**Predict survival of patients with heart failure and rank the features corresponding to the most important risk factors**

**Bakht Singh Basaram (11546730)**

**Sahithi Chindam (11617204)**

**Santhoshi Indrakanti (11610865)**

**ShanmukhaNadh Uppugandla (11550894)**

**Problem statement**

Heart failure occurs when the heart is unable to pump enough blood to the body, and it is usually caused by diabetes, high blood pressure, or other heart conditions or diseases. Heart and blood vessel problems, such as coronary heart disease (heart attacks), cerebrovascular illnesses (strokes), heart failure (HF), and other pathologies, are referred to as cardiovascular diseases (CVDs). When the heart cannot pump enough blood to meet the body's needs, results in heart failure. Annually, cardiovascular diseases cause the death of approximately 17 million people worldwide. Heart failure is the state in which muscles in the heart wall fade and enlarge, limiting heart pumping of blood. The ventricles of the heart can get inflexible and do not fill properly between beats. With the passage of time the heart fails to fulfil the proper demand of blood in the body and therefore a person starts having difficulty in breathing.

The main objective of this study is to estimate death rates due to heart failure and to investigate its link with some major risk factors by choosing a dataset of medical record which contains 299 patients having heart failure. This dataset is collected during their follow up period. Machine learning can identify the most crucial characteristics in a patient's medical records and forecast the patient's longevity from their data.

We demonstrate that machine learning techniques can achieve the best predictive performances aimed at emphasizing hidden patterns and correlations in data otherwise unidentified by medical doctors. In this study we want to predict survival of patients with heart failure symptoms and rank the features that are corresponding to the most important risk factors.

**Workflow**

**Diagram

Description automatically generated**

**Data Specification**

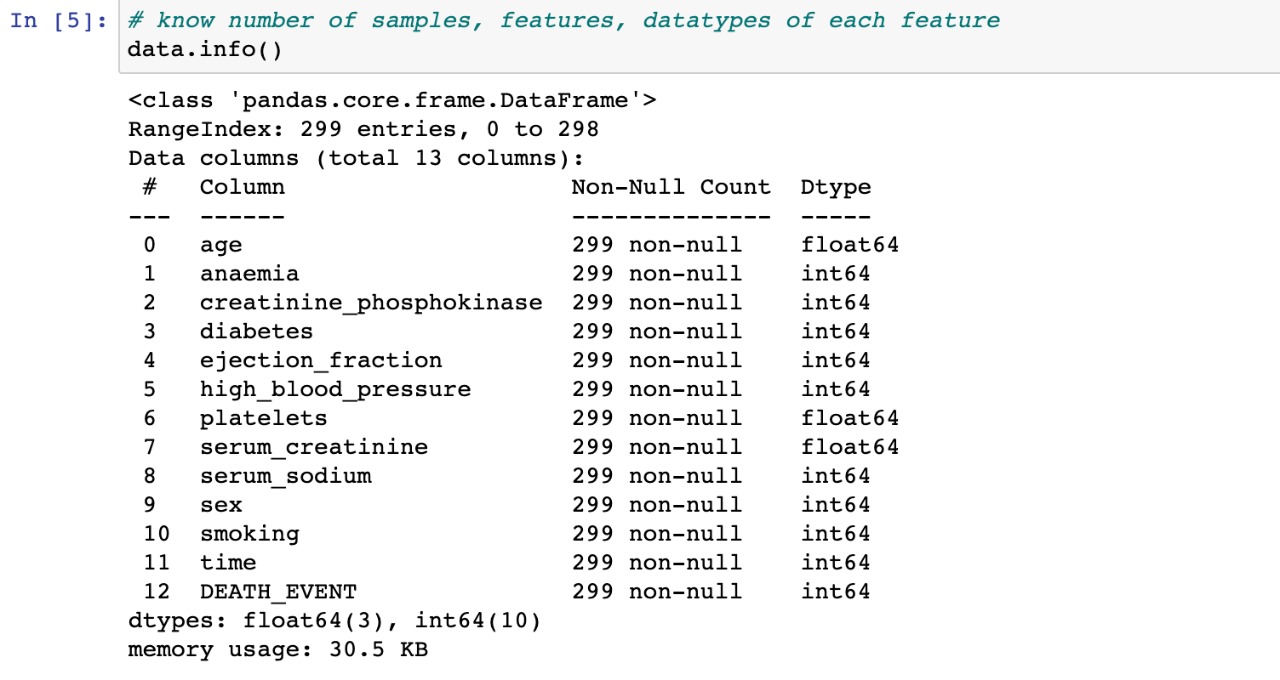
The dataset describes the medical records of 299 heart failure patients collected during the April-December 2015 at the Faisalabad Institute of Cardiology. The dataset consists of 299 observations (rows) and 13 attributes (columns) including 105 women and 194 men and with age ranges from 40 and 95 years old. There are six categorical columns (anemia, diabetes, high blood pressure, sex, smoking, death event) and seven numerical columns (age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, time). All 299 patients had prior heart failure and left ventricular systolic dysfunction.

