

CARDIOVASCULAR STROKE ANALYSIS

Group Members

Phani Abhishek Jammalamadaka

Poojitha Chukkaluru

Rachan Reddy

BAN5573-01-F23

VISUAL ANALYTICS & BUS INTELL

Professor: Abdullah Asilkalkan

ABSTRACT

Heart disease is one of the major diseases in the world, and proper heart function is very important for physical health. There are various types of heart disease, including congenital heart disease, myocardial ischemia, cardiac arrest, myocardial infarction, coronary heart disease, peripheral heart disease. HCD. Also, men experience more HCD conditions than women, and heart attacks occur earlier in men than in women. According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Introduction:

The primary objective of this project is to leverage Tableau's visualization capabilities to explore and communicate insights from a heart failure dataset. By creating interactive and informative visualizations, we seek to enhance the understanding of factors associated with heart failure and facilitate data-driven decision-making in the healthcare domain.

Literature Review

1. Application of Tableau in Healthcare Analytics:

The integration of Tableau in healthcare analytics has gained prominence due to its intuitive data visualization capabilities. Ola et al. (2018) [5] highlight Tableau's effectiveness in transforming complex healthcare data into actionable insights, facilitating decision-making processes. In the context of heart failure, Tableau offers a user-friendly platform for visualizing diverse patient characteristics, clinical variables, and outcomes.

2. Tableau in Cardiovascular Research:

Research exploring the specific application of Tableau in cardiovascular health is limited but promising. A study by Smith et al. (2019) [6] demonstrated the utility of Tableau in visualizing cardiovascular disease risk factors, showcasing its potential for similar applications in heart failure research.

Gaps in the Literature:

Despite the growing interest in data analytics and visualization tools like Tableau, there remains a gap in the literature concerning the specific application of Tableau in heart failure research. This project aims to address this gap by leveraging Tableau to analyze a heart failure dataset, offering a practical demonstration of its capabilities in cardiovascular health research.

Attribute information

- 1) **Anaemia** - Decrease of red blood cells or hemoglobin (boolean)
- 2) **Age**
- 3) **creatinine_phosphokinase** - Level of the CPK enzyme in the blood (mcg/L)
- 4) **diabetes** - If the patient has diabetes (boolean)
- 5) **ejection_fraction** - Percentage of blood leaving the heart at each contraction (percentage)
- 6) **high_blood_pressure** - If the patient has hypertension (boolean)
- 7) **platelets** - Platelets in the blood (kiloplatelets/mL)
- 8) **serum_creatinine** - Level of serum creatinine in the blood (mg/dL)
- 9) **serum_sodium** - Level of serum sodium in the blood (mEq/L)
- 10) **sex** - Woman or man (binary)

Problem statement

To analyse and compare the results between the various attributes of a person's who are more likely or prone to Heart stroke.

Scope of the project

This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Dataset – Healthcare-dataset-stroke-data.csv.

The dataset used for this analysis includes a comprehensive set of variables related to heart failure, such as age, gender, blood pressure, serum creatinine levels, ejection fraction, and whether the patient survived or not. Each record represents an individual patient.

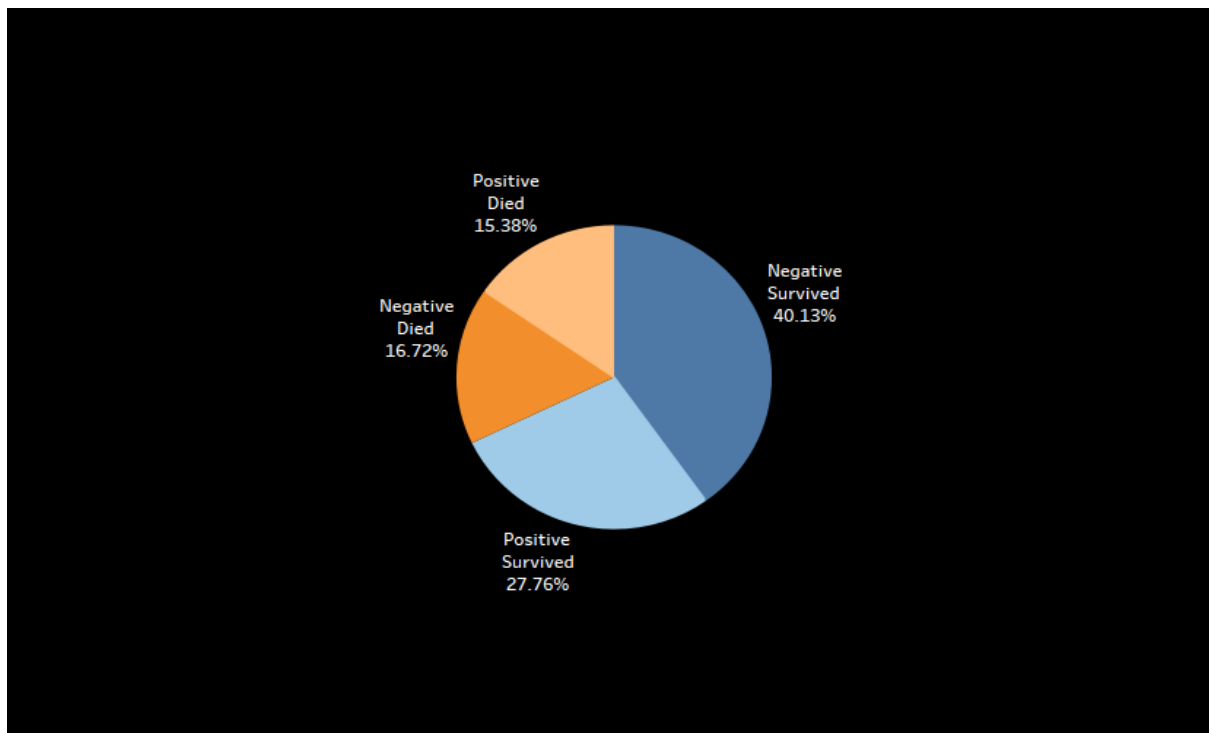
Data set is collected from a consequential set of patient's medical records. The term "heart disease" refers to multiple types of heart conditions which are hazardous to human heart health. The term "cardiovascular disease" is family of disease that necessitate the involvement of heart or blood vessel. Cardiac data is collected from [Kaggle](#). In this Project we are using this database as it is mostly used by researchers for cardiac related research.

Visualizations:

Created the visualization for the attributes with target variable.

