------Explorarory Data Analysis of Heart Failure Clinical Records Data Set------

Introduction

Cardiovascular diseases kill millions of people globally every year, and they mainly exhibit as myocardial infarctions and heart failures.When the heart cannot pump enough blood to meet the needs of the body , chances of Heart failure are extremely high.We can use the available medical data of patients quantify symptoms, body features, and clinical laboratory test values to perform statistical analysis aimed at highlighting patterns and correlations which could be undetectable by medical doctors. Machine learning can predict patients' survival from their data and can characterize the most important features among those included in their medical records. The heart_failure_clinical_records data set used in this project was sourced from the UC Irvine Machine Learning Repository published in 2020.This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period. Each patient profile has 13 clinical features and our target feature is "death event".

Data

The data set used in this project can be found here.

The data set contains 12 features and the target is 'DEATH_EVENT'.Each row contains record of one patient about the parameters causing the death event.

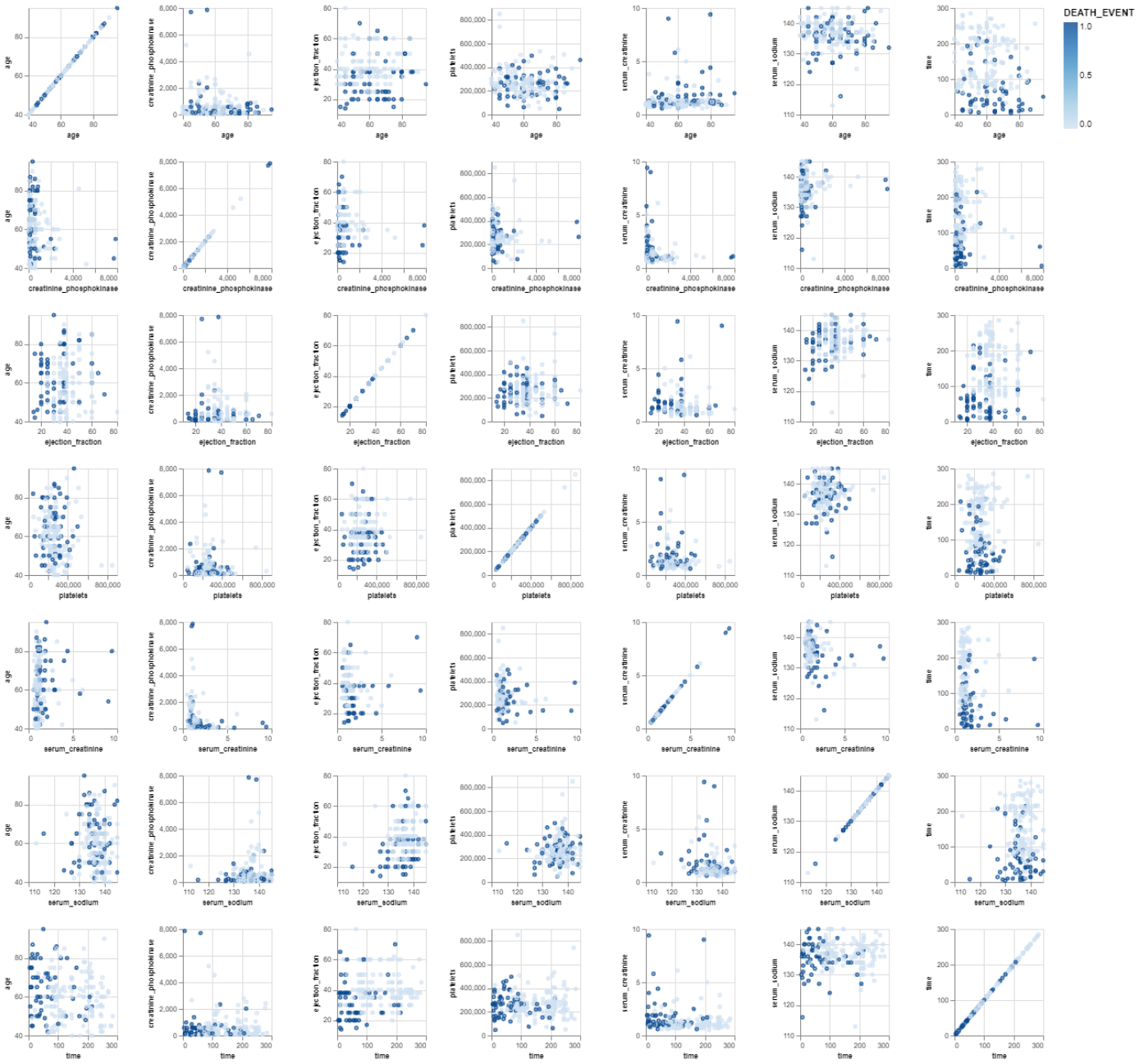
	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4	1

