CHALMERS
UNIVERSITY OF TECHNOLOGY

Statistical methods for data science

Heart failure prediction

Alejandro Corrochano
(Group 12)

Overview

3 **Data specification**

Data source, description of the data.

5 **Problem definition**

Reasoning behind the chosen topic.

6 **Descriptive analysis & probability distribution**

Exploration and analysis of the features contained within the data set.

19 **Predictive analysis**

Model building.

29 **Conclusions**

Final thoughts and results obtained from the research.

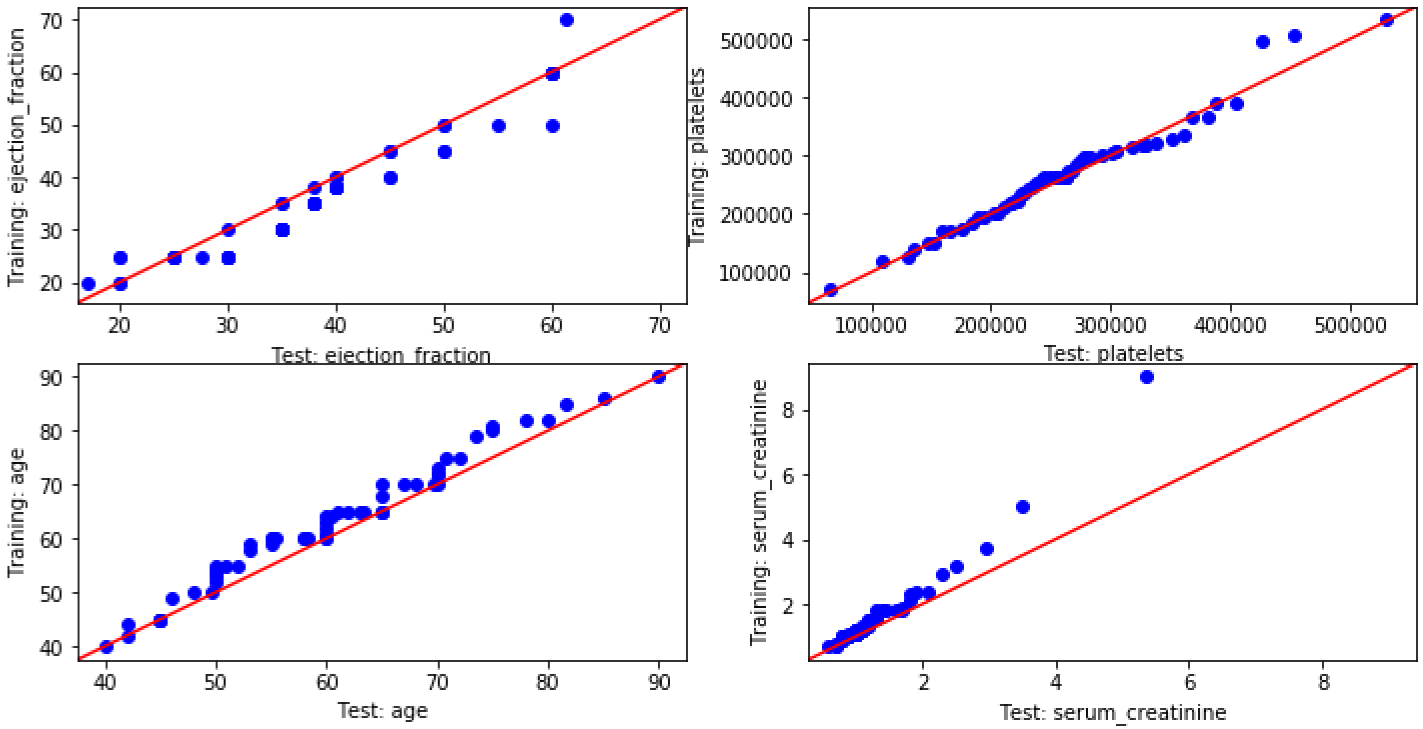
Data specification

- Data has been downloaded from UCI Machine Learning Repository, which is a collection of databases used by students, educators, and researchers from all over the world interested in machine learning and who want to put their knowledge into practice.

- The chosen data set contains the medical records of 299 patients who suffered from heart failure. This records were collected during a follow-up period, where each patient has the following 13 clinical features:

1. Age: Age of the patient (*Integer*)
2. Anaemia: Decrease of red blood cells or hemoglobin (*Boolean*)
3. High blood pressure: If patient has hypertension (*Boolean*)
4. Creatinine phosphokinase (CPK): Level of the CPK enzyme in the blood (*Float*)
5. Diabetes: If patient has diabetes (*Boolean*)
6. Ejection fraction: Percentage of blood leaving the heart in each contraction (*Float*)
7. Platelets: Platelets in blood (*Float*)
8. Sex: Woman or man (*Integer*)
9. Serum creatinine: Level of serum creatinine in the blood (*Float*)
10. Serum sodium: Level of serum sodium in the blood (*Float*)
11. Smoking: If the patient smokes or not (*Boolean*)
12. Time: Follow-up period (*Integer*)
13. Death_event (Target): If the patient deceased during the follow-up period (*Boolean*)

Data specification | Q-Q plots



We select the variables **ejection_fraction**, **platelets**, **age**, and **serum_creatinine** from training and test sets to check if the distribution of the data is similar.

Apparently, we could say that both **sets are following a similar distribution**.

Problem definition

"The ultimate goal is to build a model that **predicts** the patients survival, finding out which could be the most relevant **factors** that should be tackled first in order to prevent deceases with heart failure as the cause"

Descriptive analysis

We will start by doing some **exploratory data analysis (EDA)** to see what interesting insights we can retrieve from it. This will help us discover the possible **existing relationships** between features and to better select the ones that might have a **bigger impact** when predicting a decease.