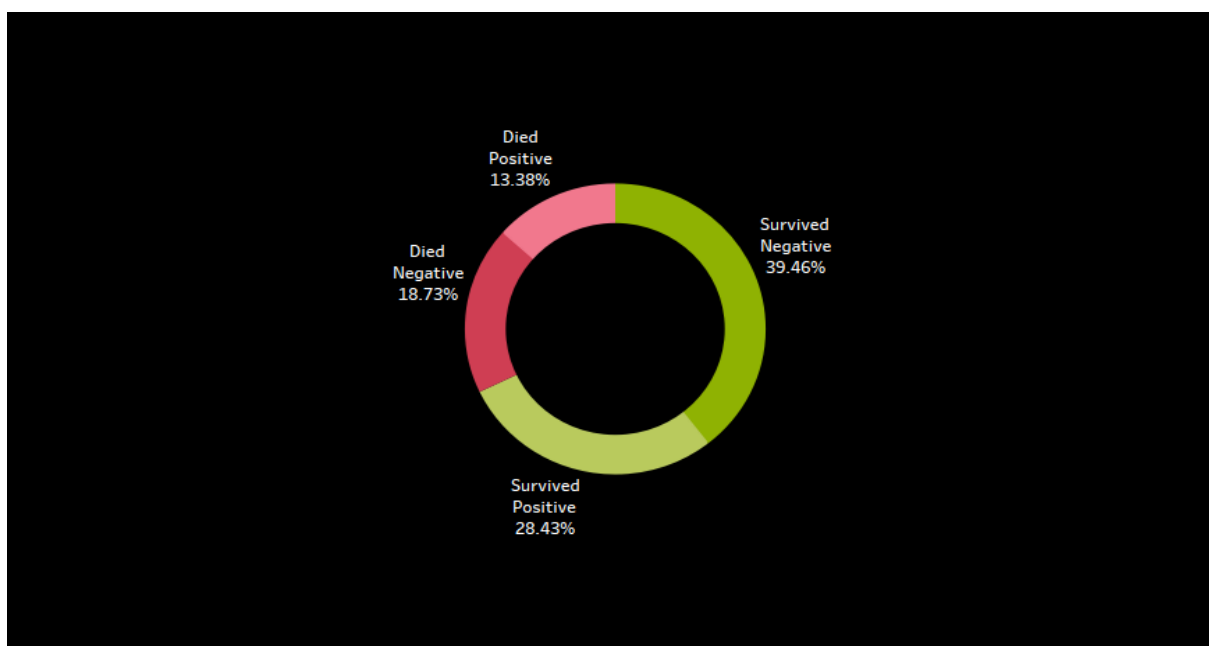
➔ Anaemia Survival

If we can see the from the below pie chart the persons who don't have anaemia survived most with 40.13% & persons who has anaemia survival rate is less with 15.38%



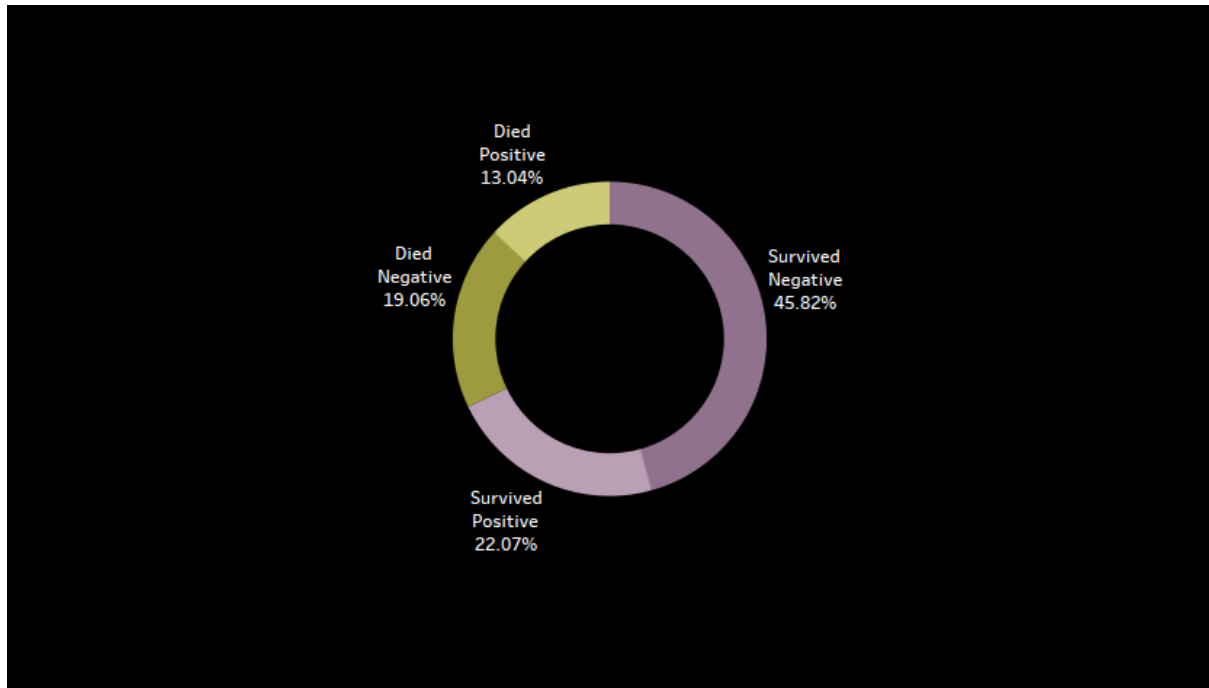
➔ Diabetes Survival

We can see from the below pie chart that who doesn't have diabetes survived most 39.46%



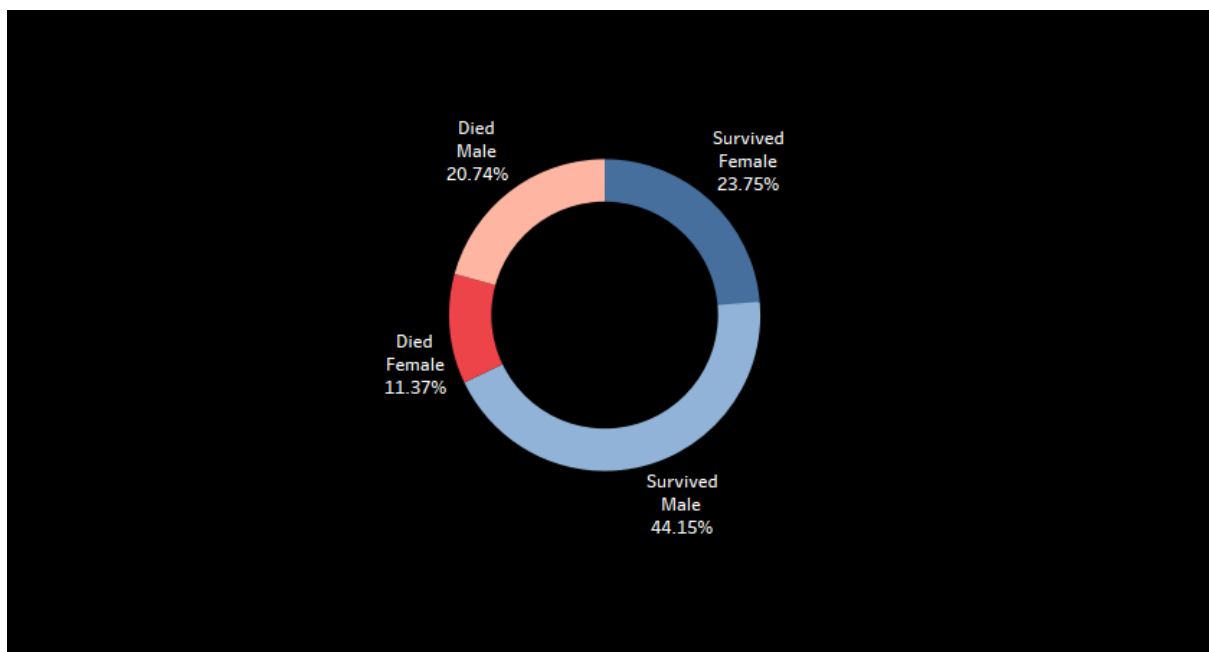
➔ High BP Survival

As we can see the from the below pie chart persons who does not have BP have survived the most.