Among these variables mentioned above 'sex', 'diabetes', 'high_blood_pressure','anaemia' have binary values i.e if person has diabetes the value will be 1 and 0 otherwise. Rest of the features are numeric and have continuous values.

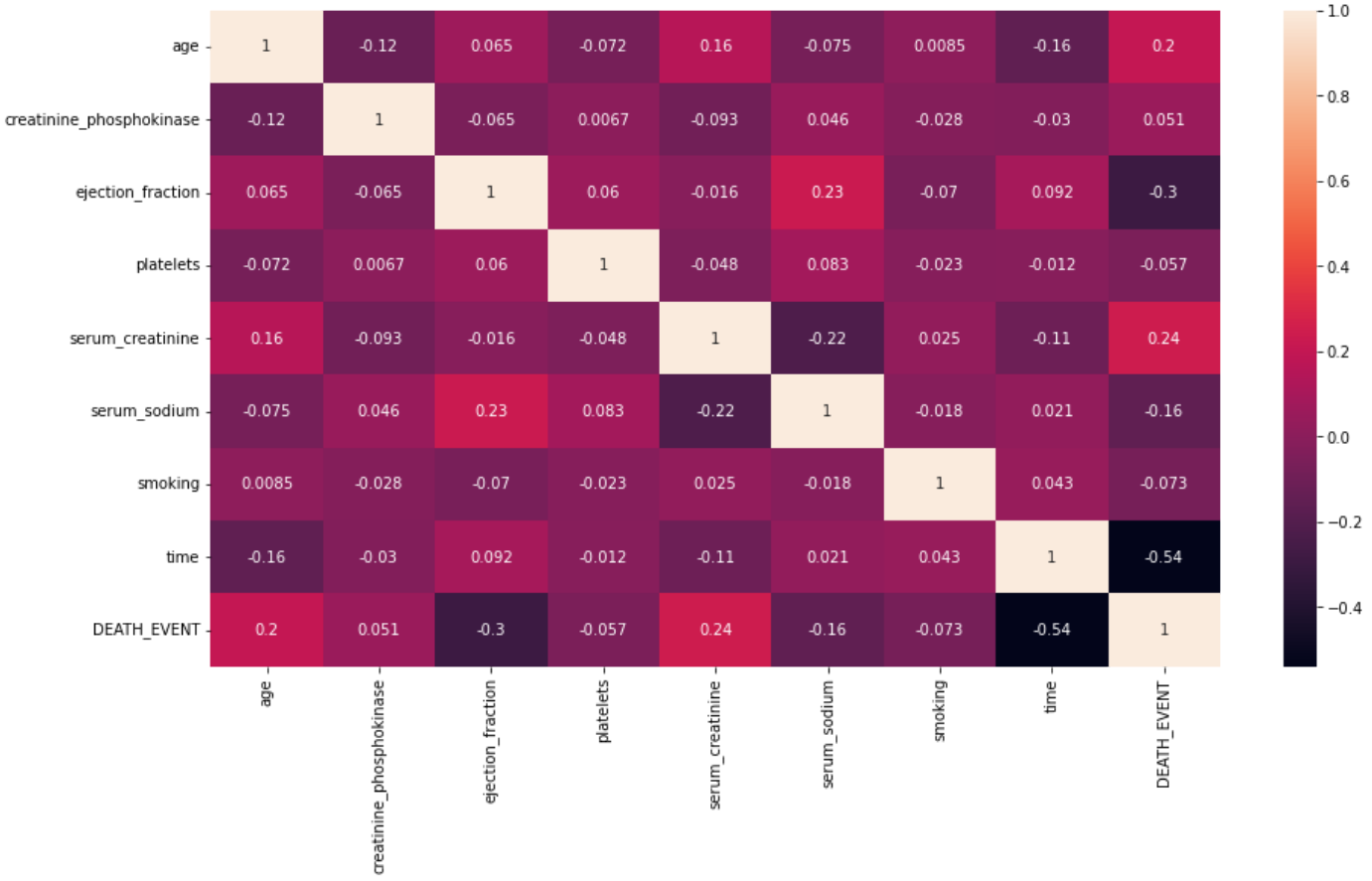
Features Exploration

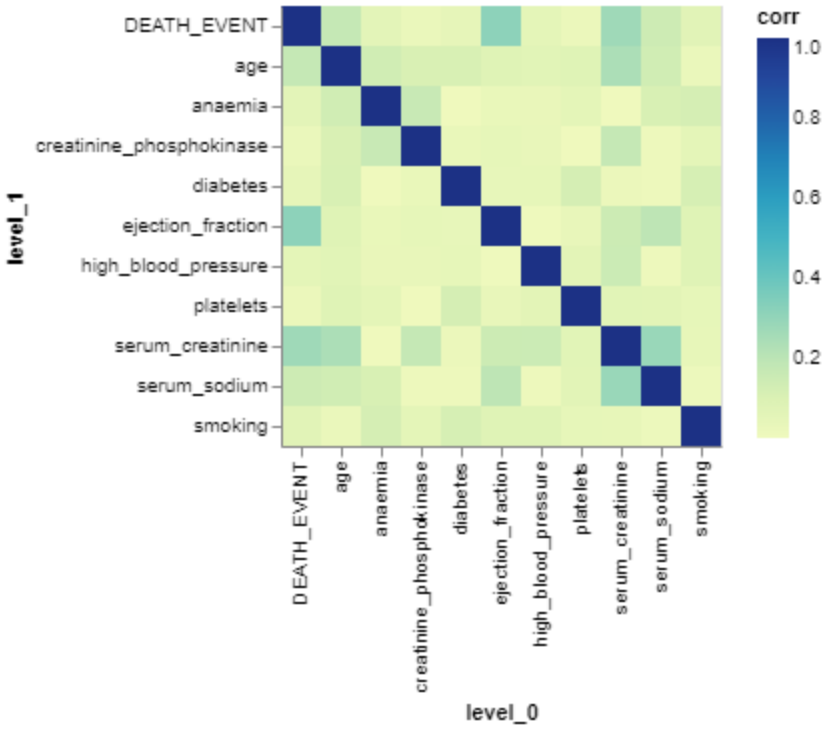
Our interest is in predicting the death of the patient by analyzing the clinical records of corresponding feature and trying to find which features affect the most in causing the death. Thus to start the analysis process we'll first compare all the features with eachother and to the death event.

Following plot gives pairwise comparison of all the features :