Follow-up period

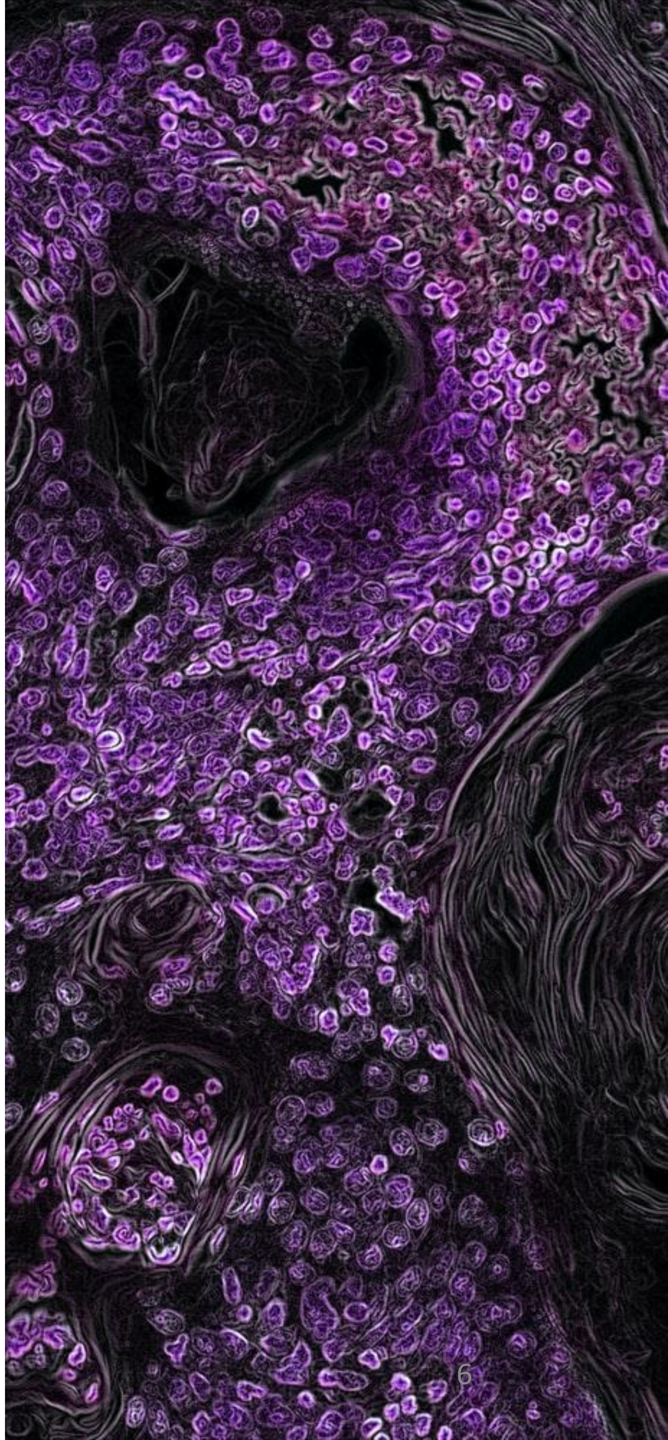
Age

Ejection-fraction

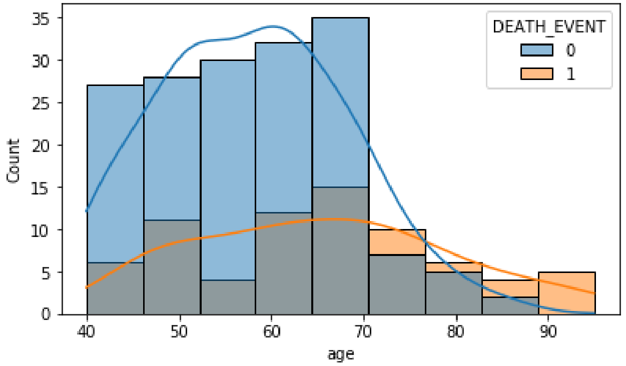
Serum creatinine

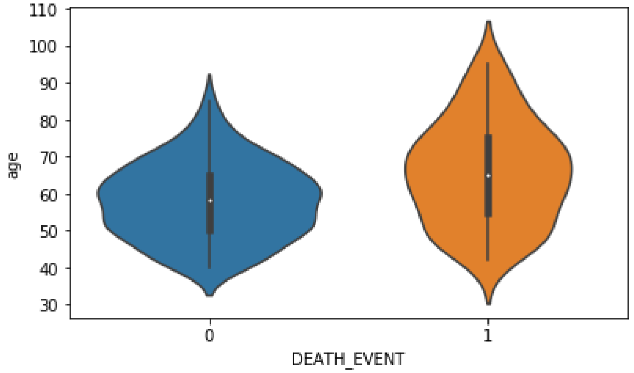
Serum sodium

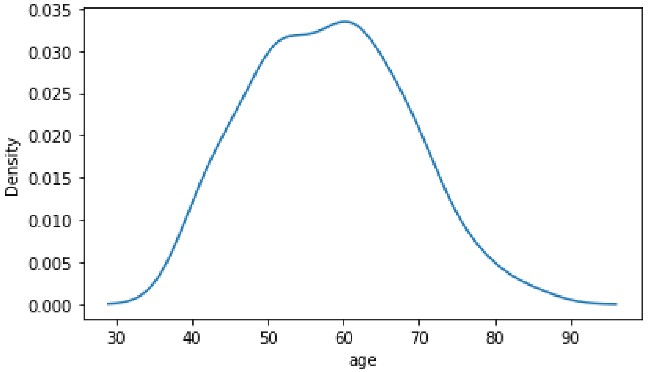
Platelets

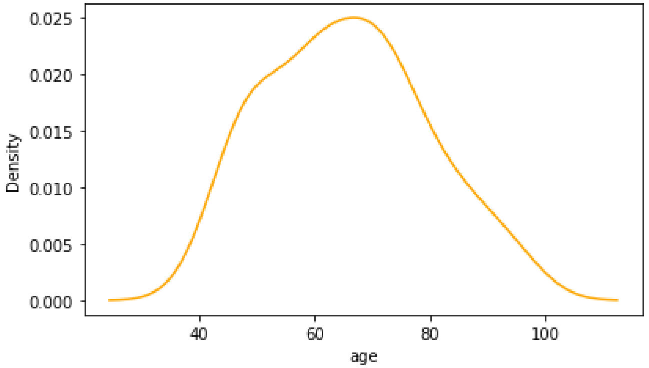


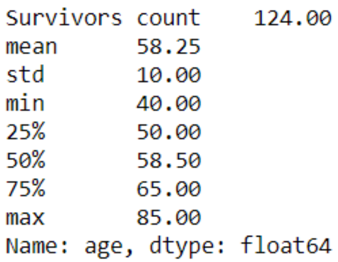
Descriptive analysis | Age



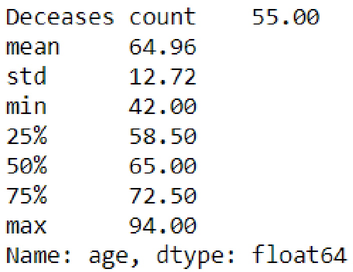








```
Survivors count    124.00
mean        58.25
std         10.00
min         40.00
25%         50.00
50%         58.50
75%         65.00
max         85.00
Name: age, dtype: float64
```

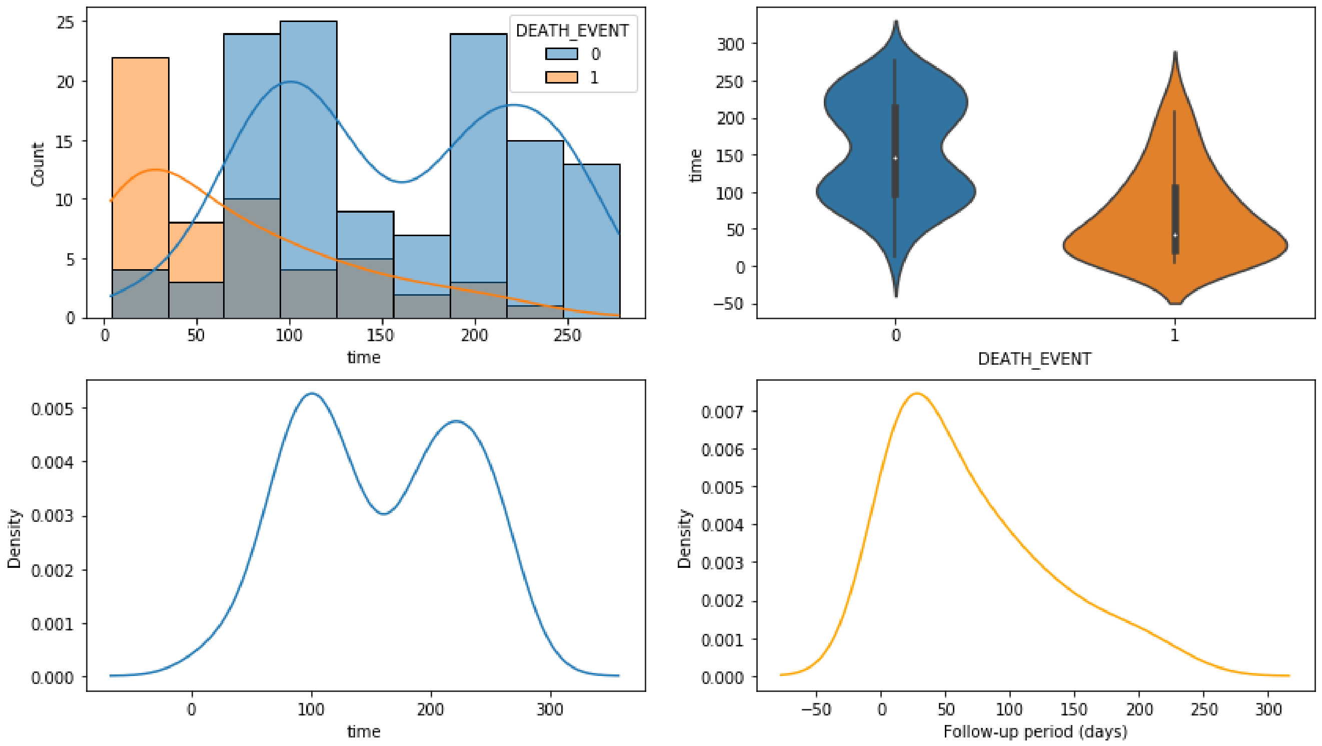


```
Deceases count    55.00
mean       64.96
std        12.72
min        42.00
25%        58.50
50%        65.00
75%        72.50
max        94.00
Name: age, dtype: float64
```

Insights

- Age of the patients oscillates between 40 and 100 years old.
- Data is revealing a higher probability of decease in those patients whose age is above the 75 years old.

Descriptive analysis | Follow-up period



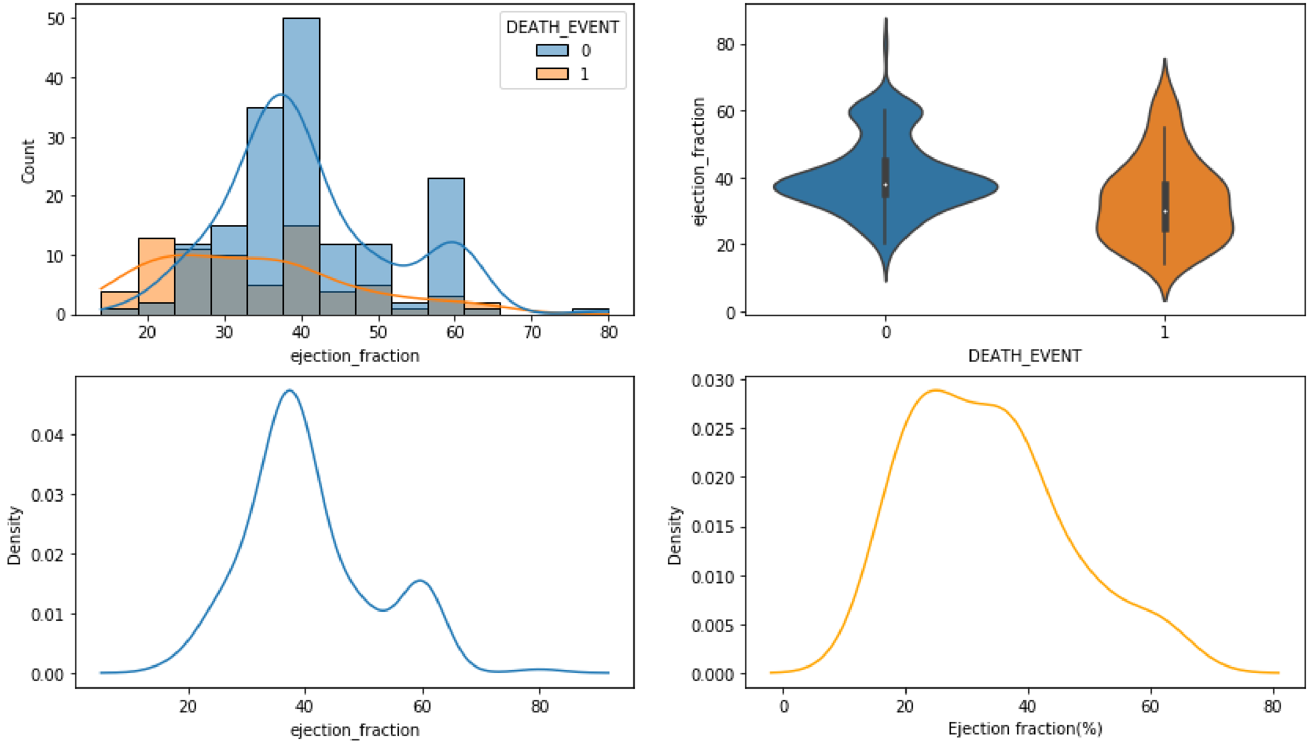
```
Survivors count    124.00
mean       157.49
std         69.09
min         12.00
25%         96.50
50%        146.50
75%        214.25
max        278.00
Name: time, dtype: float64
```

