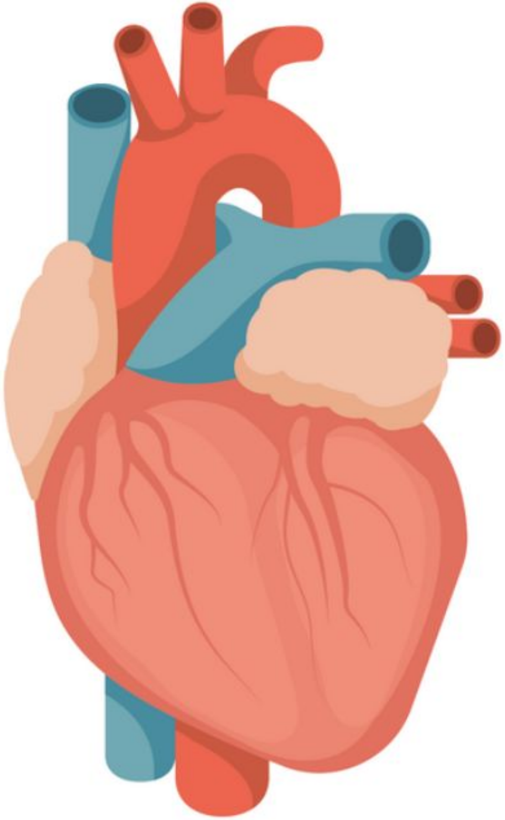
Predicting Heart Failure with ANN

Content

1. What is a heart attack?
2. The dataset
3. Numerical variable analysis
4. Categorical variable analysis
5. The 3 Neural Network models
6. Evaluation of the models
7. Conclusion
8. Appendix



Healthy heart

What is a heart attack?

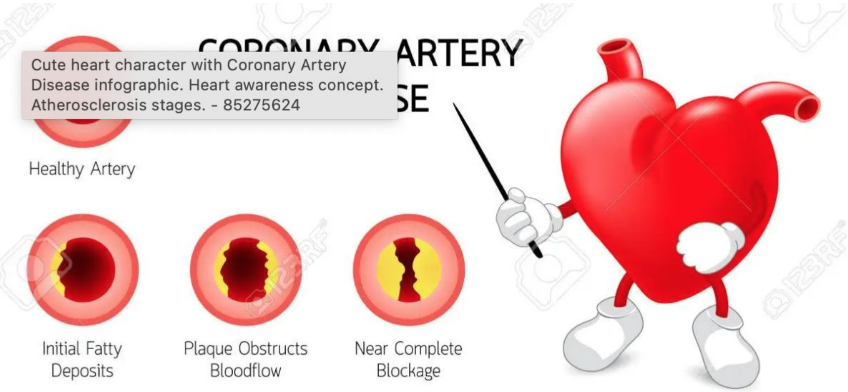
Heart failure means that the heart is unable to pump blood around the body properly. It usually happens **because the heart has become too weak or stiff**. -> lack of oxygen -> cardiac arrest

Due to many causes:

- Coronary artery disease
- Lifestyle (smoking, sedentary, nutrition)
- Genetics

Some symptoms:

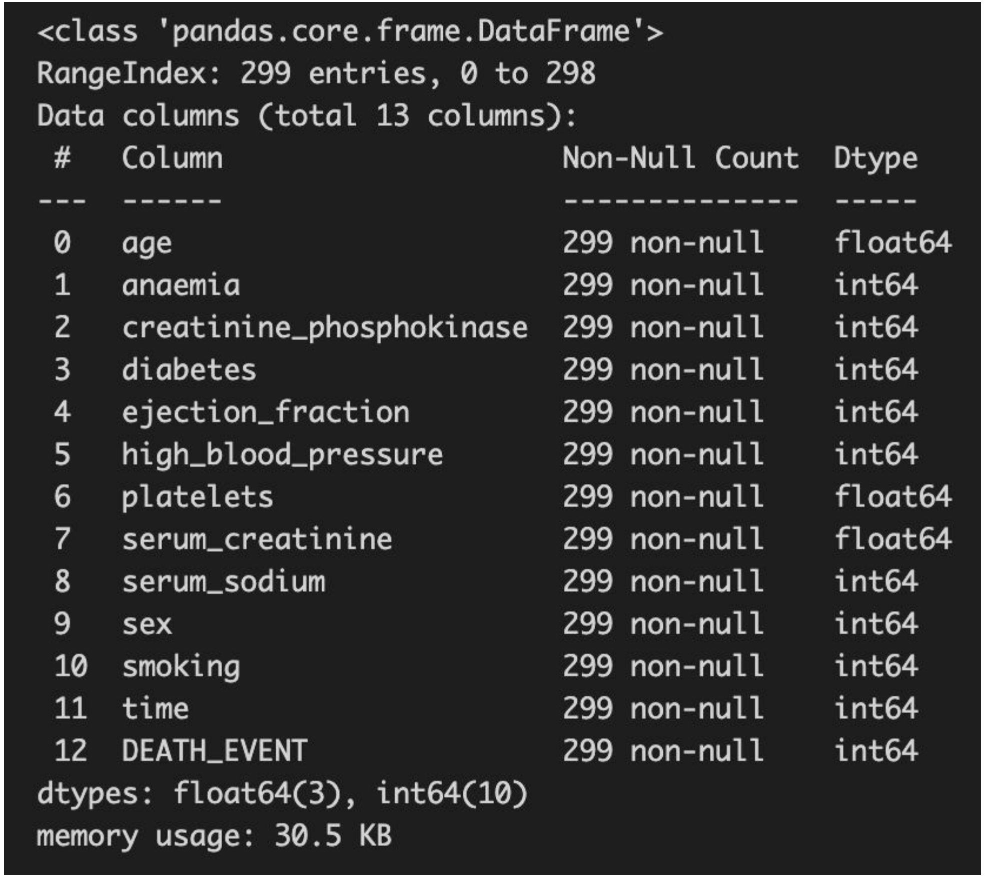
- Chest pain
- Shortness of breath
- Fatigue and weakness
- Swelling of legs, ankle, feet



