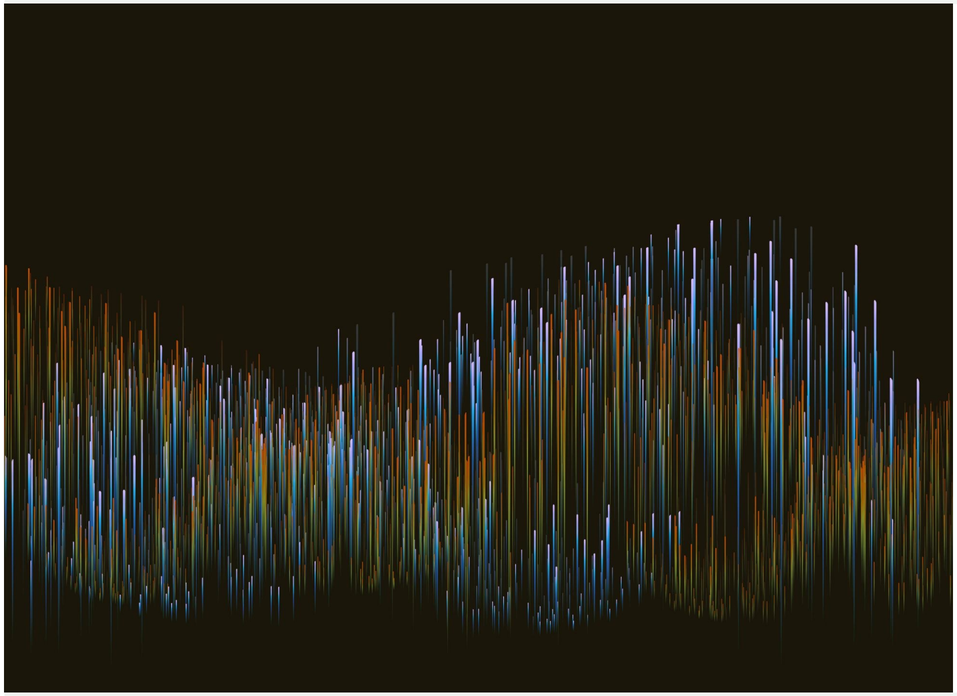
Homework Grade



Project 1: Heart Failure Prediction

DANIEL GARZA

COSC 3337 SU23

Part 1 – Background Info

1. Logistic vs Linear Regression?
   - Linear predicts dependent variables while logistic calculates probability.
2. Difference between predictors and response variables?
   - Predictors: Independent variables, x-axis, input controlled in models
   - Response: Dependent variables, y-axis, output result from predictors.
3. Benefits for preprocessing data?
   - Improve the accuracy, efficiency, and predictions. This process including handling null values, one-hot encoding, and normalization.
4. What is over/underfitting?
   - Overfitting: Occurs when you train the data too well that it becomes poor at testing the data.
   - Underfitting: Occurs when the model isn't trained enough and does not learn and therefore unable to test.
   - Both of these can lead to a bias dataset

Part 2: About the Data

- What are the features? Response variable?
  - Features: anemia, high blood pressure, diabetes, sex, age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, and time.
  - Response variable is death event. A binary results representing if the patient deceased during the follow-up period.
- Categorical vs continuous?
  - Categorical: Anemia, high blood pressure, diabetes, sex, smoking, and the target variable death event.
    - These variables are represented in binary numbers.
  - Continuous: Age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, and time
    - These variables are not limited in a range

