Project 1: Heart Failure Prediction

Regression and Classification

Part 1 - Background Info

1. What’s the difference between logistic and linear regression?
2. What’s the difference between predictors and response variables?
3. What is the purpose and benefits of preprocessing data?
4. What is overfitting and underfitting that data? What are the consequences.

Part 2: About the Data

The dataset contains the following 13 clinical features:

- 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)

Questions:

1. What are the features? Response variable? Describe each.

2. Which features would you classify as categorical vs. continuous?

   Explain.

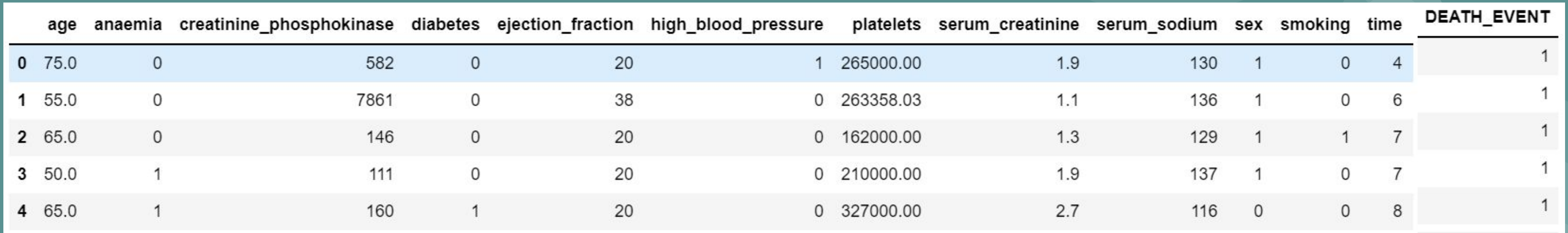
Problem Statement

Your goal is to create a model to predict mortality caused by heart failure.

Write a small paragraph (3-6 sentences) discussing the significance of addressing this issue and the benefits your model can provide.

Part 3: Data Exploration

Import the data set and print its first 5 rows



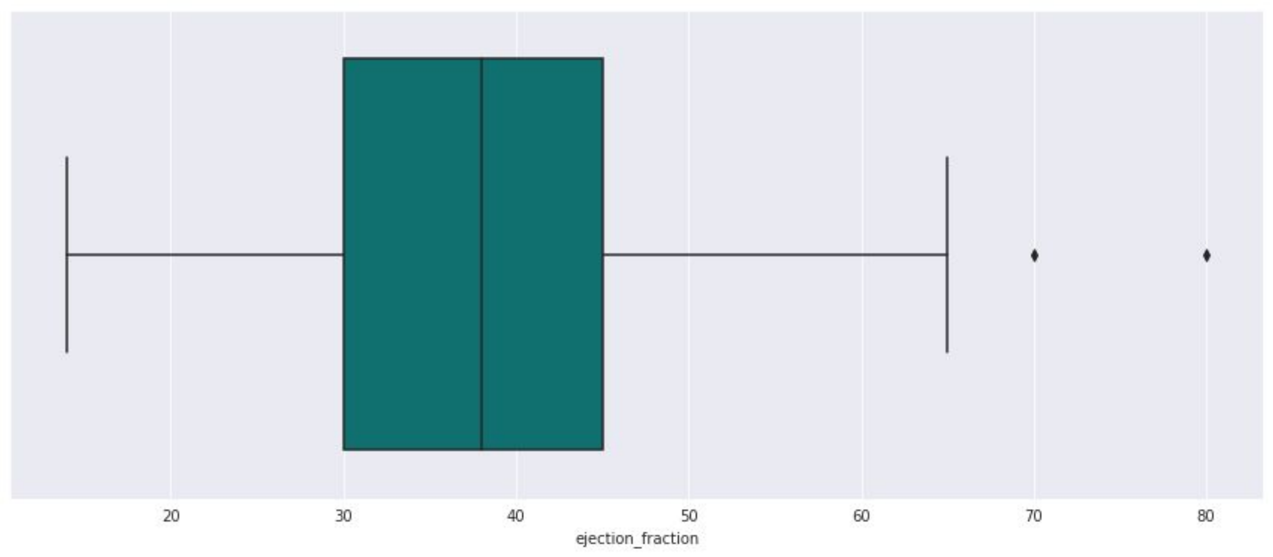
	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4	1
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7	1
3	50.0	1	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