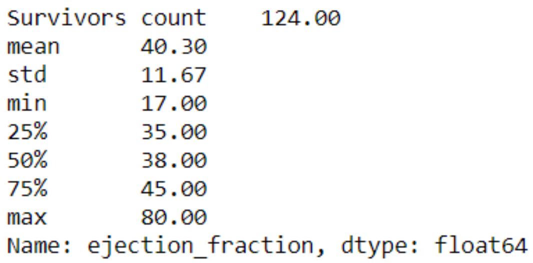
```
Deceases count     55.00
mean        69.15
std         60.40
min          4.00
25%         20.50
50%         43.00
75%        105.50
max        235.00
Name: time, dtype: float64
```

Insights

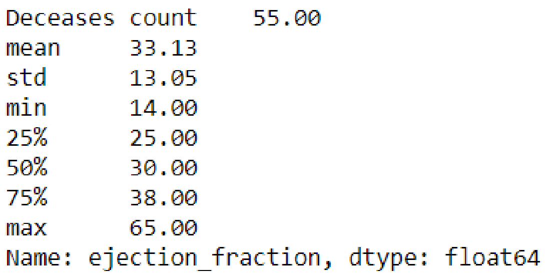
- Follow-up period indicates the **amount of days the patient has been under observation** after suffering from heart failure.
- Data reveals that patients who deceased are under observation **half the days** on average compared to those who did not.

Descriptive analysis | Ejection-fraction





```
Survivors count    124.00
mean       40.30
std        11.67
min        17.00
25%        35.00
50%        38.00
75%        45.00
max        80.00
Name: ejection_fraction, dtype: float64
```

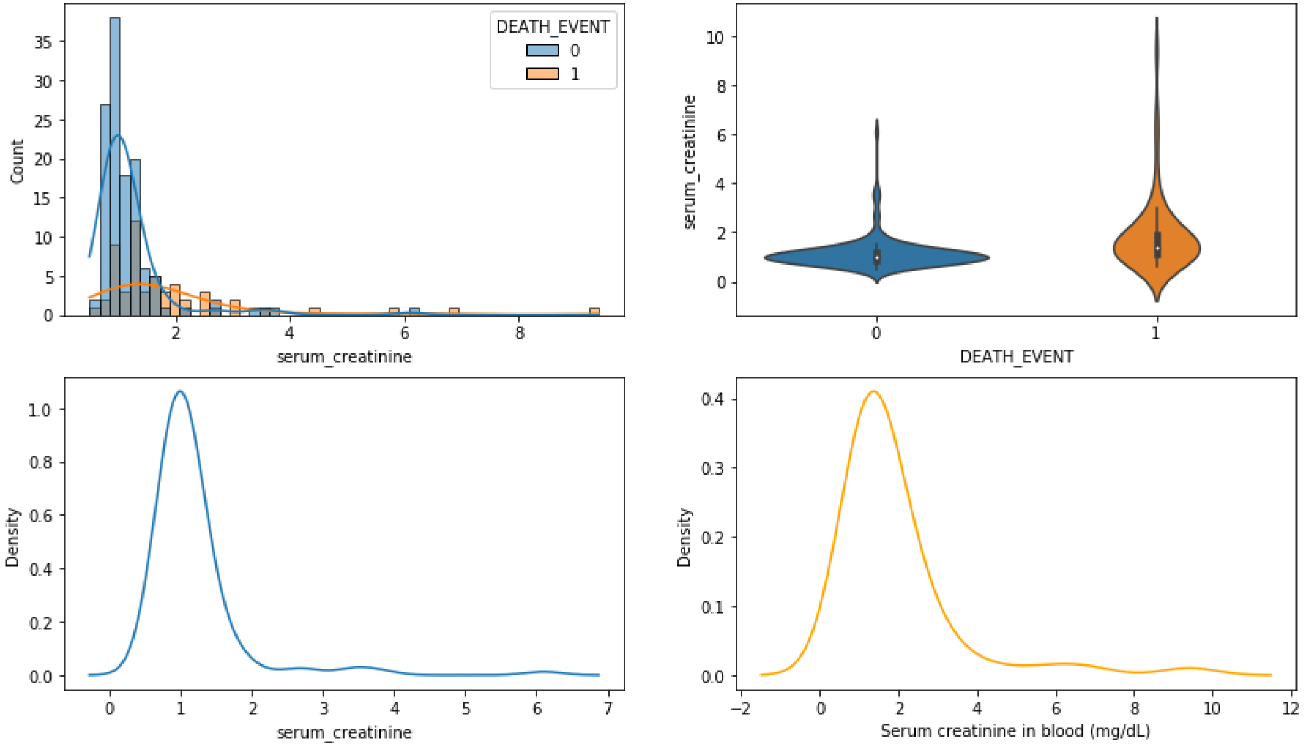


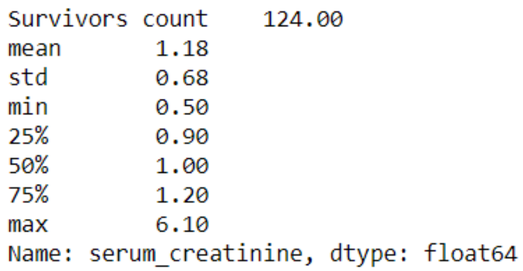
```
Deceases count     55.00
mean       33.13
std        13.05
min        14.00
25%        25.00
50%        30.00
75%        38.00
max        65.00
Name: ejection_fraction, dtype: float64
```

Insights

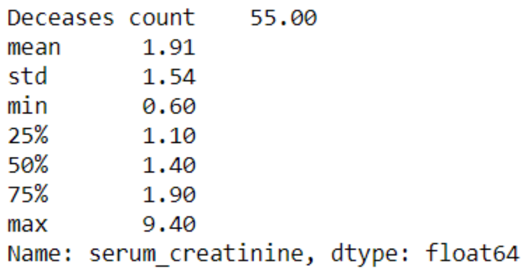
- Ejection fraction measures the amount of blood the left ventricle pumps out with each contraction, expressed in percentage.
- A normal heart's ejection fraction may be between **50-70%.**
- A ejection fraction measurement under **40%** could indicate heart failure.
- Data reveals that a lower percentage of blood was pumped in patients who deceased.

Descriptive analysis | Serum creatinine





```
Survivors count    124.00
mean           1.18
std            0.68
min            0.50
25%            0.90
50%            1.00
75%            1.20
max            6.10
Name: serum_creatinine, dtype: float64
```

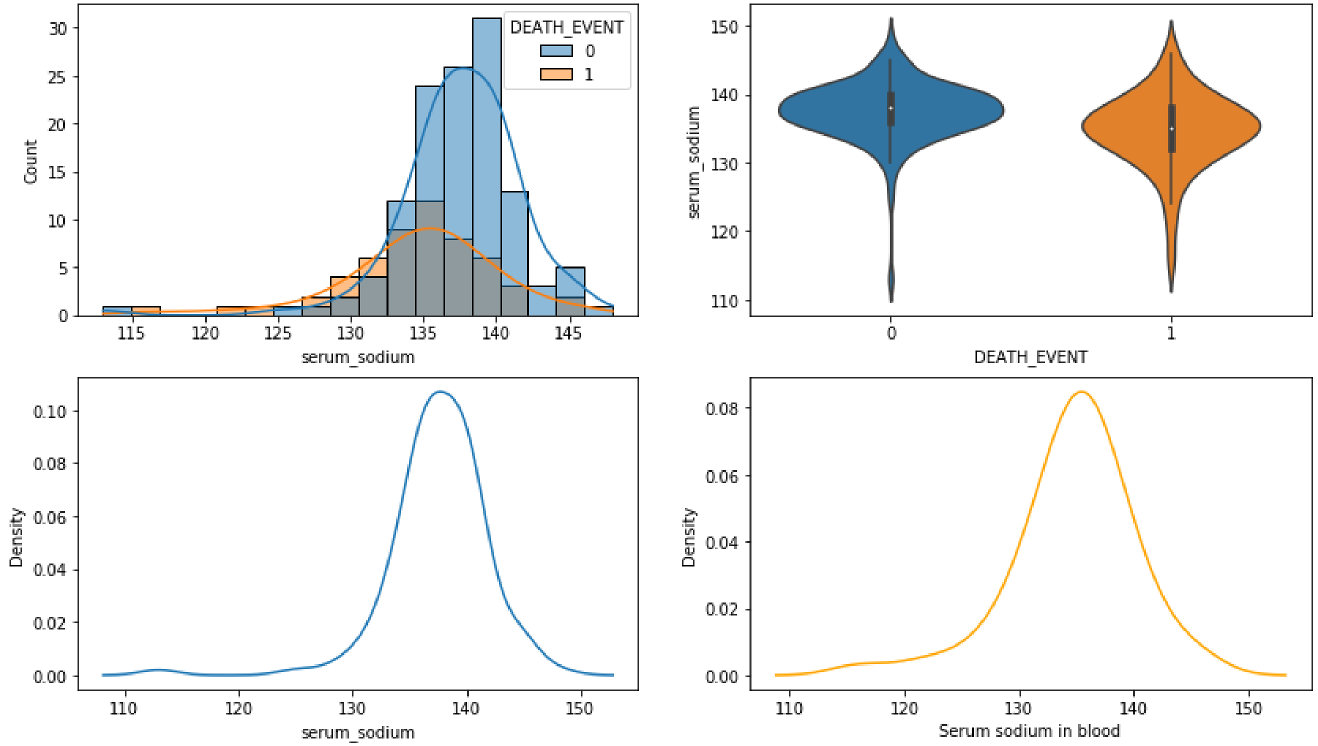


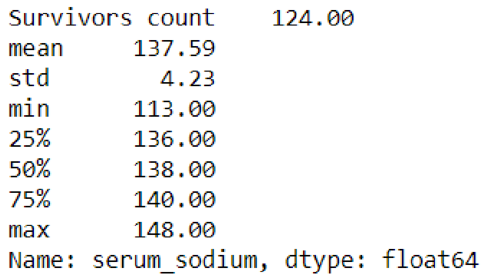
```
Deceases count    55.00
mean           1.91
std            1.54
min            0.60
25%            1.10
50%            1.40
75%            1.90
max            9.40
Name: serum_creatinine, dtype: float64
```

