Heart Failure Prediction

About this dataset

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

There are some factors that affects Death Event. This dataset contains person's information like

- age
- sex
- blood pressure
- smoke
- diabetes
- ejection fraction
- creatinine phosphokinase
- serum_creatinine
- serum_sodium
- time **we have to predict their DEATH EVENT**.

The database was collected from: Source : https://www.kaggle.com

Attributes and explanation :

Age is the age of the patient

anaemia Decrease of red blood cells or hemoglobin (0 = No, 1 = Yes)

creatinine_phosphokinase Level of the CPK enzyme in the blood (mcg/L)

diabetes If the patient has diabetes (0 = No, 1 = Yes)

ejection_fraction Percentage of blood leaving the heart at each contraction (percentage)

high_blood_pressure If the patient has hypertension (0 = No, 1 = Yes)

platelets Platelets in the blood (kiloplatelets/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 (Male = 1, Female =0)

smoking If the patient smokes or not (0 = No, 1 = Yes)

time Follow-up period (days)

DEATH_EVENT If the patient deceased during the follow-up period (0 = No, 1 = Yes)

Plan:

Importing Libraries

In [79]:

```
import os
import numpy as np
import pandas as pd
from sklearn import datasets , linear_model
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

Exploring the data

In [80]:

```
filepath = "heart_failure_clinical_records_dataset.csv"
data = pd.read_csv(filepath)
data.head(10)
```

Out[80]:

	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
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7	1
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7	1
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8	1
5	90.0	1	47	0	40	1	204000.00	2.1	132	1	1	8	1
6	75.0	1	246	0	15	0	127000.00	1.2	137	1	0	10	1
7	60.0	1	315	1	60	0	454000.00	1.1	131	1	1	10	1
8	65.0	0	157	0	65	0	263358.03	1.5	138	0	0	10	1
9	80.0	1	123	0	35	1	388000.00	9.4	133	1	1	10	1

Ceaning the data

This dataset is already processed and provided with cleaned data but lets also remove some data like time and .ejection fraction. Thus we will use this data and lets get started with Exploratory Data Analysis(EDA)

In [91]:

```
del data["time"]
del data['ejection_fraction']
data.head(5)
```

Out[91]:

	age	anaemia	creatinine_phosphokinase	diabetes	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	DEATH_EVENT
0	75.0	0	582	0	1	265000.00	1.9	130	1	0	Not Alive
1	55.0	0	7861	0	0	263358.03	1.1	136	1	0	Not Alive
2	65.0	0	146	0	0	162000.00	1.3	129	1	1	Not Alive
3	50.0	1	111	0	0	210000.00	1.9	137	1	0	Not Alive
4	65.0	1	160	1	0	327000.00	2.7	116	0	0	Not Alive

Exploratory Data Analysis(EDA)

In [92]:

```
data.describe()
```

Out[92]:

	age	anaemia	creatinine_phosphokinase	diabetes	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.00000	299.000000	299.000000	299.00000
mean	60.833893	0.431438	581.839465	0.418060	0.351171	263358.029264	1.39388	136.625418	0.648829	0.32107
std	11.894809	0.496107	970.287881	0.494067	0.478136	97804.236869	1.03451	4.412477	0.478136	0.46767
min	40.000000	0.000000	23.000000	0.000000	0.000000	25100.000000	0.50000	113.000000	0.000000	0.00000
25%	51.000000	0.000000	116.500000	0.000000	0.000000	212500.000000	0.90000	134.000000	0.000000	0.00000
50%	60.000000	0.000000	250.000000	0.000000	0.000000	262000.000000	1.10000	137.000000	1.000000	0.00000
75%	70.000000	1.000000	582.000000	1.000000	1.000000	303500.000000	1.40000	140.000000	1.000000	1.00000
max	95.000000	1.000000	7861.000000	1.000000	1.000000	850000.000000	9.40000	148.000000	1.000000	1.00000

lets go through the data frame of the data one

In [93]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64
 2   creatinine_phosphokinase  299 non-null    int64
 3   diabetes                  299 non-null    int64
 4   high_blood_pressure       299 non-null    int64
 5   platelets                 299 non-null    float64
 6   serum_creatinine          299 non-null    float64
 7   serum_sodium              299 non-null    int64
 8   sex                       299 non-null    int64
 9   smoking                   299 non-null    int64
 10  DEATH_EVENT               299 non-null    object
dtypes: float64(3), int64(7), object(1)
memory usage: 25.8+ KB
```

