Student ID: 3912063
Student Name: Sasha Nazareth

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

Title- To predict the survival of heart patients using certain features.

Affiliations:

RMIT university

Contact details:

Email ID- s3912063@student.rmit.edu.au

Date of report: 22nd May 2022

Table of Content **Page no.**

Abstract: -

The main aim of this report is to find the best model which will accurately predict the heart failure patients who would survive if we kept in check the features in the dataset. Our main goal is to predict which heart fail patients would not live during the follow up period. This analysis would help us decrease the number of deaths in patients.

I have used two classification techniques- K neighbours' classifier and decision tree classifier to train and test my model. Before modelling I have used the hill climbing approach for feature selection.

Overall, the results include eight features selected which are- age, anaemia, creatinine phosphokinase, diabetes, high blood pressure, sex, smoking and time. The features which were not selected are platelets, ejection fraction, serum creatinine, serum sodium, and death event. Death event is not selected as it is the target column.

Introduction: -

In this report I will be exploring the different features from the heart dataset and analysing which of these features are the best to be trained in our model. The features include anaemia, creatinine phosphokinase, diabetes, high blood pressure, serum creatinine, serum sodium, sex, smoking, time, age, platelets, ejection fraction and death event. Our target column is the death event. Some of the features are boolean values, those features are anaemia, high blood pressure, diabetes, sex, smoking, and death event. While the rest of the data contains numeric values.

The main purpose of this report is to explain the features well with visualisation graphs, explanations and to find a solution which is backed up with data.

Methodology: -

To get a well-trained model, before that I prepared the data. The heart dataset consists of 299 patients, and we will be analysing this dataset further.

DATA PREPARATION:

In data preparation I will explain my step-by-step process:

1. Firstly, I loaded my dataset and checked for the null values in all the columns.

   For checking the null values I used the [.isnull().sum()] function, all the columns showed that they had no null values present. I also changed the column names just for it to seem a bit more presentable as I believe the data should be in its best form while presenting. Hence, after this I moved forward.

2. Secondly, I checked for the inappropriate values

   In the paper there is a table given to us in which the range of the data is mentioned. Hence, checking those value for each column I clean the data. I used all the ranges for the columns, and all were within the range mentioned except serum sodium. One

of the values was lesser than 114, it was 113. hence, instead of dropping the column I changed the value to 114 as it is not a major difference. In my opinion it was a better option rather that dropping the column and losing the data of one patient as there were no other problems for that patient in any other feature columns.

3. Checking all the feature columns having Boolean values

The features having boolean values are anaemia, high blood pressure, diabetes, sex, smoking, and death event. I checked whether all were within the range of 0 and 1 and that they also did not have any negative values.

4. Checking the data types

All the features were in the correct data type form except for age. In my opinion, age cannot be a float type as age can only be an integer. Therefore, before that when I checked the data there are two float numbers- 60.667(both). I rounded the numbers to 61 and filled in the values using the [.at] function. After filling the appropriate numbers, I changed the data type using [.astype(int)] function.

5. Dropping the duplicate rows

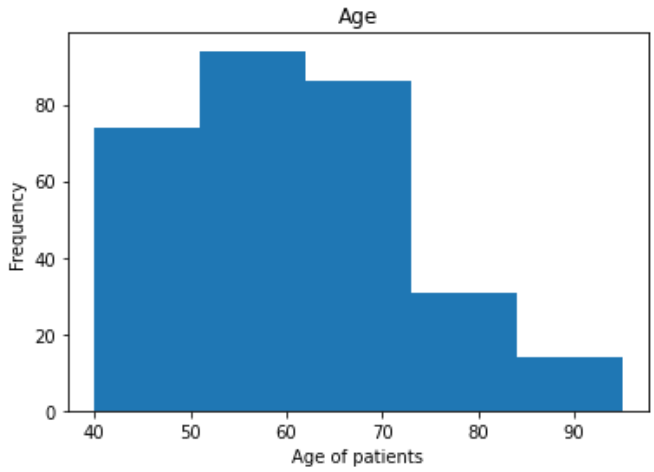
In the end, I dropped the duplicate rows using the [.drop]_duplicates function and my data set was prepared to be explored and modelled further.

DATA EXPLORATION:

In data exploration, I visualised my data individually and in pairs. I found out many interesting relationships which I will be mentioning below.