**Features of Dataset:**

1. **Age**: Age is a numerical column that tracks a patient's age, with 0 indicating a woman and 1 indicating a man.
2. **Anemia**: Anemia is the categorical column where 0 represents false and 1 represents true. When a patient's hematocrit levels fell below 36%, the hospital doctor declared them to have anemia.
3. **creatinine\_phosphokinase:** This column gives information about the level of CPK enzyme in blood. CPK enzyme enters the circulation when muscle tissue is injured. Therefore, elevated CPK levels in the patient's blood might be a result of heart failure or another harm.
4. **Diabetes:** Diabetes is a categorical column that contains information with true or false values of 0 (patient does not have diabetes) and 1 (patient has diabetes).
5. **ejection\_fraction:** The ejection fraction column states the percentage of how much blood the left ventricle pumps out with each contraction.
6. **high\_blood\_pressure:** High\_blood\_pressure is thecategorical feature which gives the information with 0’s (false) and 1’s (true).
7. **Platelets:** This column gives information about the patient’s platelets level.
8. **serum\_creatinine:** serum creatinine is numerical column. If a patient has high levels of serum creatinine, it may indicate renal dysfunction**.** When a muscle breaks down, creatine produces a waste product called serum creatinine. Doctors specifically check serum creatinine levels in blood to determine kidney function. High serum creatinine levels in a patient may be a sign of renal disease.
9. **serum**\_**sodium:** serum sodium is the numerical column**.** Sodium is a mineral that supports the healthy operation of the muscles and nerves. The serum sodium test is a common blood examination that shows whether a patient has normal sodium levels in the blood. Heart failure may be the cause of a blood sodium level that is abnormally low.
10. **Sex**: Sex is the categorical column which represents 0’s and 1’s where 0 represents woman and 1 represents man.
11. **Smoking**: Smoking is the categorical column which is indicated with 1 (true) that means patient has the habit of smoking and indicated with 0 (false) that means patient does not have the habit of smoking.
12. **Time**: time is the numerical column which gives the information about the patient time.
13. **Death event**: Death event is numerical column. It refers to whether the patient passed away or lived before the average follow-up duration of 130 days.

**EDA / Data Understanding & Data preparation**

* In our dataset we looked if there are any null values or missing values. There are no null values, and this dataset consists of 299 samples or patient records with 12 predictors and a target variable (DEATH\_EVENT). Below python code gives us the initial information.



We also verified, if all the features are of proper datatypes, we have seen age column is of ‘float’ datatype, so we have changed it to ‘int’ datatype.

* We did quick statistics of all the feature variables based on the two classes of the target variable and we found the below results.

