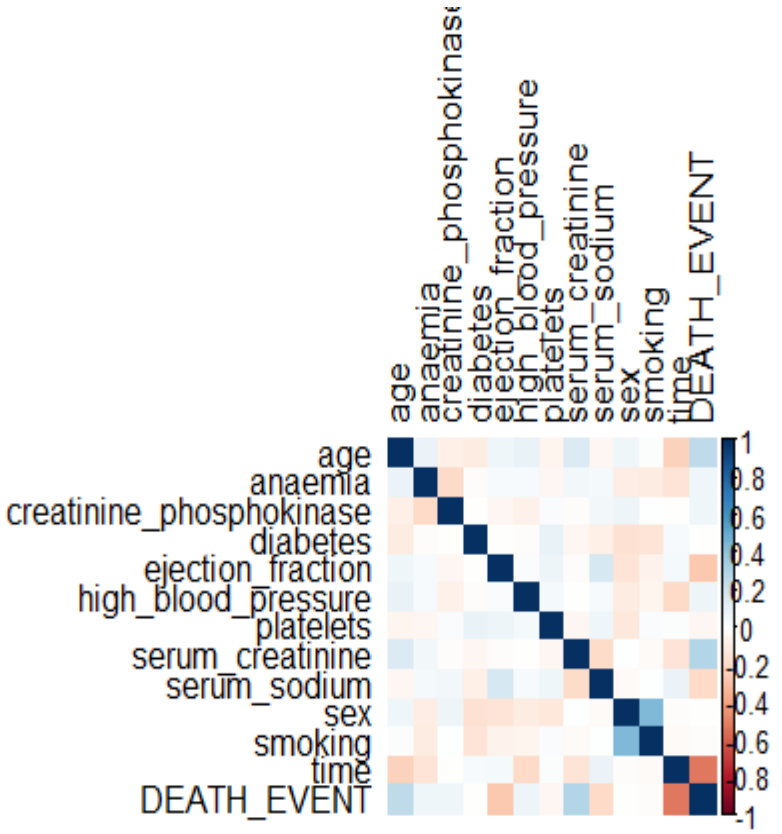

Histogram of d1$platelets
Frequency
0
50
100
150
0e+00
2e+05
4e+05
6e+05
8e+05
d1$platelets


Histogram of d1$sex
Frequency
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
d1$sex


Histogram of d1$smoking
Frequency
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
d1$smoking


Histogram of d1$serum_creatinine
Frequency
0
20
40
60
80
100
120
0
2
4
6
8
10
d1$serum_creatinine





Heat Map of Heart failure prediction data

2. Decision Trees:

Decision tree is the most powerful and popular device for type and prediction. A decision tree is a flowchart like tree shape, wherein each inner node denotes a check on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

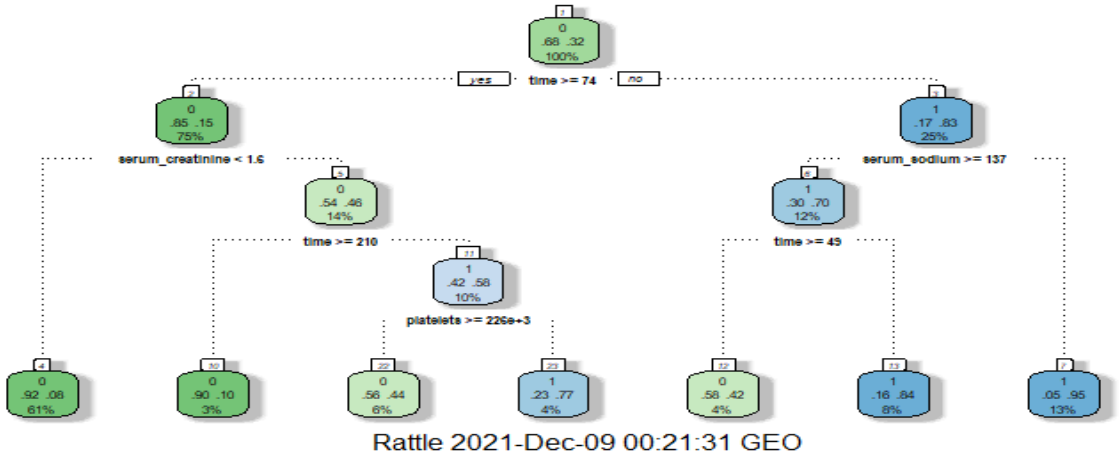
Decision tree algorithms use the training data to section the predictor space into non-overlapping regions, the nodes of the tree. Every node is described by using a set of rules that are then used to expect new responses. The predicted value for each node is the most common response within the node (classification), or mean response in the node (regression).

Advantages of decision trees:

1. Decision trees are able to generate understandable rules.
2. Decision trees are able to handle both continuous and categorical variables.
3. New features can be easily added.
4. Decision trees can be used to build larger classifiers by using ensemble methods.

Disadvantages of decision trees:

1. Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
2. Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
3. Need to be careful with parameter tuning.
4. Can create biased learned trees if some classes dominate.