1. Individual columns exploration

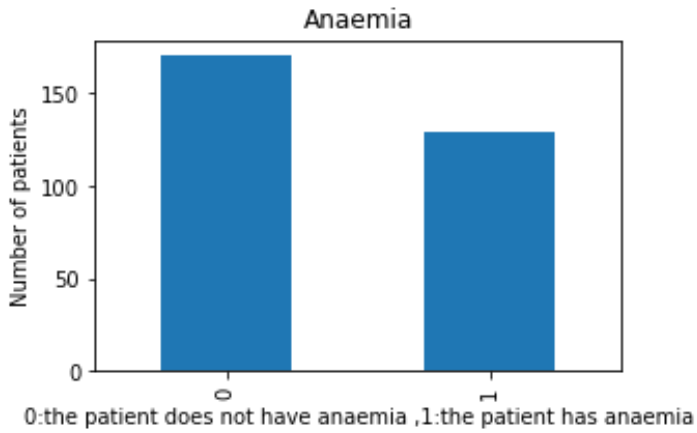
Firstly, I explored the Age column.



I used a histogram to explore the age column as I wanted to know the range of the age the patients are in and is the data left skewed or right skewed.

As we can observe by the graph the range of the patients lie between 40 years old to 95 or 100 years approximately. And the number of patients is highest from the age of 55 to 65 years.

Hence, what we can learn from this graph is that most of the patients are in the mid age range and the number of patients is high in the age range of 55 to 65 years.

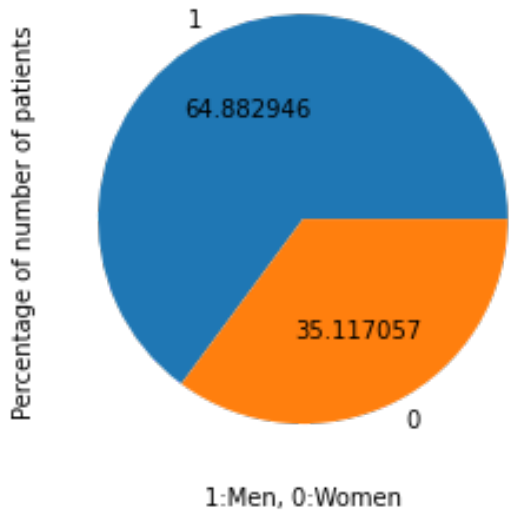


Then moving forward, I explored the boolean value columns. I used the bar graph because in my opinion to present boolean values bar graph is the best.

In this graph, we can observe that the number of patients who have anaemia, i.e., people who have less of red blood cells or haemoglobin are less. There are a larger number of patients who do not have anaemia. This is a very interesting finding. And we see this trend in high blood pressure, diabetes and smoking too.

There are a greater number of patients who do not have high blood pressure and diabetes. Also, there are a greater number of patients who do not smoke. This finding is interesting because, in the back of our minds we would think that a patient who is likely to get a heart attack would be a person who smokes and who has high blood pressure and diabetes. I also created a pie graph for each boolean value just to see the percentage difference and in my opinion, I feel it gives a deeper understanding with both the graphs.

The number of patients who do not have anaemia is 57%. It is not a drastic difference as it's a 7% difference from half of the patients. But even 57% is a value to be noted. Diabetes is similar to anaemia, 58% difference of patients who do not have diabetes. The patients who do not have high blood pressure is 64% which is a major difference. While smoking is like high blood pressure 67% of patients do not smoke. A vast difference between them.

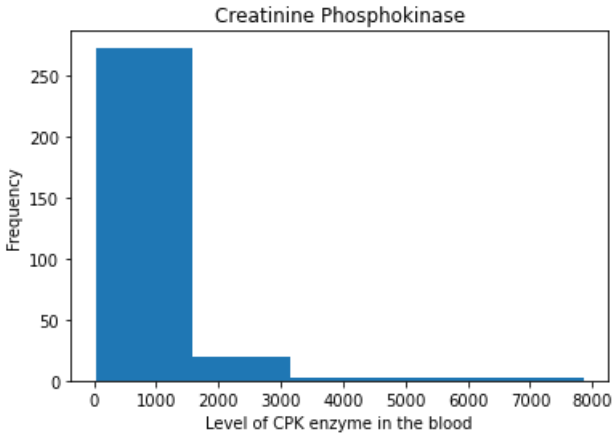


Next, I explored the sex column.

I chose to use the pie chart because it gave a clear difference between the number of men and women as patients. By this chart we can clearly see, the number of men patients is high (64%) than the number of women patients (35%).

