)

Heart Failure Prediction

Cardiovascular diseases (CVDs) are ranked to have the highest death rate globally, which takes about 17.9 millions of lives annually and accounts for about 31% of all deaths worldwide.Heart failure is a common symptom of CVDs, which brings serious consequences, such as death.

This clinical dataset is from Kaggle (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data) and contains 12 features, 300 rows of data, which can be used to predict mortality by heart failure. Most CVDs can be prevented by addressing behaviral risk factors such as smoking, obesity, lack of physical, alcohol, etc. People with or with high CVDs risk need early detection, and machine learning models might be a good choice.

1. Heart Failure Overview

Let's first look at the raw dataframe from the original dataset.

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blo
0	75.0000	0	582	0	20	1
1	55.0000	0	7861	0	38	0
2	65.0000	0	146	0	20	0
3	50.0000	1	111	0	20	0
4	65.0000	1	160	1	20	0

To avoid any confusions, note that **time** suggests follow-up period (days) and **ejection_fraction** suggests percentage of blood leaving the heart at each contraction.

After clear inspection, this dataset does not have any missing values. The 12 features can be splitted into two categoreis: quantitative and categorical. I further changed the categorical variables as type: categorical.

- Quantitative: **age**, **creatinine_phosphokinase**, **ejaction_fraction**, **platelets**, **serum_creatinine**, **serum_sodium**, **time**
- Categorical: **anaemia**, **diabetes**, **high_blood_pressure**, **sex**, **smoking**, **DEATH_EVENT**

☐ Wanna check the statistics for quantitative variables?

	count	mean	std	min	

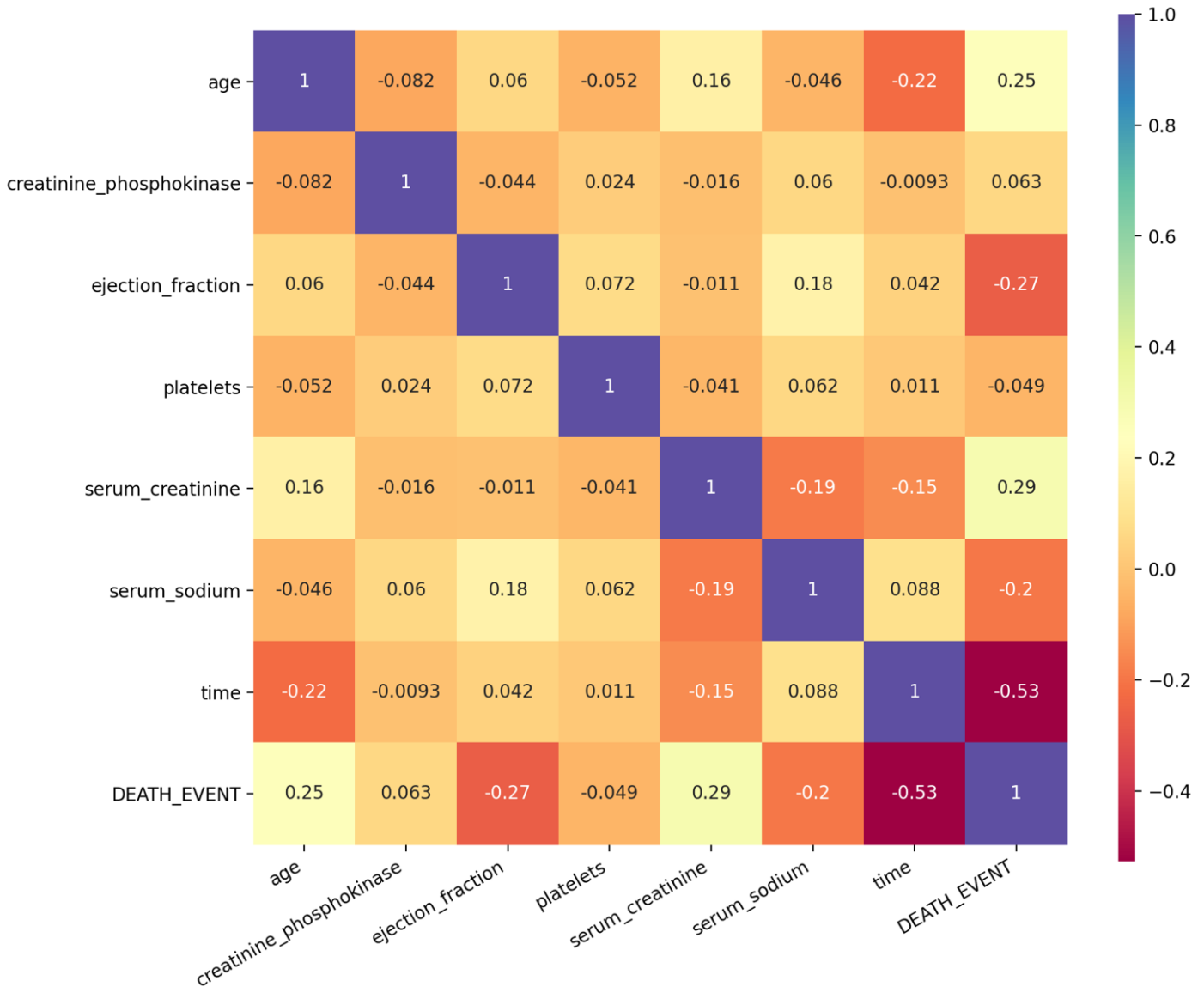
age	299.0000	60.8339	11.8948	40.0000	5
creatinine_phosphokinase	299.0000	581.8395	970.2879	23.0000	11
ejection_fraction	299.0000	38.0836	11.8348	14.0000	3
platelets	299.0000	263,358.0293	97,804.2369	25,100.0000	212,50
serum_creatinine	299.0000	1.3939	1.0345	0.5000	
serum_sodium	299.0000	136.6254	4.4125	113.0000	13
time	299.0000	130.2609	77.6142	4.0000	7