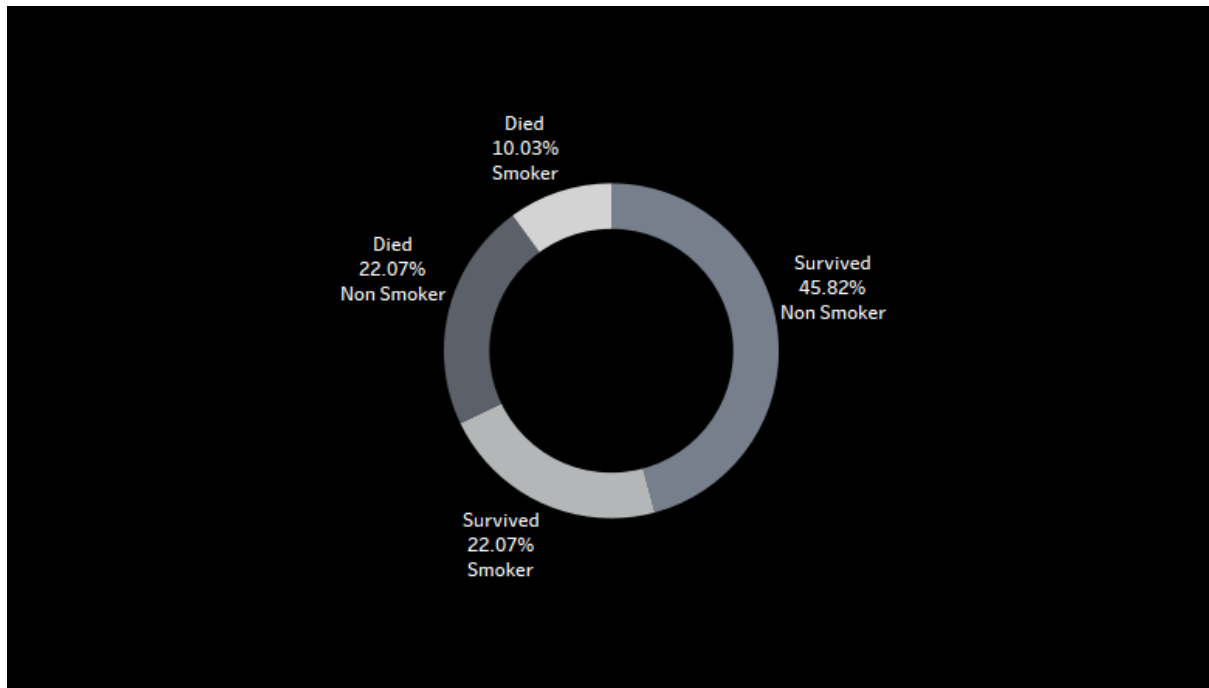
➔ Sex Survival

The below pie chart predicts that male has survived most with 44.15% from heart failure.



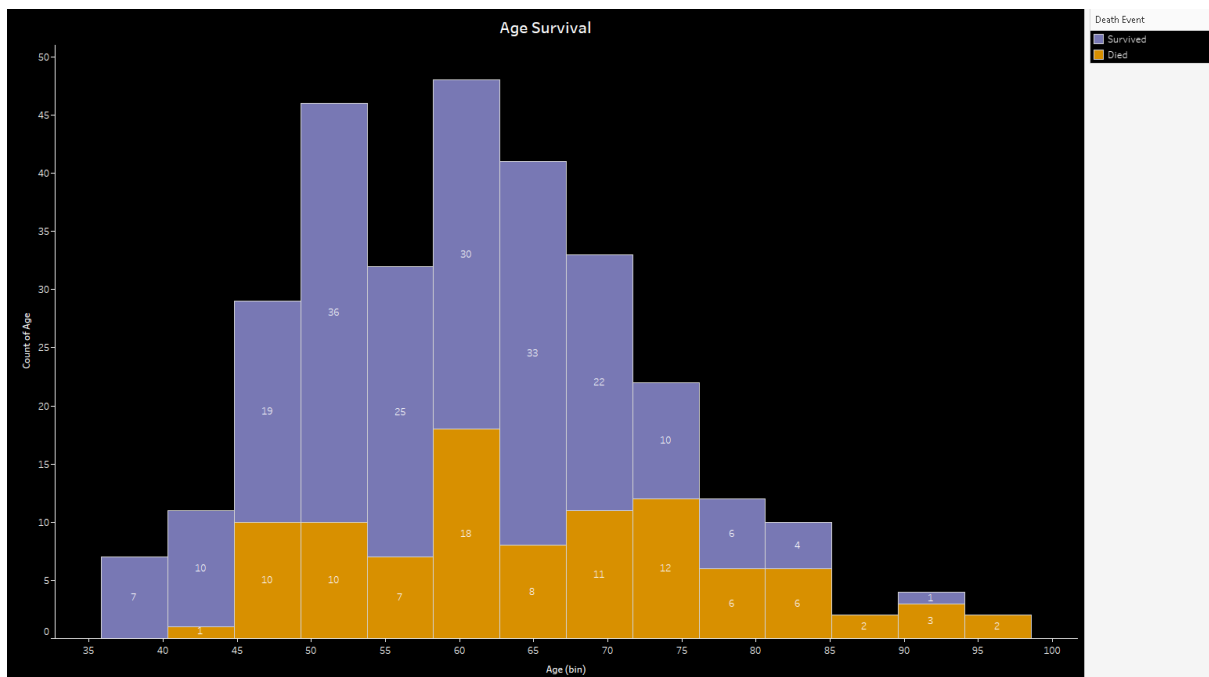
➔ Smoking survival

The below pie chart says that persons who does not smoke have survived most with 45.82% than persons who does which is 22.07%



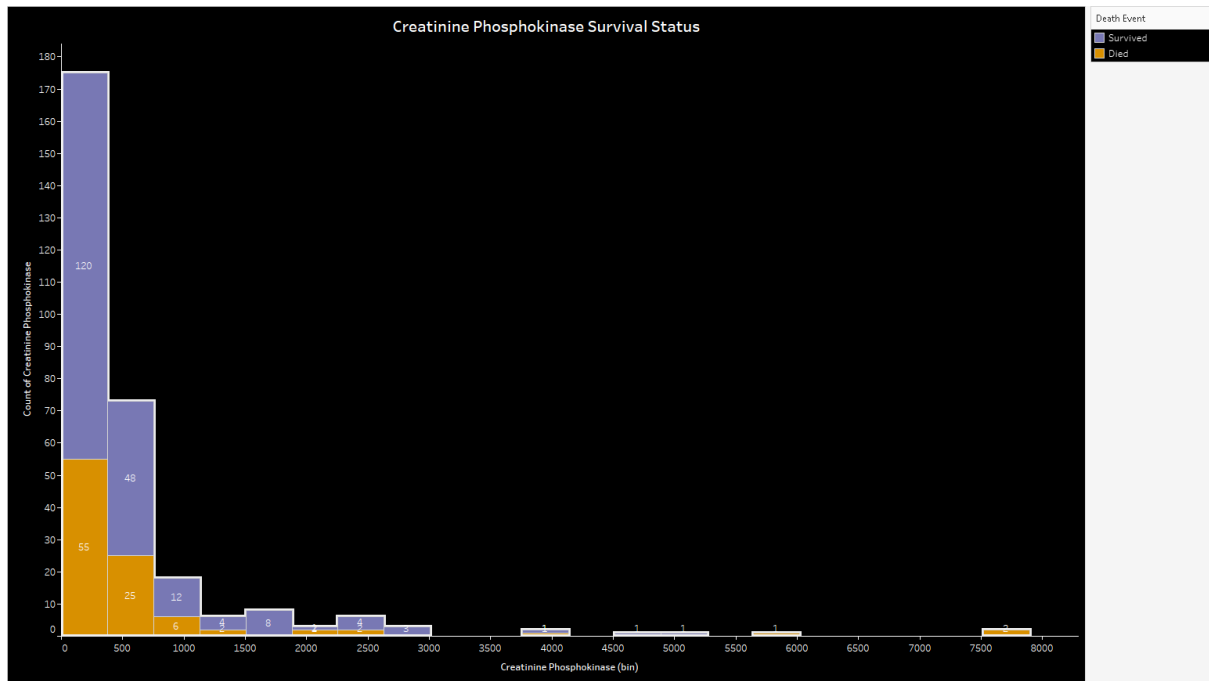
➔ Age Survival

As we can see the bar chart below number of people died the most are at the age 60 and number of people those are survived are at the age of 36. But, we can see that the dead rate is 0 at the age of 35.