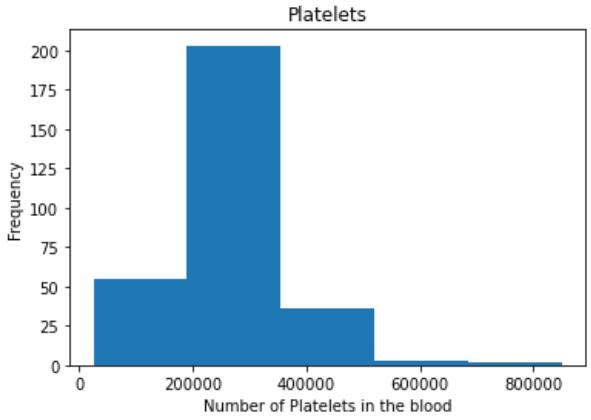
For the target column- death event I chose the pie graph too because it gave me a clear picture of the percentage of number of patients who died during the follow up period. The number of patients who died during the follow up period were 32%. While patients who are still alive were 67%. The pie chart looked like the sex pie chart.

For creatinine phosphokinase and the next four features- platelets, serum creatinine, serum sodium and time I chose to use a histogram as I wanted to observe if the data represents a bell curve or if it is skewed. In my opinion using a histogram was the best choice.



The next feature I explored is creatinine phosphokinase.

Observing this graph, we can notice it is a left skewed graph. And the level of the CPK enzyme in the blood of patients is highest between 0 to a 1050 approximately. While as the level increases the number of patients decreases. Between 1050 to 3000 there are comparatively very low number of patients with this level. And as we go on further, we can notice the number of patients decreases drastically. Hence, we can learn that more number of patients have a level of 0 to 1050 CPK enzyme in their blood.

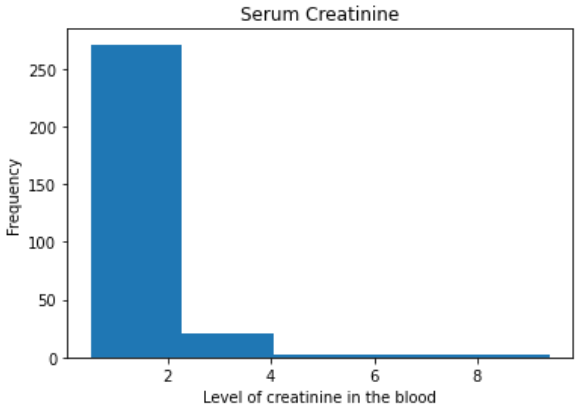


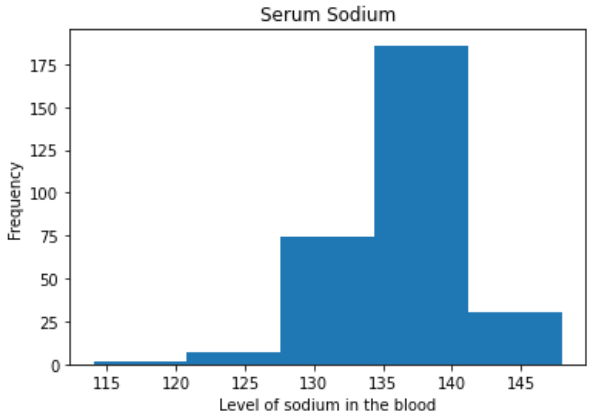
The next column is platelets.

I observed that the greatest number of patients had 200,000 to 350,000 number of platelets in their blood. The second highest would be between 0 to 200,000 number of platelets approximately. And then there are also a good number of patients who have 350,000 to 500,000 number of platelets approximately. After 500,000 number of platelets the number of patients is comparatively very less.

The next feature I explored is serum creatinine. I observed it is a left skewed histogram. And is quite clear that the number of patients have a level of 1 to 2 or 2.2 creatinine approximately in their blood. Majority patients' range is in this level.

While from 2.2 to 4 approximately, there are a few more patients whose level of creatinine is in this range. From 4 to 8 the number of patients decreases.