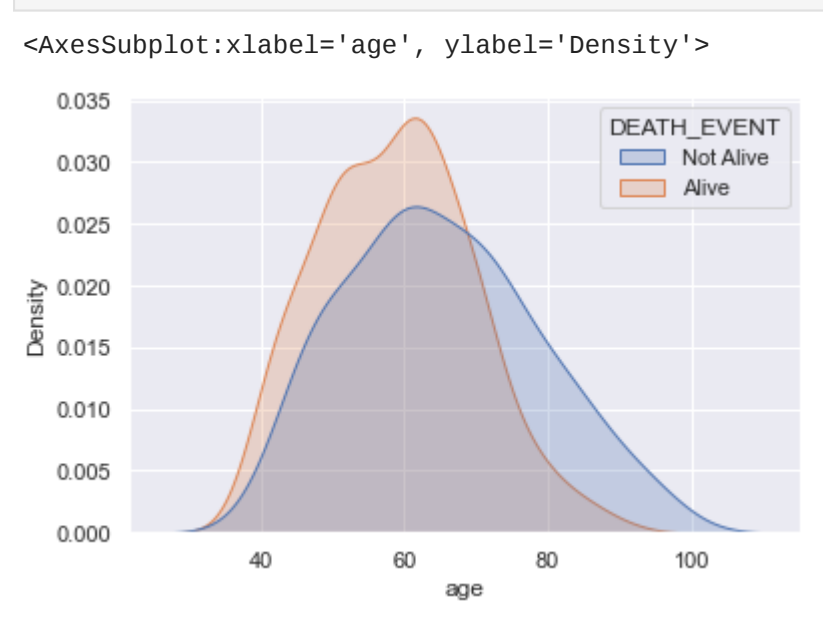
Relation between Age and death event

The relation between age and demise possibility can be now viewed.

In [94]:

```
sns.kdeplot(data = data, x = "age", hue = "DEATH_EVENT",
            common_norm = False, shade = True)
```

Out[94]: <AxesSubplot:xlabel='age', ylabel='Density'>



So here the density map shows that the death rate was high in the middle aged(60 years around) patients with their data but the possibility of non death event was high than the death event. But around the age of 70 to 100 years we can see the curve varied. Here the death event is more as we can visualise. In general we can say that the **death is more possible with same symptoms after 60 - 70 than the age below 60.**

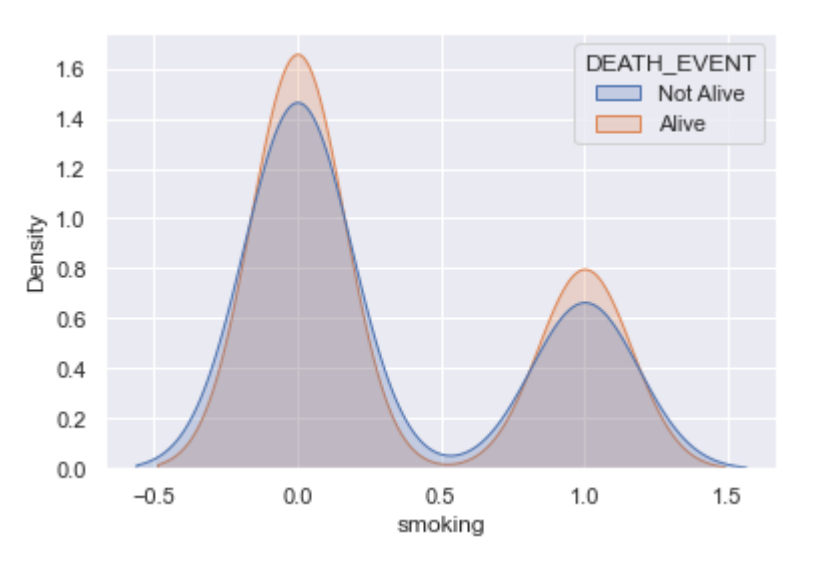
Relation between smoking and death event

The Relation between smoking and death event can be now viewed.