Questions

1. Are there any missing values?
2. Any null values?
3. How many unique values are there for each column?

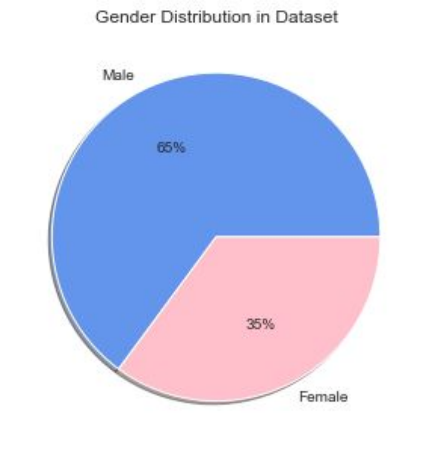
Plots

Create a boxplot for ejection_fraction. What does this plot tell you? Any outliers? Do they need to be removed, if so, why? (or why not?). Repeat this for time and serum creatinine. Report your conclusions from each plot.

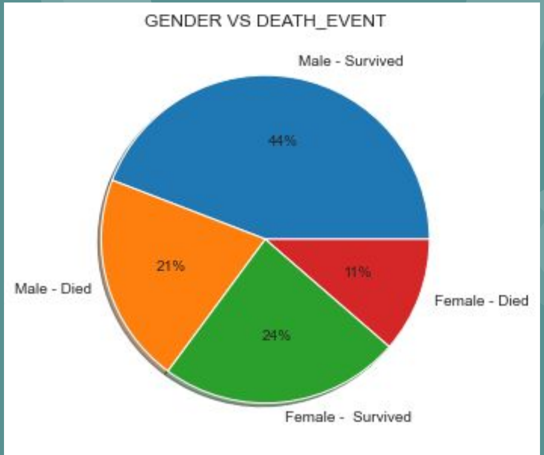


Plots

Create a pie chart to see the gender distribution in the dataset. Do this for the remaining categorical variables (there are 5 total, including gender).



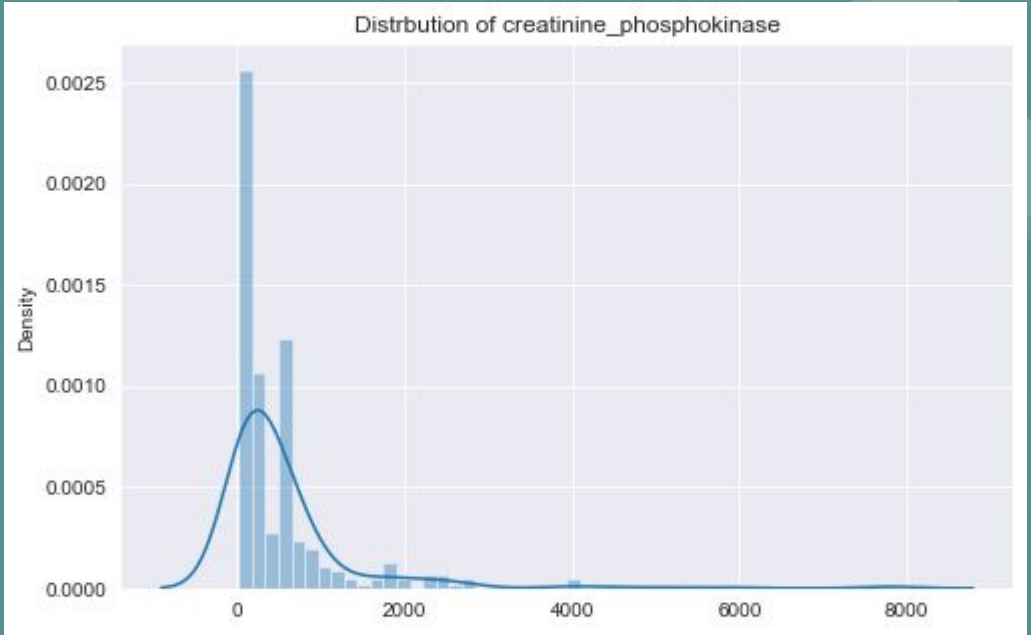
Since the data is not evenly distributed, create a plot that shows the percentage of survivors from males and females separately.