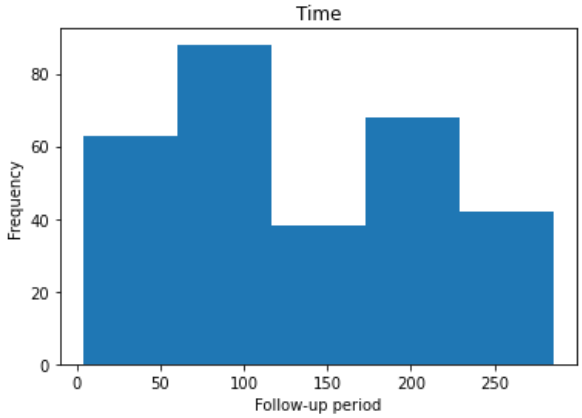


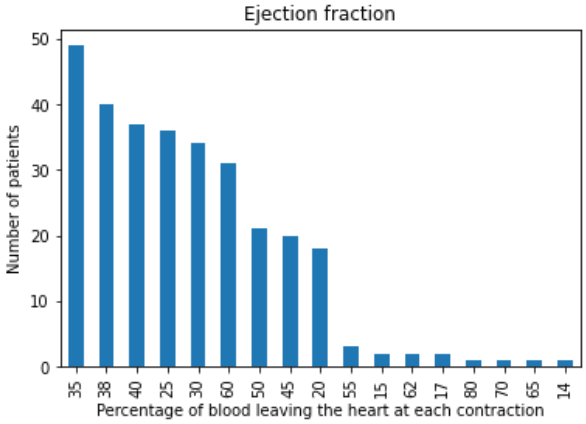
The next feature is serum sodium. The highest number of patients have a sodium level ranging between 135 to 141 or 142 approximately. While the second highest level is around 127 to 135. And then 141 to 147 is the next highest.

From 115 to 127 the number of patients having this level of sodium are very less.

For the next feature, time I chose to explore it using a histogram. The time indicates the number of days the patient needed as a follow up period.

As we can observe, the highest number of days was approximately 50 to 110 days. Which means that, more number of patients needed a follow up period of 50 to 110 days. While the least number of patients needed a 110 to 180 days for follow up approximately.

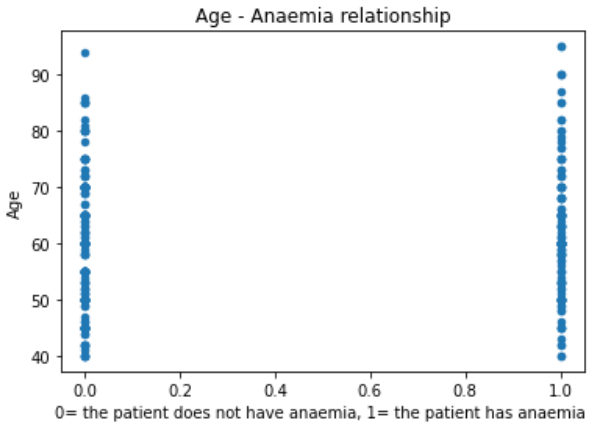




The next feature I explored is ejection fraction. I used a bar chart as the bar chart visualised the data the best and, in my opinion, will help me in exploration of the feature the best.

As we can observe, 35% is the highest percentage of blood leaving the heart at each contraction, it has the highest number of patients. And the least is 14%.

2. Exploring relationships between the features



For exploring relationships, I used the scatter plot as to show any type of a "relationship" between two features it is the best visualized by a scatter plot.

The first relationship I explored is the age and anaemia relation.

Hypothesis- Patients above the age of 85 have anaemia.

As we can observe in the graph, the difference between the patients who have anaemia and who do not have is very less. Hence, it proves to us that age does not have a relationship positive relationship with anaemia. One interesting thing we can notice is above the age of approximately 85 there are patients who have anaemia. Although there is just one patient who does not have anaemia, but I am considering that as an exception in this case.

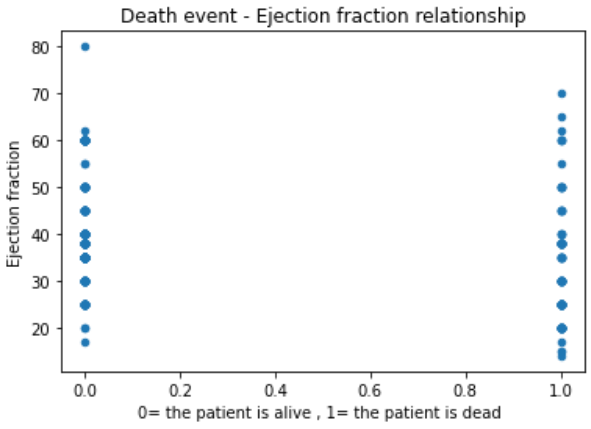
Insights- there is a lack of relationship between age and anaemia.