In [95]:

```
sns.kdeplot(data = data, x = 'smoking' , hue = "DEATH_EVENT",
            common_norm = False, shade = True)
```

Out[95]: <AxesSubplot:xlabel='smoking', ylabel='Density'>



In the above visuals we can see that the curve are very well inclined. The curve is high at two places because the values provided were discreate and binary(0 and 1) with either does'nt smoke or smokes. So, we can see that non death event is high in both the peaks. Also the peak at smokes condition is lower than the non smokers. This indicates the distribution is similar in both **non-smokers and the smokers are not directly contributing in the predictions.**

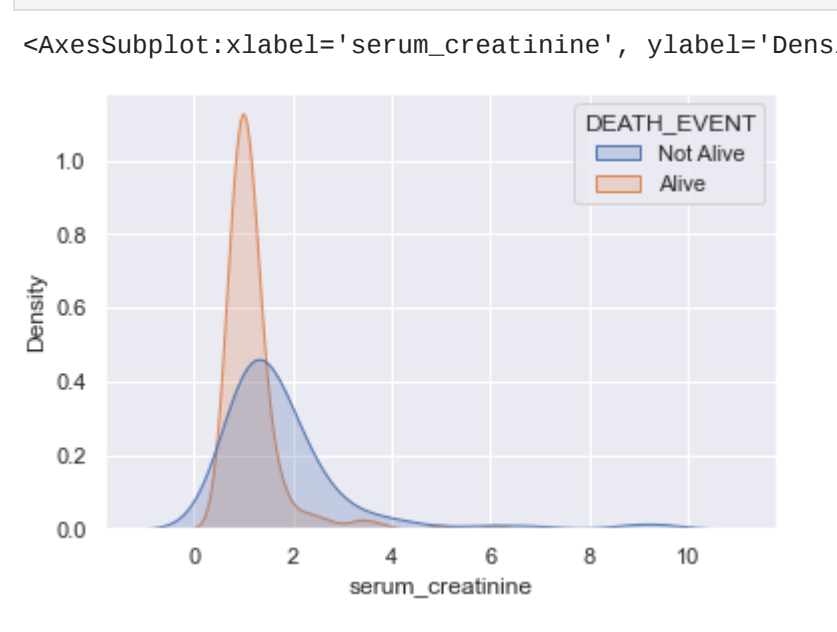
Relation between serum_creatinine and death event

The Relation between serum_creatinine and death event can be now viewed.

In [96]:

```
sns.kdeplot(data= data, x='serum_creatinine',hue ="DEATH_EVENT",
            common_norm = False, shade = True)
```

Out[96]: <AxesSubplot:xlabel='serum_creatinine', ylabel='Density'>



This shows a huge difference in death possibility with low serum creatinine but with very narrow edge. Thus if serum creatinine is 0 to 2 the possiblity of non-death event is more than the death event. Very specifically, the red curve(1) is shifted to right which indicate that the the possibility of death is more as it reaches to creainine level 2. **At high serum creatinine level the death event is high.**

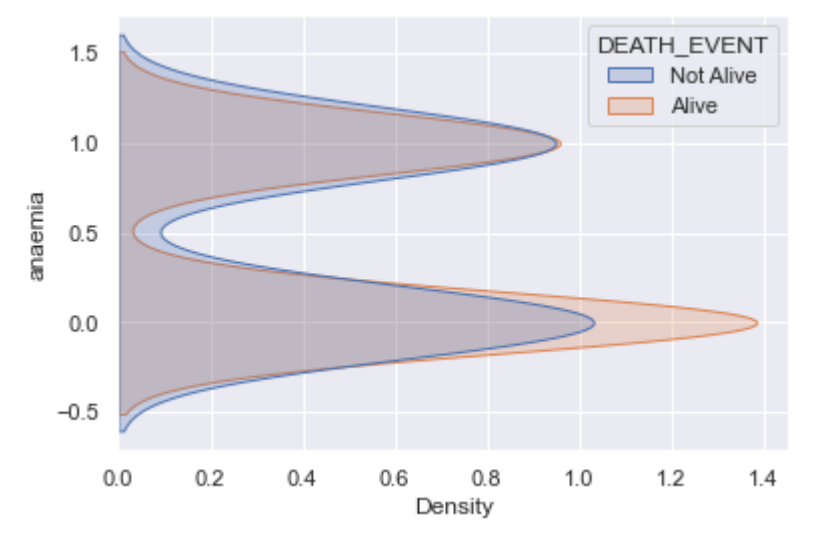
Relation between anaemia and death event

The Relation between anaemia and death event can be now viewed.