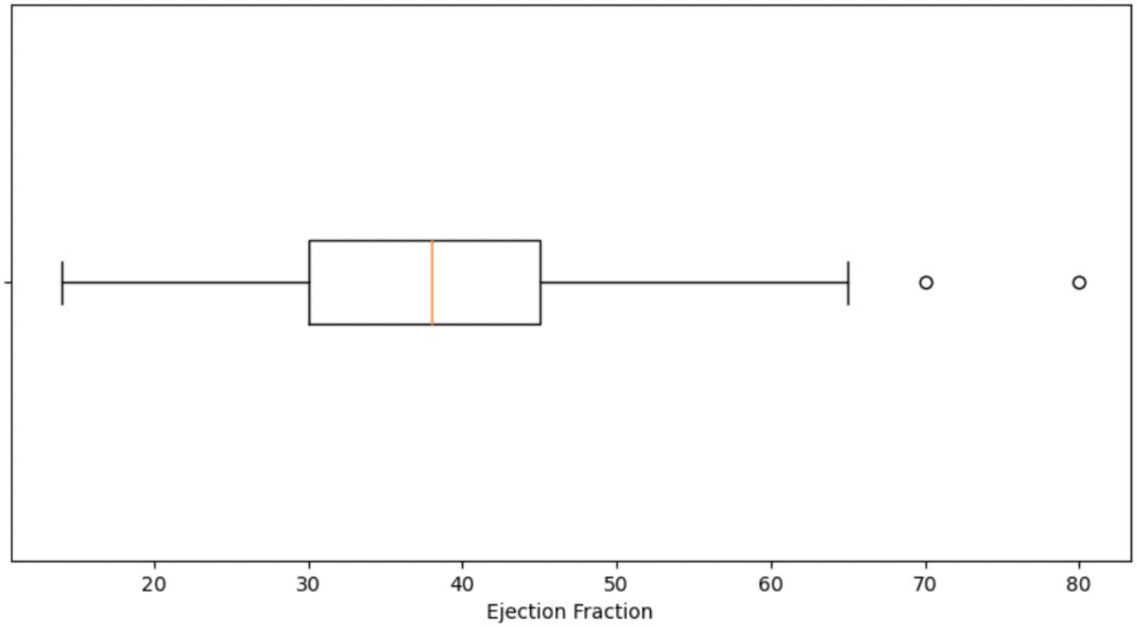
Problem Statement

- Predicting whether a person will die from a heart attack is very important and applicable to millions of people in the world. Heart attacks are one of the leading causes of death in the United States, so data research like this may help save millions of lives by informing medical professionals that certain individuals may be more at risk than others. Also, understanding the common denominators in a person that does die from heart attack may lead to prevention for others in the future.

Part 3: Data Exploration

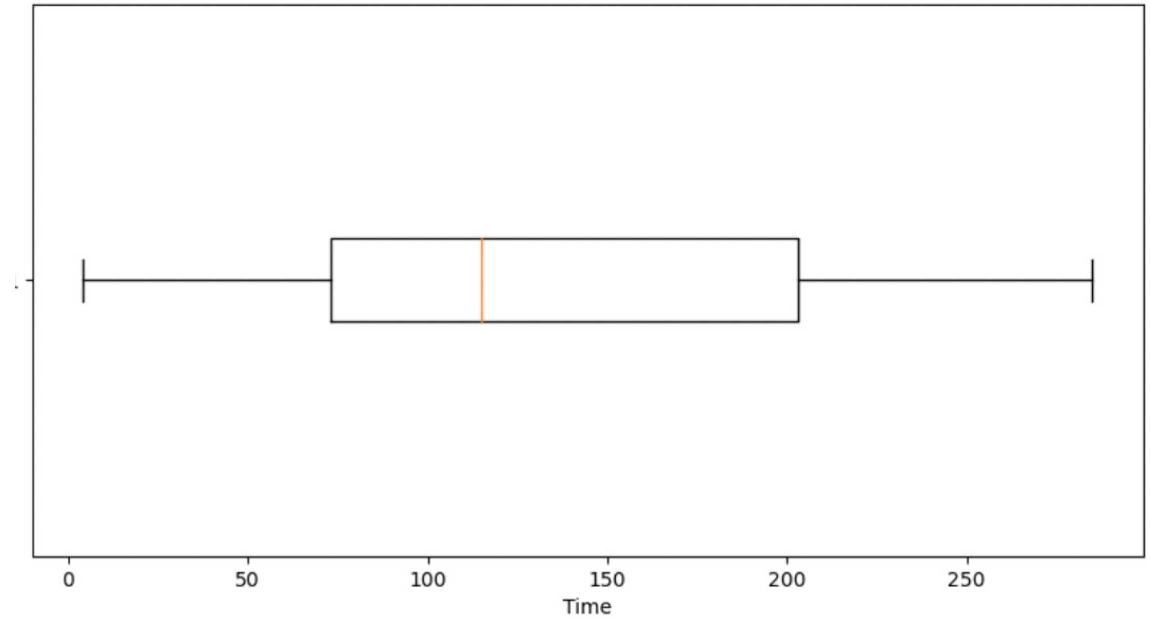
- Print first 5 rows
- Missing values?
- Null values
- Boxplots
- Gender Pie Chart
- Gender vs Death Event Pie Chart
- Distribution Charts
- Heatmap
- Scatterplots

