Heart Failure Prediction

Predicting the survival using Heart Failure clinical records dataset

Introduction

- Heart failure is an important problem and a costly one.
- It is often hard to predict what factors contribute to a heart failure.
- In this project, we will be characterizing a heart failure by obtaining the most important indicators that may lead to death.
- We are trying to know in advance which factors of the individuals who suffered a heart-attack and lead to subsequent death event.
- The factors that are indicators of heart failures can be used to diagnose a patient and take precautions.

Problem Statement

- Given a dataset of patients and their health parameter recordings predict weather or not a heart failure is probable.
- Find the best predictor variables that are best able to predict death event.
- Find the best classification algorithm that is best able to predict death event.
- Find the best clustering algorithm that is able to find groups of people who will or will not suffer a heart failure.

Dataset

- We have used the Heart Failure clinical records dataset from UCI Machine Learning Repository.
- This dataset contains different categorical and numerical values which can be used in predicting the death event using classification and clustering models.
- The dataset has very minimal/no missing or redundant data increasing the accuracy of the model performance.

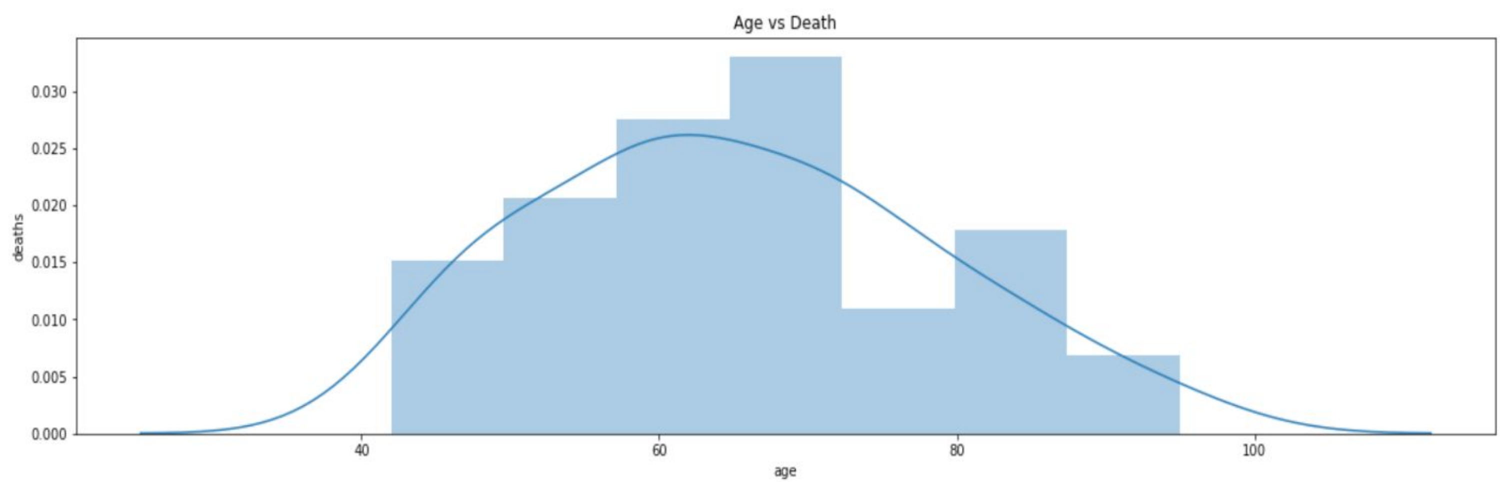
Dataset Variables

- Age: age of the patient (years)
- Anaemia: decrease of red blood cells or hemoglobin (boolean)
- High blood pressure: if the patient has hypertension (boolean)
- Creatinine phosphokinase (CPK): 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)
- Platelets: platelets in the blood (kiloplatelets/mL)
- Sex: woman or man (binary)
- Serum creatinine: level of serum creatinine in the blood (mg/dL)
- Serum sodium: level of serum sodium in the blood (mEq/L)
- 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)