In [101…

```
sns.kdeplot(data= data, y='anaemia',hue ="DEATH_EVENT",
            common_norm = False, shade = True)
```

Out[101… <AxesSubplot:xlabel='Density', ylabel='anaemia'>



This graph shows that differance of death possibility with low anaemia content. thus from the graph we can see that as the amount of anaemia in body incrarses the possiablity of increase in mortality rates are high, so **high anaemia level the death event is high**

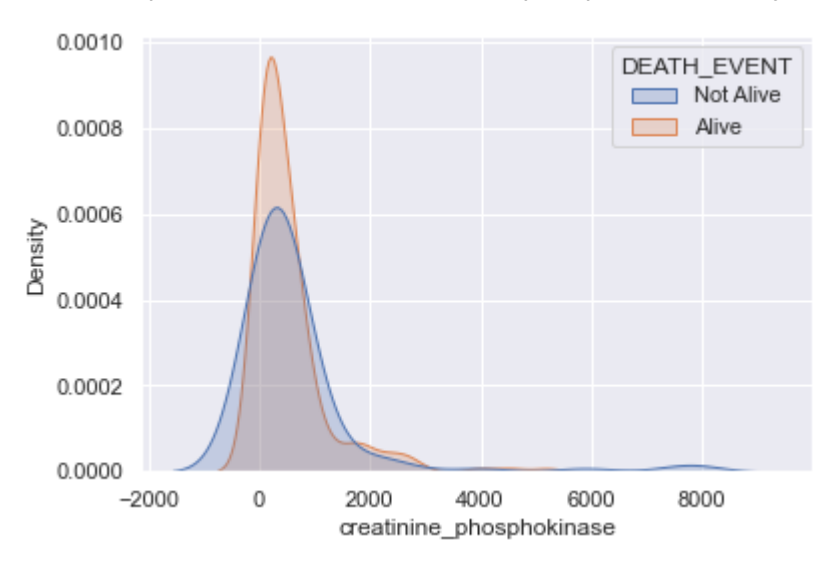
Relation between creatinine_phosphokinase and death event

The Relation between serum_phosphokinase and death event can be now viewed.

In [102…

```
sns.kdeplot(data =data, x = "creatinine_phosphokinase" , hue = "DEATH_EVENT",
            common_norm = False , shade = True )
```

Out[102… <AxesSubplot:xlabel='creatinine_phosphokinase', ylabel='Density'>



Here we can see that at low level(nearby zero) the non death event is more than the death event. The case becomes more with death events as the levels reaches to high. so **high cretinine phosphokinase level the death event is high.**

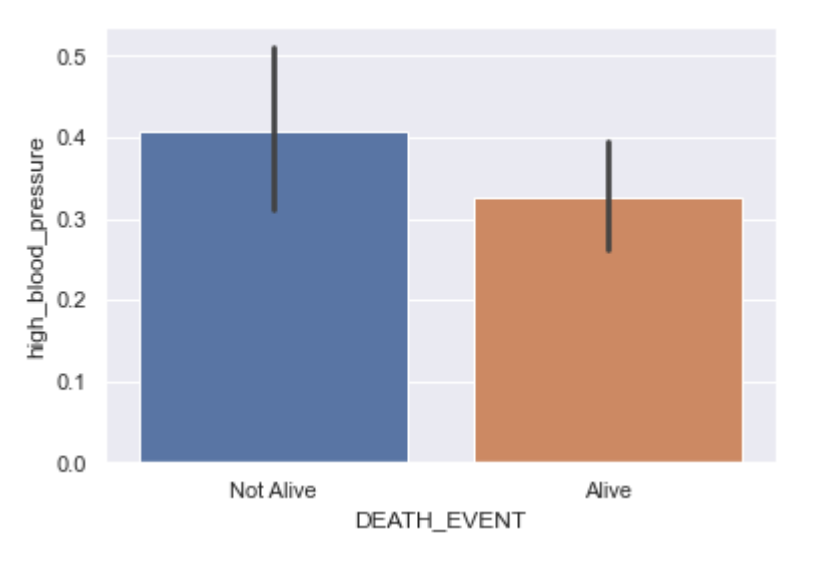
Relation between Blood Pressure and death event

The Relation between blood pressure and death event can be now viewed.

In [103…

```
sns.barplot(data=data, x= "DEATH_EVENT", y = "high_blood_pressure")
```

Out[103… <AxesSubplot:xlabel='DEATH_EVENT', ylabel='high_blood_pressure'>



This shows the **death possibility is more when the blood pressure of the patients are high.**

Lets see an pairplot:

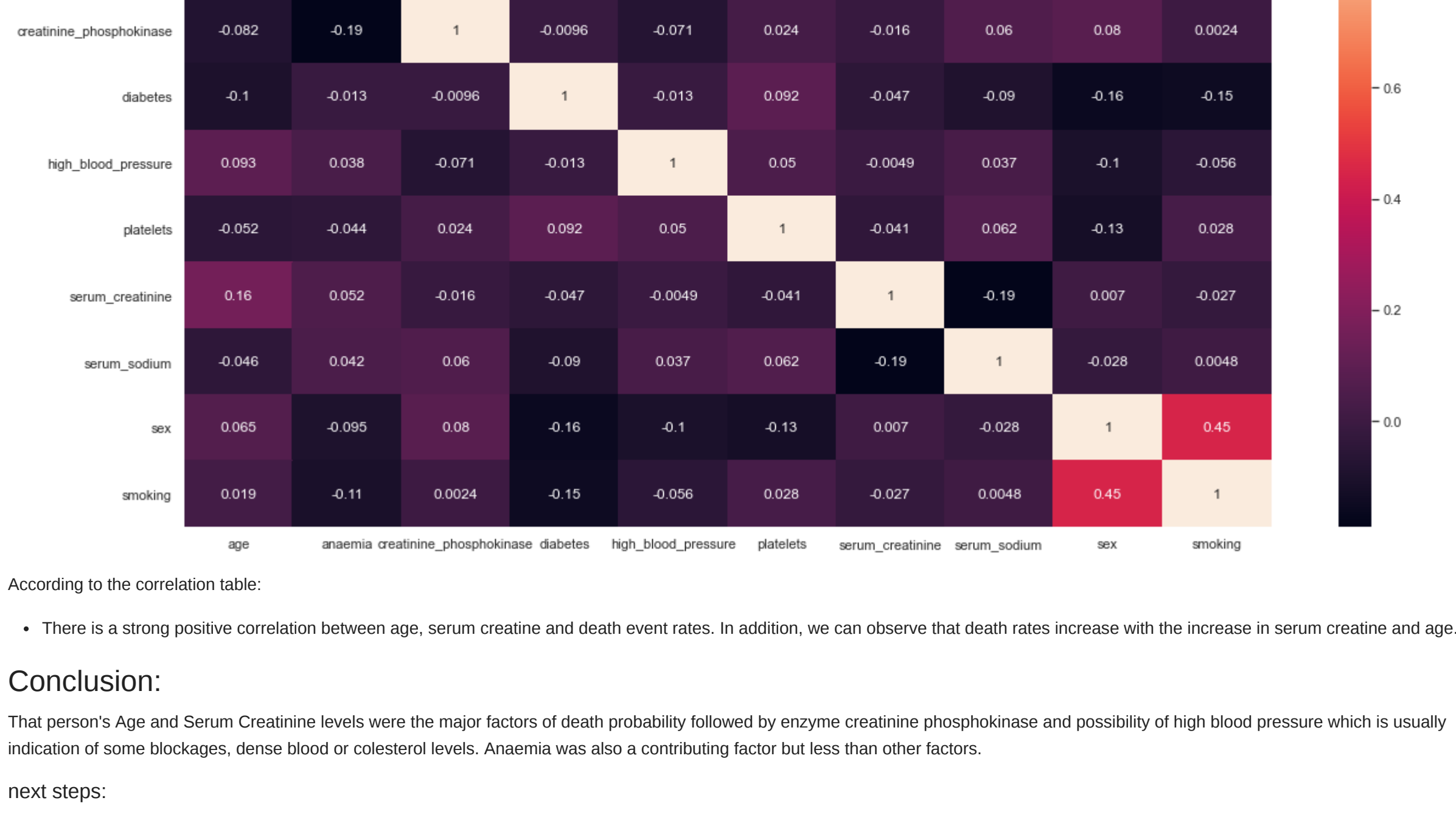
In [104…

```
data = data.copy()
data.loc[data.DEATH_EVENT == 0, 'DEATH_EVENT'] = "Alive"
data.loc[data.DEATH_EVENT == 1, 'DEATH_EVENT'] = "Not Alive"
sns.pairplot(data=data[['age','creatinine_phosphokinase','anaemia','serum_creatinine','serum_sodium','DEATH_EVENT']],
             hue='DEATH_EVENT');
```

In [105…

```
plt.figure(figsize=(20,10))
sns.heatmap(data.corr(), annot = True)
plt.title('Correlation Matrix ')
```

Out[105… Text(0.5, 1.0, 'Correlation Matrix ')



According to the correlation table:

- There is a strong positive correlation between age, serum creatine and death event rates. In addition, we can observe that death rates increase with the increase in serum creatine and age.

Conclusion:

That person's Age and Serum Creatinine levels were the major factors of death probability followed by enzyme creatinine phosphokinase and possibility of high blood pressure which is usually indication of some blockages, dense blood or colesterol levels. Anaemia was also a contributing factor but less than other factors.

next steps:

- we can predict the death rate of patients based on many other factors like time and platet count and more.
- thus improving the model to be more accurate

Done by:- **THANMAI SAI P**