Cute heart character with Coronary Artery Disease infographic. Heart awareness concept. Atherosclerosis stages. - 85275624

The dataset

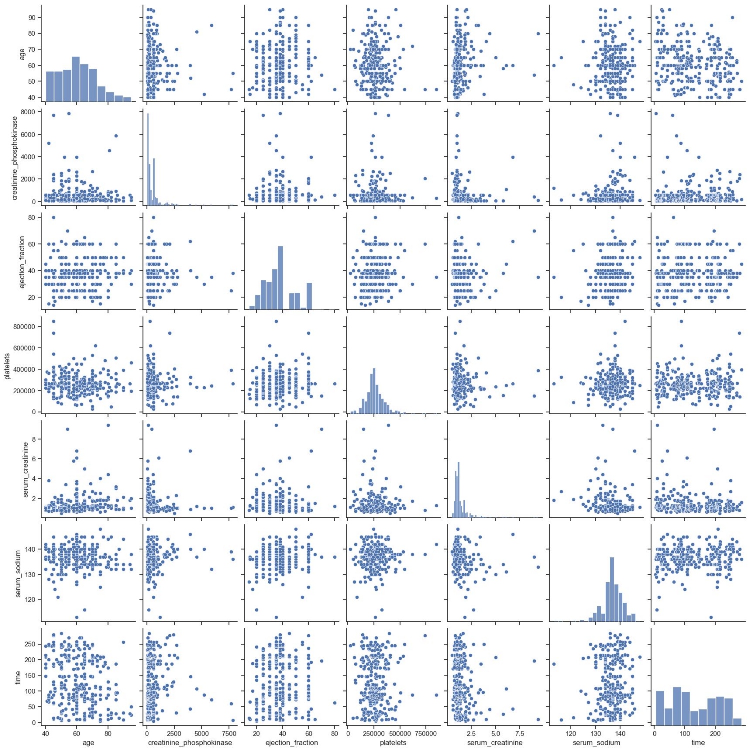
- There are 299 records
- 13 columns = 12 features + 1 target variable: DEATH_EVENT
- 6 encoded categorical variables and remaining 6 numeric variables
- 3 float columns and 9 integer columns
- No missing values