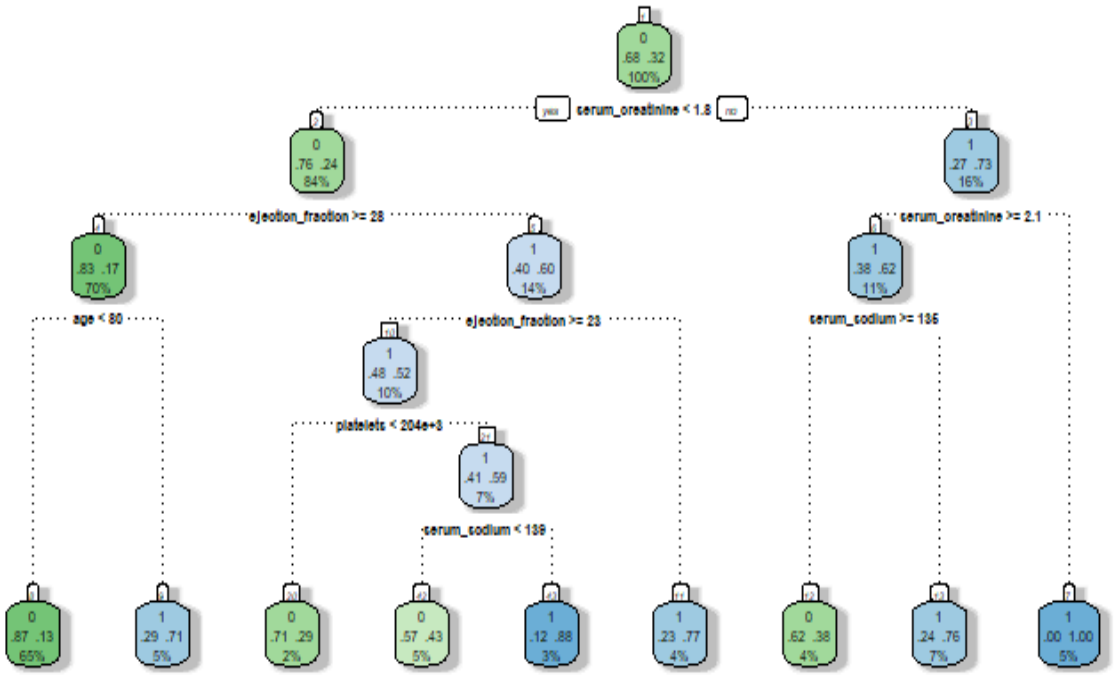


Figure of Decision Trees applying on Heart failure prediction data

Output of Decision Trees:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 37  1
         1  5 13

               Accuracy : 0.8929
                 95% CI : (0.7812, 0.9597)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : 0.006653

                  Kappa : 0.7391

 Mcnemar's Test P-Value : 0.220671

            Sensitivity : 0.8810
            Specificity : 0.9286
         Pos Pred Value : 0.9737
         Neg Pred Value : 0.7222
             Prevalence : 0.7500
         Detection Rate : 0.6607
   Detection Prevalence : 0.6786
      Balanced Accuracy : 0.9048

       'Positive' Class : 0
```

3. Random Forest:

A random forest is a machine learning technique that's used to solve regression and classification problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating.

Features of a Random Forest Algorithm:

1. it's more accurate than the decision tree algorithm.
2. It provides an effective way of handling missing data.
3. It can produce a reasonable prediction without hyper-parameter tuning.
4. It solves the issue of over fitting in decision trees.
5. In every random forest tree, a subset of features is selected randomly at the node's splitting point.

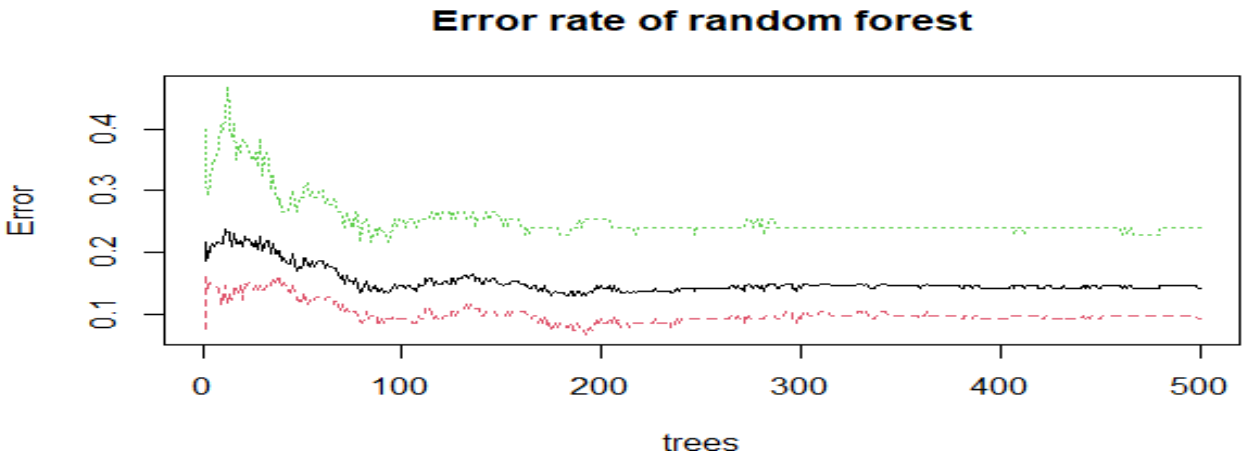
Random forest randomly selects observations, builds a decision tree and the average result is taken. It doesn't use any set of formulas.

Advantages of Random Forest:

1. It reduces over fitting in decision trees and helps to improve the accuracy.
2. It is flexible to both classification and regression problems.
3. It works well with both categorical and continuous values.

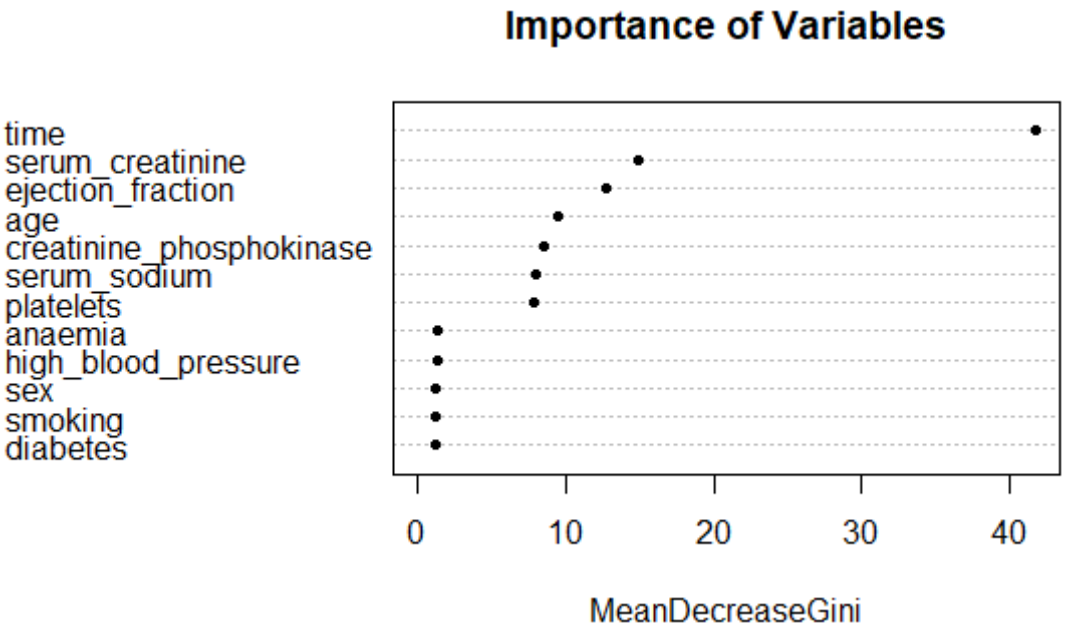
Disadvantages of Random Forest:

1. It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
2. It also require much time for training as it combines a lot of decision trees to determine the class.



Output of Random Forest:

```
Call:
 randomForest(formula = DEATH_EVENT ~ ., data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 14.11%
Confusion matrix:
    0  1 class.error
0 150 15  0.09090909
1  20 63  0.24096386
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 165   0
         1   0  83

               Accuracy : 1
                 95% CI : (0.9852, 1)
    No Information Rate : 0.6653
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.6653
         Detection Rate : 0.6653
   Detection Prevalence : 0.6653
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```



4. K- Nearest Neighbor:

K- Nearest Neighbor is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.

K- Nearest Neighbor (k-NN) is an algorithm of rules that is useful for making classifications/predictions while there are potential non-linear boundaries setting apart classes or values of interest. Conceptually, k-NN examines the training/values of the points around it (i.e., its neighbors) to determine the value of the focal point. The majority or common value can be assigned to the focus.

Algorithm:

1. Calculate the distance from x to all points in your data.
2. Sort the points in your data by increasing distance from x.
3. Predict the majority label of the k closest points.

Note that the value of k affects the results, its ideal to test the model for different values of k for better results and there by a better model.

Features of K- Nearest Neighbor Algorithm:

The KNN algorithm has the following features:

1. KNN is a Supervised Learning algorithm that uses labeled input data set to predict the output of the data points.
2. It is one of the most simple Machine learning algorithms and it can be easily implemented for a varied set of problems.
3. It is mainly based on feature similarity. KNN checks how similar a data point is to its neighbor and classifies the data point into the class it is most similar to.

Advantages of KNN:

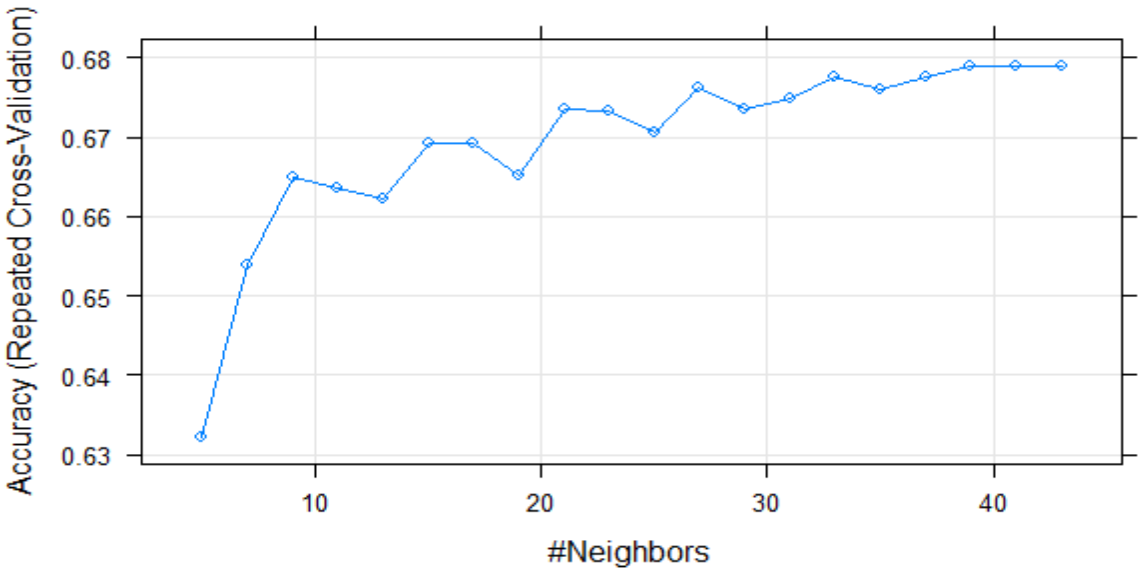
1. Simple to implement.
2. Flexible to feature/distance choices.
3. Naturally handles multi-class cases.
4. Can do well in practice with enough representative data.