Insights

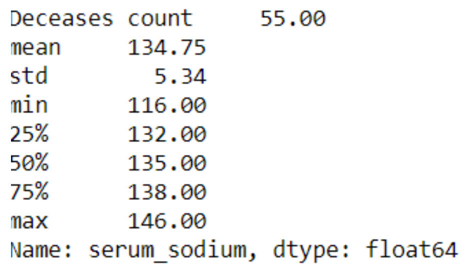
- Commonly used as an important **renal function indicator**, this blood measurement gives us information about creatinine concentration in blood, a compound generated from creatine consumption of muscles.
- Although creatinine levels vary depending on the patient's muscle mass, the typical reference ranges **are 0.5 – 1.0 mg/dL for women** and **0.7 mg/dL to 1.2 mg/dL for men**.
- Data reveals higher levels of creatinine in patients who deceased.

Descriptive analysis | Serum sodium





```
Survivors count    124.00
mean       137.59
std          4.23
min        113.00
25%        136.00
50%        138.00
75%        140.00
max        148.00
Name: serum_sodium, dtype: float64
```

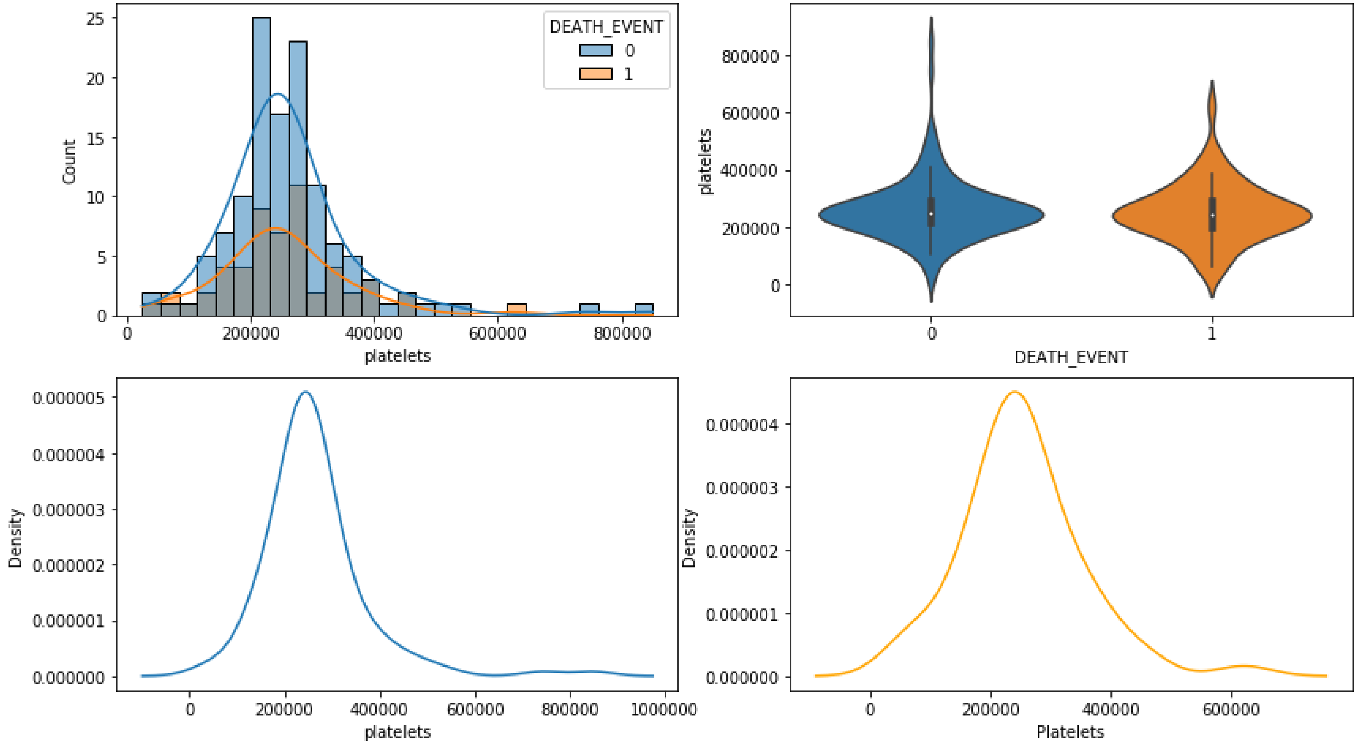


```
Deceases count     55.00
mean       134.75
std          5.34
min        116.00
25%        132.00
50%        135.00
75%        138.00
max        146.00
Name: serum_sodium, dtype: float64
```

Insights

- Sodium is important in how our nerves and muscles work.
- Sodium concentration in blood can be obtained by taking a blood test and indicates whether the patient has normal levels or not.
- Insufficient sodium in blood might be caused by heart or kidney failure among others.
- Sodium levels should be between 135 and 145 milliequivalents per litre (mEq/L).

Descriptive analysis | Platelets



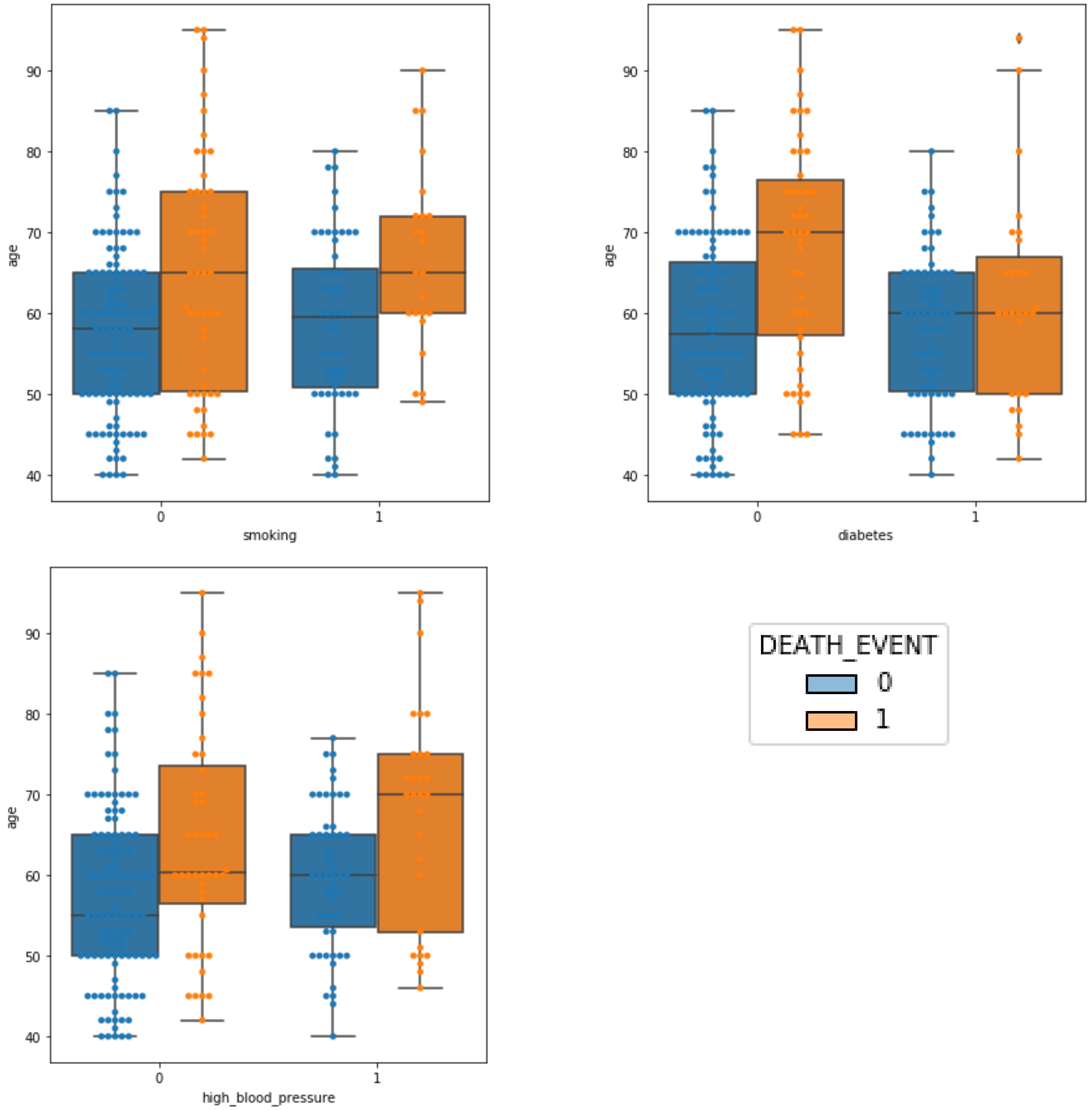
```
Survivors count       124.00
mean       261701.32
std        109282.07
min         25100.00
25%        215000.00
50%        249000.00
75%        294250.00
max        850000.00
Name: platelets, dtype: float64
```

```
Deceases count       55.00
mean       252607.86
std        101353.41
min         47000.00
25%        201000.00
50%        244000.00
75%        295000.00
max        621000.00
Name: platelets, dtype: float64
```

Insights

- Tiny blood cells that help us to stop the bleeding by forming clots.
- A normal platelet count is between **150,000 and 450,000 platelets** per microliter of blood.
- High number of platelets could lead to heart failure.
- Data does not reveal any meaningful insight regarding the number of platelets in those who deceased.