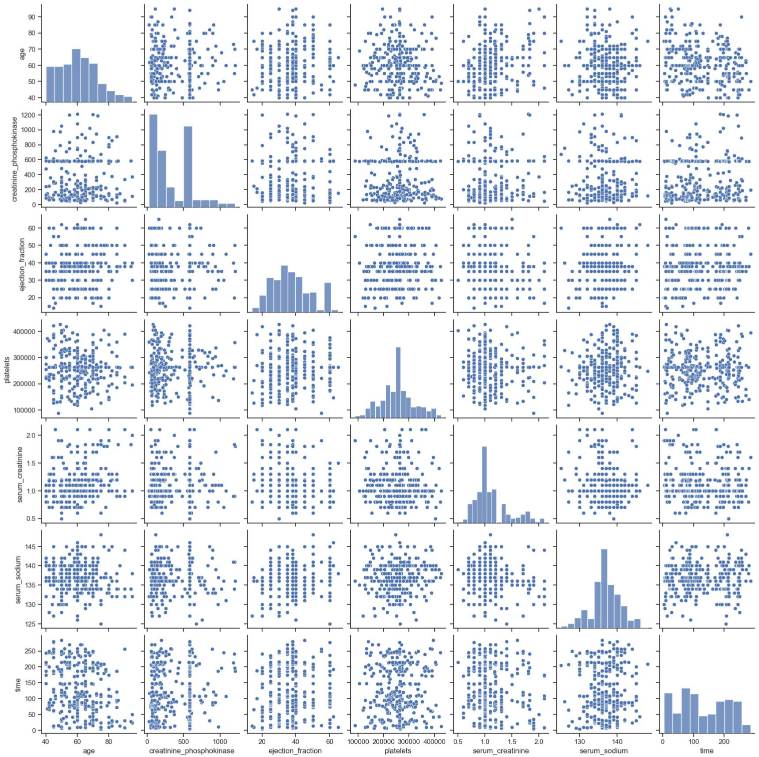


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64
 2   creatinine_phosphokinase  299 non-null    int64
 3   diabetes                  299 non-null    int64
 4   ejection_fraction         299 non-null    int64
 5   high_blood_pressure       299 non-null    int64
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64
 9   sex                       299 non-null    int64
 10  smoking                   299 non-null    int64
 11  time                      299 non-null    int64
 12  DEATH_EVENT               299 non-null    int64
dtypes: float64(3), int64(10)
memory usage: 30.5 KB
```

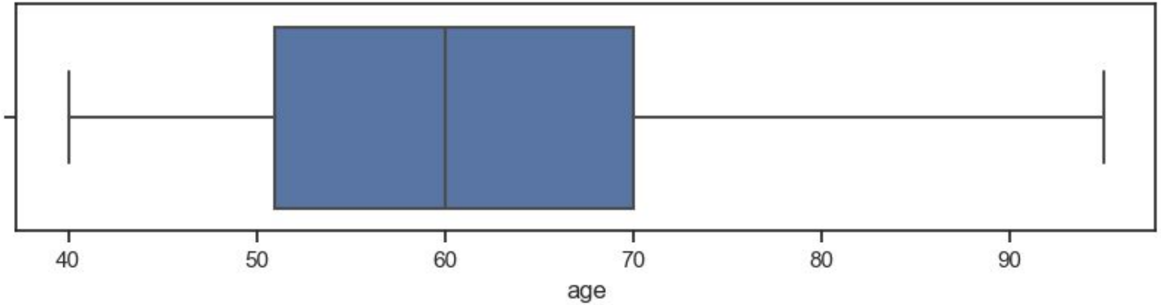
Numerical variable analysis

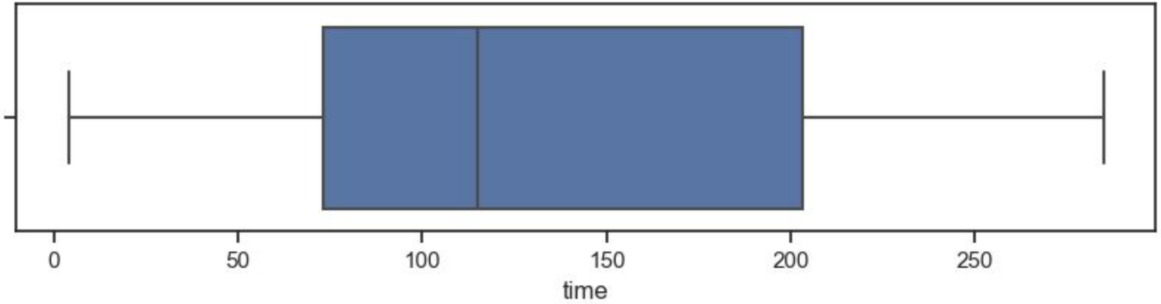
Before and after converting outliers to median



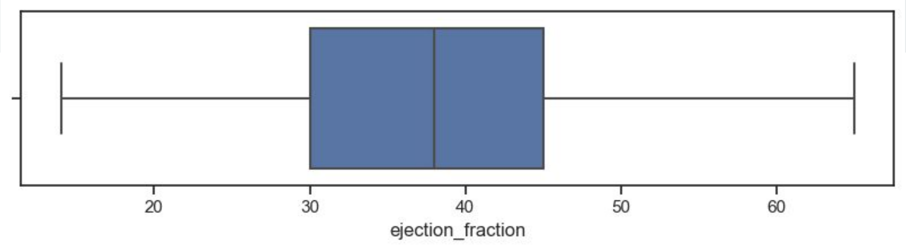


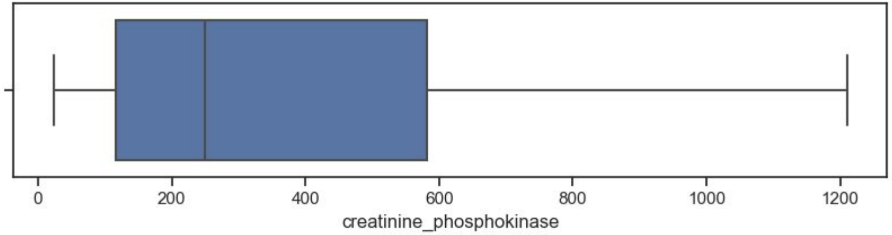
Boxplots

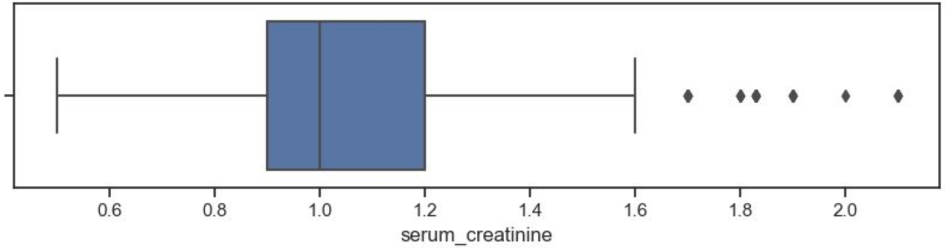


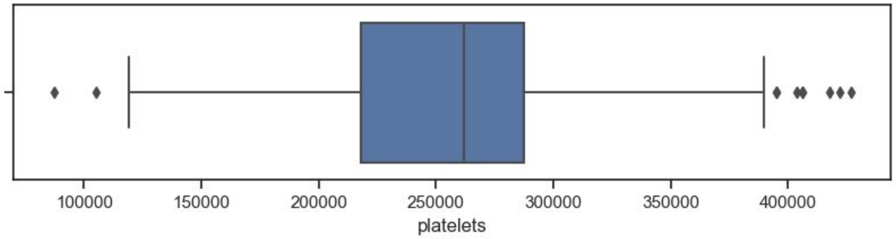


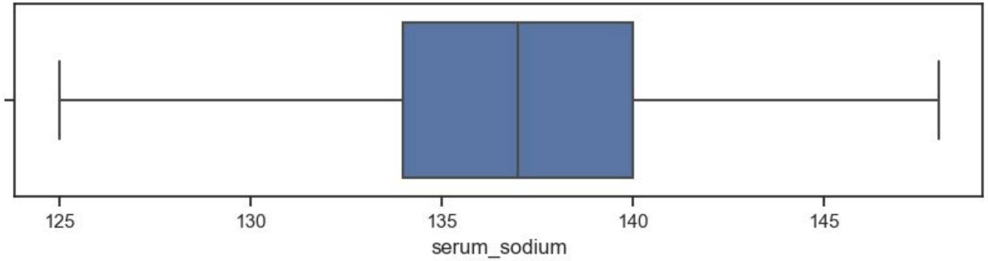
Blood content







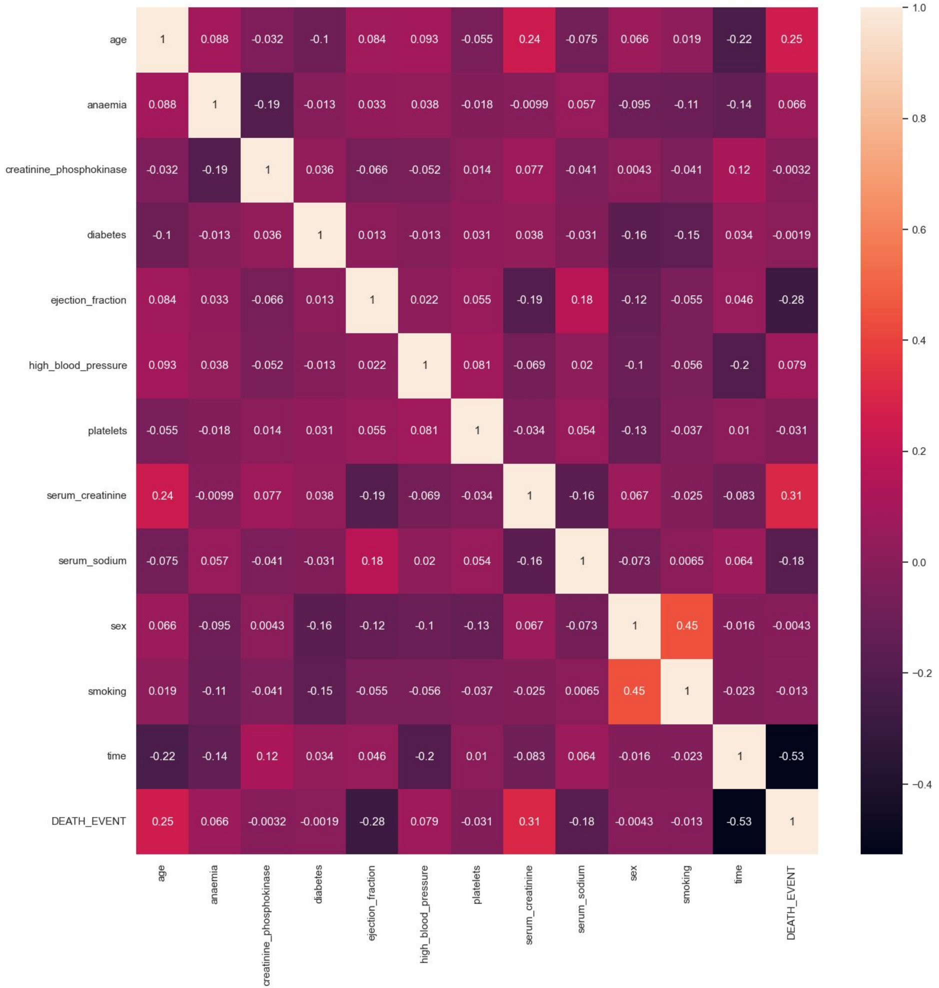




Correlation Matrix

Highest correlated variables:

