Final Project Report

Team: Ashwin Deshpande, Nandana Shimoga Prasad, Apurva Raghunath

Problem Statement:

Predicting the survival of people with heart failure using the Heart Failure clinical records dataset. By using techniques of classification and clustering in data science, our goal is to predict the death event of people who are likely to die due to heart failure influenced by various factors like age, high blood pressure, anemia and so on.

Data Collection:

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.

Different variables present in the dataset are listed below:

- 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)
- [target] death event: if the patient deceased during the follow-up period (boolean)

Data Preparation:

Following steps were used to prepare the dataset for further analysis.

1. **Partitioning the dataset for training and test** – The dataset was split such that 75% of the observations were used for training the models and 25% to test the performance of the models built.

2. **Exploring the dataset** - The data.info() shows that the dataset consists of 299 observations and 13 variables in all. The data type of three variables are float64 and ten variables are of int64.
3. **Irrelevant and Missing Attributes** - The dataset was complete and did not contain any missing/irrelevant/redundant, noise and outliers. However data.dropna() was used to ensure that any missing data are not over looked.

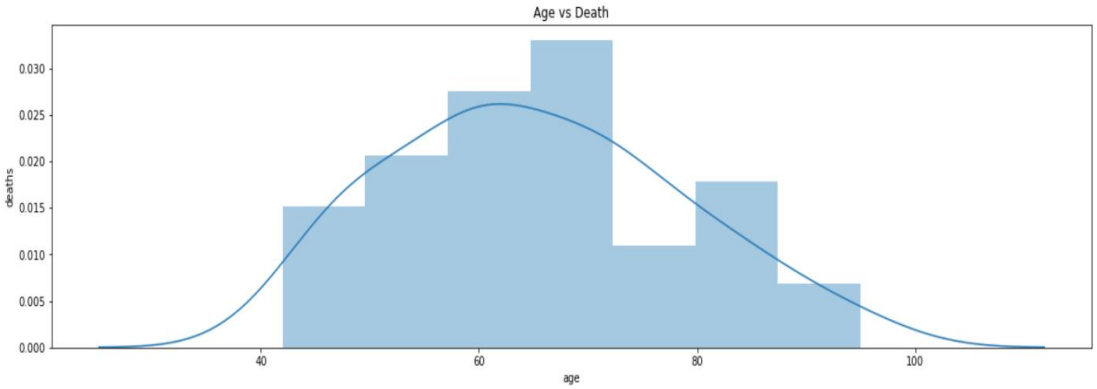
Data Exploration:

The dataset is explored in the following manner

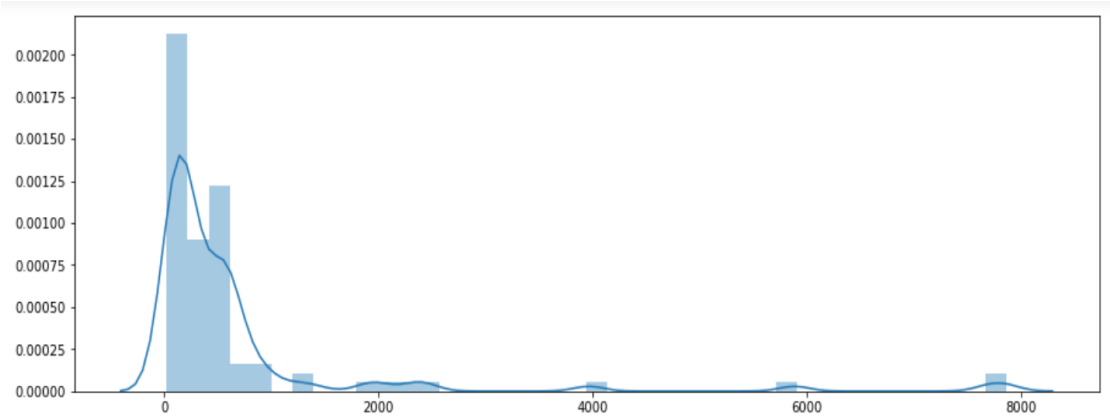
1. **Descriptive Statistics** – Various statistics such as count, median, mean and so on were computed. This helps us to better understand the dataset and predict the target variable accurately.