Part 3 Cont: Ejection Fraction Boxplot



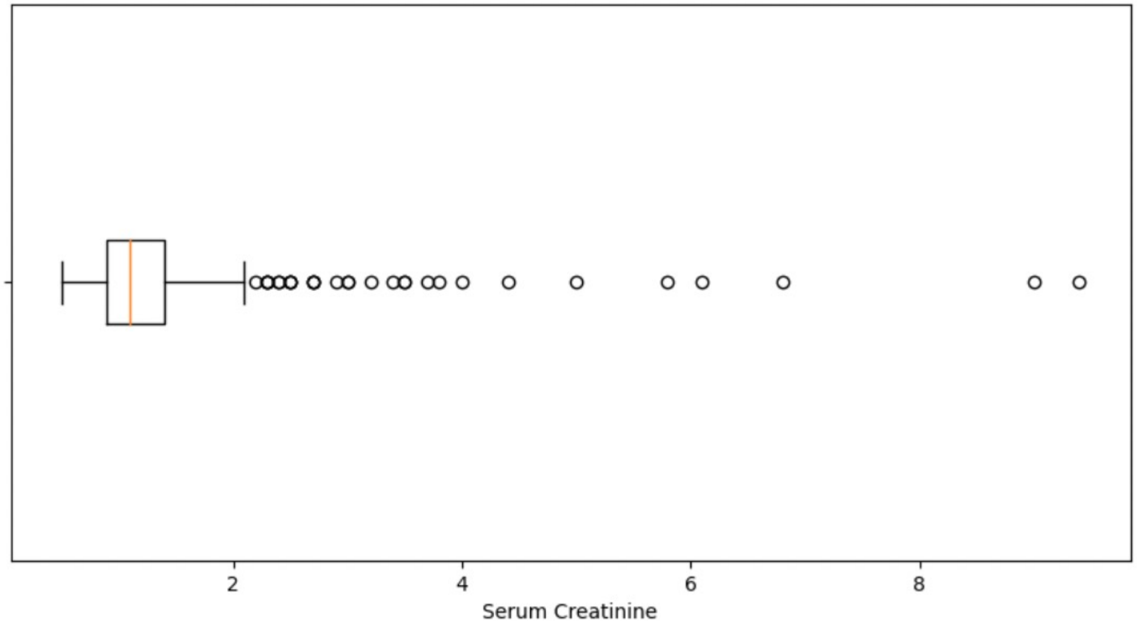
- Percentage of blood leaving the heart at each contraction
- Majority of people fall within 30-45%
- 2 outliers above the 3rd quartile
- It is appropriate to leave outliers in this dataset, assuming it is accurate, because it is important to understand these high numbers are possible and should be studied.

Part 3 Cont: Time Boxplot



- Follow-up period
- A large portion had a follow-up period between 75-200 days.
- Wide range of 281
- No outliers.
- This is a rather large range that could possibly lead to bias information.

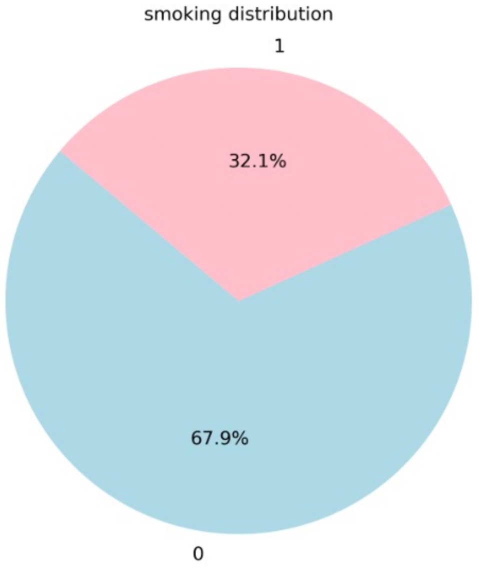
Part 3 Cont: Serum Creatinine Boxplot

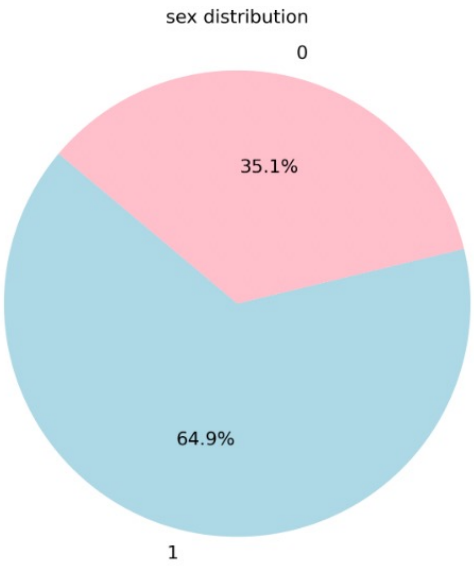


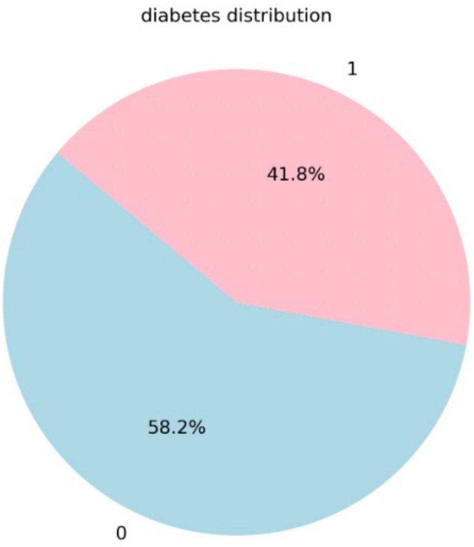
- Level of serum creatinine in the blood
- There is a wide spread of serum creatinine in individuals with many outliers.
- We can keep these outliers but must keep them in mind when using the serum creatinine variable going forward.