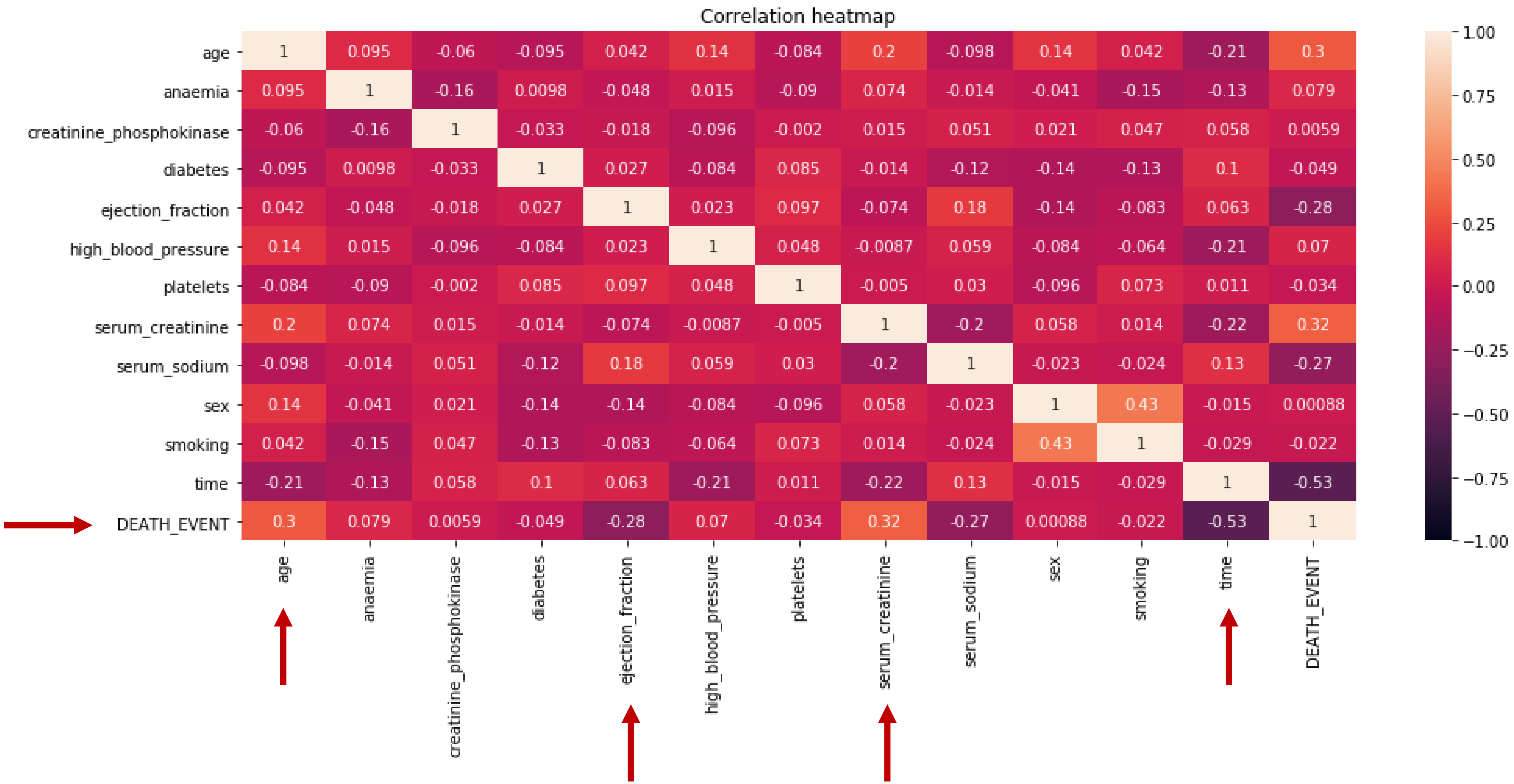
Descriptive analysis | Smoking, diabetes, and high blood pressure



Insights

- Patients who smoked **deceased at a similar age** to those who did not.
- Patients who suffered from diabetes **died at an earlier age** compared to those who did not suffered it.
- **Higher survival rate** in patients below 65 years old who suffered from high blood pressure.

Descriptive analysis | Correlation heatmap



Correlation heatmap

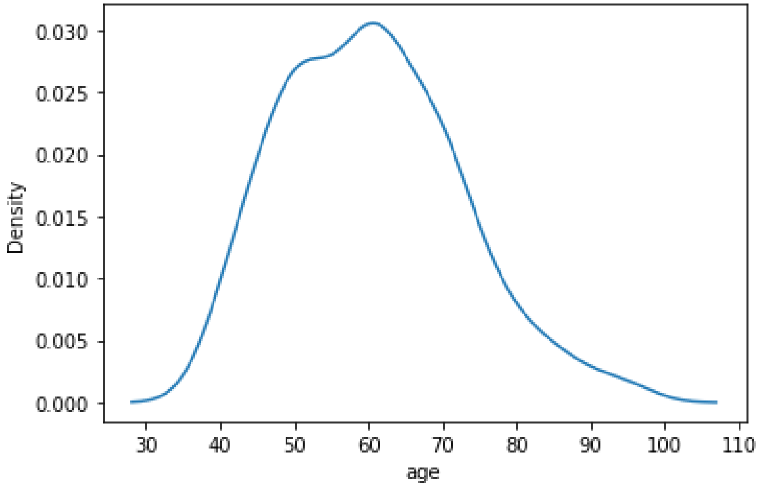
	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
age	1	0.095	-0.06	-0.095	0.042	0.14	-0.084	0.2	-0.098	0.14	0.042	-0.21	0.3
anaemia	0.095	1	-0.16	0.0098	-0.048	0.015	-0.09	0.074	-0.014	-0.041	-0.15	-0.13	0.079
creatinine_phosphokinase	-0.06	-0.16	1	-0.033	-0.018	-0.096	-0.002	0.015	0.051	0.021	0.047	0.058	0.0059
diabetes	-0.095	0.0098	-0.033	1	0.027	-0.084	0.085	-0.014	-0.12	-0.14	-0.13	0.1	-0.049
ejection_fraction	0.042	-0.048	-0.018	0.027	1	0.023	0.097	-0.074	0.18	-0.14	-0.083	0.063	-0.28
high_blood_pressure	0.14	0.015	-0.096	-0.084	0.023	1	0.048	-0.0087	0.059	-0.084	-0.064	-0.21	0.07
platelets	-0.084	-0.09	-0.002	0.085	0.097	0.048	1	-0.005	0.03	-0.096	0.073	0.011	-0.034
serum_creatinine	0.2	0.074	0.015	-0.014	-0.074	-0.0087	-0.005	1	-0.2	0.058	0.014	-0.22	0.32
serum_sodium	-0.098	-0.014	0.051	-0.12	0.18	0.059	0.03	-0.2	1	-0.023	-0.024	0.13	-0.27
sex	0.14	-0.041	0.021	-0.14	-0.14	-0.084	-0.096	0.058	-0.023	1	0.43	-0.015	0.00088
smoking	0.042	-0.15	0.047	-0.13	-0.083	-0.064	0.073	0.014	-0.024	0.43	1	-0.029	-0.022
time	-0.21	-0.13	0.058	0.1	0.063	-0.21	0.011	-0.22	0.13	-0.015	-0.029	1	-0.53
DEATH_EVENT	0.3	0.079	0.0059	-0.049	-0.28	0.07	-0.034	0.32	-0.27	0.00088	-0.022	-0.53	1

Inferential statistics | Probability distributions

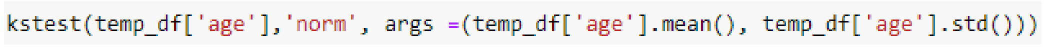
What kind of distribution are following some of the described features? Let’s try to figure out by making use of Kolmogorov-Smirnov test, commonly applied to verify whether a sample comes from a certain distribution.

Null hypothesis states that the empirical distribution will be similar to our theoretical proposal. On the other hand, alternative hypothesis will state the opposite. Significance level (α) will be set to 0.05.

Age

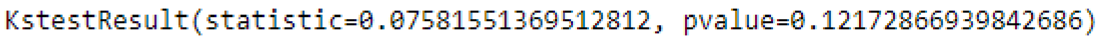


At first sight we could think Age distribution follows a Gaussian. Let’s check what KStest tells us about it.



```
kstest(temp_df['age'],'norm', args =(temp_df['age'].mean(), temp_df['age'].std()))
```

↓

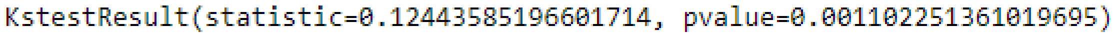


```
KstestResult(statistic=0.07581551369512812, pvalue=0.12172866939842686)
```

P-value is greater than 0.05, which means we fail to reject the null hypothesis and we can suggest that the distribution follows a Gaussian with mean 60.3 and std 11.9.