Data Exploration

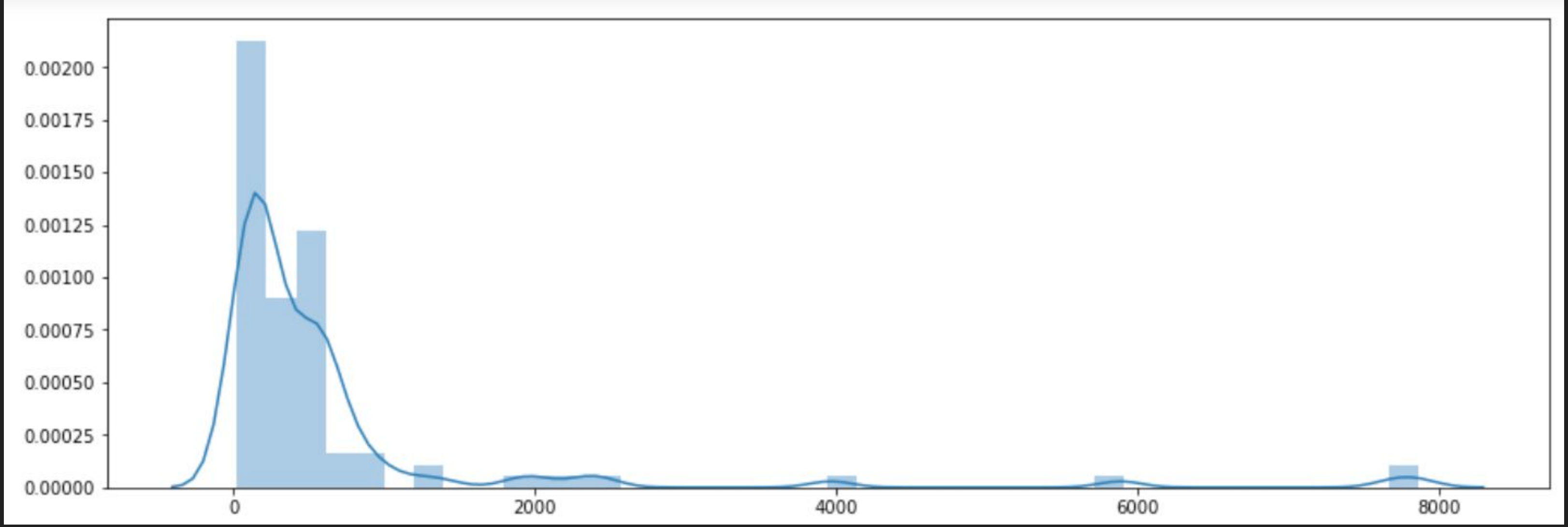
By using a number of plotting techniques we try to find out which variables are useful for our prediction task.