Disadvantages of KNN:

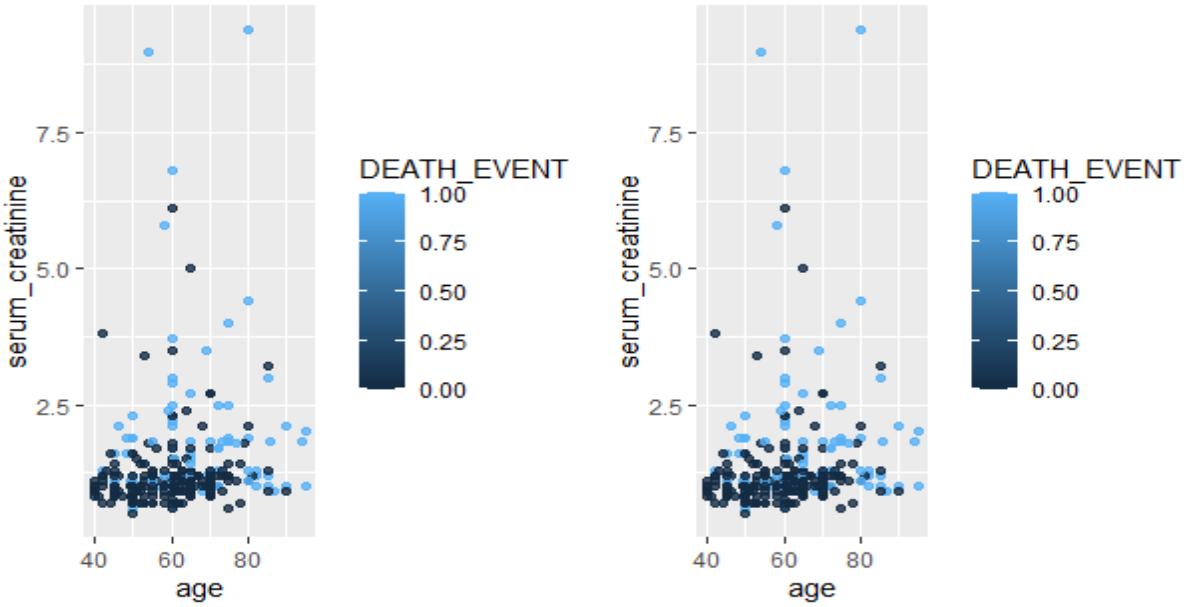
1. Need to determine the value of parameter K (number of nearest neighbors).
2. Computation cost is quite high because we need to compute the distance of each query instance to all training samples.
3. Storage of data.
4. Must know we have a meaningful distance function.

Output of K- Nearest Neighbor:

```
k-Nearest Neighbors

243 samples
 12 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 218, 219, 219, 218, 219, 219, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.6321232   0.036336378
   7  0.6537899   0.050917785
   9  0.6649614   0.055019765
  11  0.6636232   0.028776718
  13  0.6622295   0.012196010
  15  0.6691739   0.015725048
  17  0.6691739   0.015340557
  19  0.6651787  -0.006377673
  21  0.6735169  -0.002470489
  23  0.6732850  -0.004735198
  25  0.6706232  -0.015721043
  27  0.6761836  -0.005411765
  29  0.6734614  -0.010585322
  31  0.6747947  -0.008033488
  33  0.6776329  -0.002666667
  35  0.6761232  -0.005214153
  37  0.6775725  -0.002745098
  39  0.6790217   0.000000000
  41  0.6790217   0.000000000
```



Lo



5. Logistic Regression:

Logistic regression is also known as generalized linear model. As it is used as a classification technique to predict a qualitative response, Value of y ranges from 0 to 1

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most commonplace logistic regression fashions a binary outcome; some thing that can take values which include true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.

Logistic Regression is one of the supervised device learning algorithms used for class i.e. to predict discrete valued final results. It is a statistical method that is used to are expecting the final results of a established variable based totally on observations given in the training set.