Graphical user interface, text, application

Description automatically generated

Graphical user interface

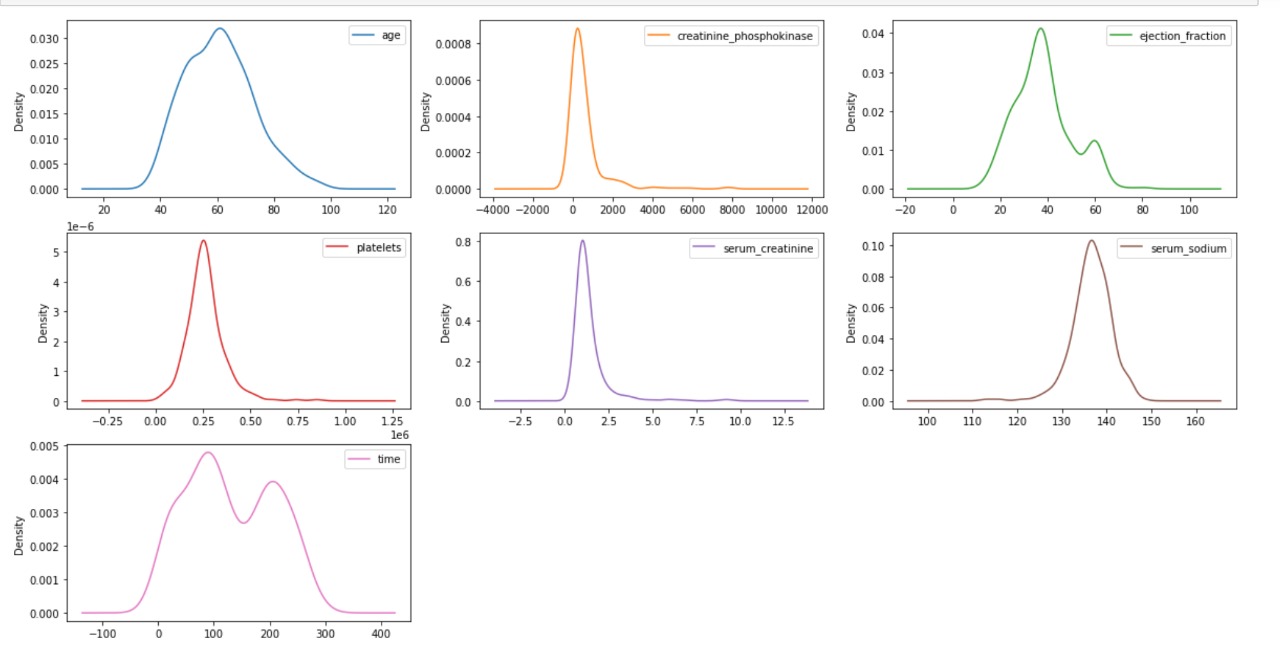
Description automatically generated

* Mean values of all features for cases of survived and dead patients.
* The average age of patients detected for survived and dead cases are 58.76 and 65.21 respectively.
* It looks like the diabetes; smoking features are not significant factors in predicting the target variable because it appears that the proportion of diabetes cases and the patients who smoke are doesn’t smoke for both survived and dead patients are equal.
* As we mentioned above ejection\_fraction is the percentage of how much blood the left ventricle pumps out with each contraction. The less percentage it is the more chance of patient’s death. So, from the above statistics it clearly shows the percentage of ejection\_fraction for dead patients is less (33.47) when compared with the survived patients (40.27).
* As we mentioned above when a patient has high levels of serum creatinine, it may indicate renal dysfunction. From the above visualization It is found that the dead patients has high levels of serum creatinine.
* We also have a column in our dataset that states if the patient died or survived before the end of the follow-up period. The average follow-up period was 130 days. From the above visualization the follow-up period of dead patients is significantly less that is 70.89 days for dead and 158.34 for survived patients.
* From these mean values we can say that some features show us the clear difference between survived and dead patients. However, other features might require more effort to find their relationship with target variable.

**Distributions:**

Diagram

Description automatically generated

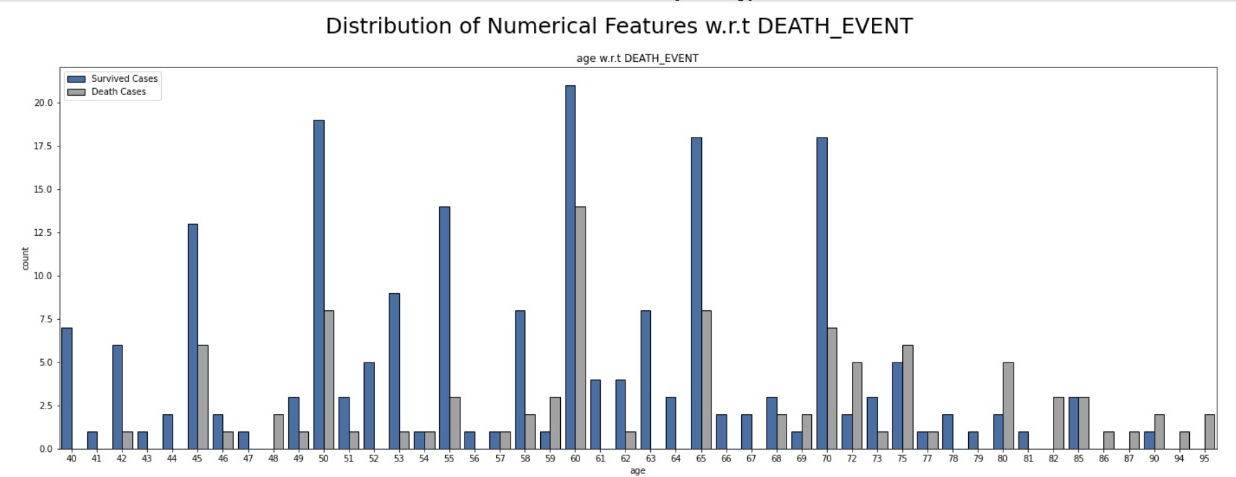
As we discussed about distributions in our classroom, we have plotted density graphs and boxplots to know the frequency distributions of feature variables.