	count	mean	std	min	25%	50%	75%	max
age	299.0	60.83	11.89	40.0	51.0	60.0	70.0	95.0
anaemia	299.0	0.43	0.50	0.0	0.0	0.0	1.0	1.0
creatinine_phosphokinase	299.0	581.84	970.29	23.0	116.5	250.0	582.0	7861.0
diabetes	299.0	0.42	0.49	0.0	0.0	0.0	1.0	1.0
ejection_fraction	299.0	38.08	11.83	14.0	30.0	38.0	45.0	80.0
high_blood_pressure	299.0	0.35	0.48	0.0	0.0	0.0	1.0	1.0
platelets	299.0	263358.03	97804.24	25100.0	212500.0	262000.0	303500.0	850000.0
serum_creatinine	299.0	1.39	1.03	0.5	0.9	1.1	1.4	9.4
serum_sodium	299.0	136.63	4.41	113.0	134.0	137.0	140.0	148.0
sex	299.0	0.65	0.48	0.0	0.0	1.0	1.0	1.0
smoking	299.0	0.32	0.47	0.0	0.0	0.0	1.0	1.0
time	299.0	130.26	77.61	4.0	73.0	115.0	203.0	285.0
DEATH_EVENT	299.0	0.32	0.47	0.0	0.0	0.0	1.0	1.0

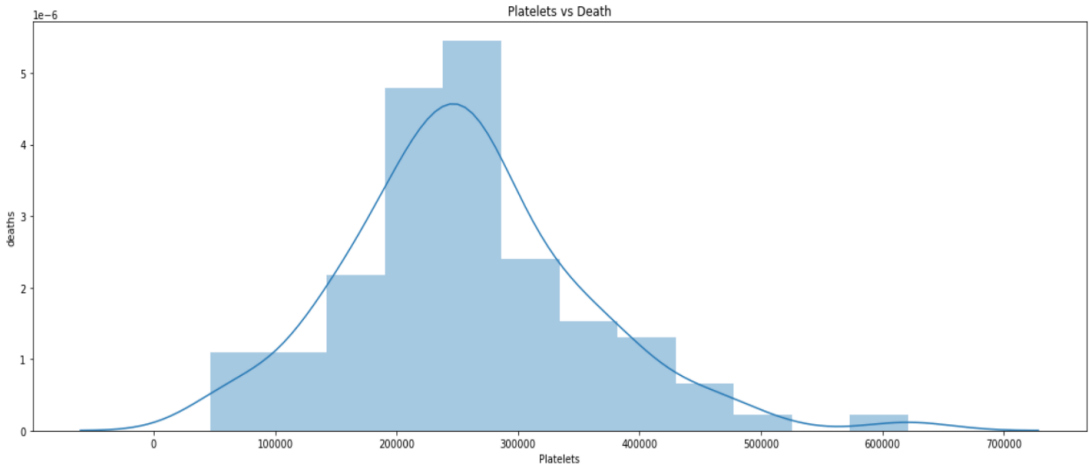
2. **Plots** - Visualized the death event vs various predictor variables using plots to understand the factors causing heart failure.