The second relationship I explored is the age and diabetes relation. The graph is very similar to this with a few differences which I will explain further.

Hypothesis- a patient having diabetes does not depend on the age of the patient.

In the graph, we can notice a lack of relationship between age and diabetes. Patients between the age of 40 to 75 have diabetes and some don't have diabetes too. There are small differences like between 81 to 87 approximately the patients do not have diabetes. There is a gap also between 75 to 78 years who do not have diabetes.

Insights- there is a lack of relationship amid age and diabetes.

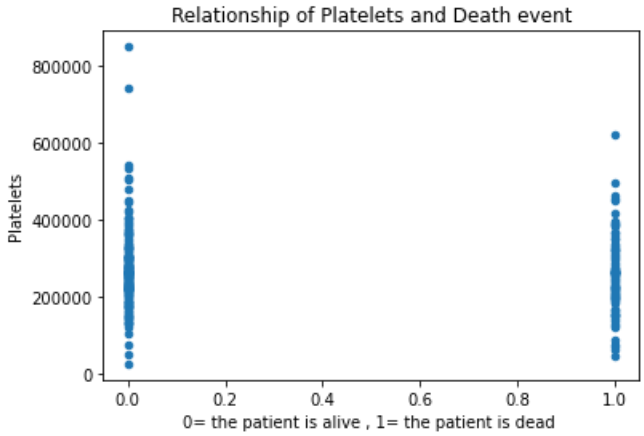


The third relationship I explored is the death event and ejection fraction relation.

Hypothesis- When the percentage of blood leaving the heart at each contraction is more than 65 the patient dies.

As we can observe there is a very interesting relation between the two features. Several patients are alive even when the ejection fraction is between 20% to 65%. But there are a greater number of patients who die after 65%. In the graph we can notice an exception that even after 65% there is still a chance for a patient to live. But I will just be considering that as a small fraction of possibility.

Insights- there is an interesting relationship amid death event and ejection fraction.



The fourth relationship I explored is of platelets and death event.

Hypothesis- Patients with a high number of platelets (above 550,000) live.

As we can notice in the graph even when the number of platelets is high in patients, they are alive. Therefore, platelets and death do not have a positive relationship. i.e., we cannot

determine if a person will die or live with the number of platelets. After 550,000 number of platelets patients are still living. And in between 0 to 550,000 platelets patients are alive and not living.
Insights- there is an interesting relationship between platelets and death event.

The fifth relationship I explored is of platelets and high blood pressure. The graph is like the graph of platelets and death event.

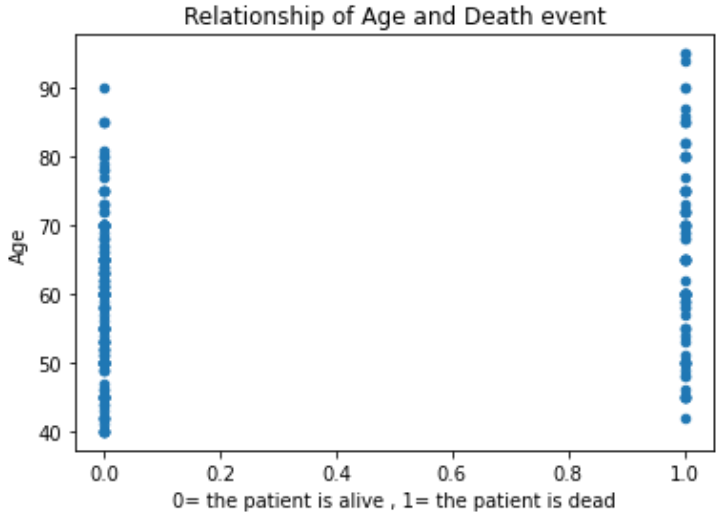
Hypothesis- when a patient has their number of platelets between 0 to 100,000 and 550,000 and above the patient does not have high blood pressure.

In the graph of platelets and high blood pressure it is very interesting as there are small gaps in which the patients do not have high blood pressure. The range is from 0 to 100,000 and 550,000 and above. While for the other number of platelets the person can have high blood pressure too.
Insights- there is an interesting relationship between platelets and high blood pressure.

The sixth relationship I explored is of platelets and diabetes.

Hypothesis- Patients with a high number of platelets (above 550,000) have diabetes.