- Age: We see that the maximum number of deaths in the age between 60 to 80



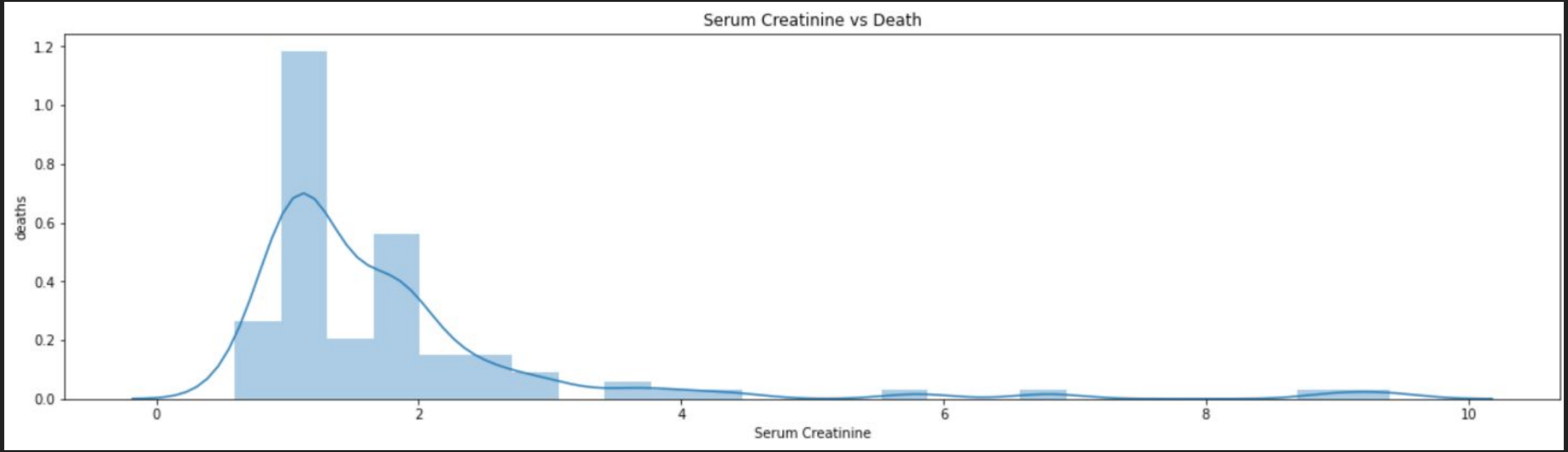
Data Exploration

- Creatinine Phosphokinase: when the level of CPK enzyme is low in the blood, the chances of heart failure is higher.



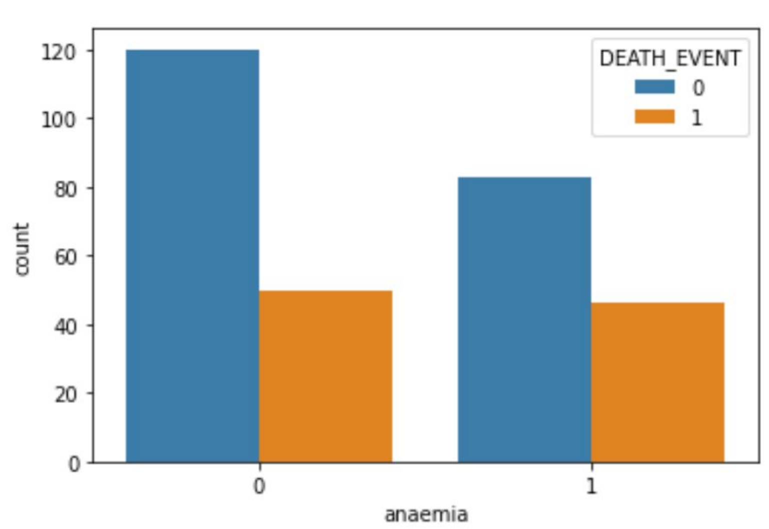
Data Exploration

- Serum creatinine: when the level of serum creatinine in the blood is low, the chances of heart failure is higher.

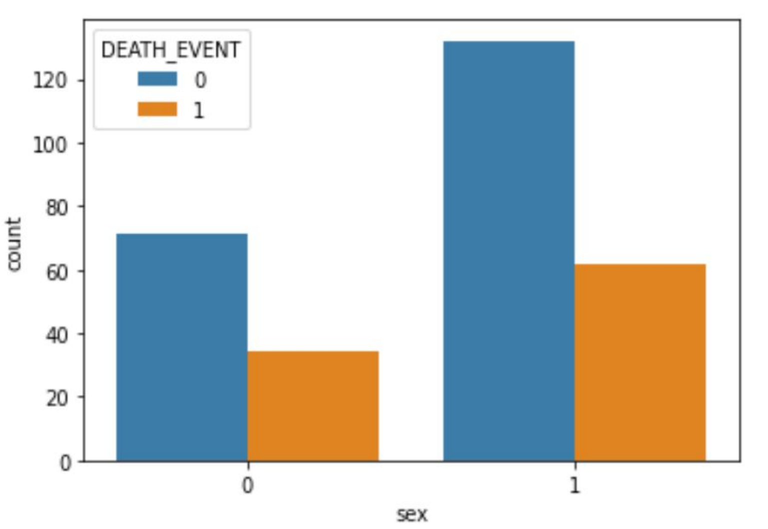


Data Exploration

- Anaemia: survivors among the ones who have Anaemia is lower.