- time: -0.53
- serum_creatinine: 0.29
- ejection_fraction: -0.27
- age: 0.25
- serum_sodium: -0.2
- others below 0.1

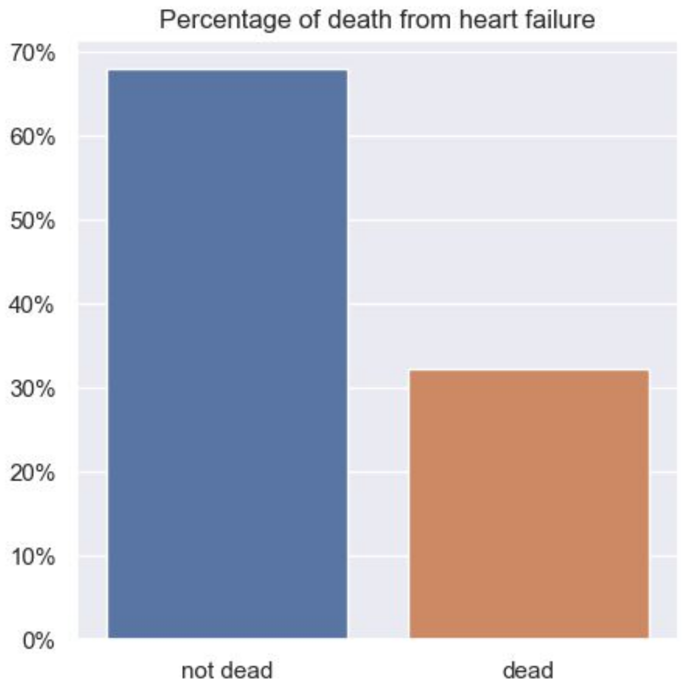


	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
age	1	0.088	-0.032	-0.1	0.084	0.093	-0.055	0.24	-0.075	0.066	0.019	-0.22	0.25
anaemia	0.088	1	-0.19	-0.013	0.033	0.038	-0.018	-0.0099	0.057	-0.095	-0.11	-0.14	0.066
creatinine_phosphokinase	-0.032	-0.19	1	0.036	-0.066	-0.052	0.014	0.077	-0.041	0.0043	-0.041	0.12	-0.0032
diabetes	-0.1	-0.013	0.036	1	0.013	-0.013	0.031	0.038	-0.031	-0.16	-0.15	0.034	-0.0019
ejection_fraction	0.084	0.033	-0.066	0.013	1	0.022	0.055	-0.19	0.18	-0.12	-0.055	0.046	-0.28
high_blood_pressure	0.093	0.038	-0.052	-0.013	0.022	1	0.081	-0.069	0.02	-0.1	-0.056	-0.2	0.079
platelets	-0.055	-0.018	0.014	0.031	0.055	0.081	1	-0.034	0.054	-0.13	-0.037	0.01	-0.031
serum_creatinine	0.24	-0.0099	0.077	0.038	-0.19	-0.069	-0.034	1	-0.16	0.067	-0.025	-0.083	0.31
serum_sodium	-0.075	0.057	-0.041	-0.031	0.18	0.02	0.054	-0.16	1	-0.073	0.0065	0.064	-0.18
sex	0.066	-0.095	0.0043	-0.16	-0.12	-0.1	-0.13	0.067	-0.073	1	0.45	-0.016	-0.0043