The relationship of platelets and diabetes is very similar to platelets and death event.
The higher number of platelets above 550,000 patients have diabetes, hence it is a positive relationship between platelets and diabetes.
Insights- there is a positive relationship between platelets and death event.

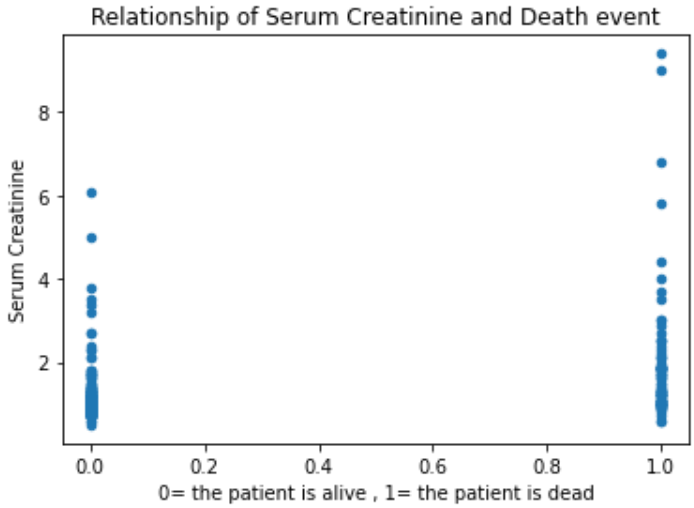


The seventh relationship I explored is of age and death event.

Hypothesis- After the age of 85, a greater number of patients die during the follow up period.

In this graph as we can observe, after the age of 85, there are a greater number of patients who die. For the patients alive, there is a few scatter plots above the age of 85 but it is not as dense as we can see in the segment where patients die. We can also consider this as a positive relationship, as we go above 85 the death rate increases.

Insights- there is an interesting relationship between age and death event.

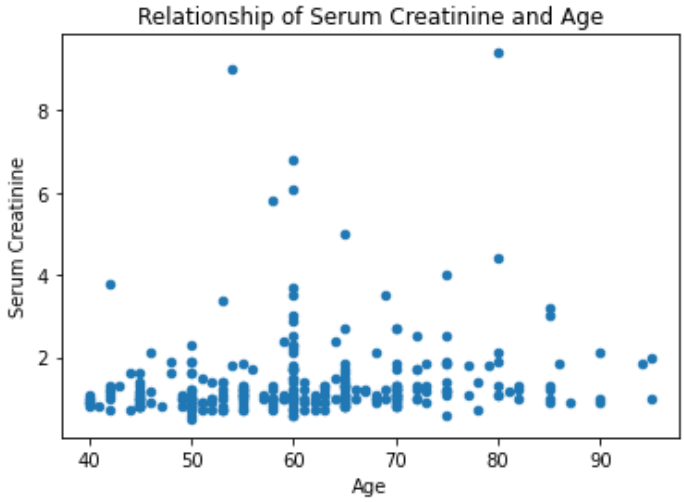


The eight relationship I explored is of serum creatinine and death event.

Hypothesis- As the level of creatinine in blood increases, there are higher chances of the patient to not live.

This is a very straightforward relationship. As the level of creatinine in blood increases above 6 the patient dies. Hence, this is a positive relationship between the features.

Insights- there is an interesting relationship between serum creatinine and death event.



The ninth relationship I explored is of serum creatinine and age.

Hypothesis- A greater number of patients age is in the range of 40 to 70 years old whose level of creatinine in blood is 2.

In this scatter plot, it is quite evident that most of the patients have a creatinine level of 2 in their blood.

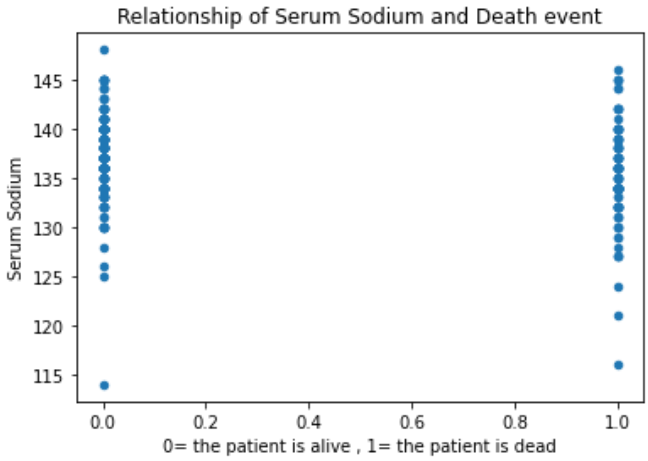
Insights- there is an interesting relationship between serum creatinine and age.