2. Corellation Visualization

You can choose quantitative or categorical values to inspect the correclation matrix. Since we want to predict mortality, I included DEATH_EVENT in both cases.

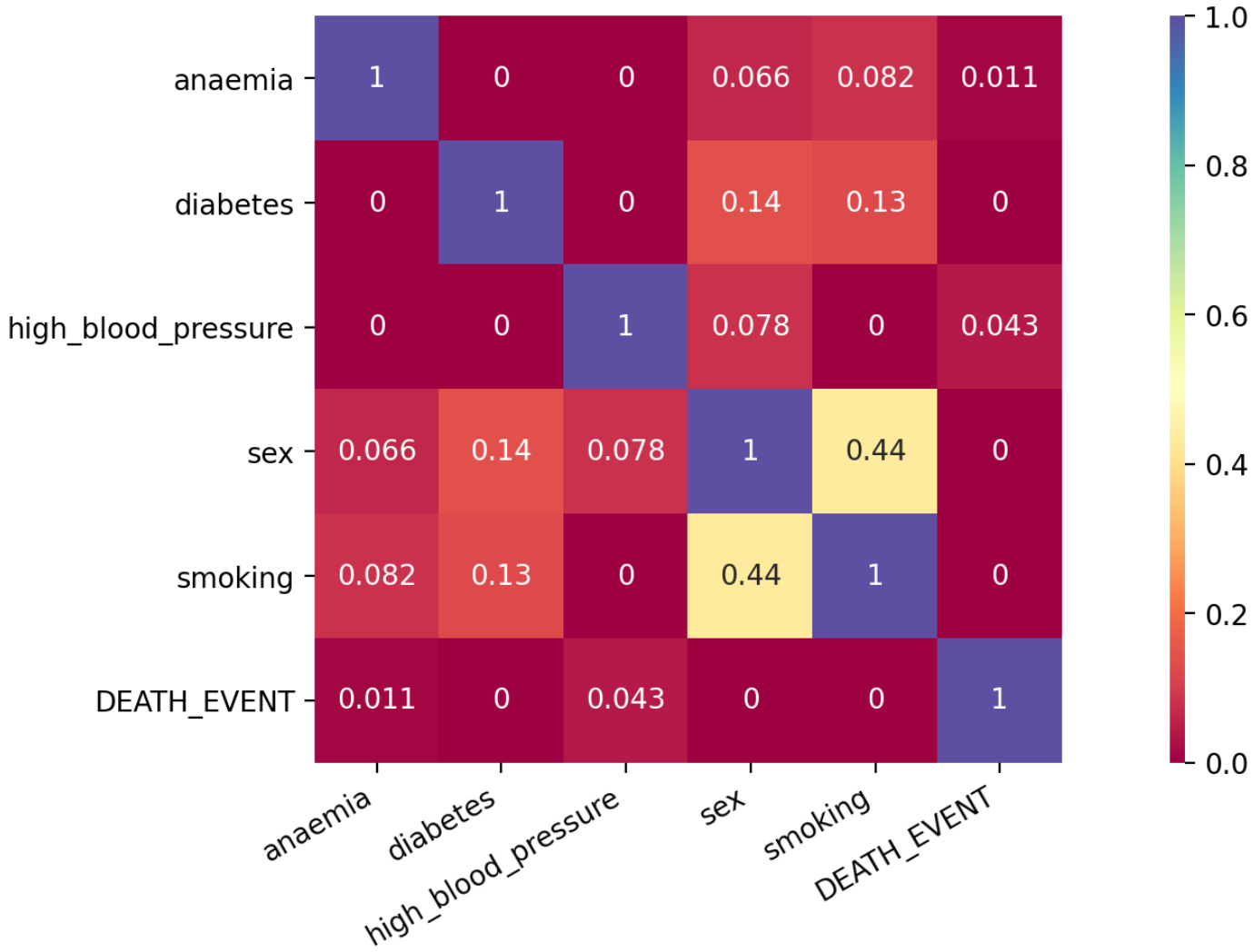
Quantitative or Categorical?

Quantitative Categorical



For quantitative variables, I used Pearson correlation. We can see that time (folllow-up period),

serum_creatinine, ejaction_fraction, and are are the four most correlated factors.