* Creatinine\_phosphokinase, serum\_creatinine has positively or rightly skewed data distribution as there are many outliers in those columns, we can see them in the above boxplots.
* Ejection\_fraction and time display a bimodal data distribution.
* Serum\_sodium has negatived or left skewed data distribution.
* Data distribution of age and platelets are a bit tricky. This is because they are almost like a normal distribution or bimodal distribution.

****

From the above graph, it looks like dataset is unbalanced and is nearly in a 2:1 ratio for survived and dead patient cases, due to these predictions will be biased towards survived cases.

**Distribution of numerical features with respect to DEATH\_EVENT**

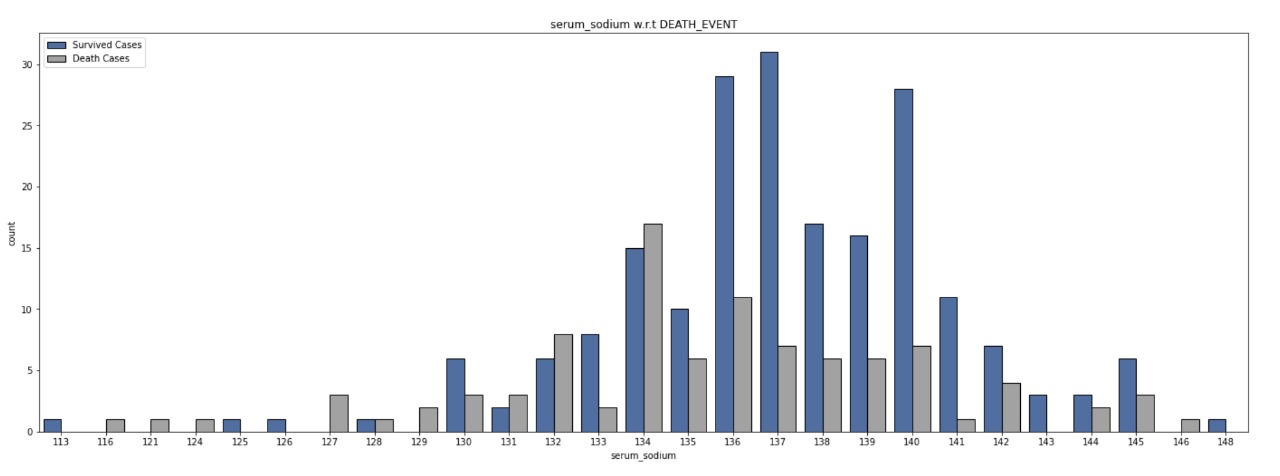
****

From the above distribution, death cases are more for the patients whose age has crossed 68.