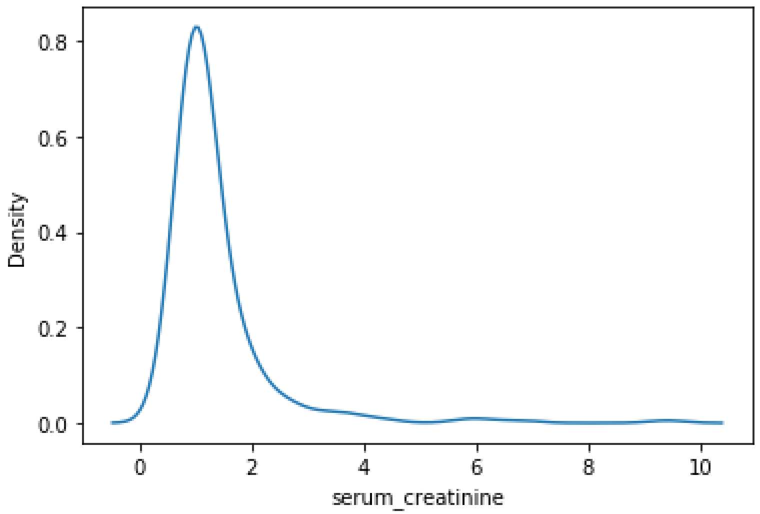
Inferential statistics| Probability distributions

Serum creatinine reminded us of a lognormal distribution. However, we rejected that possibility after executing the KStest.

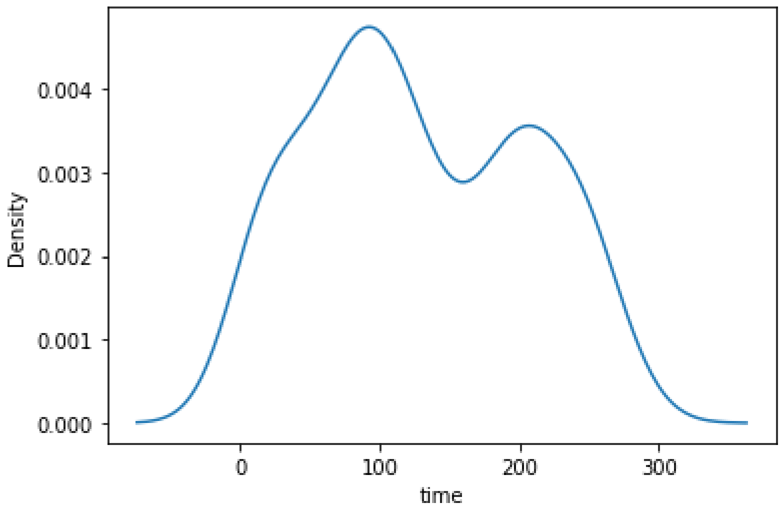


```
KstestResult(statistic=0.12443585196601714, pvalue=0.001102251361019695)
```

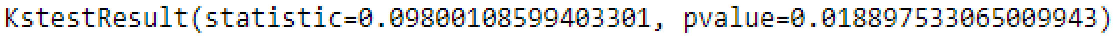
Serum creatinine



Follow-up period



For the follow-up period, the idea of a Gaussian distribution was discarded, as the p-value was lower than the significance level.



```
KstestResult(statistic=0.09800108599403301, pvalue=0.01889753306500943)
```

Inferential statistics| Diabetes and smoking

During the exploration of the data, we noticed that the average age of the patients who deceased that smoked and had diabetes was slightly lower than those who had diabetes but did not smoke.

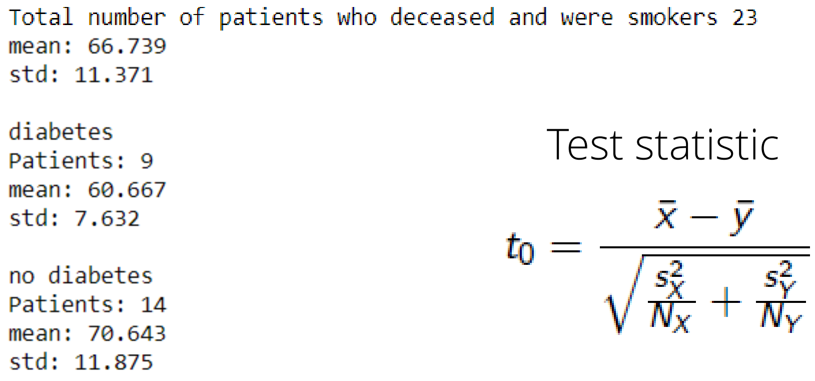
Is it plausible to think that people with diabetes who smoke have a lower life expectancy?

For this to be proven, we will carry out a two-sample t-test and see if we can confirm our hypothesis.

Let N_x be the number of patients who deceased and suffered from diabetes and were smokers, which sums up to 9 in total. On the other hand, N_y will be the patients who also deceased but did not suffer from diabetes, which are a total of 14 patients.

We calculate the mean value and standard deviation for each sample to let the reader appreciate the fact mentioned above. The test will be conducted by making use of the **ttest_ind** function, hence it will not be done manually.

Significance level (α) is set to **0.05**



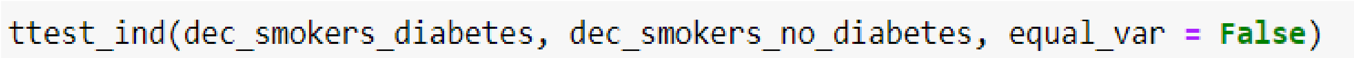
```
Total number of patients who deceased and were smokers 23
mean: 66.739
std: 11.371

diabetes
Patients: 9
mean: 60.667
std: 7.632

no diabetes
Patients: 14
mean: 70.643
std: 11.875
```

Test statistic

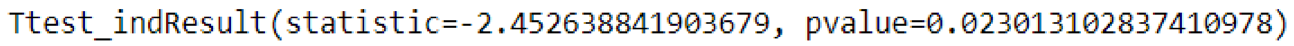
$$t_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y}}}$$

Inferential statistics| Diabetes and smoking

We can partially confirm our hypothesis, as apparently, we have some evidence to suggest that the smoking effect in people who suffer from diabetes might worsen their health and reduce life expectancy.



```
ttest_ind(dec_smokers_diabetes, dec_smokers_no_diabetes, equal_var = False)
```



```
Ttest_indResult(statistic=-2.452638841903679, pvalue=0.023013102837410978)
```

The p-value obtained equals to **0.02**. Considering that the significance level was set to 0.05 we can confirm our alternative hypothesis, although we would need much more information to firmly state it.

Predictive analysis

For this project we will apply **two machine learning algorithms** that are related with the concepts that have been acquired during the course.

Gaussian Naive Bayes

K-nearest neighbours

Predictive analysis| Gaussian Naive Bayes

- Prediction y: Patient will decease or not.
- Variables X = $[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}]$: As seen in the previous step, the variables that could help us build a more accurate model according to their correlation against the target are the following:
  - Age ($\mathbf{X_1}$)
  - Ejection-fraction ($\mathbf{X_2}$)
  - Serum creatinine ($\mathbf{X_3}$)
  - Follow-up period ($\mathbf{X_4}$)

- Assumptions:
  - Variables are independent
  - Data is generated from a Gaussian distribution

$$f_i(x_i \mid c) = \frac{1}{\sqrt{2\pi\sigma^2_{c,i}}} e^{-\frac{(x_i-\mu_{c,i})^{2}}{2\sigma^{2}_{c,i}}}$$

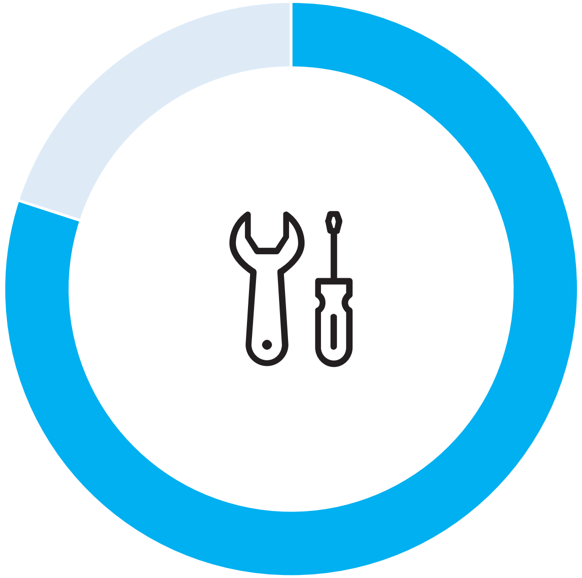
- Model g: Gaussian naive Bayes classifier

$$\hat{y} = \underset{c\in\{\text{No heart failure, heart failure}\}}{\arg\max} P(c) \prod_{i=1}^{d} f_i(x_i \mid c)$$

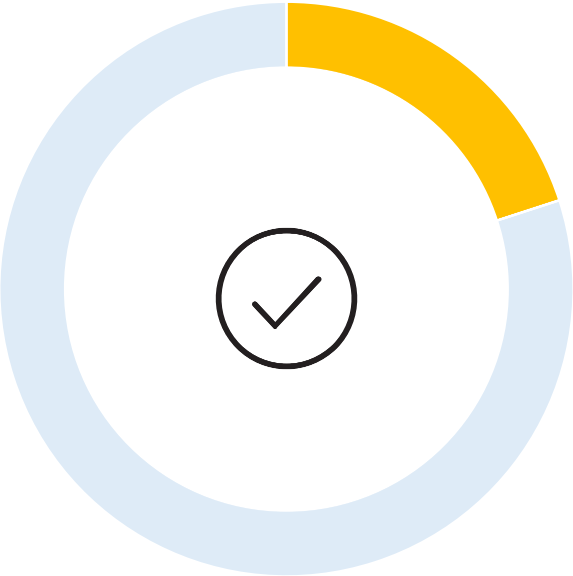
- Parameters θ : P(c) and $\mu_{c,i}$, $\sigma_{c,i}$ in the likelihood $f_i\,(x_i| \, c)$ for every variable i and class c

Predictive analysis| Gaussian Naive Bayes

We will divide data into 2 sets



Training set
80%



Test set
20%

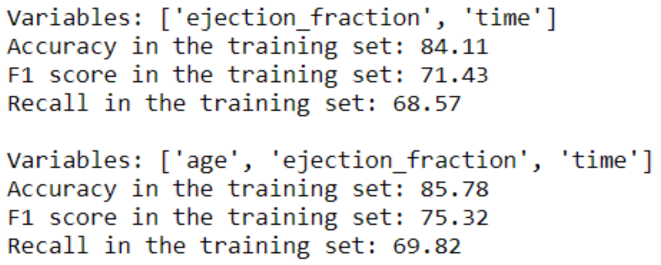
Predictive analysis| Gaussian Naive Bayes

Although we know which features hold a higher correlation against the target, we are not sure if this will ensure us the best performance.