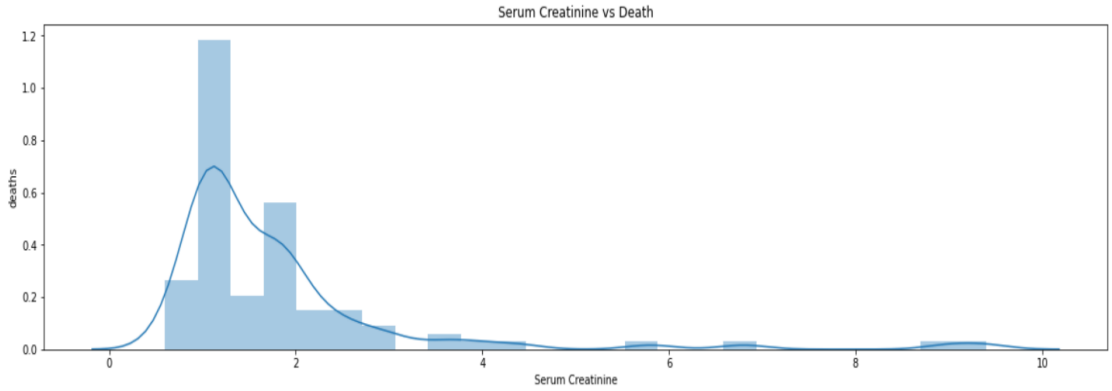
1. **Plot - Age vs Death**



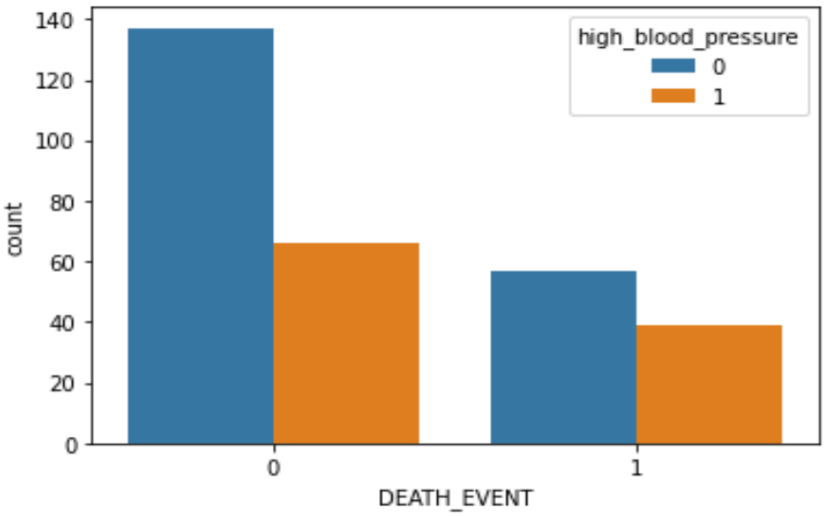
2. **Plot - Creatinine_Phosphokinase vs Death**