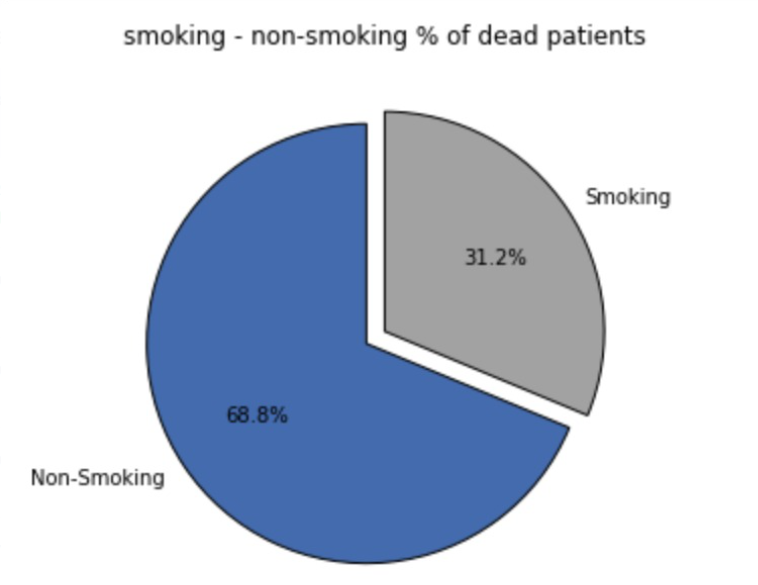
**Chart

Description automatically generated**

In the above graph, we can see that there are more death cases when a patient’s ejection\_fraction decreases. which means, as ejection fraction increases percentage of death cases decreases and vice versa and hence death cases are inversely proportional to ejection fraction.



Above graph illustrates that majority of data is distributed around the mean, the normal range of serum sodium level in blood is 135 to 145 (mEg/L) so it looks like, most of the patients have normal serum sodium levels. We can also say that even the patients that have normal serum sodium levels have also faced heart failures.



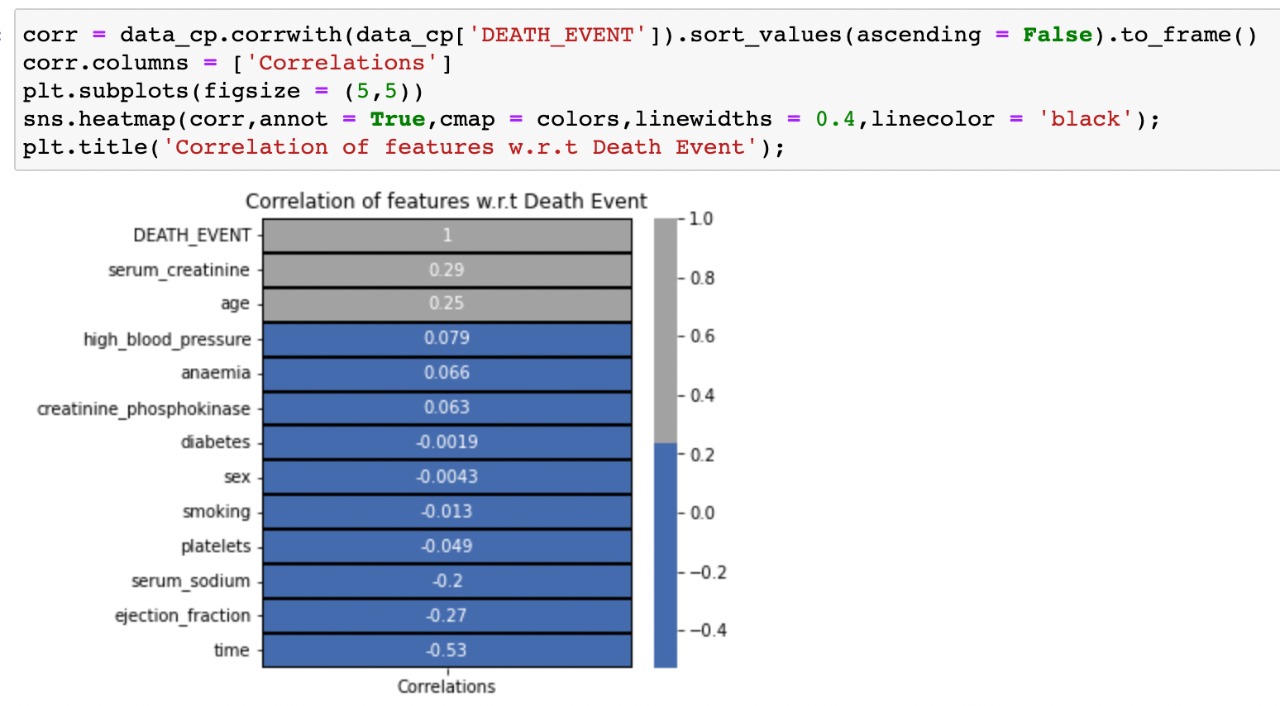
Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

We just wanted to see whether the smoking, diabetes and high blood pressure impacts the count of death cases. So, we have plotted percentages of smoking vs non-smoking, patients with diabetes and without diabetes, and patients with high blood pressure and without high blood pressure of dead patients.