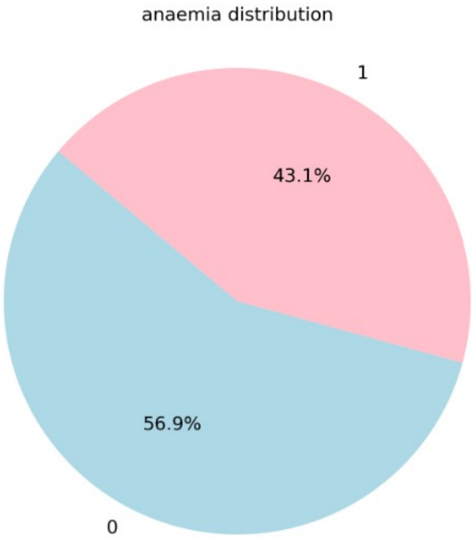
8/31/23

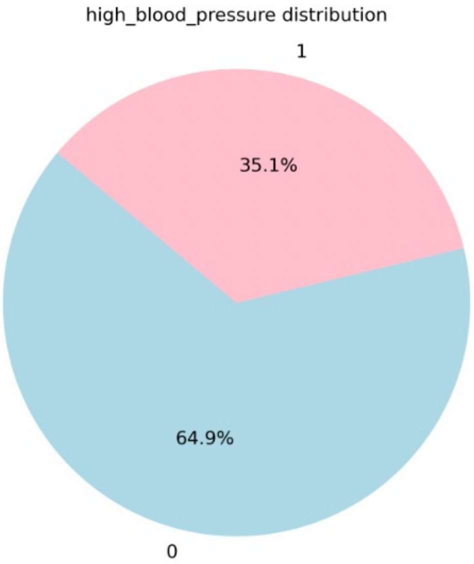
Part 3 Cont: Categorical Pie Charts


smoking distribution

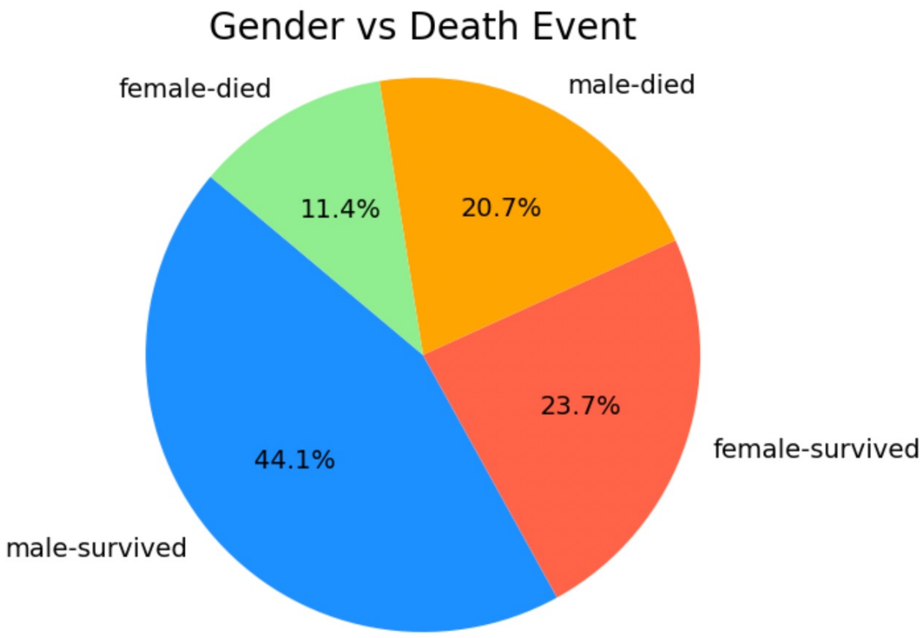

sex distribution


diabetes distribution


anaemia distribution


high_blood_pressure distribution

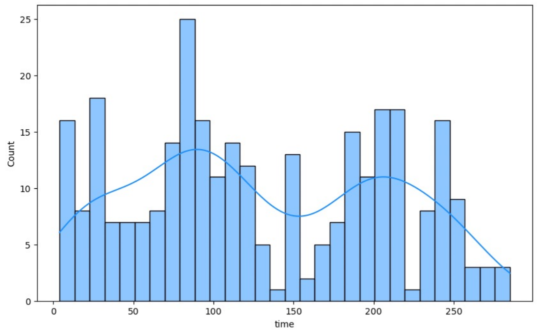
Part 3 Cont: Gender vs Death Event



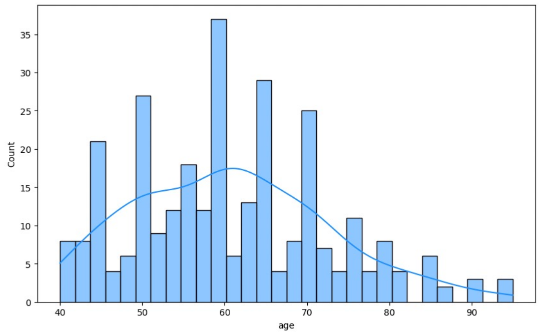
- Yes, it is important to work with a balanced dataset to prevent bias in your results.
- In this study, unbalanced data sets, may inaccurately predict whether a person will die from a heart attack
- To handle imbalanced datasets, we can use techniques such as resampling.