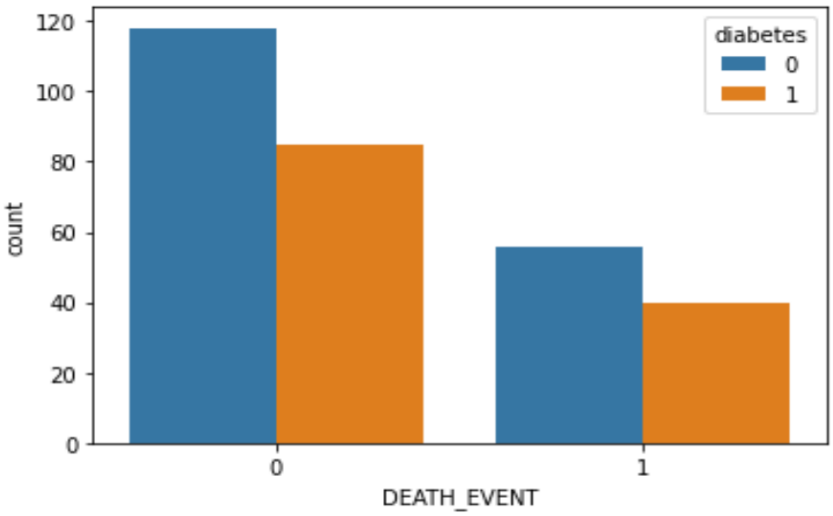
3. **Plot – Platelets vs Death**



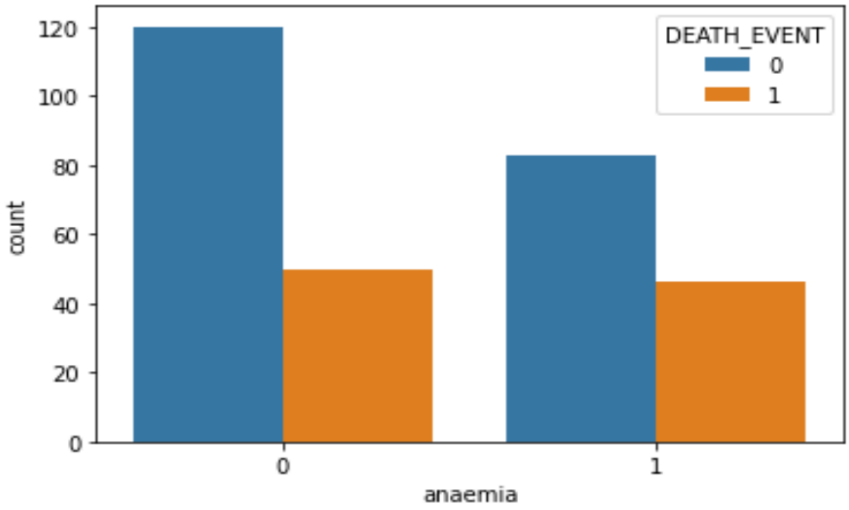
4. **Plot – Serum Creatinine vs Death**



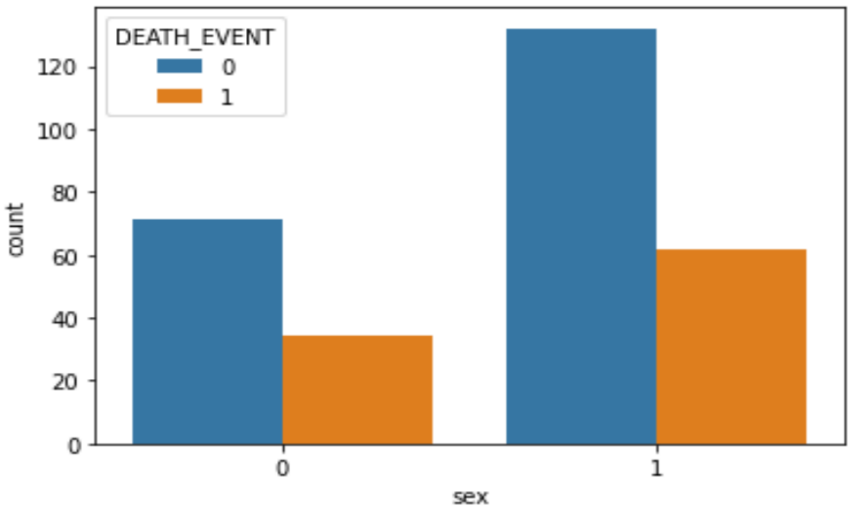
5. **Plot – High Blood Pressure vs Death**



6. **Plot – Diabetes vs Death**



7. **Plot – Anemia vs Death**



8. **Plot – Gender vs Death**

After analyzing these plots we consider age, creatinine phosphokinase, anaemia, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking as the important variables for prediction while the variable 'time' is not considered as it is not prominent.

Data Modelling:

i. Classification

Several classifications techniques were employed to find the best model that predicts death event.

Model 1 - Decision Tree classifier

Variables	Accuracy	Precision	Recall	F1 Score
All	0.81	0.72	0.78	0.75
Age & Gender	0.65	0.53	0.30	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

Model 2 - K Nearest Neighbors classifier

Variables	Accuracy	Precision	Recall	F1 Score
All	0.69	0.67	0.30	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	0.50	0.44	0.47

Model 3 - Naïve Bayes classifier

Variables	Accuracy	Precision	Recall	F1 Score
All	0.67	0.57	0.30	0.39
Age & Gender	0.65	0.57	0.15	0.24
Diabetes, High BP, Platelets, Anaemia & Smoking	0.64	0.00	0.00	0.00
creatinine_phosphokinase, serum_creatinine,serum_sodium and ejection_fraction	0.64	0.50	0.19	0.27
serum_sodium and ejection_fraction	0.67	0.67	0.15	0.24

Model 4 - Support Vector Machine classifier

Variables	Accuracy	Precision	Recall	F1 Score
All	0.81	0.84	0.59	0.70
Age & Gender	0.65	0.57	0.15	0.24
Diabetes, High BP, Platelets, Anaemia & Smoking	0.64	0.00	0.00	0.00
creatinine_phosphokinase, serum_creatinine,serum_sodium and ejection_fraction	0.73	0.68	0.48	0.57
serum_sodium and ejection_fraction	0.75	0.70	0.52	0.60

Best performing classification model

Upon trial of many different combinations we understand that few groups of variables do not contribute to the best performance of a model. So, we chose all predictor variables for the classification as they give the best F1 score and performance.

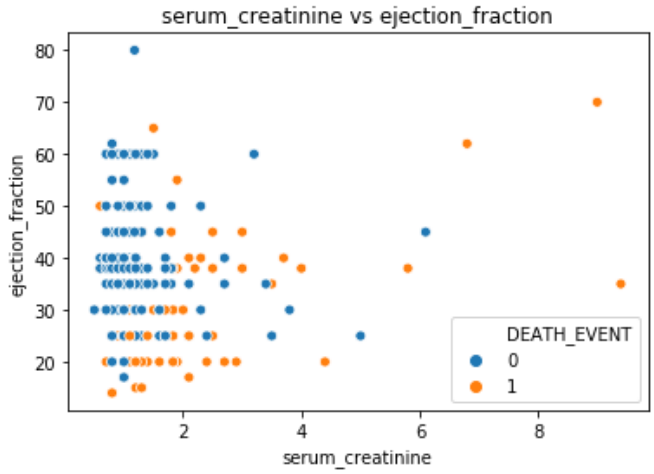
We can conclude that 'Decision Tree' classifier gives the best F1 Score when trained with all predictor variables.