Advantages of Logistic Regression:

1. Logistic regression is easier to implement, interpret, and very efficient to train.

2. It makes no assumptions about distributions of classes in feature space.

3. It can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions.

Disadvantages of Logistic Regression:

1. If the number of observations is lesser than the number of features, Logistic Regression should not be used; otherwise, it may lead to over fitting.
2. It constructs linear boundaries.
3. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

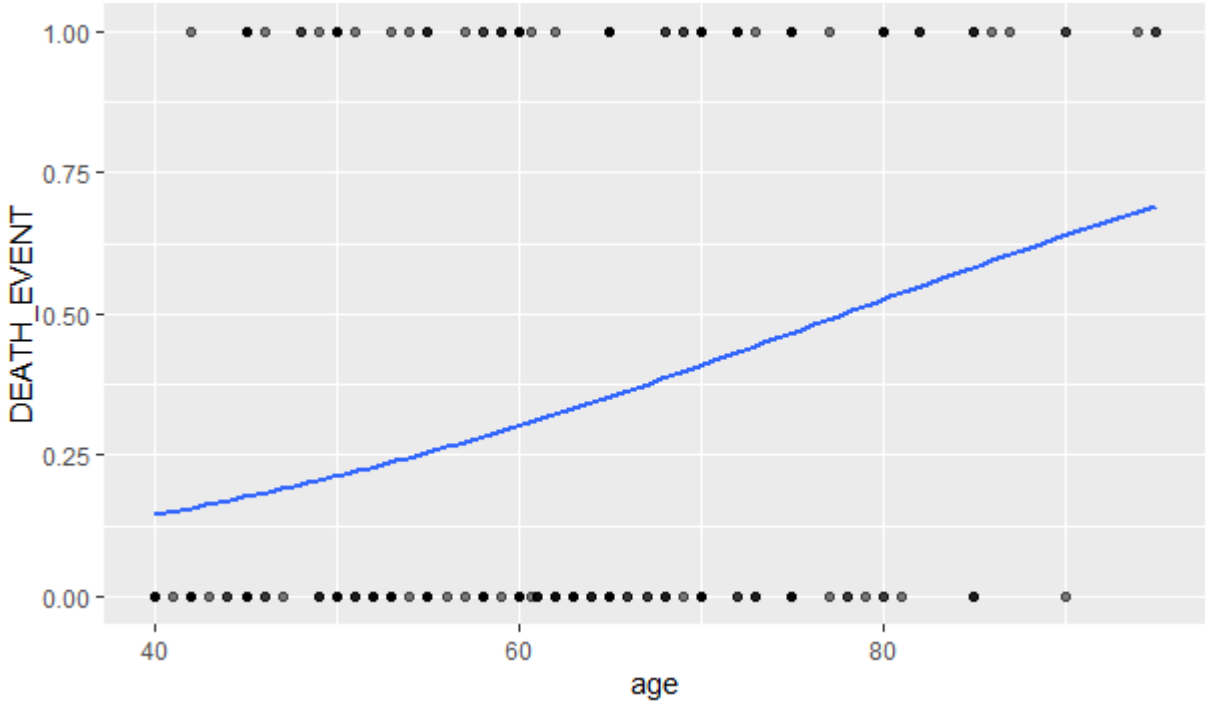
Features of Logistic Regression:

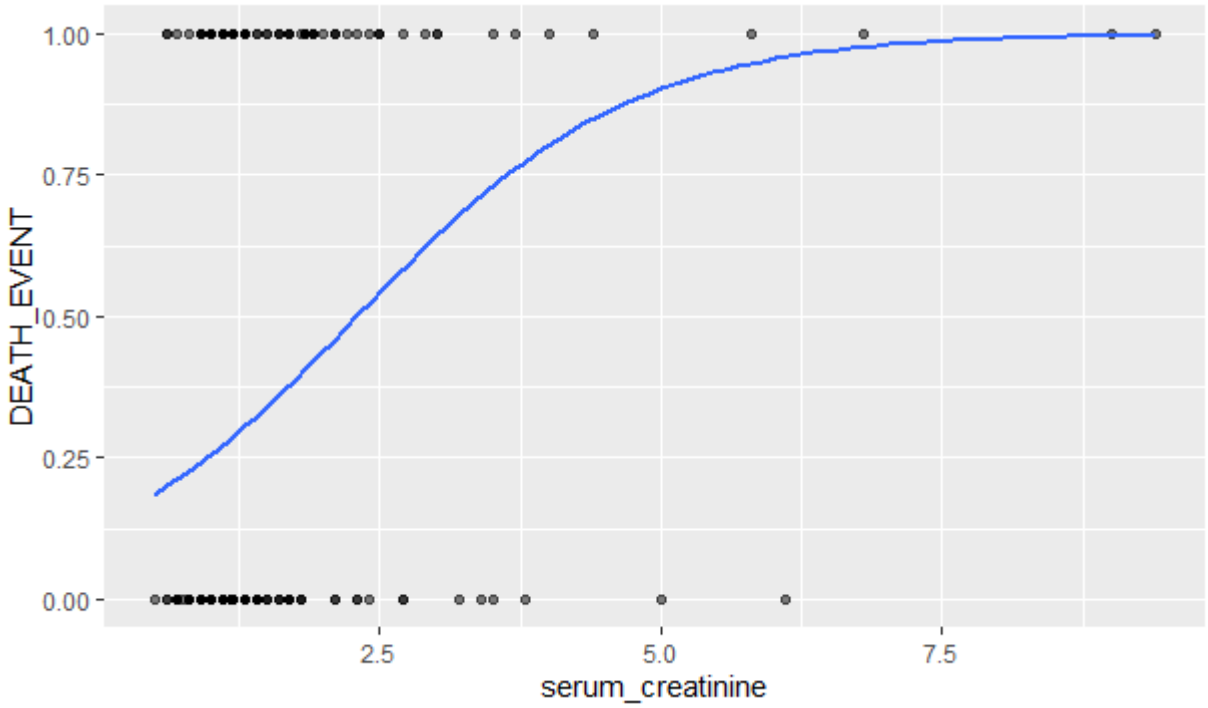
1. The dependent variable in logistic regression follows Bernoulli distribution.
2. Estimation is done through maximum likelihood.
3. No R Square, Model fitness is calculated through Concordance, KS-Statistics.'