For categorical variables, I used Cramer's V correlation. We can see that smoking and and high_blood_pressure are two most correlated factors.

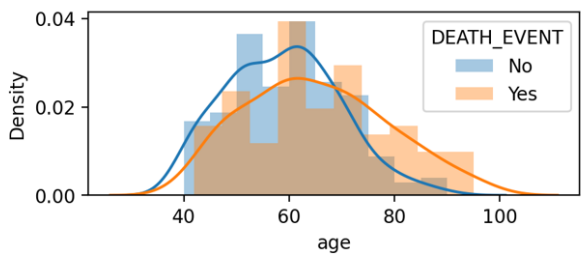
3. Data Visualization

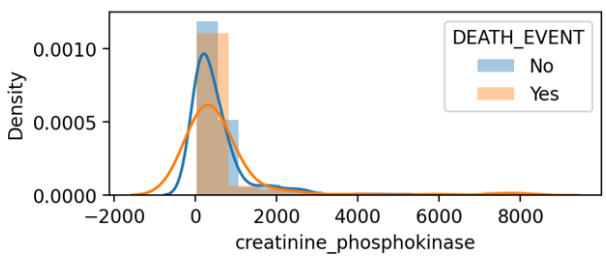
3.1 How long do they plan to close

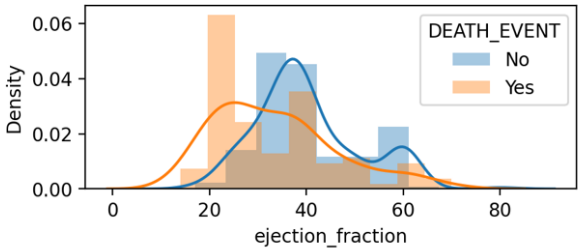
You can choose quantitative or categorical values to see the its distribution.