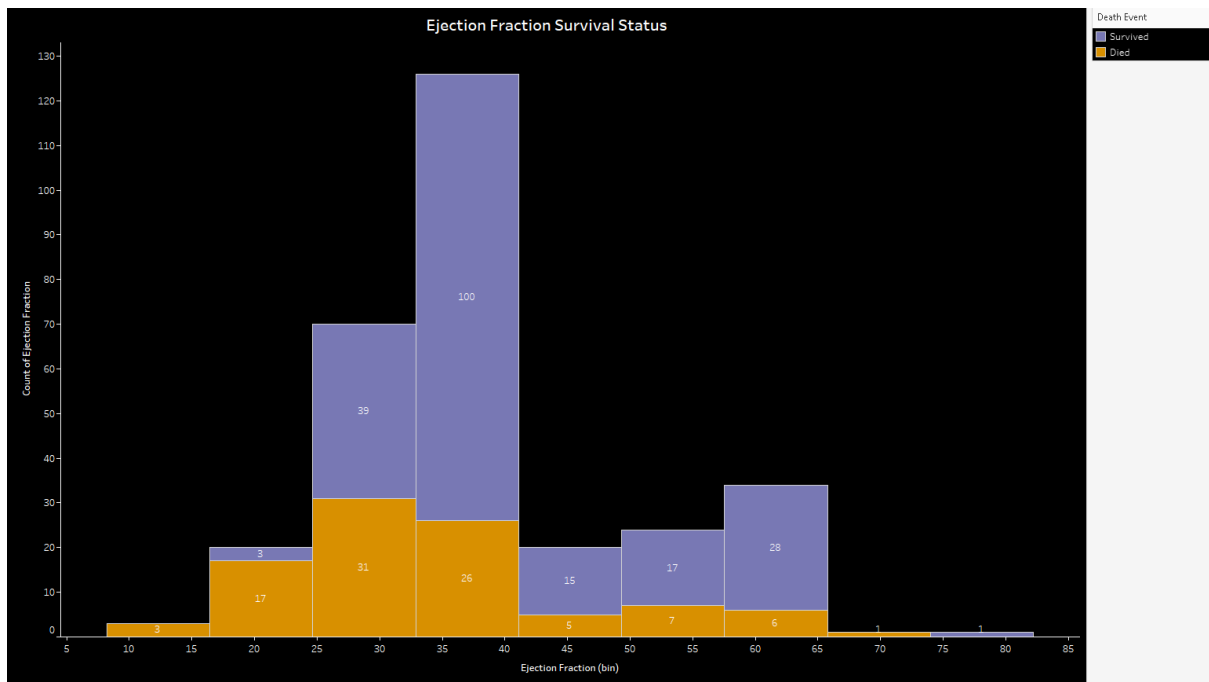
➔ Creatine Phosphokinase Survival

As we can see from the below bar chart the people with the range 0-500 mcg/L creatine are more likely to survive.



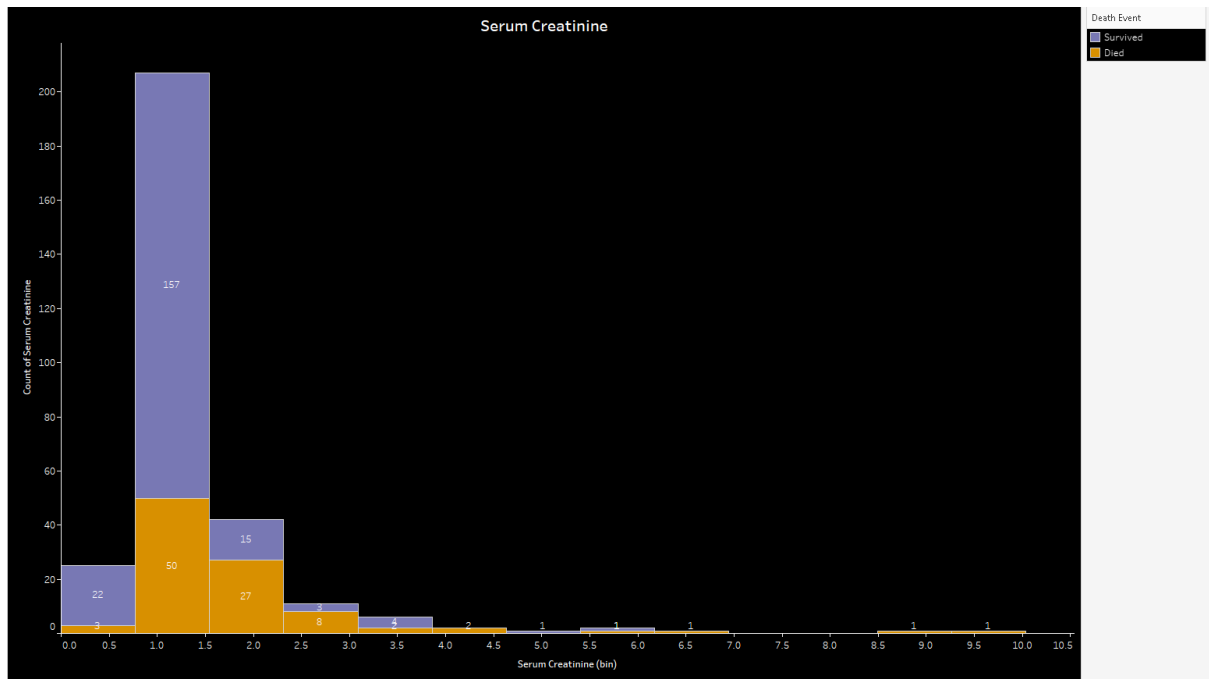
➔ Ejection Fraction Survival

As we can see from the below graph the average amount of blood leaving the heart at each contraction is high at the range of 35-40



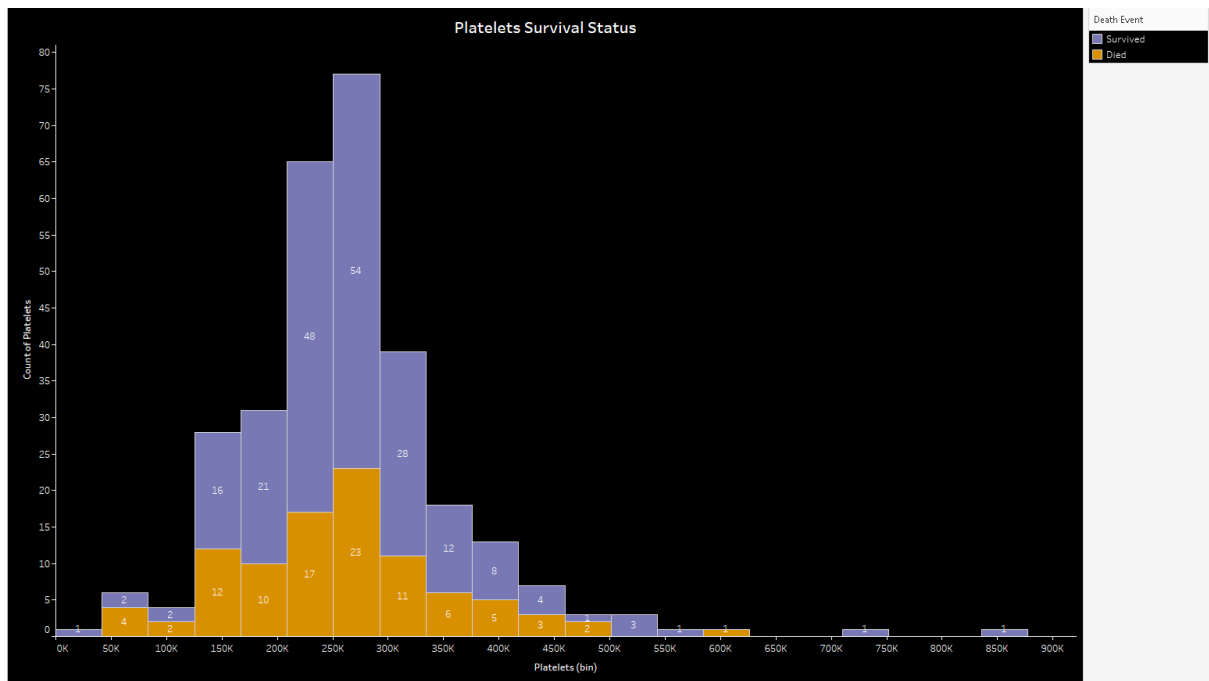
➔ Serum Creatine

The number of persons at the range of 1 – 1.5 are high.