- Sex: number of heart failures among men is higher.



Classification

We try several different classification algorithms:

- Decision Tree

Variables	Accuracy	Precision	Recall	F1 Score
All	0.81	0.72	0.78	0.75
Age & Gender	0.65	0.53	0.3	0.38
Diabetes, High BP, Platelets, Anaemia & Smoking	0.51	0.29	0.26	0.27
creatinine_phosphokinase, serum_creatinine,serum_sodium and ejection_fraction	0.61	0.46	0.48	0.47
serum_sodium and ejection_fraction	0.72	0.64	0.52	0.57

Classification

We try several different classification algorithms:

- K Nearest Neighbors

Variables	Accuracy	Precision	Recall	F1 Score
All	0.67	0.57	0.3	0.39
Age & Gender	0.65	0.57	0.15	0.24
Diabetes, High BP, Platelets, Anaemia & Smoking	0.64	0	0	0
creatinine_phosphokinase, serum_creatinine,serum_sodium and ejection_fraction	0.64	0.5	0.19	0.27
serum_sodium and ejection_fraction	0.67	0.67	0.15	0.24

Classification

We try several different classification algorithms:

- Naive Bayes

Variables	Accuracy	Precision	Recall	F1 Score
All	0.69	0.67	0.3	0.41
Age & Gender	0.59	0.42	0.37	0.39
Diabetes, High BP, Platelets, Anaemia & Smoking	0.55	0.33	0.26	0.29
creatinine_phosphokinase, serum_creatinine,serum_sodium and ejection_fraction	0.67	0.56	0.37	0.44
serum_sodium and ejection_fraction	0.64	0.5	0.44	0.47

Classification

We try several different classification algorithms:

- Support Vector Machines

Variables	Accuracy	Precision	Recall	F1 Score
All	0.81	0.84	0.59	0.7
Age & Gender	0.65	0.57	0.15	0.24
Diabetes, High BP, Platelets, Anaemia & Smoking	0.64	0	0	0
creatinine_phosphokinase, serum_creatinine,serum_sodium and ejection_fraction	0.73	0.68	0.48	0.57
serum_sodium and ejection_fraction	0.75	0.7	0.52	0.6

Clustering

We try several different clustering algorithms:

- Hierarchical Clustering - Single Linkage

Variables	Rand Index	Silhouette Coefficient
All	0.007	0.416
Age & Gender	-0.001	0.552
Health Problem	-0.003	0.543
Body Parameters	0.007	0.666
Anaemia	0.007	1

Clustering

We try several different clustering algorithms:

- Hierarchical Clustering - Complete Linkage

Variables	Rand Index	Silhouette Coefficient
All	0.022	0.469
Age & Gender	0.111	0.342
Health Problem	-0.007	0.11
Body Parameters	0.026	0.657
Anaemia	0.007	1

Clustering

We try several different clustering algorithms:

- K Means Clustering

Variables	Rand Index	Silhouette Coefficient
All	-0.003	0.117
Age & Gender	-0.001	0.552
Health Problem	-0.005	0.246
Body Parameters	0.154	0.24
Anaemia	0.007	1

Conclusion

After performing data science techniques on Heart Failure clinical records dataset, we can conclude that ejection fraction, serum creatinine, anaemia, Diabetes and High Blood Pressure are prominent variables in determining the survival rate of people with heart failure.