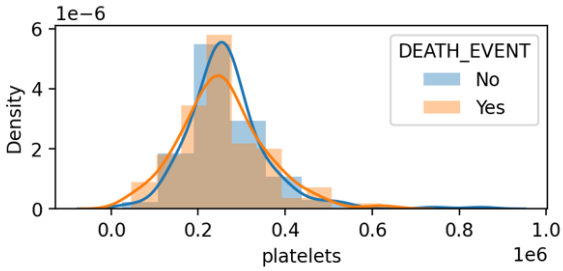
Quantitative or Categorical?

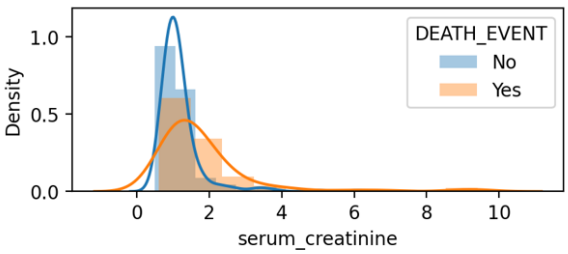
Quantitative Variab… Categorical Variables

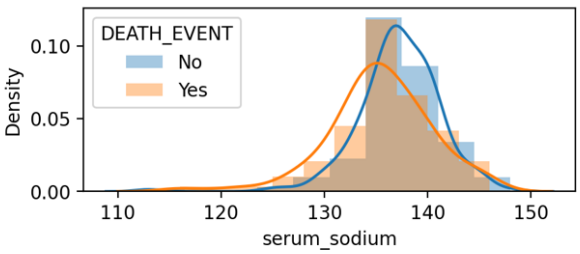


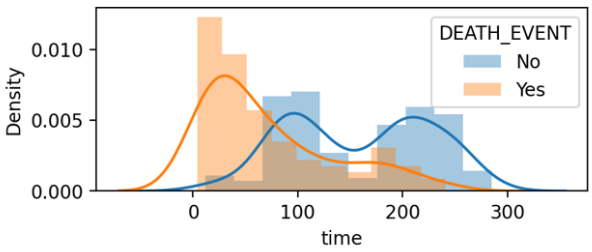




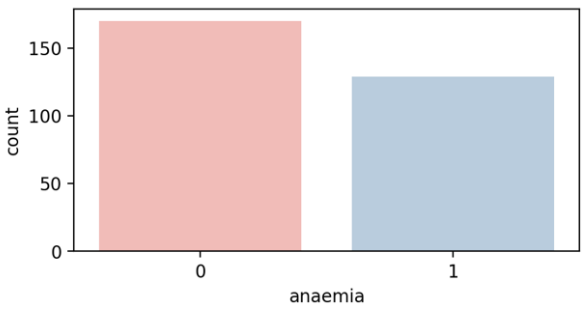


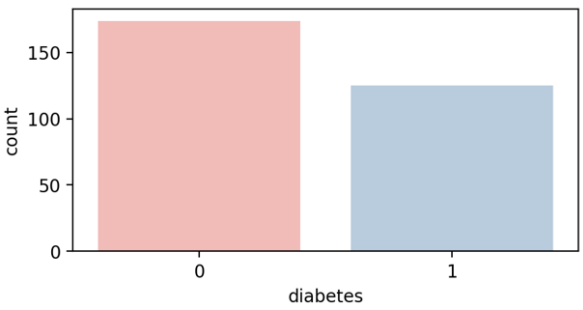


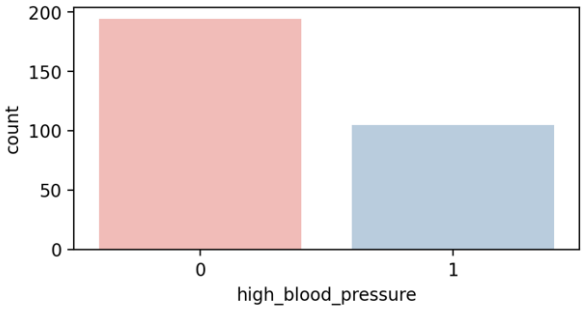


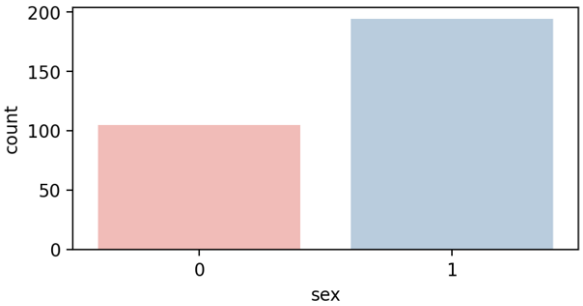


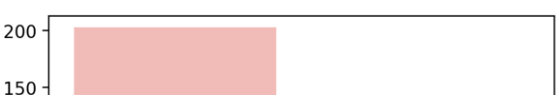
For continuous variables, I seperated the data by DEATH_EVENT. We can see that for serum_sodium, ejaction_fraction, and density, the distribution are quite different: the mean are apprantly different.