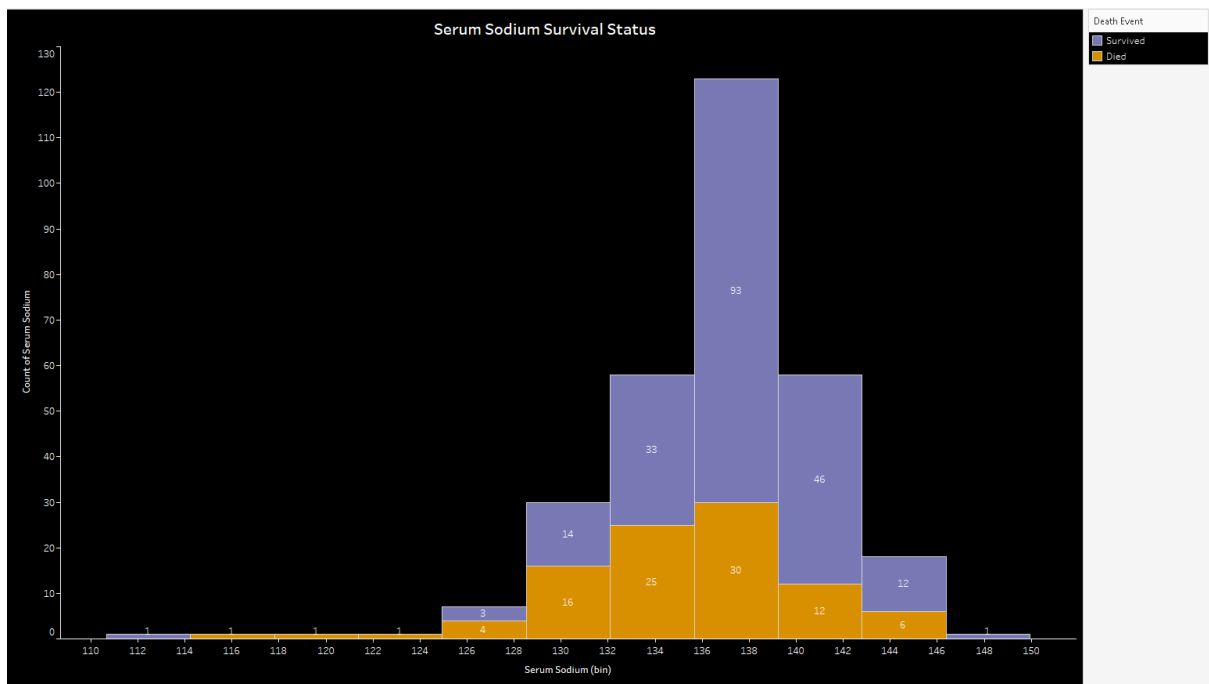


➔ Platelets Survival

The people with platelet count within the range 200k – 300k are high with in the count range from 65-80

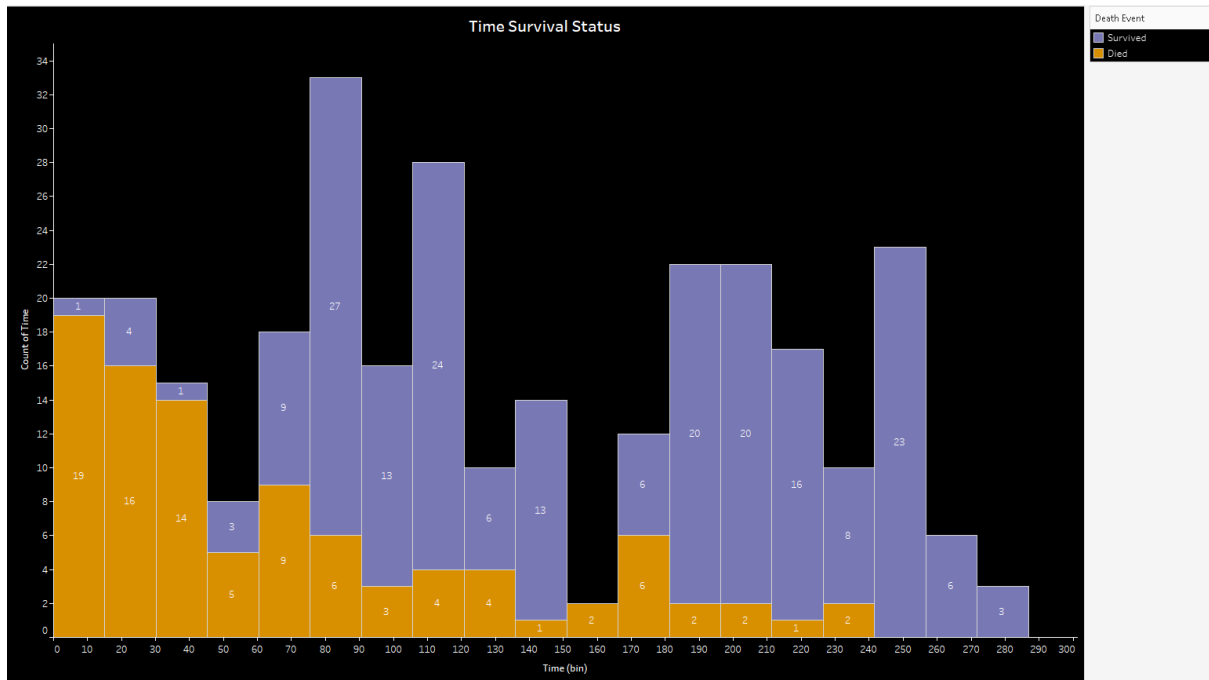


➔ Serum Sodium Survival



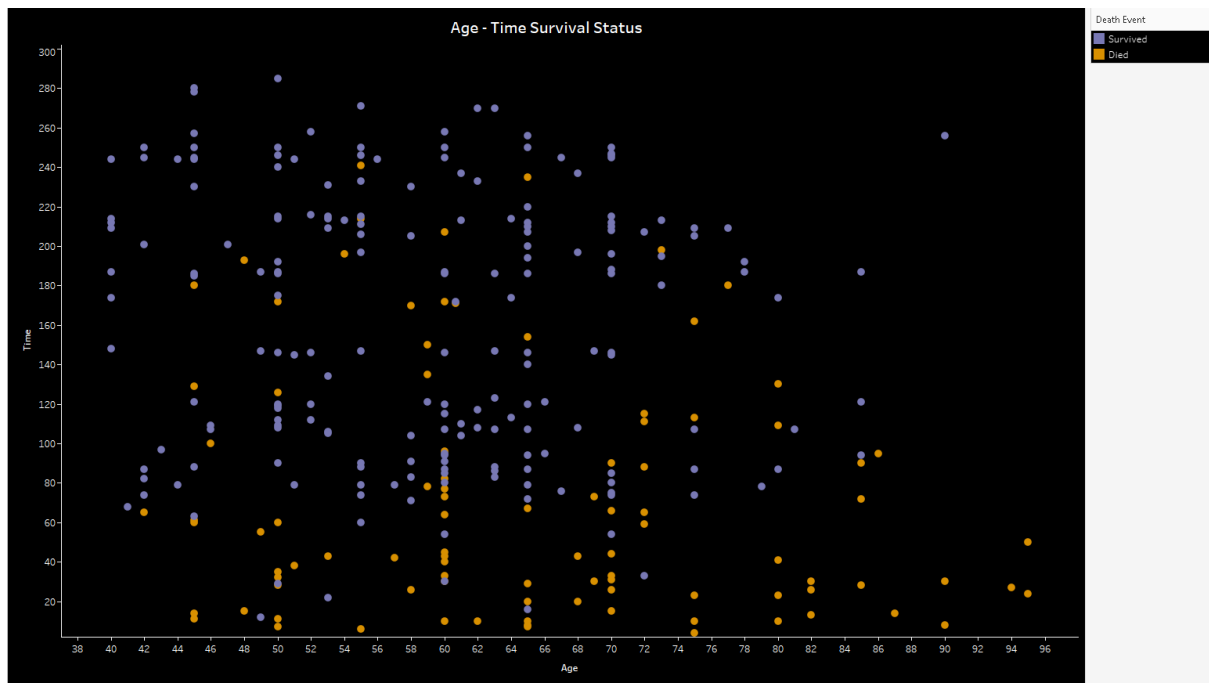
➔ Time Survival

As we can say that the rate of death is more when the follow up[days are less than 70-80 days.