We plotted correlation of features with respect to target variable which helps in feature selection so from the above visualization serum creatinine, age have positive correlation with target variable whereas, ejection\_fraction, serum\_sodium, time columns have negative correlation and remaining features are not related to output variable.

**Methods & Modelling**

We have divided this modelling into following steps

1. **Data scaling using standardization and normalization techniques**

It is important for us to scale the data before applying any machine learning techniques, because the units of feature values are not properly interpreted by the machine learning models, it treats the input values as just a simple number but does not understand true meaning of the value. For example, it cannot identify age, year, blood pressure it considers all of them as just numbers.

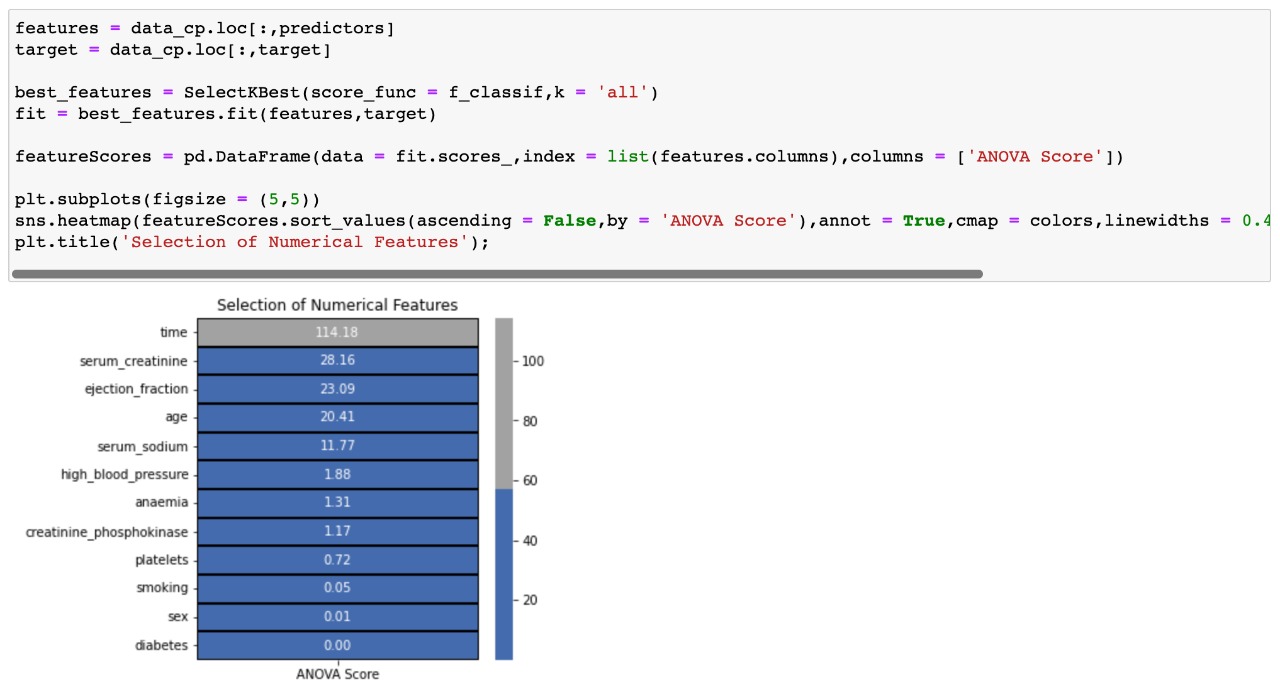
So, the solution for this is data scaling, we can standardize or normalize the data based on the values of feature variables, generally we use standardization for the data that is normally distributed and normalization for the data that does not display normal (Gaussian) distribution. So, we have used both normalization and standardization for our feature variables based on the distribution.

Graphical user interface, text, application, email

Description automatically generated

1. **Feature selection**

This is one of the important steps before training the model. So, reducing the dimension of the dataset, reduces the computational cost. The more noise in the data, the less is the efficiency of the model. For example, if there are two columns in the data that are highly correlated then we can drop one of the columns and we can also drop the columns that doesn’t display any correlation with the target variable. Here in our study, we used ANOVA technique for feature selection. According to ANOVA test, the higher the ANOVA score more is the importance of feature.