Report your conclusions from each plot. Is working with a balanced dataset important? Why or why not? How can we deal with an imbalanced dataset?

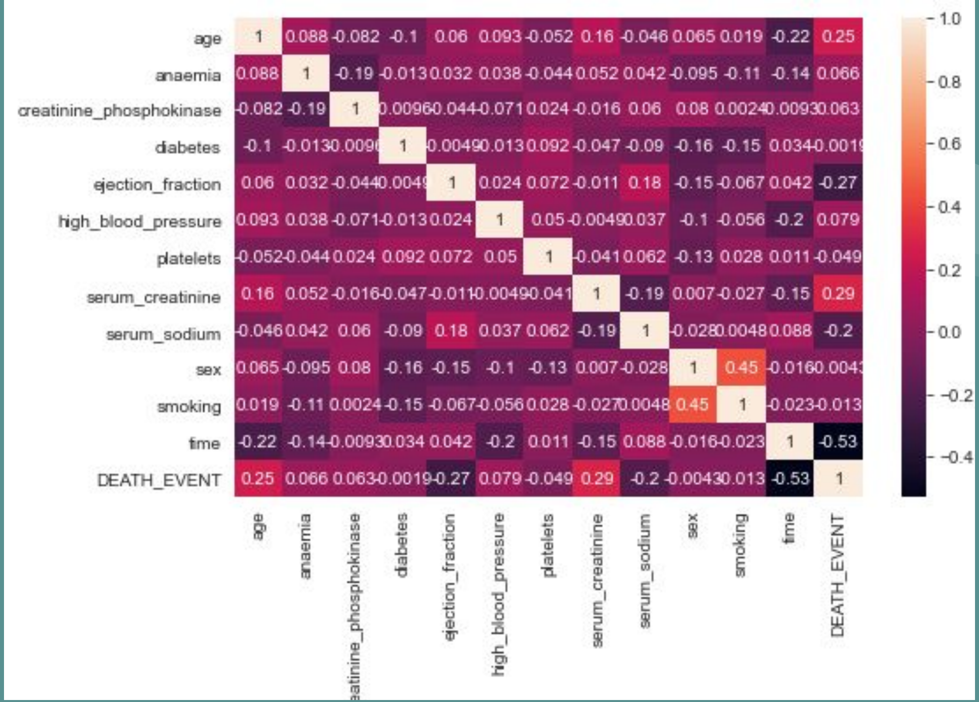
Plots

Create distribution plots of the numerical variables. Below is an example. Report your conclusions from each plot.