smoking	0.019	-0.11	-0.041	-0.15	-0.055	-0.056	-0.037	-0.025	0.0065	0.45	1	-0.023	-0.013
time	-0.22	-0.14	0.12	0.034	0.046	-0.2	0.01	-0.083	0.064	-0.016	-0.023	1	-0.53
DEATH_EVENT	0.25	0.066	-0.0032	-0.0019	-0.28	0.079	-0.031	0.31	-0.18	-0.0043	-0.013	-0.53	1

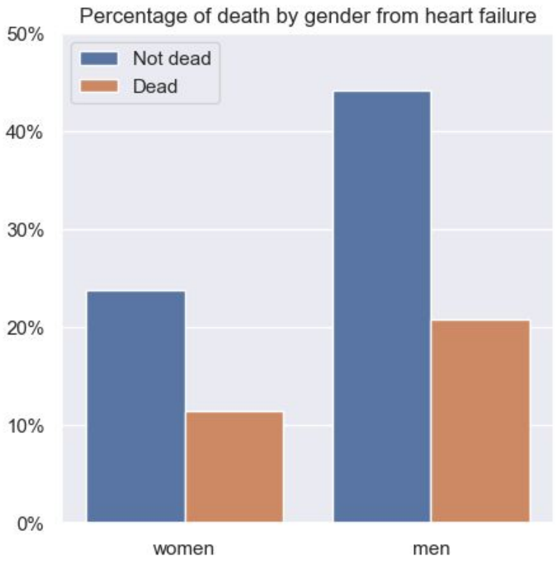
Categorical variable analysis

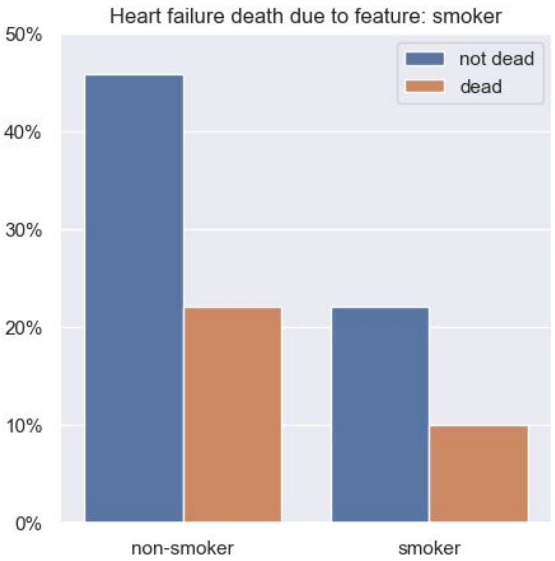
imbalance of dataset :

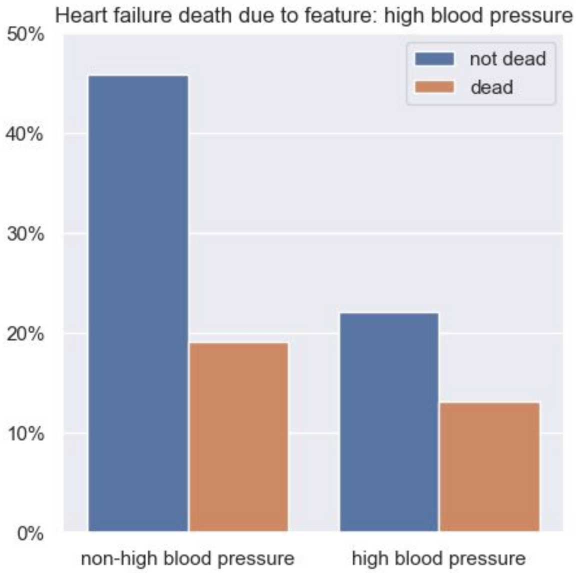
-->model might be better in predicting the outcome for people who didn't die

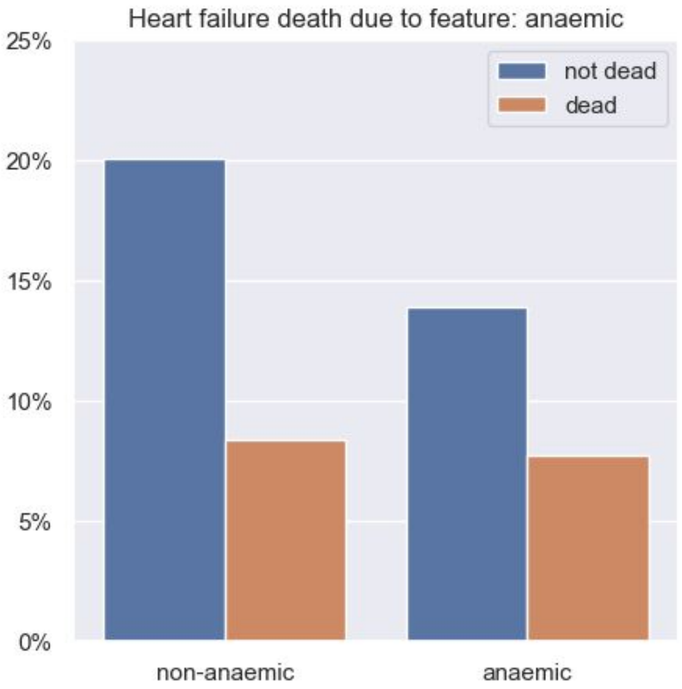


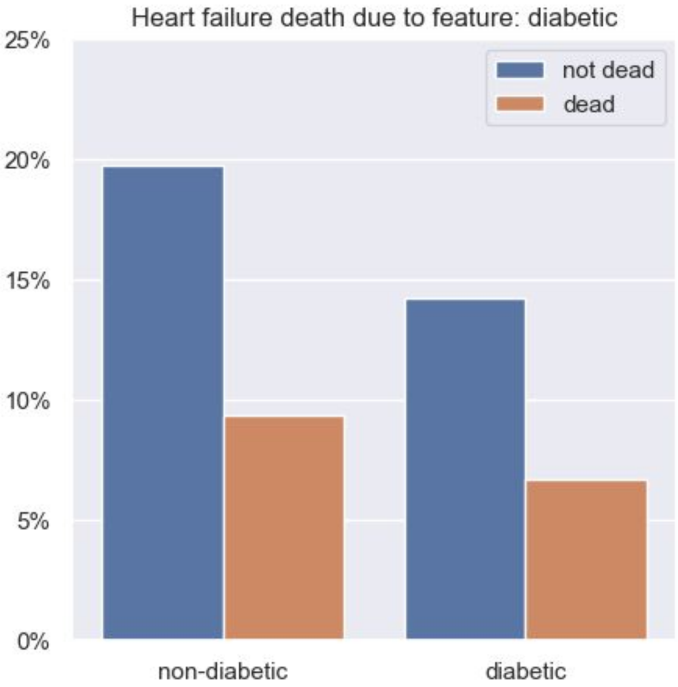
Profile of person with likelihood of heart failure





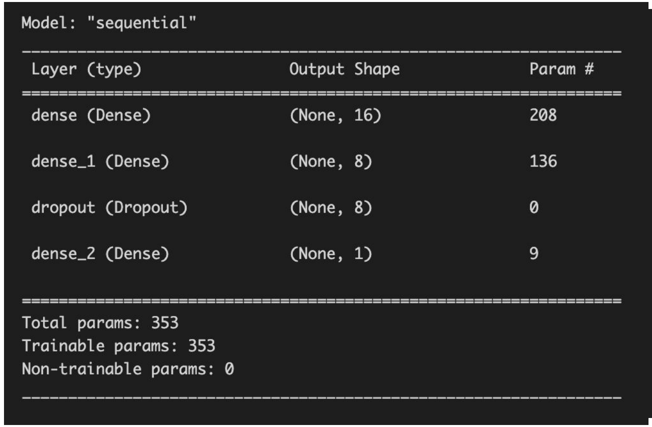




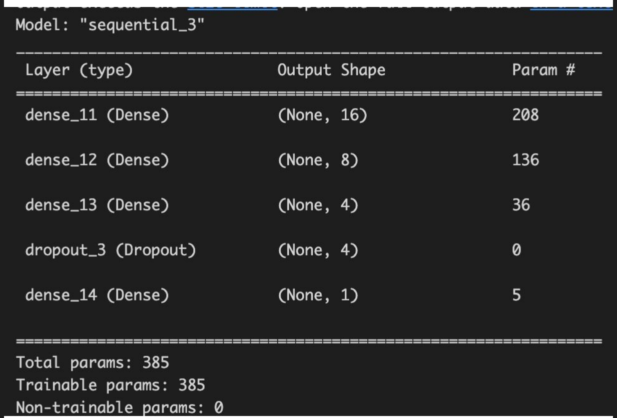


Non-conclusive: as the imbalance of data could affect

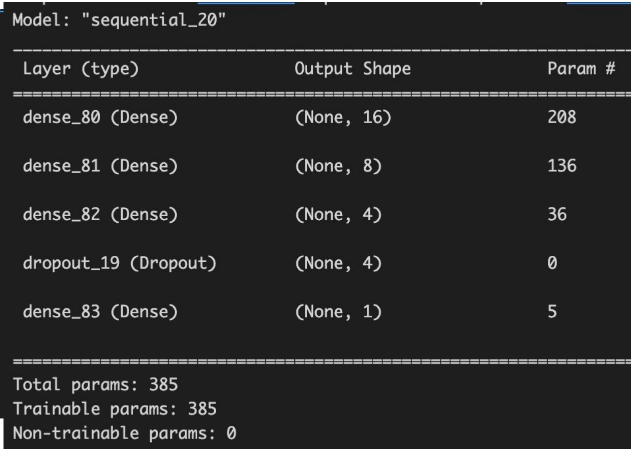
The 3 Neural Network models



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 16)                208

 dense_1 (Dense)             (None, 8)                 136

 dropout (Dropout)           (None, 8)                 0

 dense_2 (Dense)             (None, 1)                 9

=================================================================
Total params: 353
Trainable params: 353
Non-trainable params: 0
_________________________________________________________________
```



```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_11 (Dense)            (None, 16)                208

 dense_12 (Dense)            (None, 8)                 136

 dense_13 (Dense)            (None, 4)                 36

 dropout_3 (Dropout)         (None, 4)                 0

 dense_14 (Dense)            (None, 1)                 5

=================================================================
Total params: 385
Trainable params: 385
Non-trainable params: 0
```



```
Model: "sequential_20"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_80 (Dense)            (None, 16)                208

 dense_81 (Dense)            (None, 8)                 136

 dense_82 (Dense)            (None, 4)                 36

 dropout_19 (Dropout)        (None, 4)                 0

 dense_83 (Dense)            (None, 1)                 5

=================================================================
Total params: 385
Trainable params: 385
Non-trainable params: 0
```