Part 3 Cont: Numerical Distribution Charts

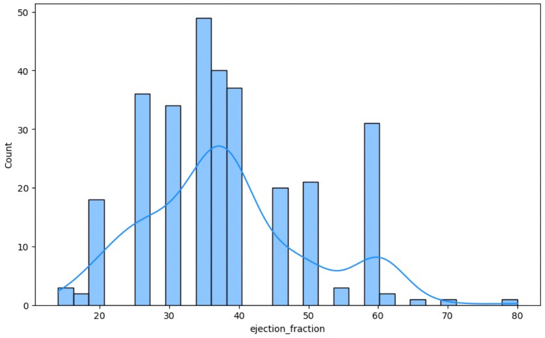
time distribution



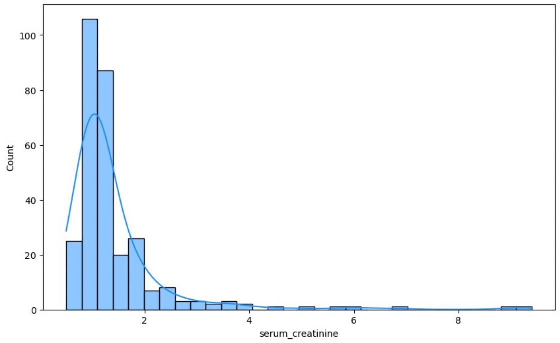
age distribution



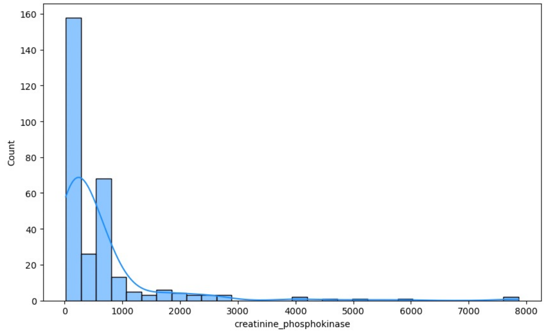
ejection_fraction distribution



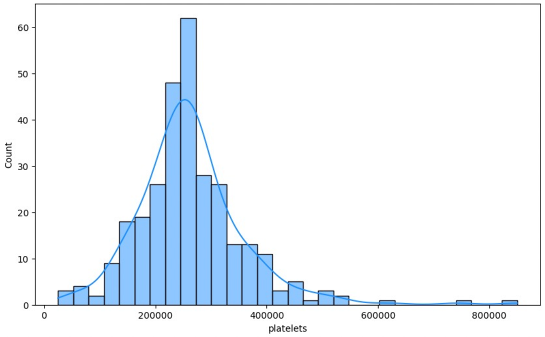
serum_creatinine distribution



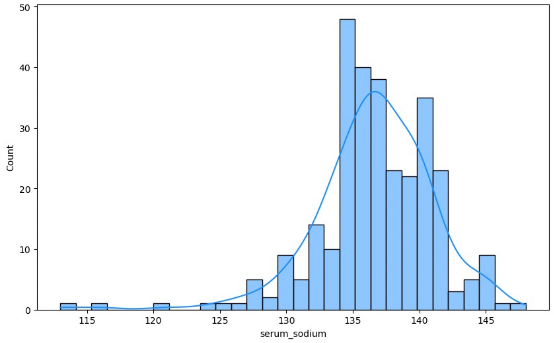
creatinine_phosphokinase distribution



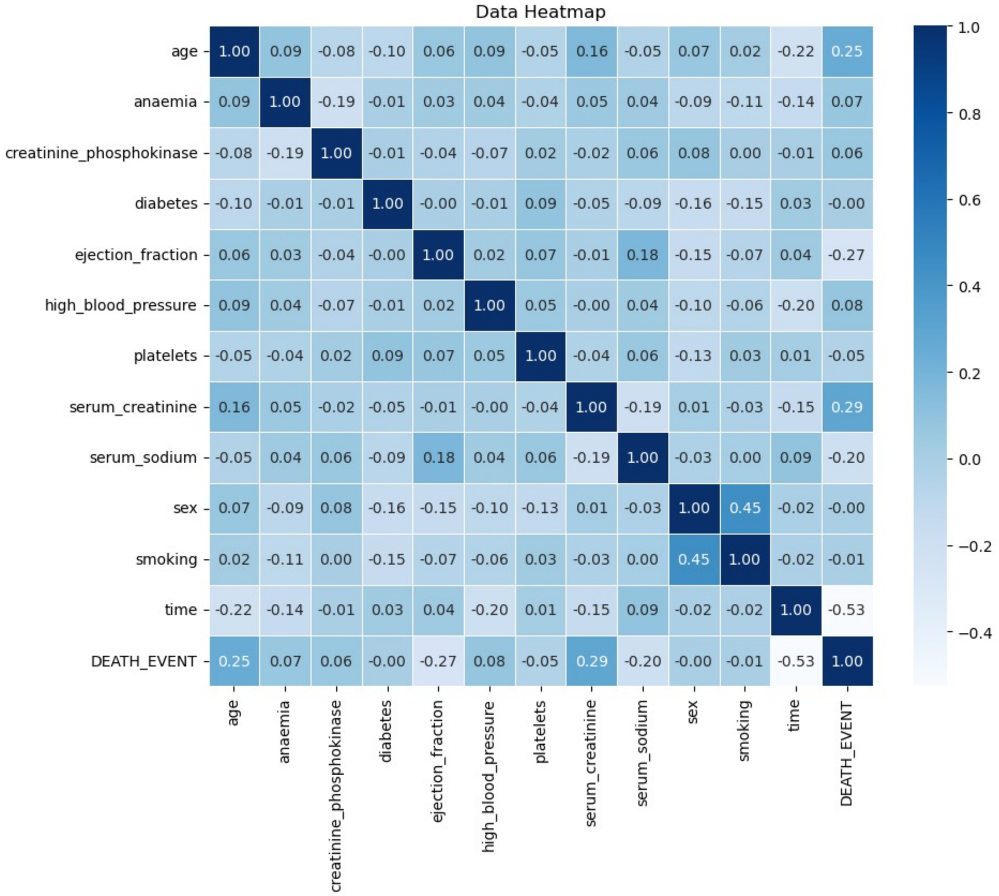
platelets distribution



serum_sodium distribution



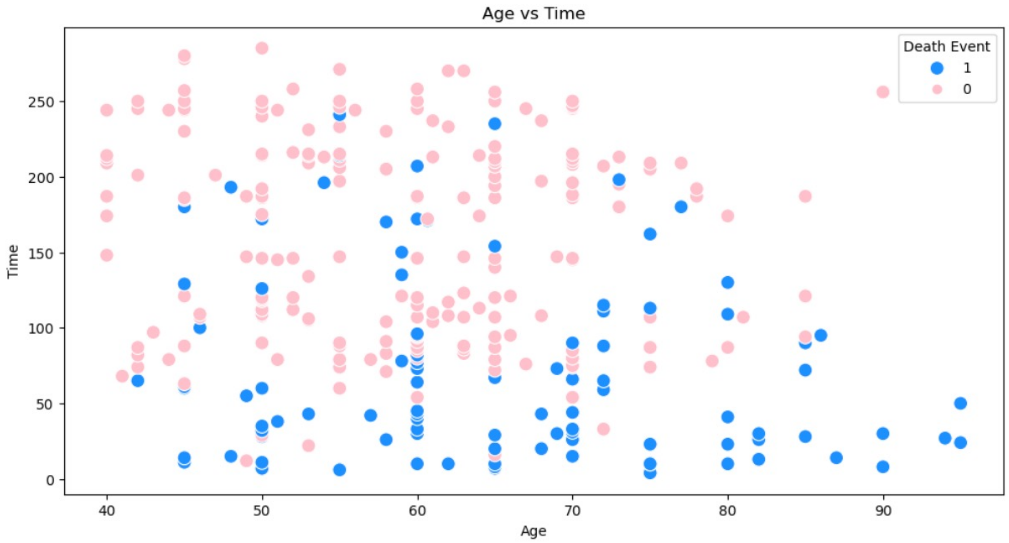
Part 3 Cont: Heatmap



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
age	1.00	0.09	-0.08	-0.10	0.06	0.09	-0.05	0.16	-0.05	0.07	0.02	-0.22	0.25
anaemia	0.09	1.00	-0.19	-0.01	0.03	0.04	-0.04	0.05	0.04	-0.09	-0.11	-0.14	0.07
creatinine_phosphokinase	-0.08	-0.19	1.00	-0.01	-0.04	-0.07	0.02	-0.02	0.06	0.08	0.00	-0.01	0.06