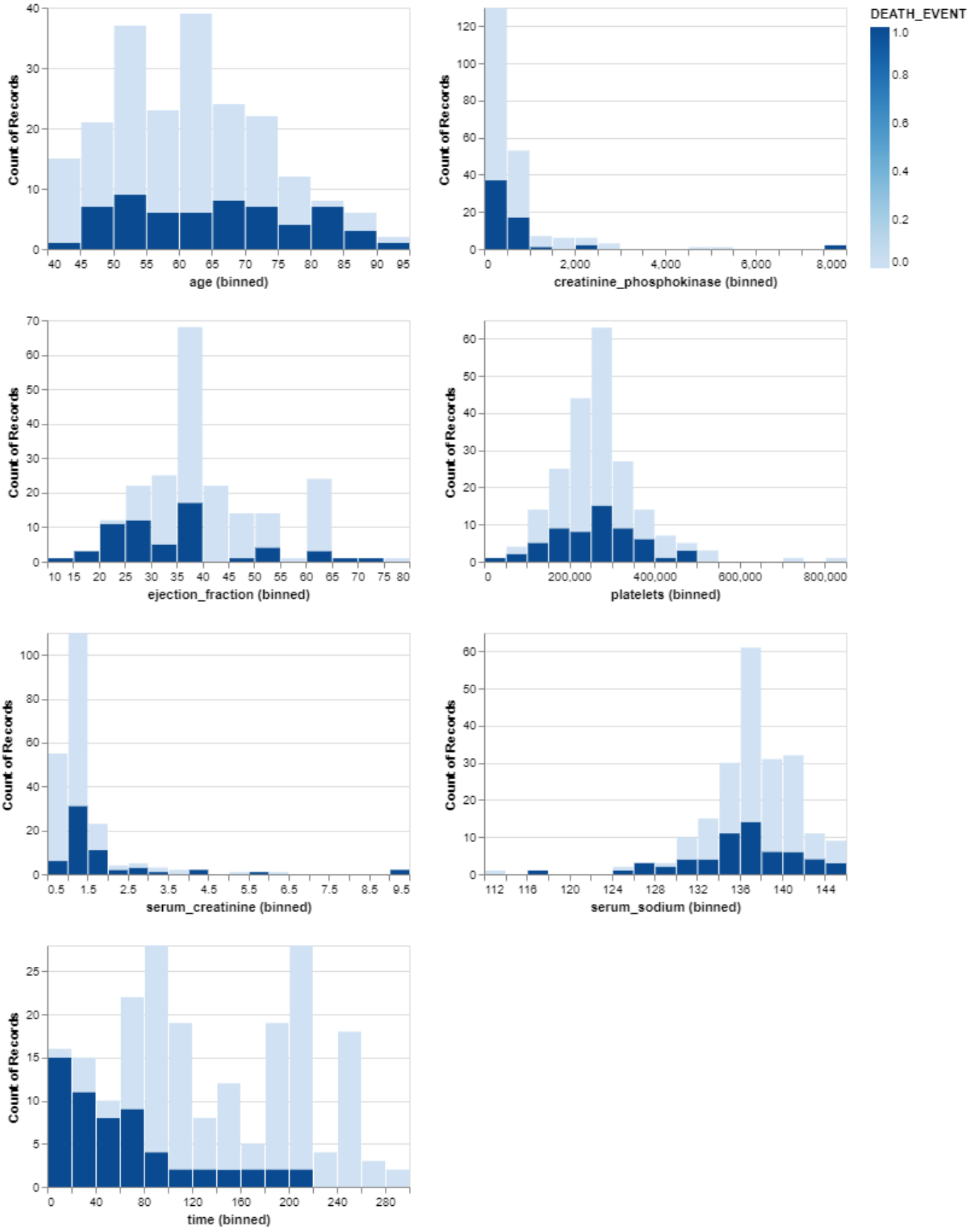


Correlation heatmap of the features :