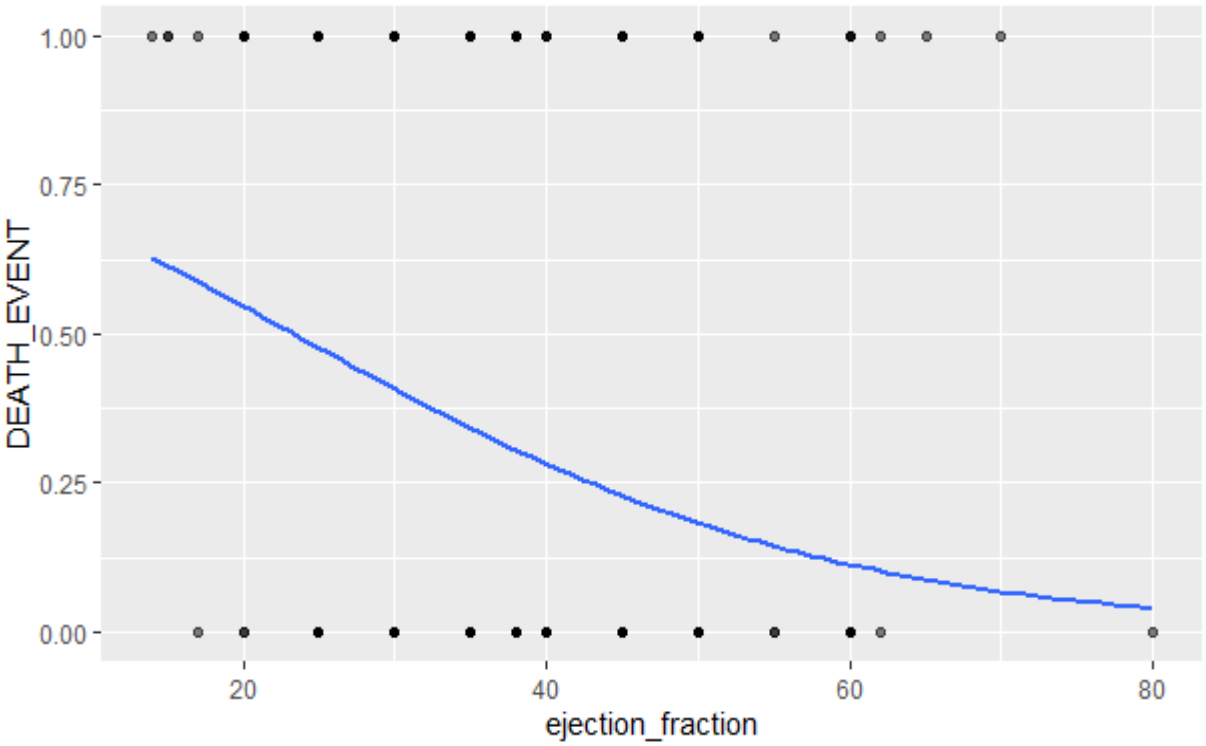
Output of Logistic Regression:

```
Call:
glm(formula = DEATH_EVENT ~ ejection_fraction + age + serum_creatinine +
    time, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0819  -0.6290  -0.2472   0.5134   2.9108

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.696706   1.186177   0.587   0.5570
ejection_fraction -0.081556   0.018777  -4.343 1.40e-05 ***
age                0.040843   0.016635   2.455   0.0141 *
serum_creatinine   0.841623   0.202542   4.155 3.25e-05 ***
time              -0.019697   0.003137  -6.279 3.40e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 305.02  on 242  degrees of freedom
Residual deviance: 187.32  on 238  degrees of freedom
AIC: 197.32

Number of Fisher Scoring iterations: 5
```

Plot of Logistic Regression:







6. Fuzzy C means Clustering:

Fuzzy C-Means clustering is a soft clustering approach, where each data point is assigned a likelihood or probability score to belong to that cluster. In fuzzy c-means clustering, we find out the centroid of the data points and then calculate the distance of each data point from the given centroids until the clusters formed becomes constant. More the data is near to the cluster center more is itsmembership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to at least one.

Fuzzy C means Clustering Algorithm steps:

Step 1: Initialize the membership matrix (Clusters).