diabetes	-0.10	-0.01	-0.01	1.00	-0.00	-0.01	0.09	-0.05	-0.09	-0.16	-0.15	0.03	-0.00
ejection_fraction	0.06	0.03	-0.04	-0.00	1.00	0.02	0.07	-0.01	0.18	-0.15	-0.07	0.04	-0.27
high_blood_pressure	0.09	0.04	-0.07	-0.01	0.02	1.00	0.05	-0.00	0.04	-0.10	-0.06	-0.20	0.08
platelets	-0.05	-0.04	0.02	0.09	0.07	0.05	1.00	-0.04	0.06	-0.13	0.03	0.01	-0.05
serum_creatinine	0.16	0.05	-0.02	-0.05	-0.01	-0.00	-0.04	1.00	-0.19	0.01	-0.03	-0.15	0.29
serum_sodium	-0.05	0.04	0.06	-0.09	0.18	0.04	0.06	-0.19	1.00	-0.03	0.00	0.09	-0.20
sex	0.07	-0.09	0.08	-0.16	-0.15	-0.10	-0.13	0.01	-0.03	1.00	0.45	-0.02	-0.00
smoking	0.02	-0.11	0.00	-0.15	-0.07	-0.06	0.03	-0.03	0.00	0.45	1.00	-0.02	-0.01
time	-0.22	-0.14	-0.01	0.03	0.04	-0.20	0.01	-0.15	0.09	-0.02	-0.02	1.00	-0.53
DEATH_EVENT	0.25	0.07	0.06	-0.00	-0.27	0.08	-0.05	0.29	-0.20	-0.00	-0.01	-0.53	1.00