The tenth relationship I explored is of creatinine phosphokinase and age. The graph was very similar to the graph above.

Hypothesis- From the age range of 40 to 80 years the level of CPK enzyme in the blood is 2000.

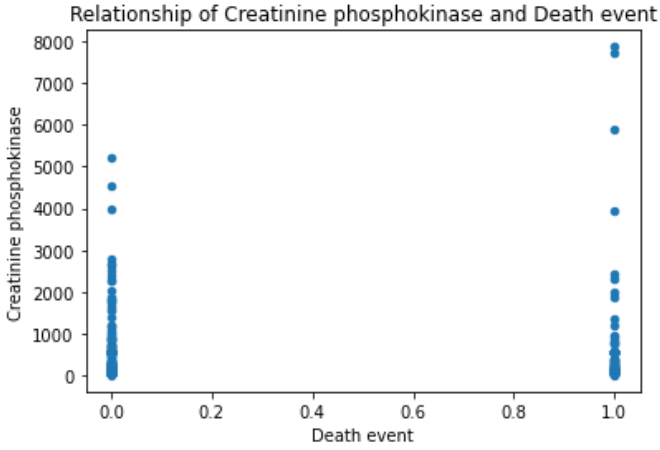
In the graph we can notice that between 40 to 80 years it is the densest and after the 2000 level it is highly scattered.

Insights- there is an interesting and clear relationship between serum creatinine and age.



The tenth relationship I explored is of serum sodium and death event.

Hypothesis- when the level of sodium in the blood is below 125 there are more chances for a patient to not live.
As we can observe by this graph it is quite straightforward that the sodium levels above 125 patients are living too. Hence, just below this level it is more likely for a patient to die. Even though it is quite scattered we can assume it to be true.
Insights- there is an interesting and clear relationship between serum sodium and death event.

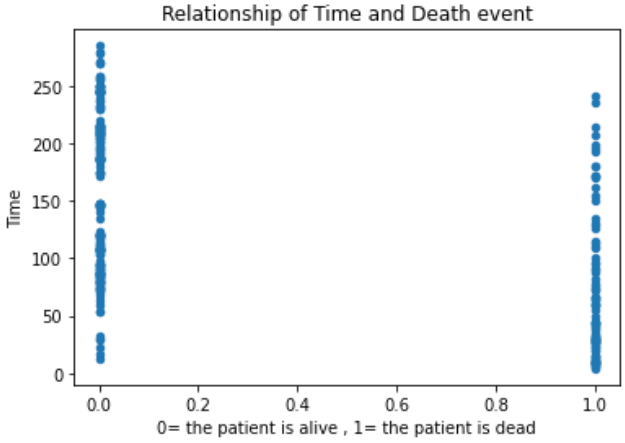


The eleventh relationship I explored is of creatinine phosphokinase and death event.

Hypothesis- after the level of 3000 CPK enzyme in the blood, there are more chances for the patient to die.

As we can notice, after the level of 3000 the graph is quite scattered. But if we observe in the area where the patients die, there it has more dense scatter plots. Hence, we can consider 3000 as the highest level of CPK enzyme a patient can have in their blood, above this the chances of the patient to die increases.

Insights- there is an interesting relationship between creatinine phosphokinase and death event.



The last relationship I explored is of time and death event.
Hypothesis- As the follow up period increases after 220 days, the patient lives.
As we can notice from 0 to 220 follow up days, it is not clear if the patient would live or die, either is possible just after 220 days we can see the patient lives.
Insights- there is an interesting relationship between time and death event.

The next step is modelling the data, I chose to train and test my data with classification. As in my opinion, I had labelled data and supervised learning therefore, it seemed best for me to go ahead with classification. I used the K- neighbours classification and the decision tree classification.

Results: -

At first, I selected all the columns and tested my model with K- neighbours and the accuracy came to 70%. I selected all the features because that seemed the best fit in the beginning. When I trained it with decision tree classifier the accuracy came up to 76%. But later I understood that the decision tree classifier came with better results just because we have categorical data. And categorical data does not have the best results with K- neighbours classifier model. This approach was the wrong approach which I used to train my model in my opinion. Therefore, I went ahead with my second approach.