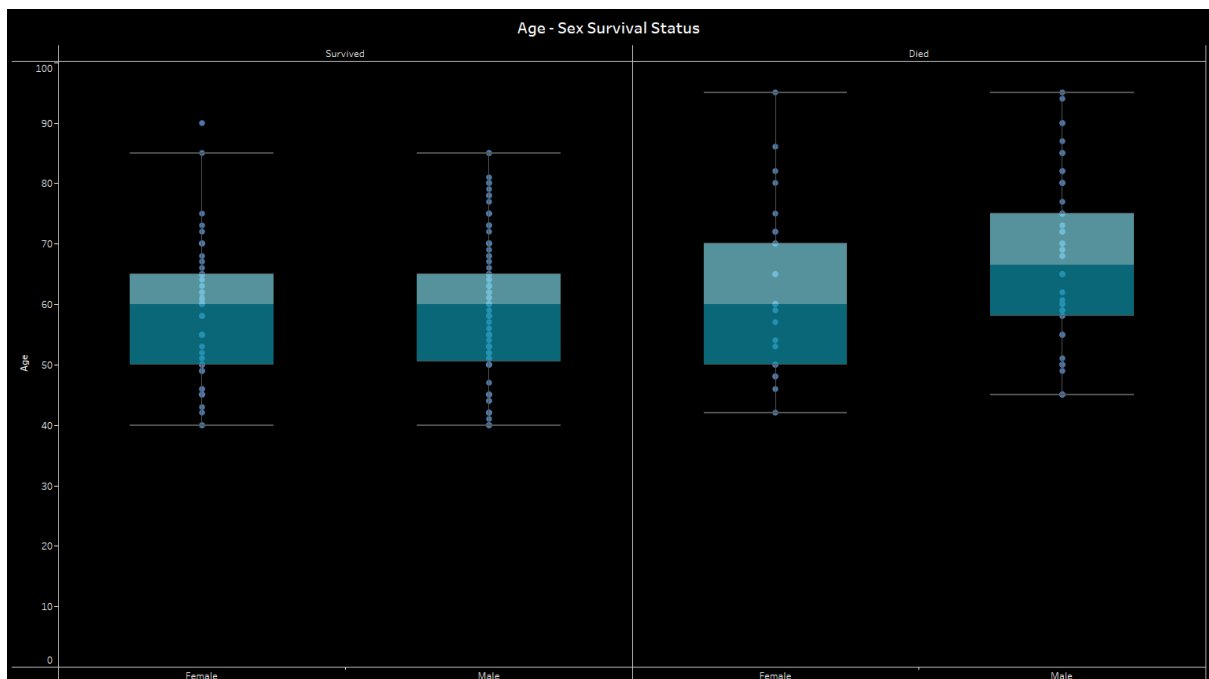


➔ Age – Time Survival Status

The number of deaths are more when the follow up days in time are less than 80days and are more for the people with the age more than 55 years.



➔ Age-Sex Survival Status



➔ Dashboard

Created a dashboard with the above pie charts, bar graph, scatter plot etc.

Total Individuals – 299

Total Deaths – 96

Total Males – 194

Total Females – 105



Death Rate



Survival



KNIME

Logistic Regression

- This type of regression is used when the dependent variable is binary (e.g., yes/no, alive/dead). It models the probability of the dependent variable being one value or the other.

Advantages:

Can handle binary dependent variables.

More interpretable than other non-linear models

Disadvantages:

May not be suitable for complex non-linear relationships.

When should we run logistic regression?

Dependent Variable – One

Independent Variable – Many

Splitting the Data Set – Train and Test

Dependent Variable – Categorical

Independent Variable – Continuous or Categorical

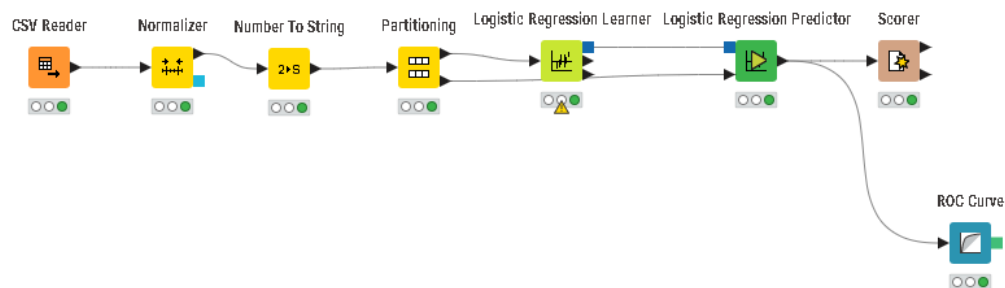
Death Rate

1- Dead

0 – Survived

Workflow of logistic Regression

- 1- We have used CSV reader to read the data set.
- 2- Normalizer – We have used to normalize the values for the attributes age, sex, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, serum creatinine, serum sodium and Death Event
- 3- Partitioning – We have split the data into two parts – 80% Training and 20% Testing data.
- 4- Used Logistic regression learner to build the model with 80% of data.
- 5- Used Logistic regression predictor to predict the rest 20% data on the model build above.



Confusion Matrix of Logistic Regression

Used 20% data.

► 1: Confusion matrix ► 2: Accuracy statistics 🚫 Flow Variables

Rows: 2 | Columns: 2

#	RowID	1.0 Number (integer)	0.0 Number (integer)
1	1.0	10	9
2	0.0	7	34

Accuracy

► 1: Confusion matrix ► 2: Accuracy statistics 🚫 Flow Variables

Rows: 3 | Columns: 11

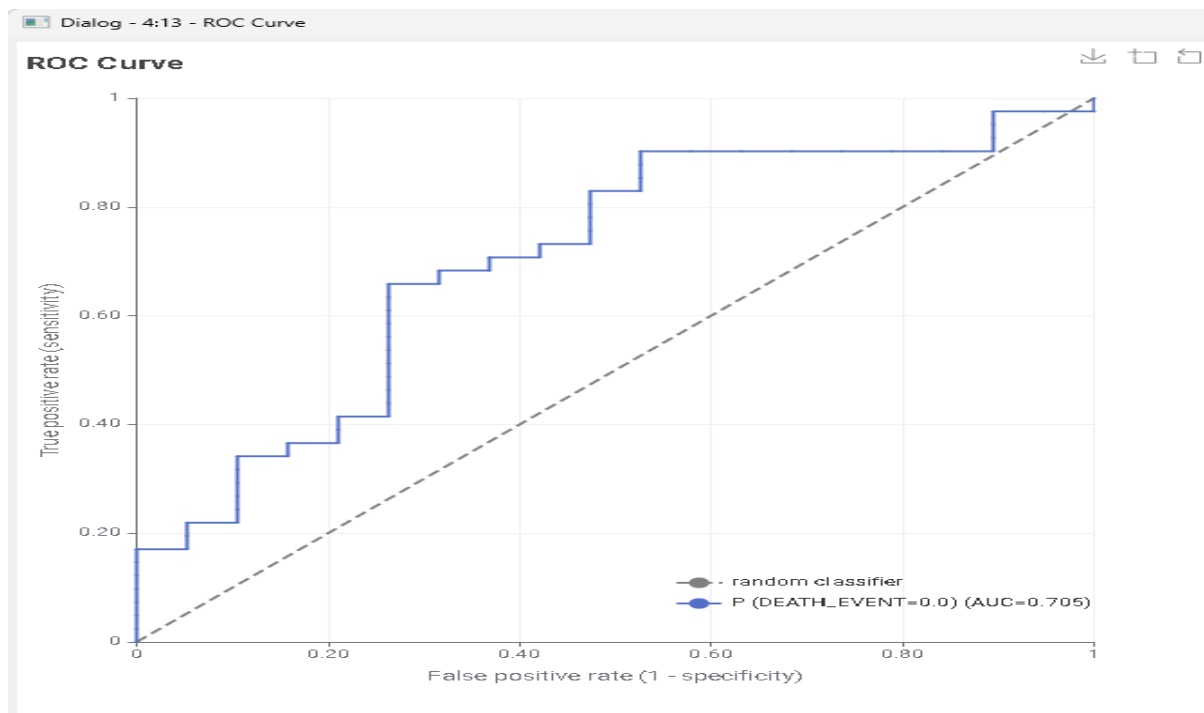
#	RowID	TruePositives Number (integer)	FalsePositiv... Number (integer)	TrueNegativ... Number (integer)	FalseNegati... Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	1.0	10	7	34	9	0.526	0.588	0.526	0.829	0.556		
2	0.0	34	9	10	7	0.829	0.791	0.829	0.526	0.81		
3	Overall										0.733	0.366