We have plotted a graph, that shows the ANOVA scores of each feature. We have dropped last four attributes i.e platelets, smoking, sex, and diabetes as they have no impact on the target variable which is almost 0. It looks like serum\_creatinine and ejection\_fraction are the most important features for prediction of target variables.

1. **Balancing the dataset using SMOTE:**

As we mentioned above, our dataset is unbalanced which results in predictions that will be biased towards survived patients, so we have used ‘SMOTE’ technique to balance the dataset. This method uses KNN algorithm in creating new synthetic samples to balance the class distribution of the dataset.

Chart, pie chart

Description automatically generated

**Evaluation / Discussion**

The binary prediction of the patient’s survival during the follow-up period is the focus of this section of our analysis. To predict patient survival, we have used the below machine learning algorithms.

1. Logistic Regression
2. Gaussian Naive Bayes Classifier
3. Random Forest Classifier
4. KNN Classifier

We measured the prediction using the metrics for common confusion matrix, receiver operating characteristic (ROC) area under the curve, and precision-recall (PR) area under the curve we were able to find out the accuracy of our predictions.

An indicator of performance for classification tasks at several threshold levels is the AUC-ROC curve. AUC refers for the level or measure of separability, and ROC is a probability curve. It reveals how well the model can differentiate across classes. The model is more accurate at classifying 0 classes as 0, and classifying 1 class as 1, the higher the AUC, the model is more effective at differentiating between patients with the condition and those without condition.

To avoid the problem of overfitting, we used cross validation technique because in machine learning, we cannot know how well our model will perform on new data until we test it, cross validation technique will address the following problems.

1. The concept is to partition the data into K number of subsets
2. Holdout a set and train on the remaining set
3. Test the model on the holdout set
4. Repeat this process for K subset

Chart, bar chart

Description automatically generated

**Results**

**Logistic Regression**

Chart

Description automatically generated

**Gaussian Naive Bayes**

Chart

Description automatically generated with medium confidence

**Random Forest Classifier**

A picture containing graphical user interface

Description automatically generated

**KNN Classifier**

Chart

Description automatically generated

**Feature Importance Graph:**

Chart

Description automatically generated

Our prediction results show that random forest significantly outperforms when compared with all other methods by obtaining top accuracy i.e., 0.92 with ROC\_AUC\_Score of 92.14% and a cross validation score of 93.24%.

Our feature importance graph using random forest classifier and a shape library shows that it can be possible to predict the survival of heart failure patients with only two attributes, they are mainly serum creatinine and ejection fraction. Follow up period or time also a major factor but not needed for biostatics analysis.

This is encouraging because, in hospital settings sometimes there is a chance of clinical features missing from the electronic health records so doctors could still predict patient’s survival, just by analysing the values of serum creatinine and ejection fraction. But before it is clinically practiced additional confirmatory studies are needed.

**Conclusion:**

In conclusion, these findings could have an impact on clinical practice by giving doctors a new tool which helps them to determine whether a patient with heart failure will survive or not. Serum creatinine and ejection fraction may be the key areas of concentration for medical practitioners that are trying to determine whether a patient will survive after heart failure.

**Limitations:**

* Using a small size dataset of only 299 patients. The results might be more reliable when large dataset is used.
* Additional features of the patients like their body mass index, other health related conditions would have been useful to detect cardiovascular health diseases.

**Tools & Frameworks**

We used Jupiter notebook to execute our python code.

**References**

[**https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5**](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5)

[**https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records**](https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records)

[**https://www.kaggle.com/code/tanmay111999/diabetes-classification-xgb-lgbm-stack-smote**](https://www.kaggle.com/code/tanmay111999/diabetes-classification-xgb-lgbm-stack-smote)

[**https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5#:~:text=the%20multiclass%20model%3F-,What%20is%20the%20AUC%20%2D%20ROC%20Curve%3F,capable%20of%20distinguishing%20between%20classes**](https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5#:~:text=the%20multiclass%20model%3F-,What%20is%20the%20AUC%20%2D%20ROC%20Curve%3F,capable%20of%20distinguishing%20between%20classes)**.**

[**https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85**](https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85)