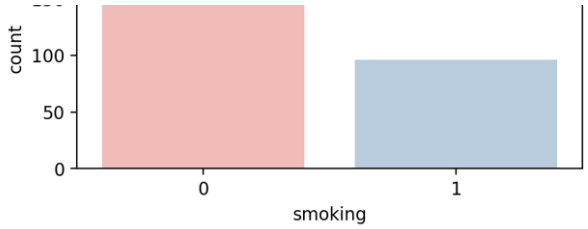


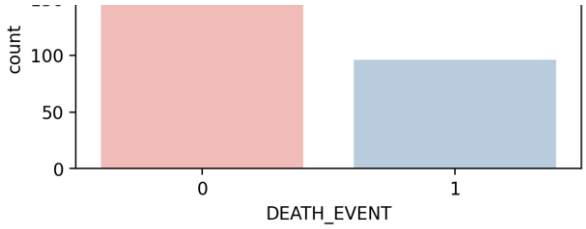








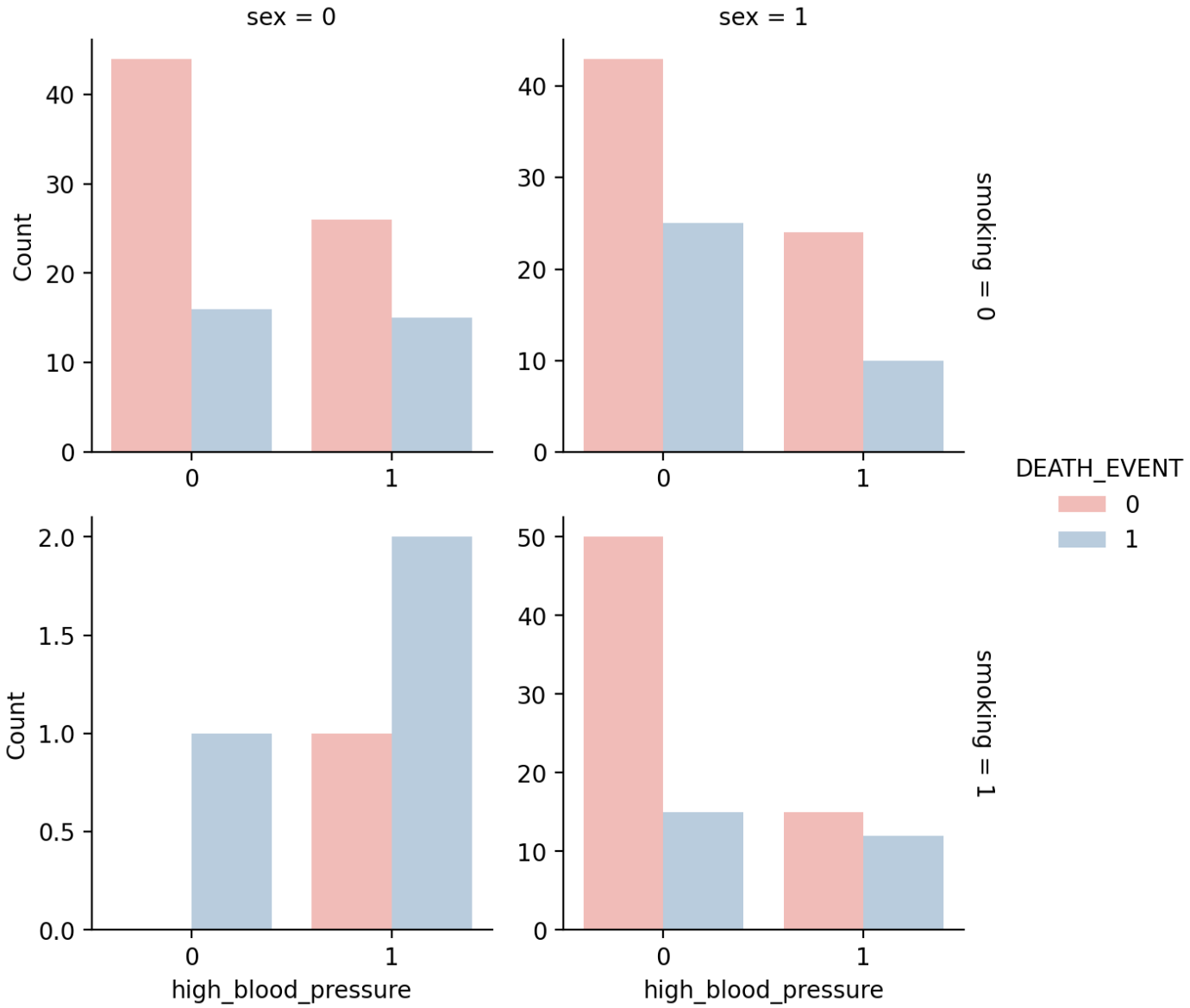




For categorical variables, we can see that there is some imbalance of data: for DEATH_EVENT, there are about 200 people dead but only 100 alive. Anaemia and diabetes relatively equat number of data. For other factors, the count between two categories are off by about half. A very interesting thing to notice is that diabetes, sex, and smoking seem to be uncorrelated to DEATH_EVENT at all.