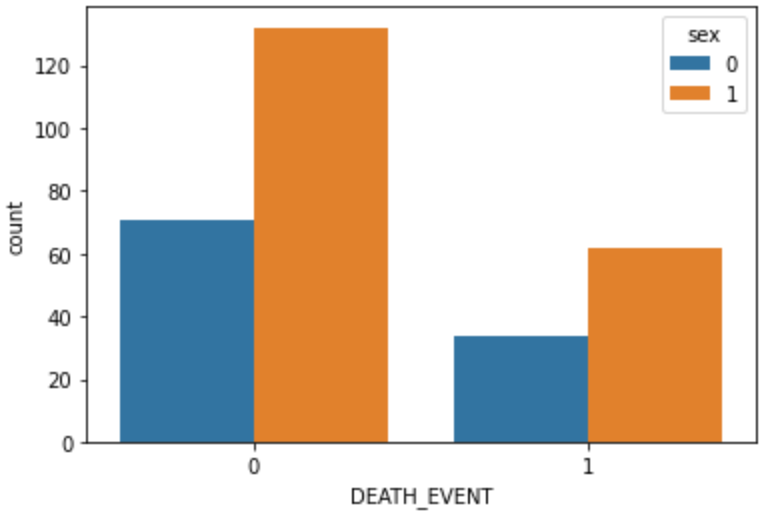


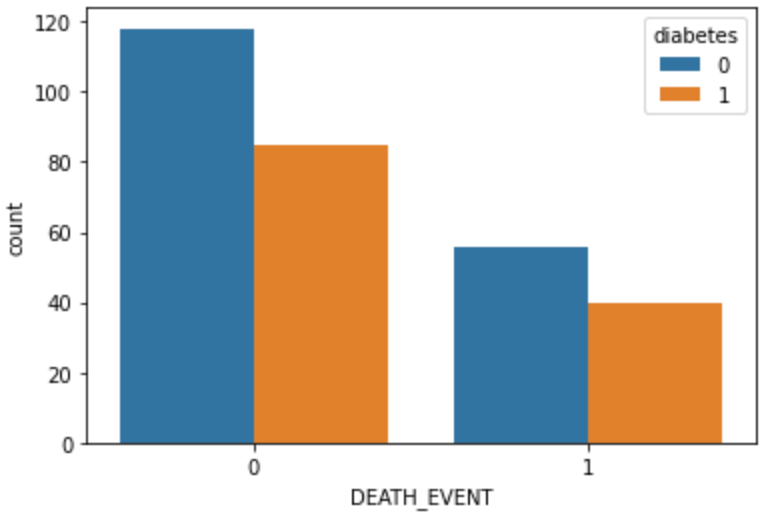
Comparison of each feature with the DEATH_EVENT :

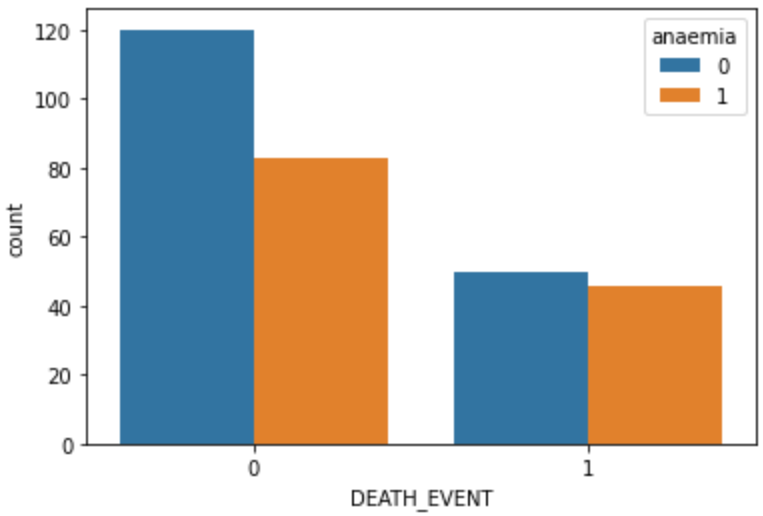


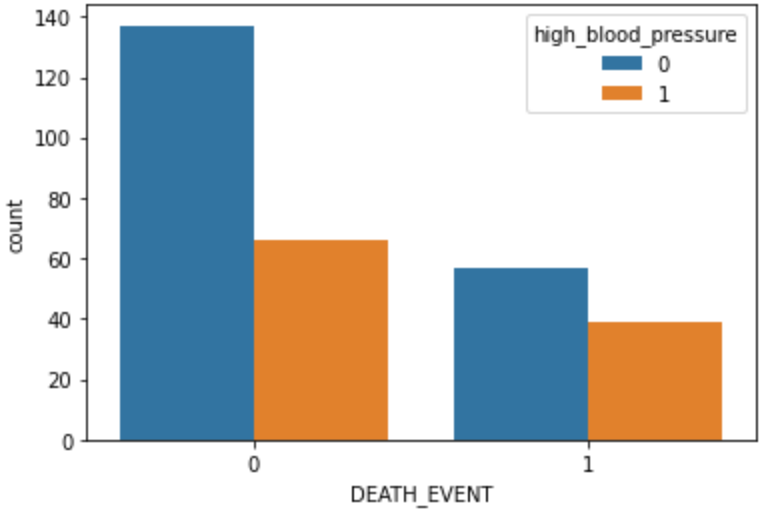
Binary Features:

Let's explore binary features one by one and try to find the association -

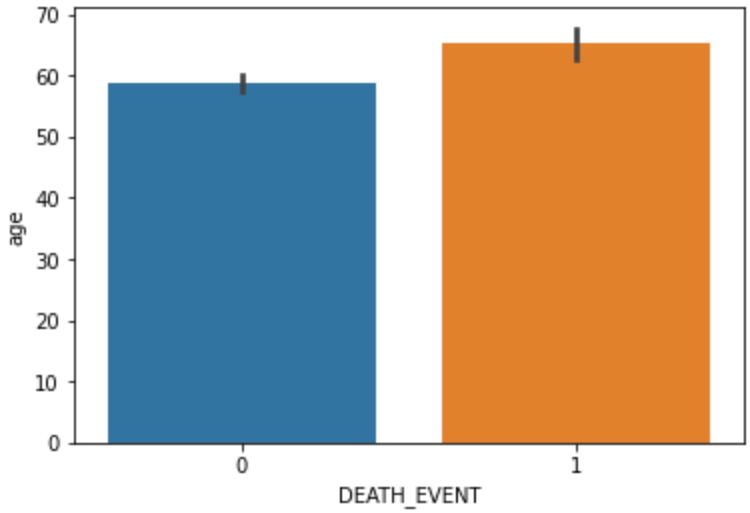

sex
0
1
count
120
100
80
60
40
20
0
0
1
DEATH_EVENT

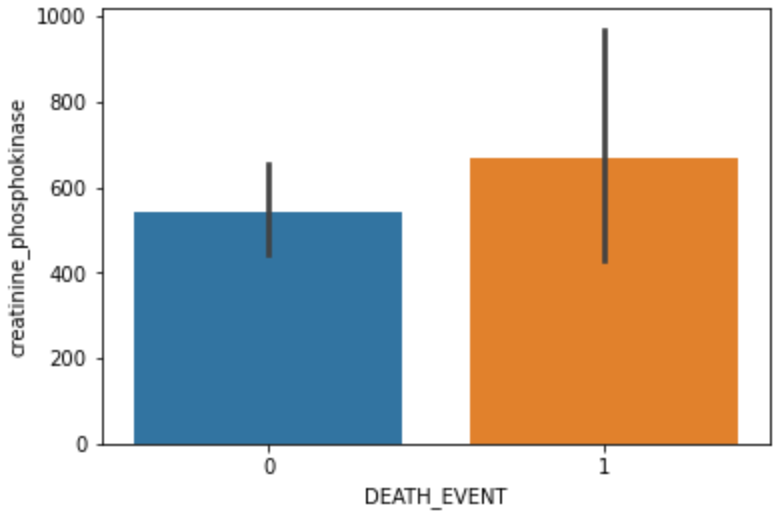

diabetes
0
1
count
120