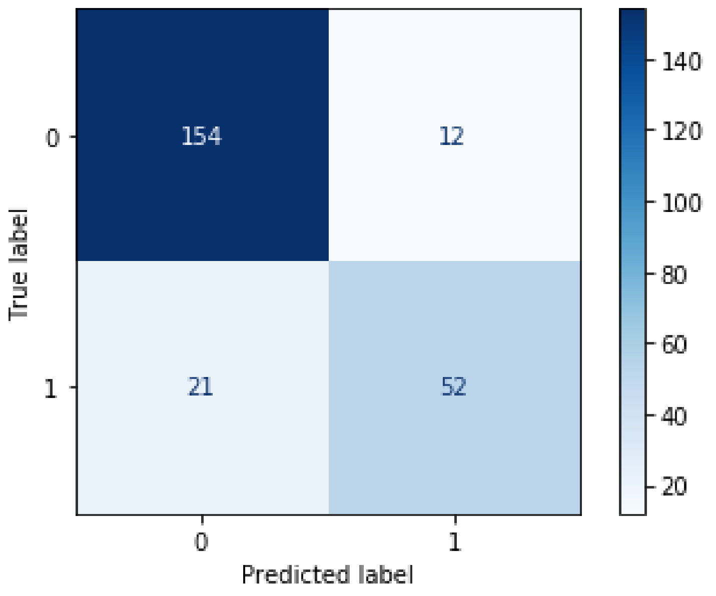
This reasoning leads us to evaluate all the possible combinations of the input features and check which model works better. It is computationally more expensive than other methods, but considering **the number of features is small**, we can accept it.

Among all the possible outcomes, we find out that there are two combinations whose scores are very close to each other.

We end up selecting the **age of the patient**, **ejection fraction indicator**, and **follow-up period**.

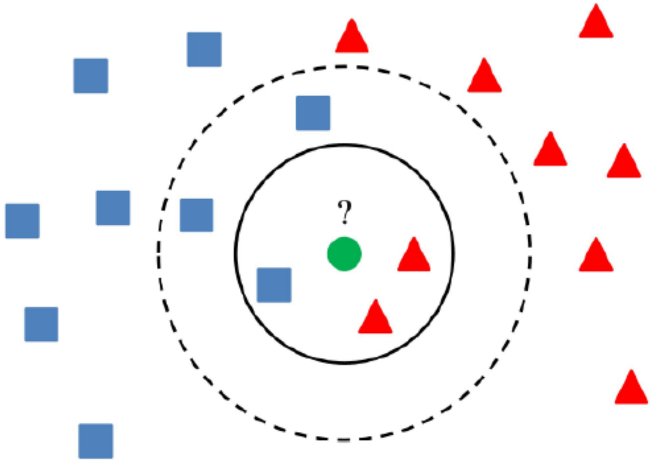


```
Variables: ['ejection_fraction', 'time']
Accuracy in the training set: 84.11
F1 score in the training set: 71.43
Recall in the training set: 68.57

Variables: ['age', 'ejection_fraction', 'time']
Accuracy in the training set: 85.78
F1 score in the training set: 75.32
Recall in the training set: 69.82
```



Predictive analysis| K-nearest neighbours

- Prediction y: Whether the patient will decease or not.
- Variables X = $[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}]$: Same variables as in the previous model:
    - Age ($\mathbf{X_1}$)
    - Ejection-fraction ($\mathbf{X_2}$)
    - Follow-up period ($\mathbf{X_3}$)



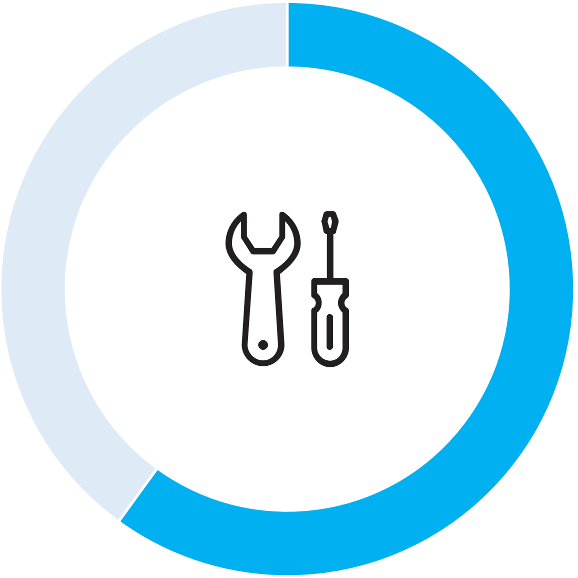
- Model g: K-nearest neighbours. A distance-based algorithm that picks the K closest points to the new one by computing the Euclidean distance. Each of these are labelled with one of the existing classes, and depending on which is more repeated within the K sized subset, will be the label assigned to the new point.

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (p_k - q_k)^2} = \sqrt{\sum_{i=1}^{k} (q_i - p_i)^2}$$

- Hyperparameters k: The number of neighbours, K.

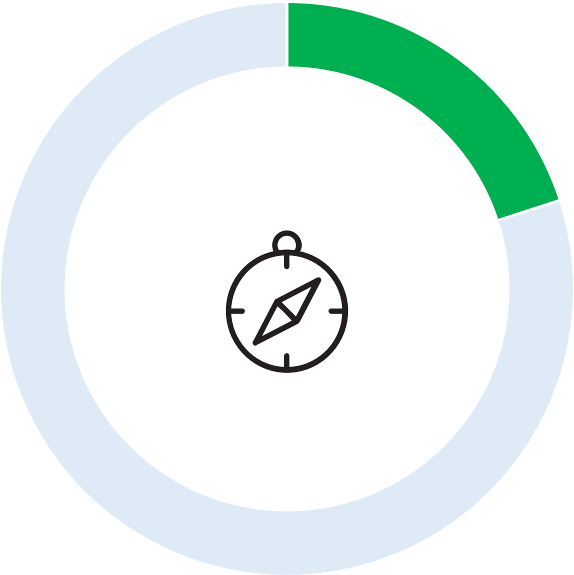
Predictive analysis| K-nearest neighbours

This time we will divide data into 3 different sets.



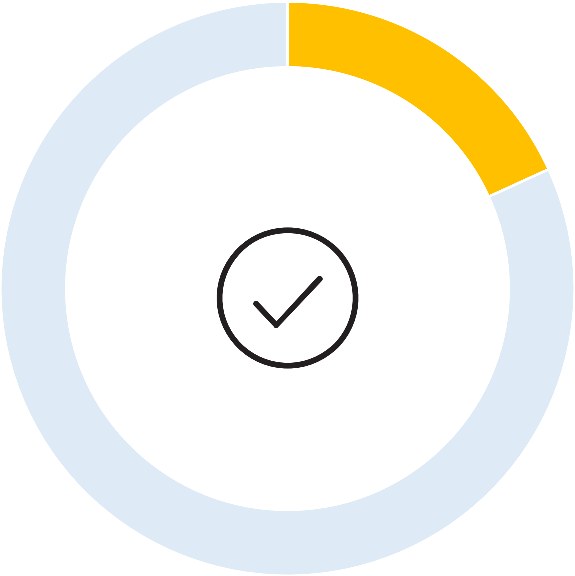
Training set

60%



Validation set

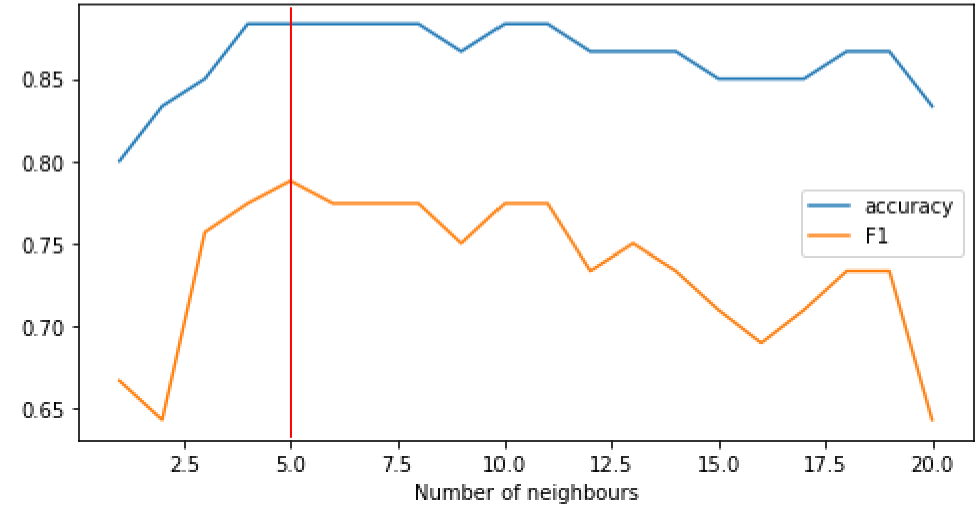
20%

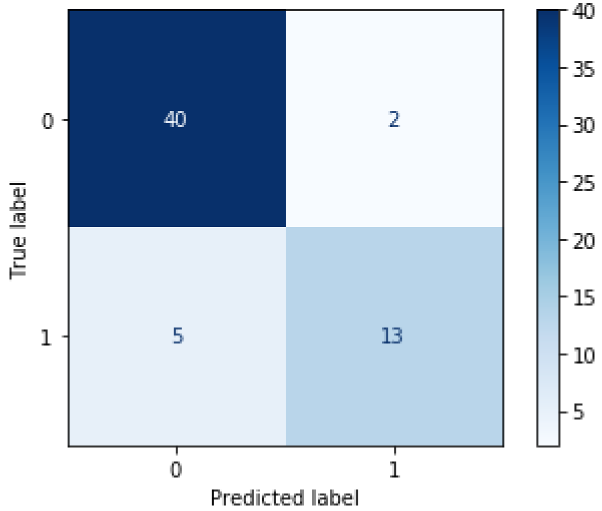


Test set

20%

Predictive analysis| K-nearest neighbours





```
Accuracy: 88.0
F1 score: 79.0
```