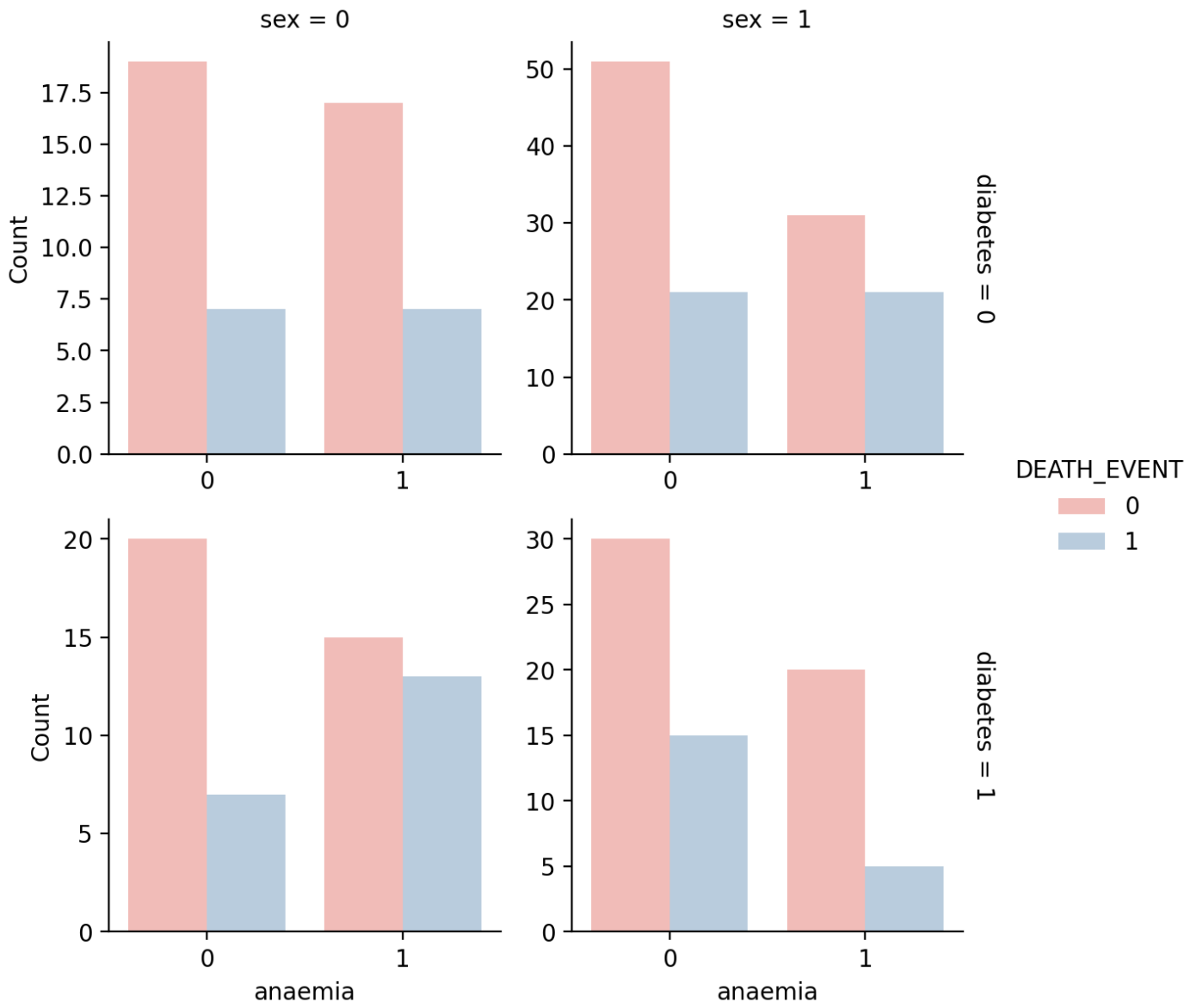
3.2 How does gender influence heart failure based on unhealthy habits?

☐ Wanna see how does heart failure among gender based on smoking and blood pressure?



We can see that combining smoking and high blood pressure make many male die from heart failure, while not creating a significant case for female. Gender does make some interesing contrasts between some variables.

☐ Wanna see how does heart failure among gender based on anemia and diabetes?



We don't notice anyting much significant here, but it's interesting that having both of diabetes and anaemia for male have a higher chance of dying from heart failure than female. Gender does make some interesing contrasts between some variables.

4. Machine Learning Models

Now, we would like to explore which machine learning model can predict mortality by heart failure the best. You can choose the following ML models to inspect the performances. You can also choose different combinations of variables to compare their performances. We used cross validations for all five models and computed the average accuracy: 0.8 for traning, and 0.2 for validation. This is a simple binary classification problem. To deal with the data imbalance problem, I used upsampling method.

Select the variables that you would like to explore.

time smoking ejection_fraction age high_blood_pressure

ML Model to choose:

KNN Random Forest Ridge Regression MLP Gaussian NB

mean validation accuracy for K Nearest Neighbors: `0.5822598870056497`

mean validation accuracy for Ridge Regression: `0.7889265536723163`

mean validation accuracy for Random Forest: `0.655593220338983`

mean validation accuracy for MLP: `0.6255932203389831`

mean validation accuracy for GaussianNB: `0.7722598870056496`