Firstly, I used the hill climbing approach for feature selection. I did not want to model my data with assumptions. I needed numbers to prove my model prediction. Hence, when I used the hill climbing approach, I got a result of 90% score for 8 selected features with using the K-neighbours classifier as 21. I concluded to use 21 because, after 21 the accuracy does not increase but decreases. Hence, then I chose to go ahead with the selected 8 features. The features selected are- age, anaemia, creatinine phosphokinase, diabetes, high blood pressure, sex, smoking and time. While the features which are not selected are platelets, ejection fraction, serum creatinine, serum sodium, and death event (target column). After a lot of trials with testing this result fit perfect in my opinion.

The features I chose are the right ones because as I explored the data, I found out that platelets, ejection fraction, serum creatinine, serum sodium do not have a positive relationship with the target column death event. i.e., if the values increase it is not seen that the heart patient dies in the follow up period as a result. Hence, I found it be to fit to remove these features as it would not affect the prediction of patients who would die during the follow up period which is our main goal. We require the features which will predict correctly.

Then I went ahead to train and test my model with the help of K- neighbours classifier. In K-neighbours classifier I had to keep my parameters correct to find the best accuracy for my model. The parameters are the test size, random state, and the K- neighbours classifier. In my opinion, while testing the best values for these parameters are the test size to be 0.20, random state- 1 and the K- neighbours classifier as 4. This depicts I am training 80% of my features and testing 25%. The result of this model is 85% accuracy. And the precision which is the fraction of correctly predicted instances is 85% for class 0 and 86% for class 1. While the recall values which are the fraction of relevant instances which are successfully predicted are, 98% for class 0 and 43% for class 1. Lastly the f1 score which is the combination of precision and recall is 91% for class 0 and 57% for class 1. These scores give us a very good understanding for the features we have selected to predict the main goal of our research.

The next step, I trained the model via the decision tree classifier. In this model I set the parameters test size to be 0.20, random state- 1 and the decision tree classifier max depth as 3. This implies, I am training my model 80% and testing the rest 20%. The max depth defines the different levels within the tree, i.e., in my case I chose the max depth to be 3 which means the tree will go onto three levels and apply classification on these levels only. The result of this model is 87% accuracy. Which is 2% better accuracy than the K-neighbours classifier model. The precision is 90% for class 0 and 75% for class 1. While the recall values are, 93% for class 0 and 64% for class 1. Lastly the f1 score which is the combination of precision and recall is 91% for class 0 and 69% for class 1. In the end when

visualising the decision tree, in the results the root node is time and sub nodes are age and time.

Discussion: -

When we compare both the models based on accuracy there is just a 2% difference. But as we observe the classification report with all the precision, recall and f1 score we can see major differences. All the values in decision tree for both the classes 0 and 1 are above 50% but if we see the K- neighbours classifier precision, recall and f1 score the scores for the class 1 is only 43% (which is below 50%) and 57%. Hence, the classification report of the decision tree model gives us better results than the k neighbours classifier's classification report

Another important point is, the K- neighbours classifier approach does not work best on categorical values hence, the accuracy score is lesser compared to the decision tree classifier model. In decision tree it works on categorical and continuous input and output variables from a dataset, it also splits the data into two or more homogenous sets and gives us results. In the heart dataset there are both categorical and continuous input and output variables hence, using decision tree model is the right decision. Also, the decision tree method is more precise and reliable as compared to the k neighbour's model. Hence, the decision tree classifier model is better in my opinion. It has a good accuracy score and is perfect for the dataset.
The results answer our goal as it shows us a good predictive power. With the features selected it will help in predicting the number of heart patients who will not live in the future. This model is a steppingstone to achieving our goal.

Conclusion: -

To conclude, with my data model it can help in the medical field to support doctors and nurses in forecasting the survival of a heart failure patient with the features I have chosen above- age, anaemia, creatinine phosphokinase, diabetes, high blood pressure, sex, smoking and time.

In the future to get better results, a deeper analysis can be done using this model. A few interesting things I found while analysing was, earlier researchers have said to just focus on two features- serum creatinine and ejection fraction but when I modelled my data, I used 8 features and in those 8 features these two features are not included. Which shows us the more we analyse with different point of views the results vary and we cannot come to conclusions on which model is better. Hence, in my opinion the more research we would do with different perspectives in the future we may get an outstanding model to predict the survival rate, which would be very helpful for the patients and their families.

References: -

Chicco, D. & Jurman, G., 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone - BMC Medical Informatics and decision making. *BioMed Central*. Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5 [Accessed May 18, 2022].