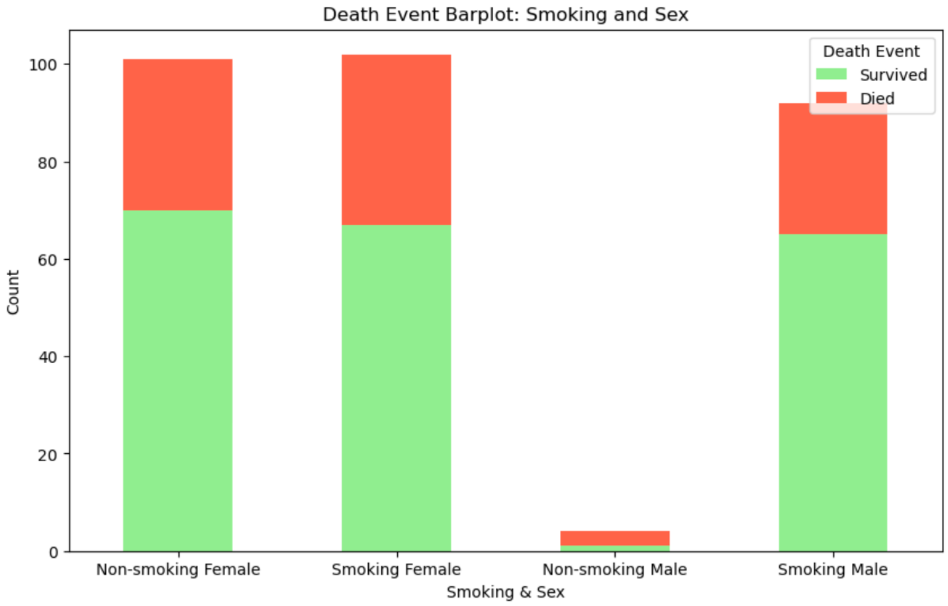
- Multicollinearity is when two or more independent variables in a model are highly correlated with each other.
- By reading the heatmap, the highest correlation is only .45.
- There is very little correlation between data values therefore we have no multicollinearity.
- When multicollinearity is present, it may improve our prediction accuracy and vice verse when it is not present.

Part 3 Cont: Plots



- We are comparing time vs age and their correlation to a death event.
- Reading our heatmap, these two values have a correlation of -.22
- We expect a very slight negative linear curve and that is what we see indicating these values are very loosely correlated.
- On average, the longer the follow-up period, the more likely the member is to survive.

Part 3 Cont: Plots



- We are comparing two binary values using smoker and sex.
- The highest correlation on the heatmap at 0.45
- With the exception, of the non-smoking male, the number is almost even across.
- About 70% of smoking or non-smoking individuals survive.
- Smoking does not have effect on a death occurring.

Scaling

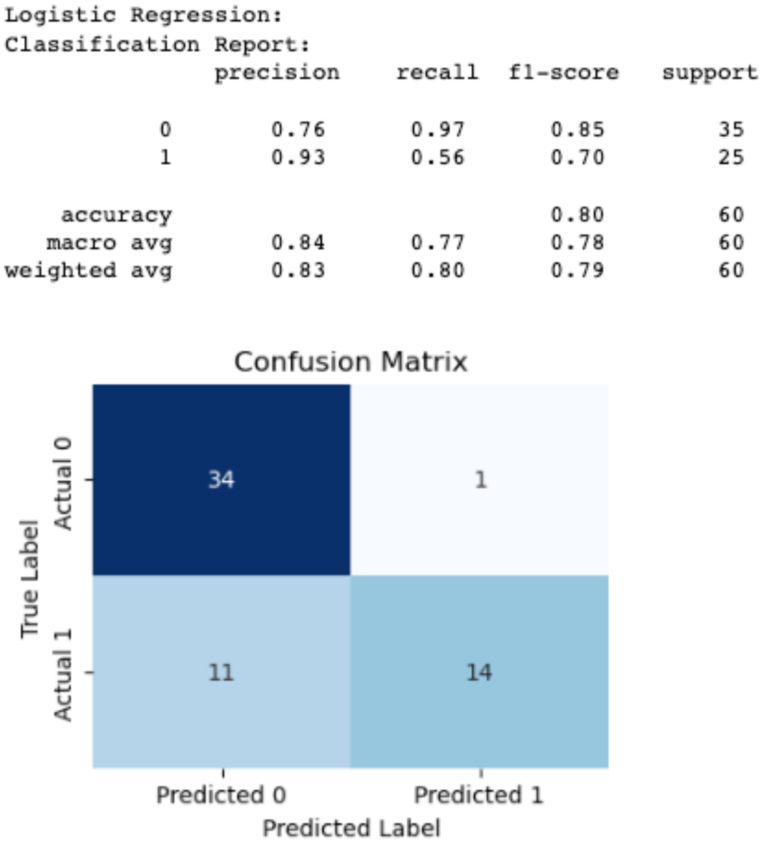
1. Necessary to scale the data?
    - Since we have a large amount of categorical data, we may not have to scale the data, but it is still a good idea to scale the numerical data. This is beneficial to improve convergence and overall better performance.
2. Which scaler will you use for this data set?
    - Since this dataset has both numerical and categorical variables, we can use the standard scaler.
3. Are the features or the response variables scaled?
    - Only the features variables are scaled. Not the response or target variables.