The performance of the model of this classifier is listed below

- Accuracy: 0.81
- Precision: 0.72
- Recall: 0.78
- F1 Score: 0.75

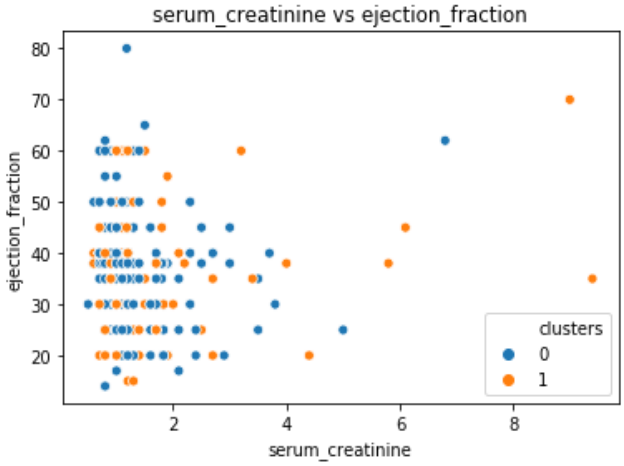
ii. **<u>Clustering</u>**

Gold Label scatter plot based on variable1 & variable2:



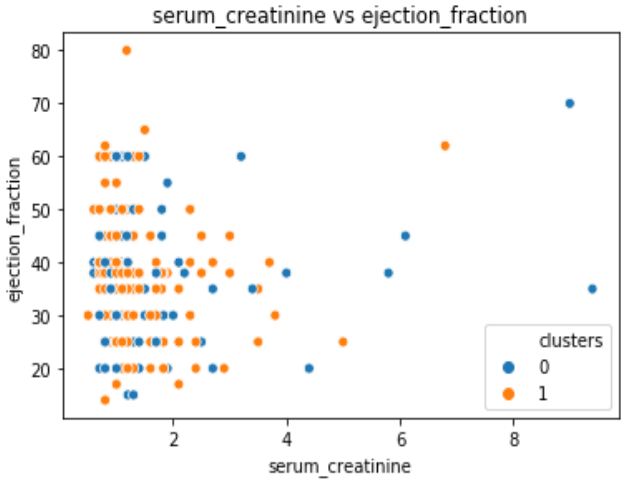
Model 1 - Hierarchical Clustering – Single Linkage

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	1.0



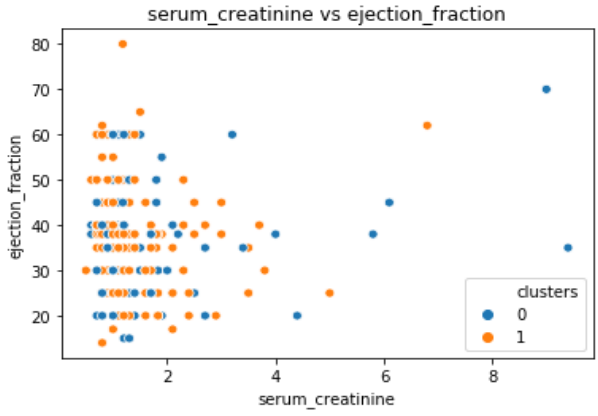
Model 2 - Hierarchical Clustering – Complete Linkage

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	1.0



Model 3 - KMeans 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	0.240
Anaemia	0.007	1.0



Best performing clustering model

When we explore data to try and explain what the causes are of death we find that health issues such as Anaemia, Diabetes and High Blood Pressure are major causes of heart attack. In addition, we find that the chances of heart attack is higher in men.

We can conclude that 'Hierarchical Clustering' clustering algorithm gives the best Silhouette Coefficient when trained on the variable 'anaemia'

The performance of the model of this classifier is listed below

- Silhouette Coefficient: 1.0
- Rand Index: 0.007

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.