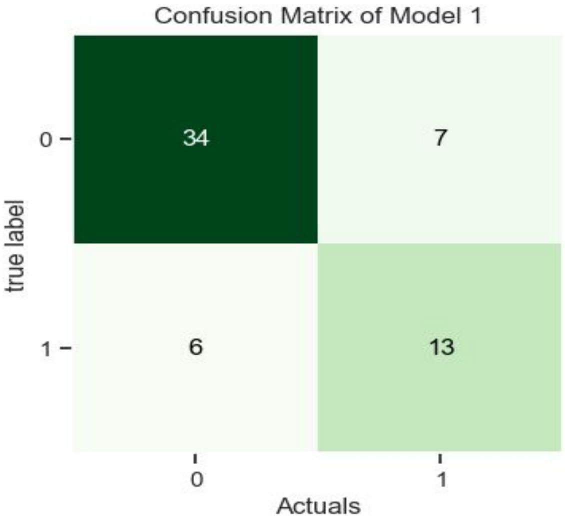
Results from models:

Model 1: loss: 0.3402 - accuracy: 0.8436 - val_loss: 0.3690 - val_accuracy: 0.8500

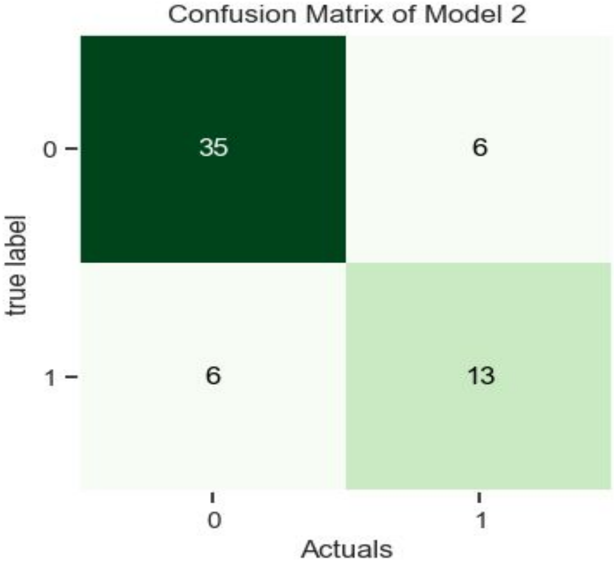
Model 2: loss: 0.4240 - accuracy: 0.8492 - val_loss: 0.4184 - val_accuracy: 0.8667

Model 3: loss: 0.4021 - accuracy: 0.8715 - val_loss: 0.4017 - val_accuracy: 0.8667

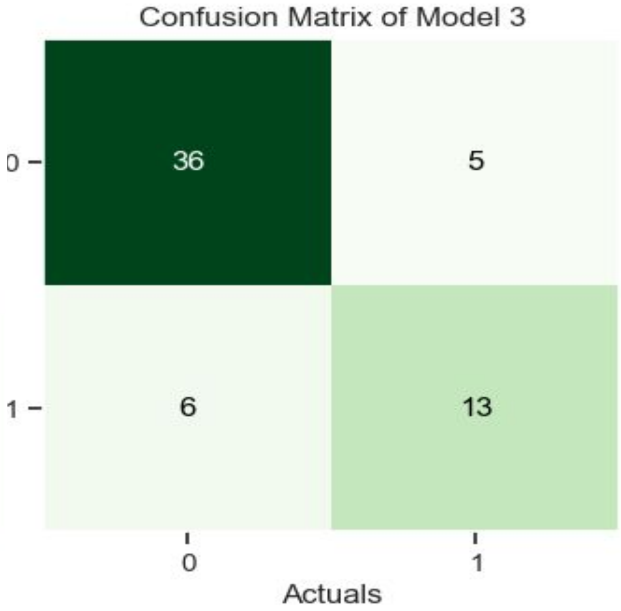
Evaluation of the models: confusion matrix