Preprocessing

1. Which columns needed to be modified?
   - Smoking and sex are two columns that needed to be modified and and changed to dummy variables represented as binary values.
2. What are parametric and nonparametric learning algos? For the models you are choosing- are they parametric or nonparametric? Explain.
   - The main difference is that parametric models make assumptions about the data and can make predictions based on those assumptions. Non-parametric models do not make assumptions. Rather, their goal is to learn about the data.
3. Define label encoding and one hot encoding and compare them.
   - Both label encoding and one-hot encoding convert data into categorical values represented by numbers. The difference is that one-hot only assigns binary values while label encoding assigned values in numerical order.

Part 4: Model

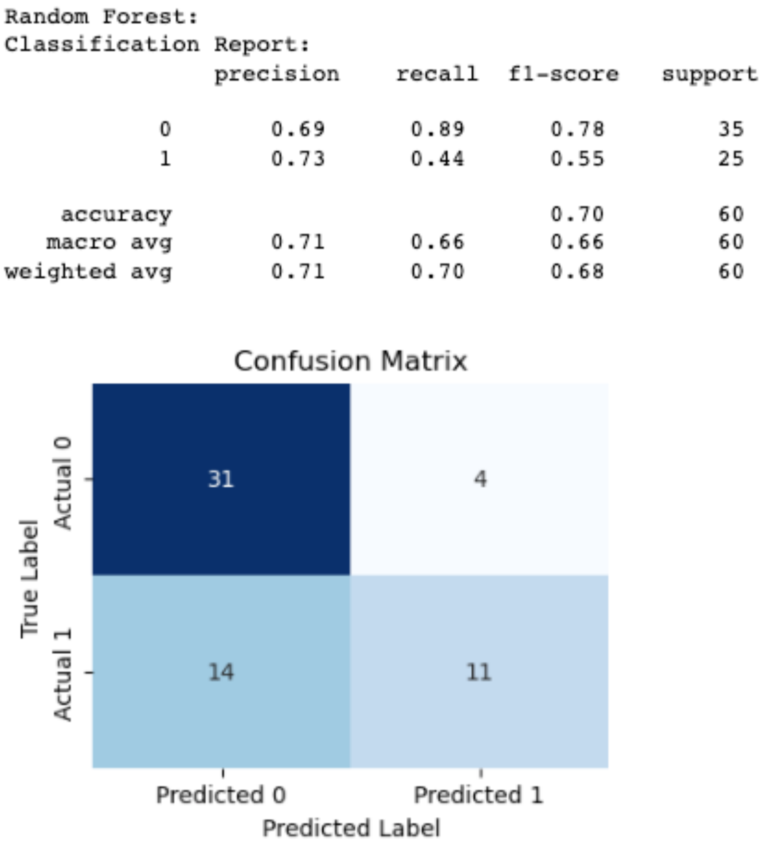
- I selected logistic regression and random forest because they are good for classification.
- Applied one-hot encoding to handle categorical variables.
- Applied the standard scaler
- Printed classification report and confusion matrix
- Used cross validation to compare models



```
Logistic Regression:
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.97      0.85        35
           1       0.93      0.56      0.70        25

    accuracy                           0.80        60
   macro avg       0.84      0.77      0.78        60
weighted avg       0.83      0.80      0.79        60
```



```
Random Forest:
Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.89      0.78        35
           1       0.73      0.44      0.55        25

    accuracy                           0.70        60
   macro avg       0.71      0.66      0.66        60
weighted avg       0.71      0.70      0.68        60
```

```
Logistic Regression cross validation scores: [0.65       0.78333333 0.88333333 0.85       0.6779661 ]

Random Forest cross validation scores): [0.48333333 0.8        0.81666667 0.7        0.6779661 ]
```

Part 5: Conclusion and Analysis

1. What do your models show?
   - Logistic regression preformed the best of the two. Proven due to higher accuracy and 93% precision when predicting deaths and higher cross validation score
2. Why is it significant?
   - Precision critical when predicting deaths
   - Cross validation score indicates that the model is making accurate predictions on the test data. Provides more confidence in the model's performance and its ability to make accurate predictions on new data.
3. How accurate was your model?
   - Logistic regression has an accuracy of 0.80 and 0.85 cross validation score.
4. How can you expand upon your work?
   - Clean data before testing, increase data size to improve confidence in numbers, continue to test for best parameters