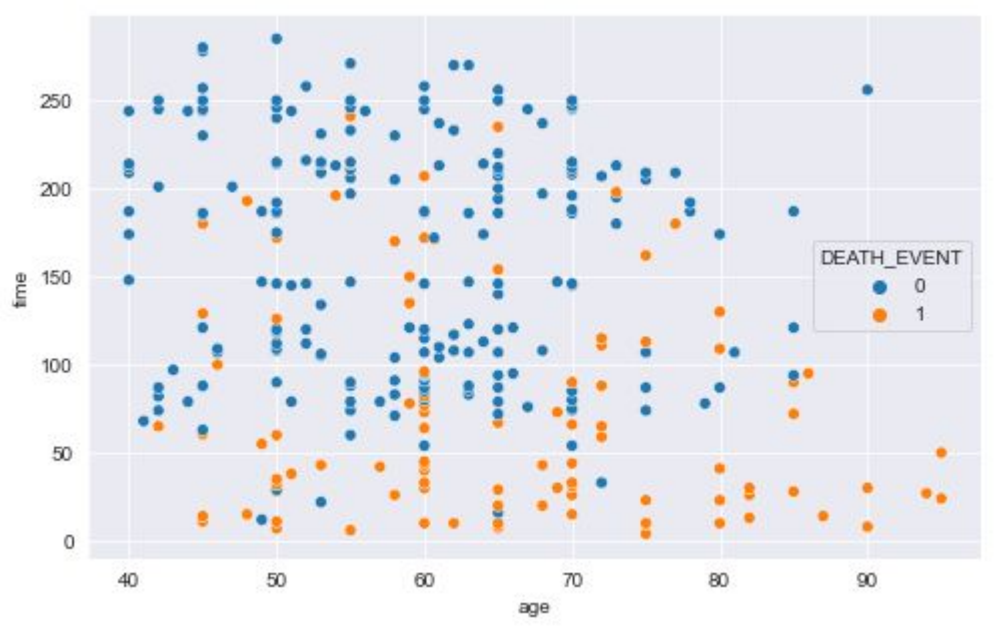
Plots

Create a heatmap of the data. What does it tell you?
Define multicollinearity. Is there multicollinearity in our dataset? Why do we care? How did you check?



Plots

Create at least 2 plots of your choice and give a brief summary of your conclusions from each plot. Example: scatterplot.



Scaling

1. Is it necessary to scale the data? What benefits would it provide?
2. Which scaler will you use for this data set? Min Max, Standard, Robust, etc.
3. Are the features or the response variables scaled?

*Don't forget to split your data into test-train splits before scaling!

Preprocessing

1. Which columns needed to be modified (dummy variables)?
2. What are parametric and nonparametric learning algos? For the models you are choosing- are they parametric or nonparametric? Explain.
3. Define label encoding and one hot encoding and compare them.

Model

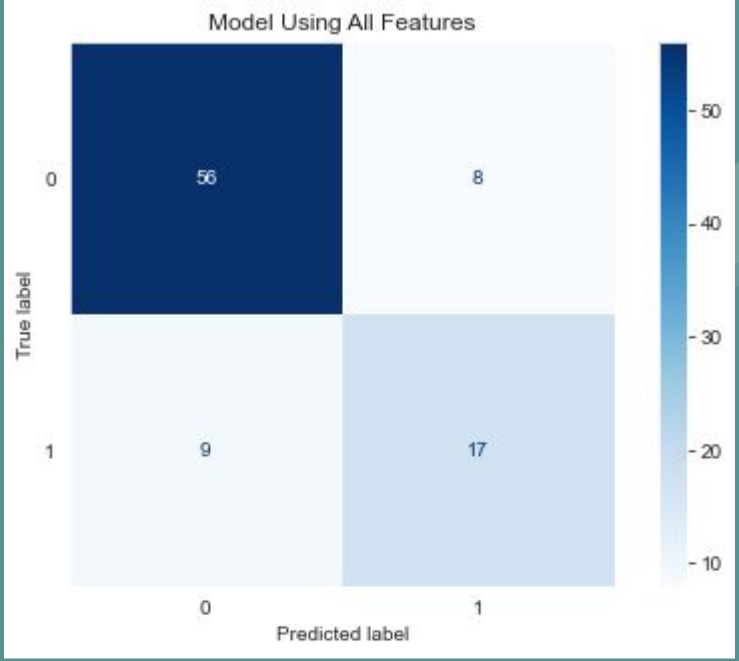
Perform logistic regression and one other model you've learned in class to solve the problem statement.

Apply any necessary encoding, scaling, and train test splits to your data and construct the 2 models you selected. Provide a classification report and confusion matrix for both models.

Explain why you picked the model you chose. Use cross validation (see sklearn's cross_val_score) to compare the models, show accuracy scores, etc...

Part 4- Model

Confusion matrix examples: Also, show your classification report, AUC plot.



Part 5- Conclusion & Analysis

Your conclusion should answer (at least) these questions. Give a summary of your results!

1. What do your models show?
2. Why is it significant?
3. How accurate was your model?
4. How can you expand upon your work?