ROC Curve

A ROC curve is a graphical representation of the performance of a binary classifier system, such as logistic regression, as its discrimination threshold is varied. The ROC curve plots the true positive rate against the false positive rate at various threshold settings.

An AUC value of 0.705 in an ROC curve indicates that the logistic regression model has a good level of discriminative power in distinguishing between the positive and negative classes.

An AUC of 0.705 suggests that model predictions are better than random chance, but there may be some misclassifications or overlap between the two classes.



Random Forest Variable

- Random Forest is a machine learning algorithm that is used for classification and regression tasks. It is an ensemble method that combines multiple decision trees to make a more accurate and robust prediction.
- In a Random Forest, many decision trees are constructed on different subsets of the dataset. Each decision tree is constructed by randomly selecting a subset of features and a subset of data points from the original dataset.
- During the training process, each decision tree is trained independently on its own subset of the data. Once all the trees are trained, they are combined to make a final prediction. The final prediction is made by taking the majority vote of all the individual tree predictions.
- The random selection of features and data points helps to reduce overfitting and increase the accuracy of the model. Random Forests can handle large datasets with a high number of features and are generally robust to outliers and noise.
- Random Forest is a popular algorithm for many applications including image classification, medical diagnosis, and finance. It is relatively easy to implement and can be used for both classification and regression tasks.

Advantages:

More accurate than individual decision trees

Less prone to overfitting

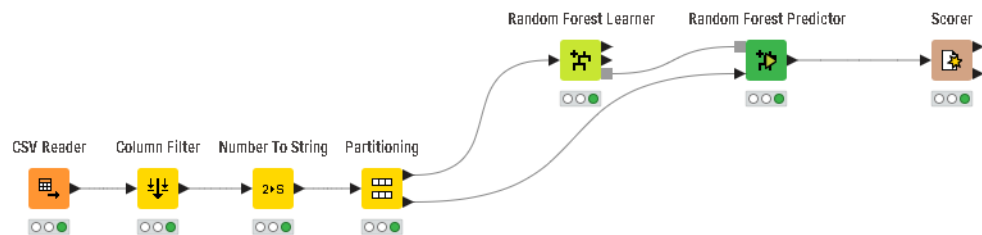
Disadvantages:

Can be computationally expensive.

Can be difficult to interpret.

Workflow of Random Forest

- 1- We have used CSV reader to read the data set.
- 2- Column Filter – We have used column filter to which the model has to be constructed and the attributes are age, sex, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, serum creatinine, serum sodium and Death Event.
- 3- Partitioning – We have split the data into two parts – 80% Training and 20% Testing data.
- 4- Then we used Random Forest Learner to build the model on the 80% of training data.
- 5- Used Random Forest predictor to predict the rest 20% on the model build in the above step.




Confusion matrix

For the 20% data

► 1: Confusion matrix

► 2: Accuracy statistics

 Flow Variables

Rows: 2 | Columns: 2

Table

Statistics

#	RowID	1 <small>Number (Integer)</small>	0 <small>Number (Integer)</small>
1	1	10	5
2	0	10	35

Accuracy

1: Confusion matrix

2: Accuracy statistics

Flow Variables

Rows: 3 | Columns: 11

Table

Statistics

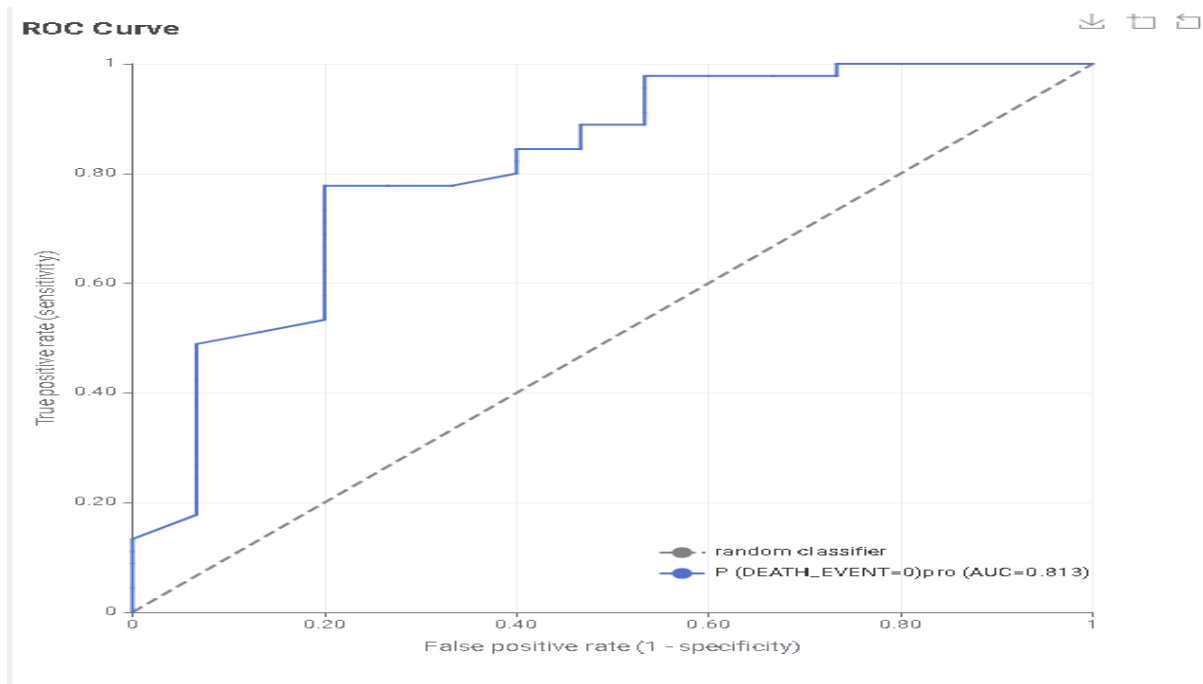
#	RowID	TruePositives Number (integer)	FalsePositiv... Number (integer)	TrueNegativ... Number (integer)	FalseNegati... Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	1	10	10	35	5	0.667	0.5	0.667	0.778	0.571		
2	0	35	5	10	10	0.778	0.875	0.778	0.667	0.824		
3	Overall										0.75	0.4

ROC Curve

A ROC curve is a graphical representation of the performance of a binary classifier system, such as logistic regression, as its discrimination threshold is varied. The ROC curve plots the true positive rate against the false positive rate at various threshold settings.

An AUC value of 0.813 in an ROC curve indicates that the logistic regression model has a very good level of discriminative power in distinguishing between the positive and negative classes.