	0	1
0	34	7
1	6	13



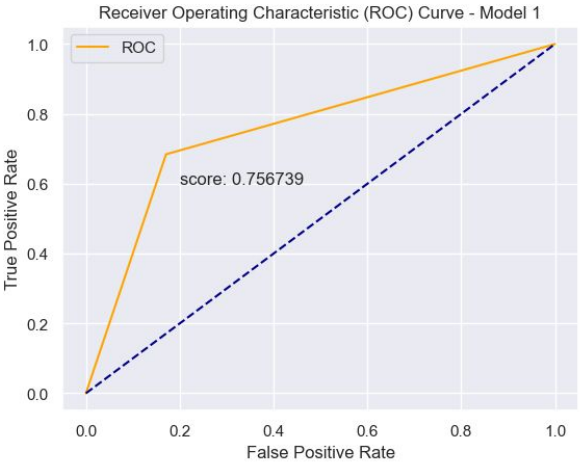
	0	1
0	35	6
1	6	13

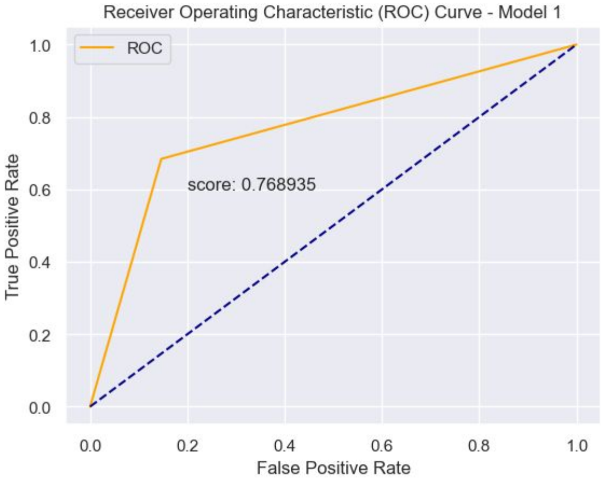


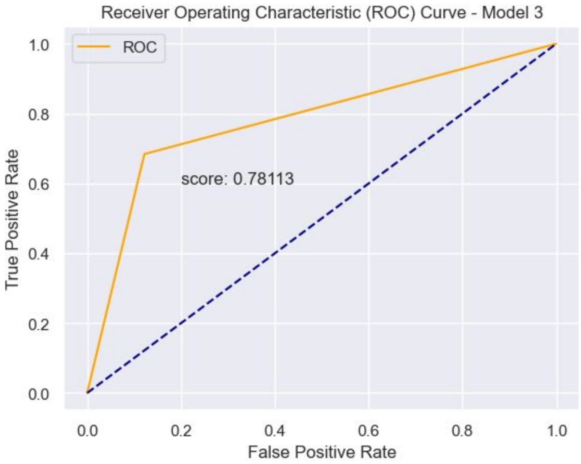
	0	1
0	36	5
1	6	13

The confusion matrices of the 3 models predicted the True Negatives the most well and this could be due to the imbalance of data having more data about people who didn't die from heart attack.

ROC


Receiver Operating Characteristic (ROC) Curve - Model 1


Receiver Operating Characteristic (ROC) Curve - Model 1


Receiver Operating Characteristic (ROC) Curve - Model 3

Conclusion

Model 3 is best.

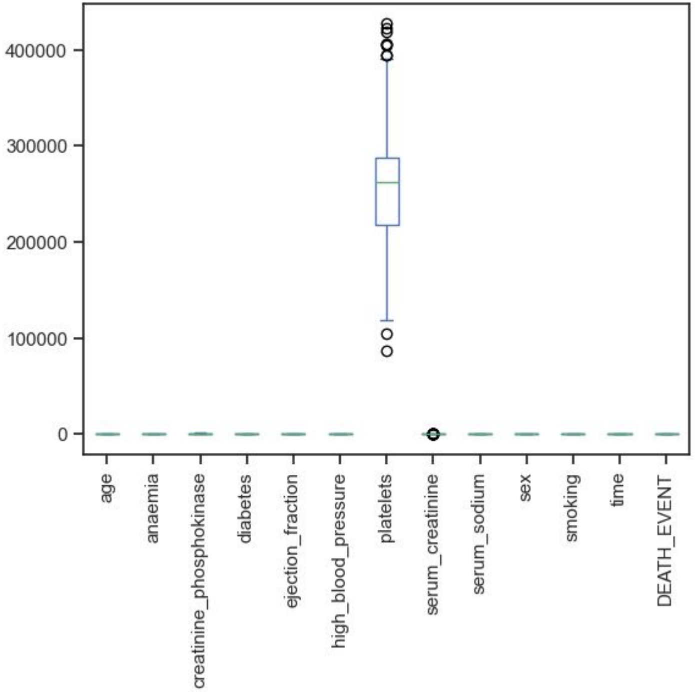
As confusion matrix and ROC are both better.

Next step:

-remove all outliers from platelets that still lingered even after converting

-remove some data about the people who didn't die so more balanced data.

-experiment with hyperparatuning the models



THANK
YOU

Appendix

Github:

Date: 16/06/2023