100
80
60
40
20
0
0
1
DEATH_EVENT

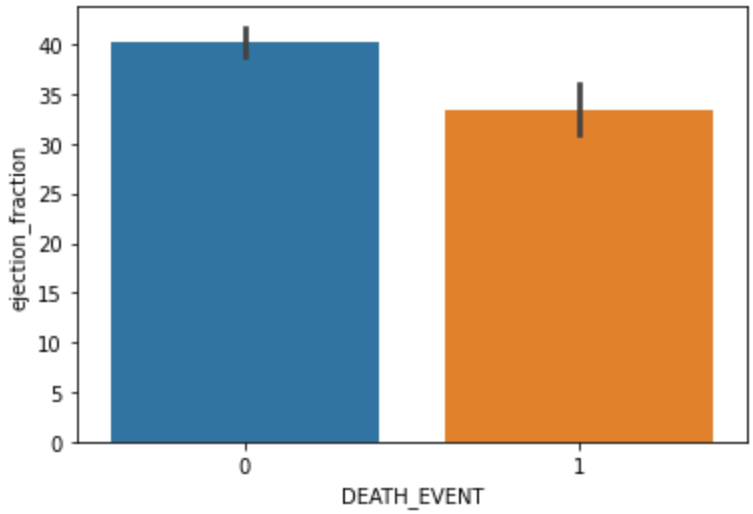

anaemia
0
1
count
120
100
80
60
40
20
0
0
1
DEATH_EVENT

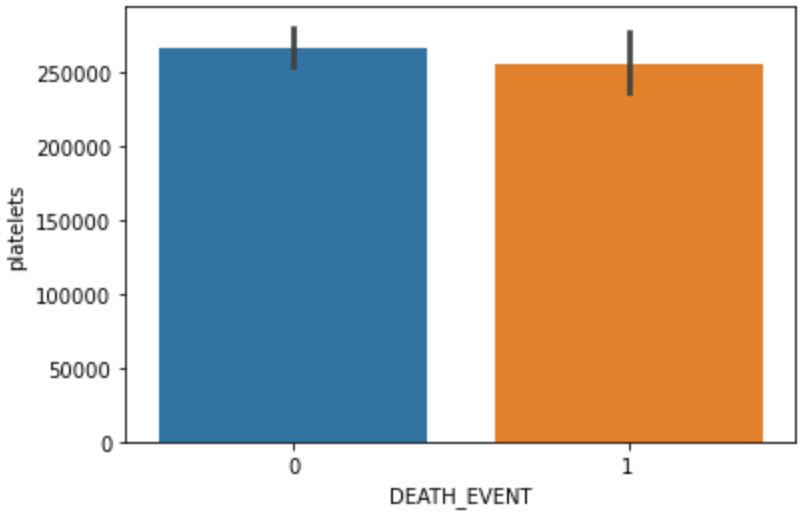


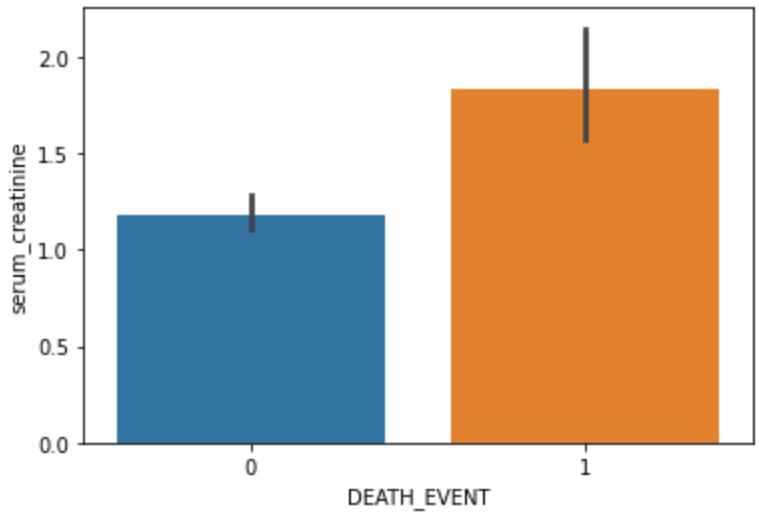
Numeric features:

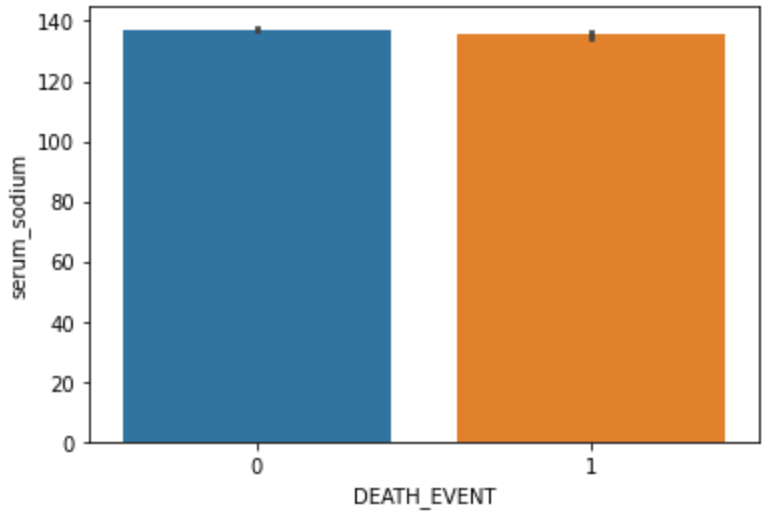


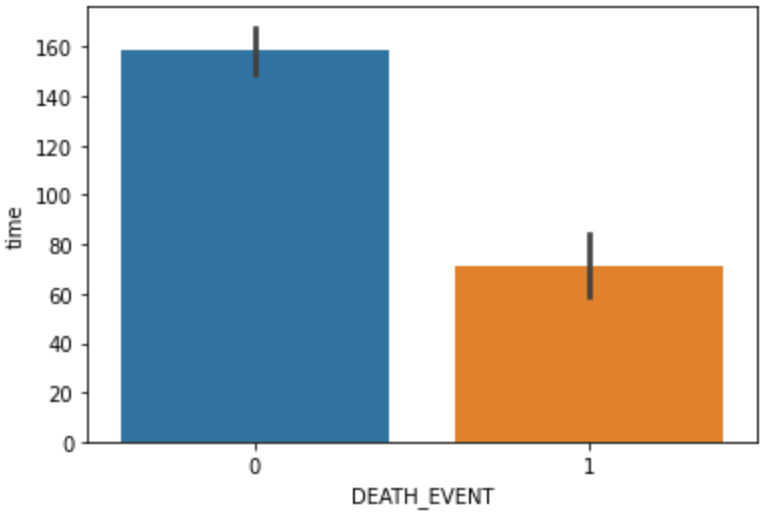



ejection_fraction
40
35
30
25
20
15
10
5
0
0
1
DEATH_EVENT


platelets
250000
200000
150000
100000
50000
0
0
1
DEATH_EVENT


serum_creatinine
2.0
1.5
1.0
0.5
0.0
0
1
DEATH_EVENT





Interpretation:

1. The mean age of people who experieced death by heart failure is higher than those who survived.
2. Average amount of creatinine_phosphokinase in heart_failure_deaths is greater.
3. The Ejection Fraction is higher in survivors.
4. The platelet counts are observed to be slightly lower in heart_failure patients.
5. Those with higher level of creatin are more likely to face heart failure.
6. Level of sodium is almost the same in survivors and dead patients.
7. People with more followup period are more likely to survive.

Model Building :

We used two classification models :

- KNeighborsClassifier
- RandomForestClassifier

- The scoring metrics used to cross validate are :

```
'accuracy'
'precision'
'recall'
 'f1'
```

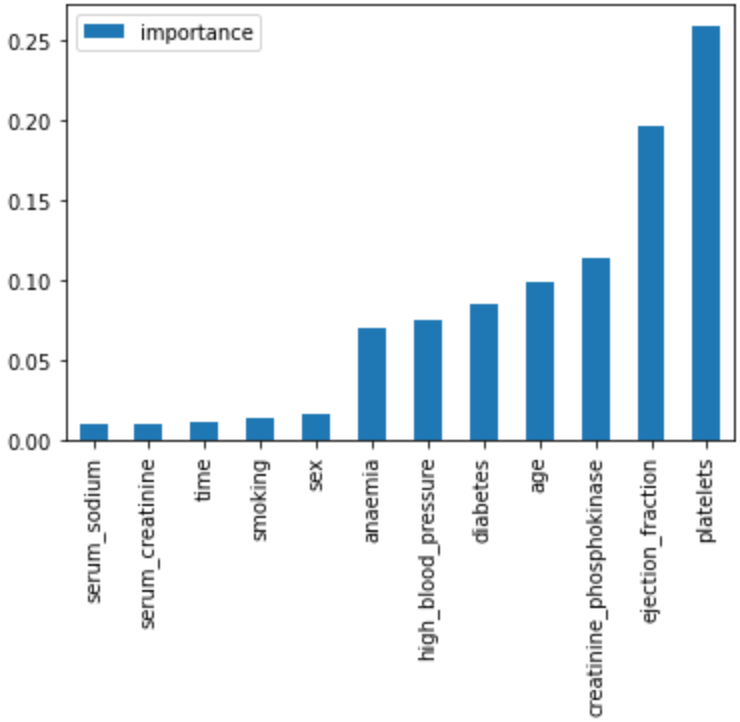
- Preprocessing:
  - Data was split based on 80% for training and 20% for testing.

    - Since the data is clinical records of the patients, all the features need to be given equal importance hence no features were dropped in the preprocessing.
- Hyper-parameters tuning.
    - The hyperparameter optimization was performed for the following three parameters of RandomForestClassifier :
        1. max_depth
        2. class_weight
        3. n_estimators

Detailed steps of building the model for this dataset can be found here

Feature Importance :

With the trained model we got the following plot for the importance of each feature in the dataset



Limitations :

- The dataset we are using is smaller: a larger dataset would give more reliable results.
- Additional information about the physical features of the patients lik height, weight, body mass index, etc. would have been useful to detect additional risk factors for cardiovascular health diseases.