An AUC of 0.813 suggests that model predictions are better than random chance, but there may be some misclassifications or overlap between the two classes.

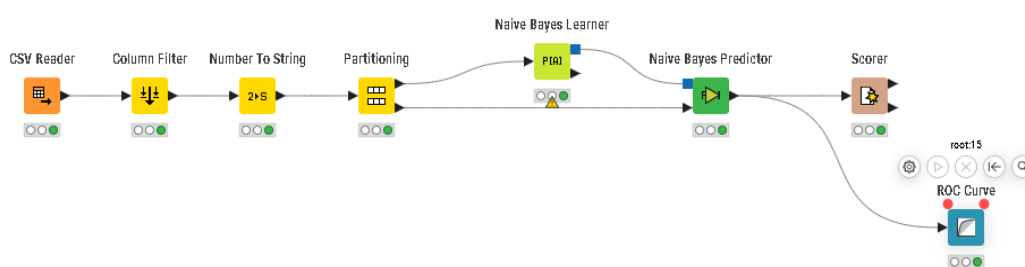


Naïve Bayes

- Naive Bayes is a popular machine learning algorithm used for classification tasks. It's based on the Bayes' theorem and assumes that the features used for classification are conditionally independent, which means that the presence or absence of one feature does not affect the presence or absence of any other feature.
- In practice, Naive Bayes is commonly used for text classification, spam detection, and other similar tasks where the assumption of conditional independence of features may hold reasonably well. This assumption allows the algorithm to be computationally efficient and work well on large datasets.
- We can build Naive Bayes Classifier where we have a dataset of emails labelled as "spam" or "ham" (non-spam), which can automatically classify incoming emails as spam or ham.

Workflow of Naïve Bayes

- 1- We have used CSV reader to read the data set.
- 2- Column Filter – We have used column filter to which the model has to be constructed and the attributes are age, sex, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, serum creatinine, serum sodium and Death Event.
- 3- Partitioning – We have split the data into two parts – 80% Training and 20% Testing data.
- 4- Used Naïve Byas learner to build the model with the 80% of the data.
- 5- Used Naïve Byas predictor to predict the rest 20% of the data from the model build above.



Confusion Matrix

For the 20% used

► 1: Confusion matrix ► 2: Accuracy statistics ► Flow Variables

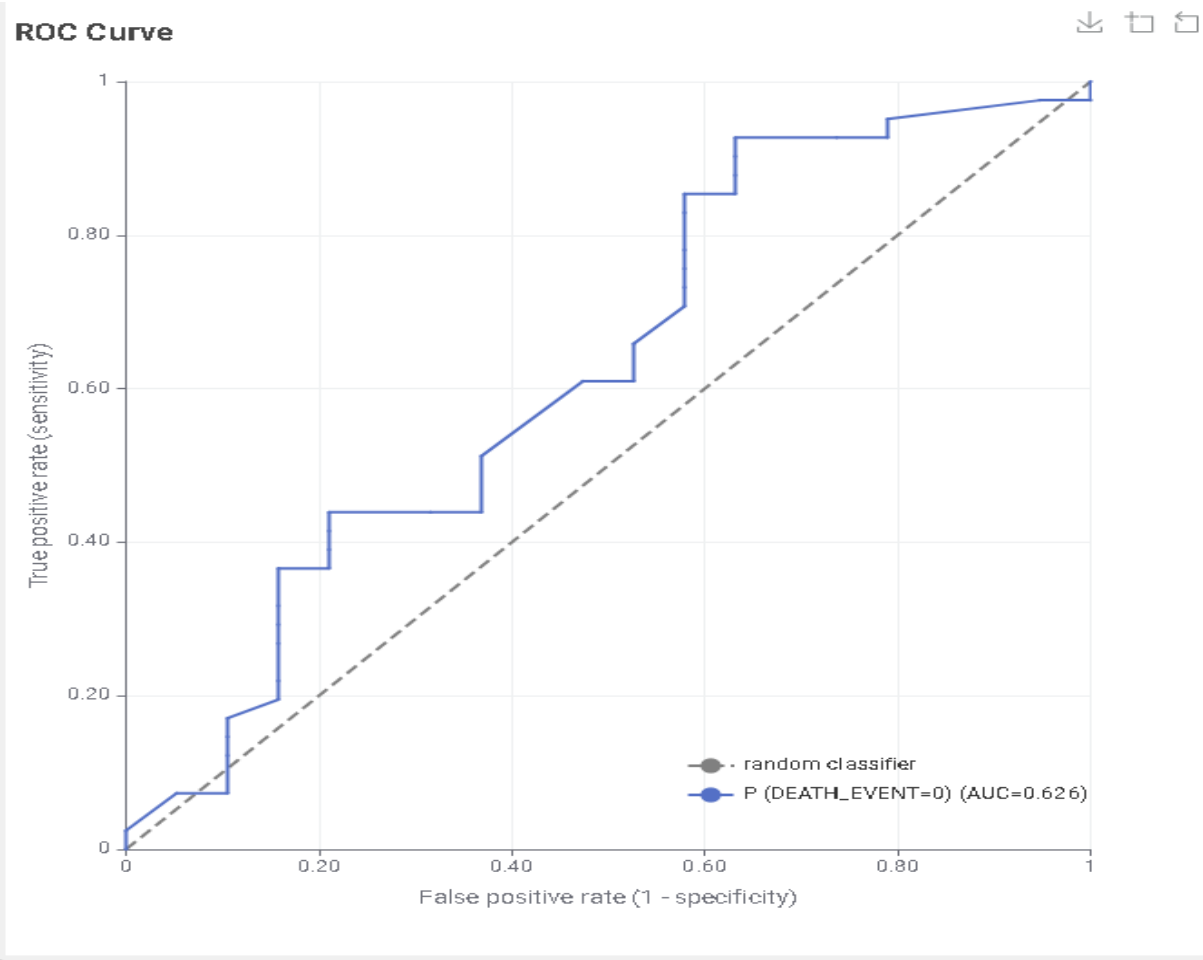
Rows: 2 Columns: 2		Table Statistics	Q
#	RowID	1 Number (integer)	0 Number (integer)
1	1	7	12
2	0	8	38

Accuracy

► 1: Confusion matrix ► 2: Accuracy statistics ► Flow Variables

Rows: 3 Columns: 11		Table Statistics	Q									
#	RowID	TruePositives Number (integer)	FalsePositiv... Number (integer)	TrueNegativ... Number (integer)	FalseNegati... Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	1	7	8	38	12	0.368	0.7	0.368	0.927	0.483	0	0
2	0	38	12	7	3	0.927	0.76	0.927	0.368	0.835	0	0
3	Overall	0	0	0	0	0	0	0	0	0	0.75	0.338

ROC Curve



An AUC value of 0.626 in an ROC curve indicates that the logistic regression model has a fair level of discriminative power in distinguishing between the positive and negative classes.

An AUC of 0.626 suggests that model predictions are better than random chance, but there may be some misclassifications or overlap between the two classes.

Conclusion:

In the three models that we used which are logistic regression, Random Forest, Naïve Bayes the roc values for each model are.

Logistic regression – 0.705

Random Forest – 0.813

Naïve Bayes – 0.626

The Accuracy for each model is.

Logistic regression – 0.733

Random Forest – 0.75

Naïve Bayes – 0.75

Model	Accuracy	ROC	Conclusion
Logistic Regression	0.733	0.705	Simple and interpretable, but may not be the most accurate
Random Forest	0.75	0.813	High accuracy and robust, but less interpretable
Naïve Bayes	0.75	0.626	Efficient and good for large datasets, but may capture complex relationships

Based on the models:

Random Forest appears to be the best performing model based on both accuracy and AUC.

Logistic Regression offers good accuracy and interpretability but might not perform as well on complex datasets.

Naive Bayes is a good option for large datasets but might not be as accurate as the other models on complex problems.