Step 2: Find the centroids.

$$V_j = \frac{\sum_{i=1}^{n} (u_{ij})^m \, x_i}{\sum_{i=1}^{n} (u_{ij})^m}, \forall_j = 1,2, \dots c$$

Step 3: Find out the Euclidean distance of each point from centroids.

$$Distance = \sqrt{(x_i - \mu_x)^2 + \left(y_i - \mu_y\right)^2}$$

Step 4: Update the membership table.

$$u_{ij} = 1/\sum_{k=1}^{c}(d_{ij}/d_{ik})^{(\frac{2}{m}-1)}$$

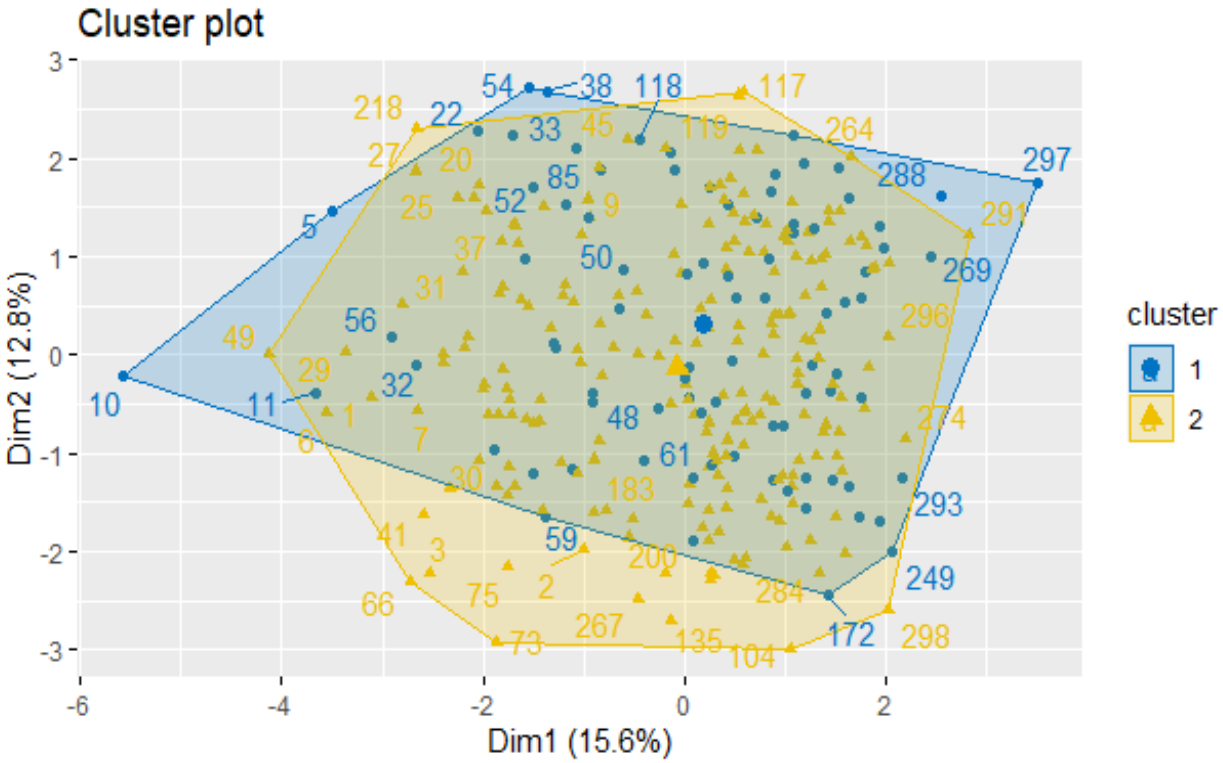
Step 5: Repeat step 2 to 4 until the constant values are obtained for the membership values or the difference is less than the tolerance value.

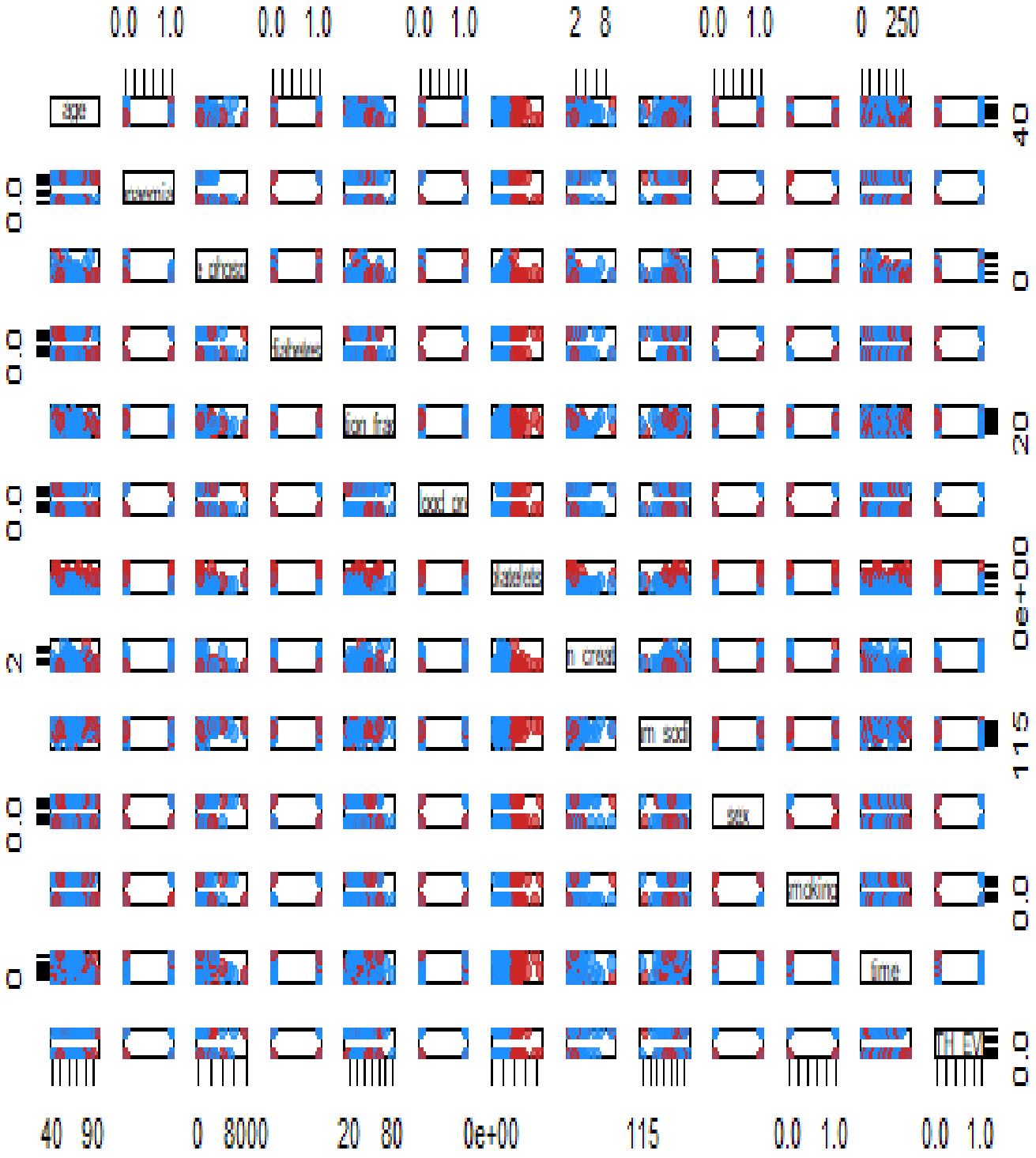
Advantages:

1. Gives best result for overlapped data set and comparatively better than k-means algorithm.
2. Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.

Disadvantages:

1. Apriori specification of the number of clusters.
2. With lower value of β we get the better result but at the expense of more number of iteration.
3. Euclidean distance measures can unequally weight underlying factors.




0.0 1.0
0.0 1.0
0.0 1.0
2 8
0.0 1.0
0 250
40
0
20
0e+00
115
0.0
0
40 90
0 8000
20 80
0e+00
115
0.0 1.0
0.0 1.0

4. Conclusion:

Decision trees: The decision tree is a key challenge in R and the strength of the tree is they are easy to understand and examine when compared with other models. They are being popularly used in data science problems. These are the tool produces the hierarchy of decisions implemented in statistical analysis. Statistical expertise is needed to recognize the logical interpretations of the decision tree. As we've got seen the choice tree is straightforward to recognize and the results are efficient when it has fewer magnificence labels and the alternative disadvantage a part of them is when there are extra magnificence labels calculations come to be complicated.. This post makes one become proficient to build predictive and tree-based learning models. We used libraries caret, rpart, rattle, rpart.plot, RColorBrewer and use fancyRpartplot for representing our decision tree and by way of the use of caret library we built confusion matrix, we get 89% accuracy.

Random Forest: We worked on RStudio for Random Forest, where we went over different commands, packages, and data visualization methods in R. We used package Random Forest and caret for visualization. For checking error rate of random forest we use plot function. We get OOB estimate of error rate 14.11% and accuracy is 1.

K-Nearest Neighbor: KNN algorithm is one of the simplest classification algorithms. Even with such simplicity, it can give highly competitive results. KNN algorithm can also be used for regression problems the best difference from the discussed technique might be the usage of averages of nearest neighbors rather than voting from nearest neighbors. We use library caret to apply KNN by this we get k=43 on the premise of higher accuracy.

Logistic Regression: Logistic regression is an efficient and powerful way to evaluate independent variable contributions to a binary outcome; however its accuracy depends in large part on careful variable selection with satisfaction of basic assumptions, as well as appropriate choice of model building strategy and validation of results. We use glm characteristics for making model and remove those variables which have higher p-value. And use predicts function on train and test data to locate accuracy.

Fuzzy C means Clustering: We have proposed an FCM clustering algorithm for locate the threat of heart attack of a patient using the information collected from the patients. With the right adjustment of FCM classifies, In order to find the abnormal and normal case efficiently we produce superlative variety of clusters with the help of FCM algorithm. FCM classifier is used to categories the information as heart sickness present or not. In this we use ppclust, factoextra, cluster, fclust libraries. The consequences of category experiment, preformed over records units received from 299 patients, indicate that the classifier has accomplished better accuracy than maximum of the existing algorithms. The performance of the proposed FCM is proved to be a well-known method in phrases of accuracy.

5.References:

https://www.researchgate.net/publication/324162326_Prediction_of_Heart_Diseases_Using_Data_Mining_and_Machine_Learning_Algorithms_and_Tools
https://iopscience.iop.org/article/10.1088/1742-6596/2031/1/012068/pdf
https://www.geeksforgeeks.org/decision-tree/
https://rpubs.com/mpfoley73/529130
https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/
https://rpubs.com/sukeshpabba/KNN#:~:text=K%20nearest%20neighbors%20is%20a,by%20increasing%20distance%20from%20x.
https://www.edureka.co/blog/knn-algorithm-in-r/#Features%20Of%20KNN%20Algorithm
https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/
https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm
https://researchhubs.com/post/ai/fundamentals/fuzzy-c-means.html