For the KNN model, we decided to work with the same features as the ones used in the gaussian naive bayes model, which were **age**, **ejection fraction**, and **follow-up period.**

Data has been normalized beforehand, avoiding this way the algorithm to be biased towards a feature with a higher magnitude. In this particular case, there is no such a big difference between variables, but it is always appropriate when we work with distance-based models.

Hyperparameter K has been tuned by iterating through the 20% of the total data we kept for the validation set.

Every accuracy and F1 score has been plotted for each number of neighbours, obtaining the best result when this value equals to 5.

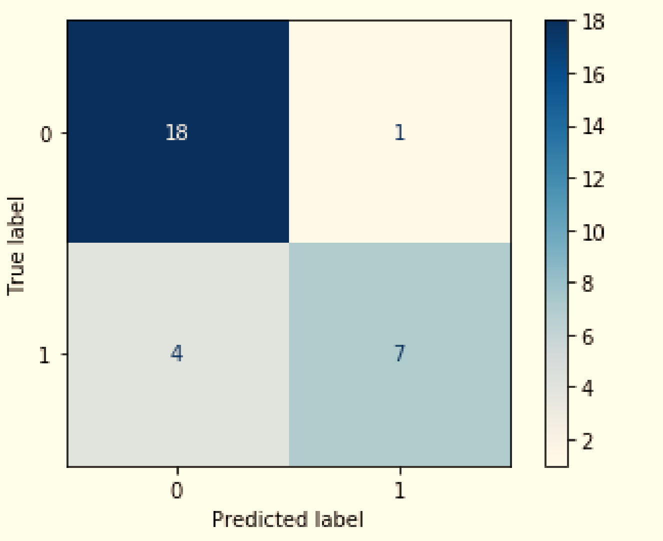
Classifier selection | Evaluation metric

Now that we know how each classifier performs with the data, we need to decide which one will be our final model.

Something that will help us making the decision is to understand which evaluation metric is more relevant to us. In this case, F1 score is the answer, but why is that?

- There is an uneven class distribution: There are much more survivals than deceases.
- We need to seek balance: It is great to identify a lot of True Negatives (TN: Patient is not deceasing), but it could mislead us by largely contributing to a high accuracy rate. This is due to the imbalance we mentioned previously. As we are trying to build a classifier that predicts the survival or decease of a patient, we need to avoid predicting as many False Negatives as possible.

Considering what has been mentioned above, let's check the following example:

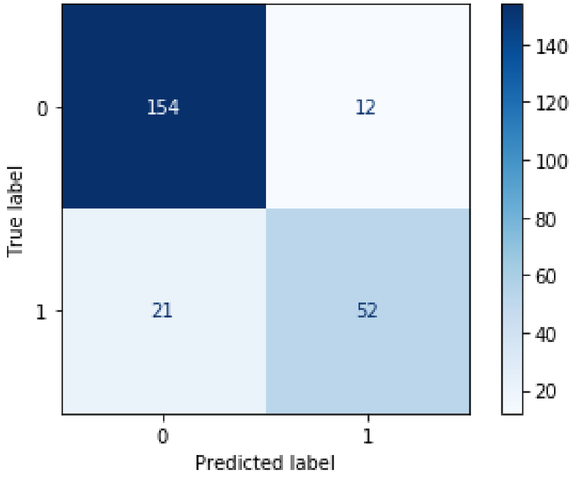


Based on the predictions that are displayed on the confusion matrix, we would say the model performs relatively well, as we predicted 25 out of 30 inputs correctly. This gives us a 83.33% accuracy. However, if we look closely, we can see that we are failing to predict 4 cases in which patients are likely to decease. This traduces to huge tangible and intangible costs, as our model is stating that these patients will remain alive.

Classifier selection | Performance

Considering the criteria described previously, we will test the two models that have been built in order to define which will better perform with future data.

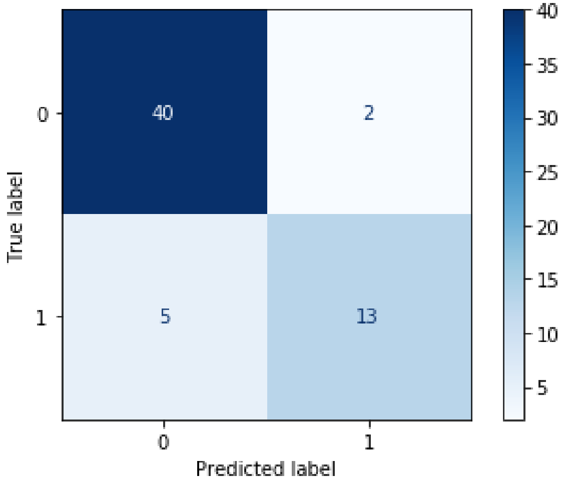
Gaussian Naive Bayes



Accuracy: 85.78%

F1 score : 75.32%

K-nearest neighbours

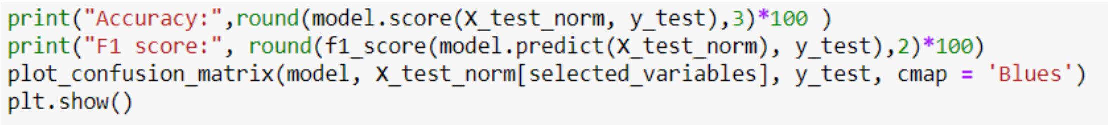


Accuracy: 88.0%

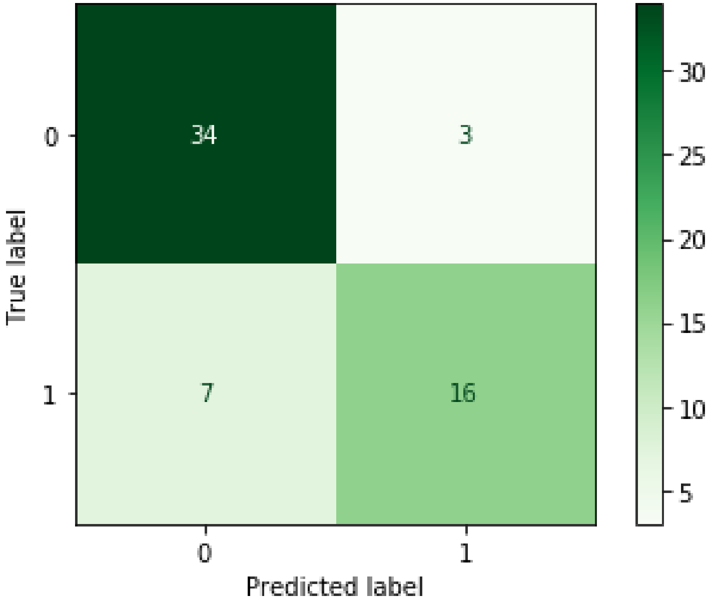
F1 score : 79.0%

Classifier selection | Test set

The classifier based on K-nearest neighbours algorithm will be the chosen one, as both metrics we were looking into are higher. Hyperparameter K will be set to 5.



```
print("Accuracy:",round(model.score(X_test_norm, y_test),3)*100 )
print("F1 score:", round(f1_score(model.predict(X_test_norm), y_test),2)*100)
plot_confusion_matrix(model, X_test_norm[selected_variables], y_test, cmap = 'Blues')
plt.show()
```



Accuracy: **83.3%**

F1 score : **76.0%**

Conclusions

Our goal was to build a classifier that could **predict the survival of a patient** who had suffered from heart failure and had been under a follow-up period.

The resulting classifier has performed relatively well with the test data, with an accuracy of 83.3% and a F1 score of 76%. However, we consider it to be insufficient due to the effects bad predictions would have in a real environment. We are still far away from a decent True Positive Rate, and the real costs that derive from predicting False Negatives are enormous.

We obtained **very relevant insights** during the exploratory data analysis that helped us understand what features are important and should be looked at when facing this type of issue within the field of cardiology. If we focus on physical aspects exclusively, serum creatinine and ejection fraction are the features doctors must be more aware of and which should be tracked constantly.

A more balanced and larger data collection would definitely allow us improving and refining our model. Nevertheless, **the result is considered to be satisfactory.**

Next steps would be taken towards the **conversion of our binary classifier into a multiclassifer.** This would help us to not only predict survival/decease but also to specify the cause.

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITAS GOTHOBURGENSIS
1891