Predicting survival of patients with heart failure from serum creatinine and ejection fraction

October 2020

ML tool

Predicting patients risk of heart failure based on easy to obtain medical data

Boris Henriksen
bh@dimaps.com

Contents

1. Introduction

1.1 Background

According to the World Health Organization (WHO) cardiovascular diseases (CVD) are the number 1 cause of death globally. An estimated 17.9 million people died from CVDs in 2016, accounting for 31% of all global deaths. [1]

Most cardiovascular diseases can be prevented by addressing risk factors. One type of heart failure is based on the ejection fraction value, which is expressed as a percentage, of how much blood is pump out with each contraction [2]. A normal heart's ejection fraction may be between 50% and 70%. A value below 40% is a heart failure. Creatinine is a waste product that forms during metabolism and is removed from the body by the kidneys. Normal levels range from 0.8 to 1.4 mg/dL. When kidney function decrease, creatinine levels increase. The level of creatinine is important in determining the risk for CVD. Detecting high levels of creatinine is important as high levels alone may cause no symptoms.[3][4]

1.2 Problem

The heart is a vital organ and medical teams may fail to see a patient's risk of heart failure.

1.3 Audience

Medical doctors could benefit highly from accurate predictions of future heart failure related events from a patient's medical record.

Machine learning applied to medical records can be an effective tool to predict the survival of patients with heart failure symptoms. An important task is to identify the most important risk factors that may lead to heart failure. To use patient data from electronic medical records could help doctors detect symptoms otherwise undetected.

2. Data

2.1 Data source

Heart failure clinical data set can be found on UCIs machine learning repository.
https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

2.2 Description

The dataset contains 299 medical records of heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad, Pakistan, during April-December 2015. Dataset is provided from Davide Chicco, Giuseppe Jurman.[5]

First line in dataset contains labels.

Filname: heart_failure_clinical_records_dataset.csv
Filtype: CSV

The dataset contains 13 features:

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40,..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
(CPK)			
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving	Percentage	[14,..., 80]
	the heart at each contraction		
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01,..., 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50,..., 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114,..., 148]
Smoking	If the patient smokes	Boolean	0, 1
Time	Follow-up period	Days	[4,...,285]
(target) death event	If the patient died during the follow-up period	Boolean	0, 1

Feature description with label names and position in file:

age: Ranges from 40 to 95.
anaemia: Decrease of red blood cells or hemoblobin. (0=No, 1=Yes).
creatine_phosphokinase: Levels of creatine phosphokinase enzyme in blood. High levels might indicate injury or heart failure
diabetes: Does patient have diabetes (0=No,1=Yes)
ejection_fraction: Percentage of how much blood is pumped with each contraction
high_blood_pressure: High blood pressure (0=No, 1=Yes). Actual values not in dataset.
platelets: Level of platelets, also called thrombocytes, in blood. This is important to prevent bleeding.
serum_creatinine: Level of creatinine in blood. Waste product when muscle breaks down.
serum_sodium: Level of sodium in blood. A mineral that is important for correct functioning of muscles and nerves
sex: Woman og man. (0=Woman, 1=Man)
smoking: Is patient a smoker (0=No, 1=Yes)
time: Number of days for follow-up
DEATH_EVENT: This is the result we are seeking.

2.3 Data used

The dataset contains several features that exist in a patient's medical record. Features relevant for predicting that a patient may die are serum_creatinine and ejection_fraction values. In addition it is relevant to investigate other features in the dataset have and if they correlate.

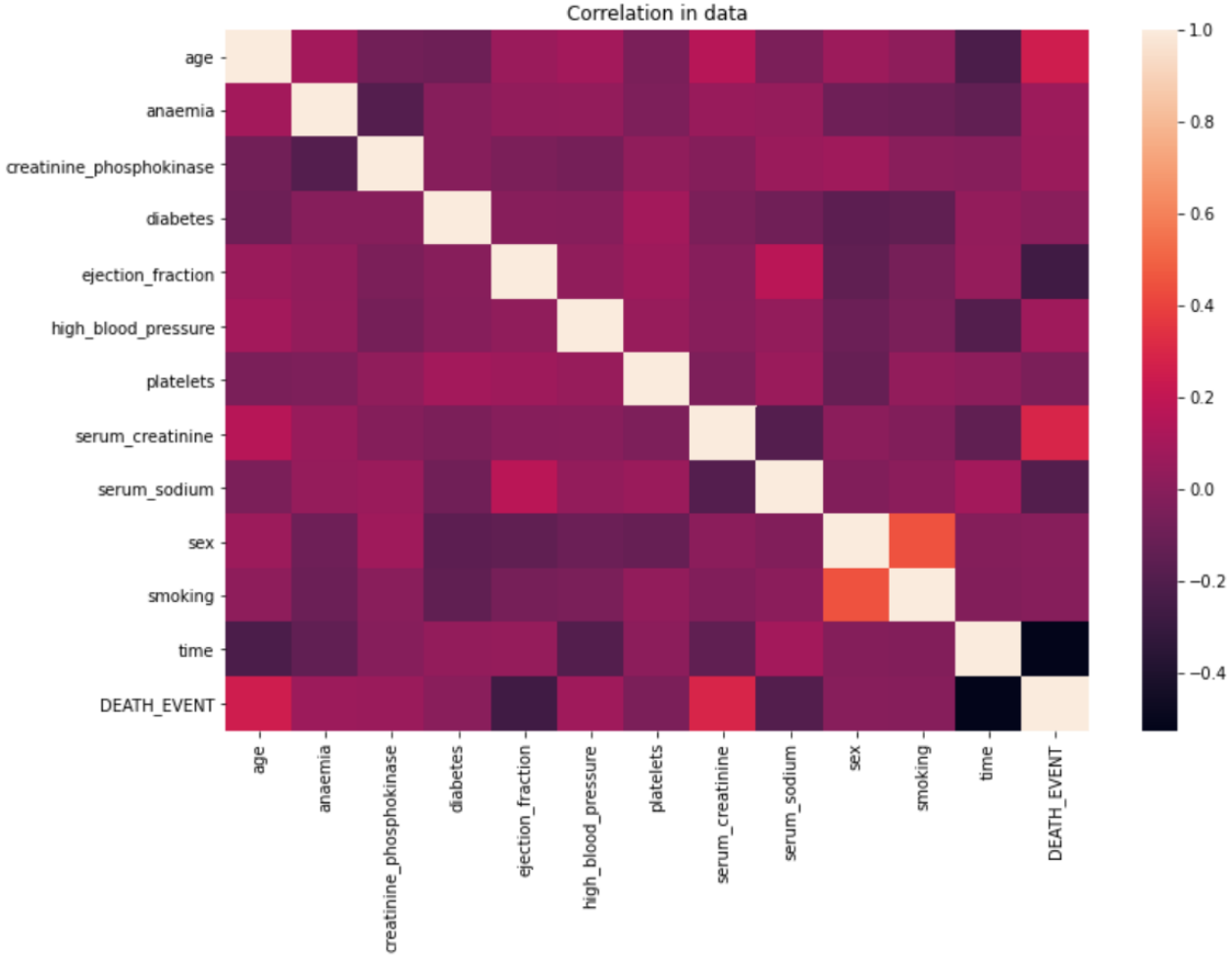
The dataset is checked for correct number of records and features. All data columns have correct data type and there are no null values.

3. Methodology

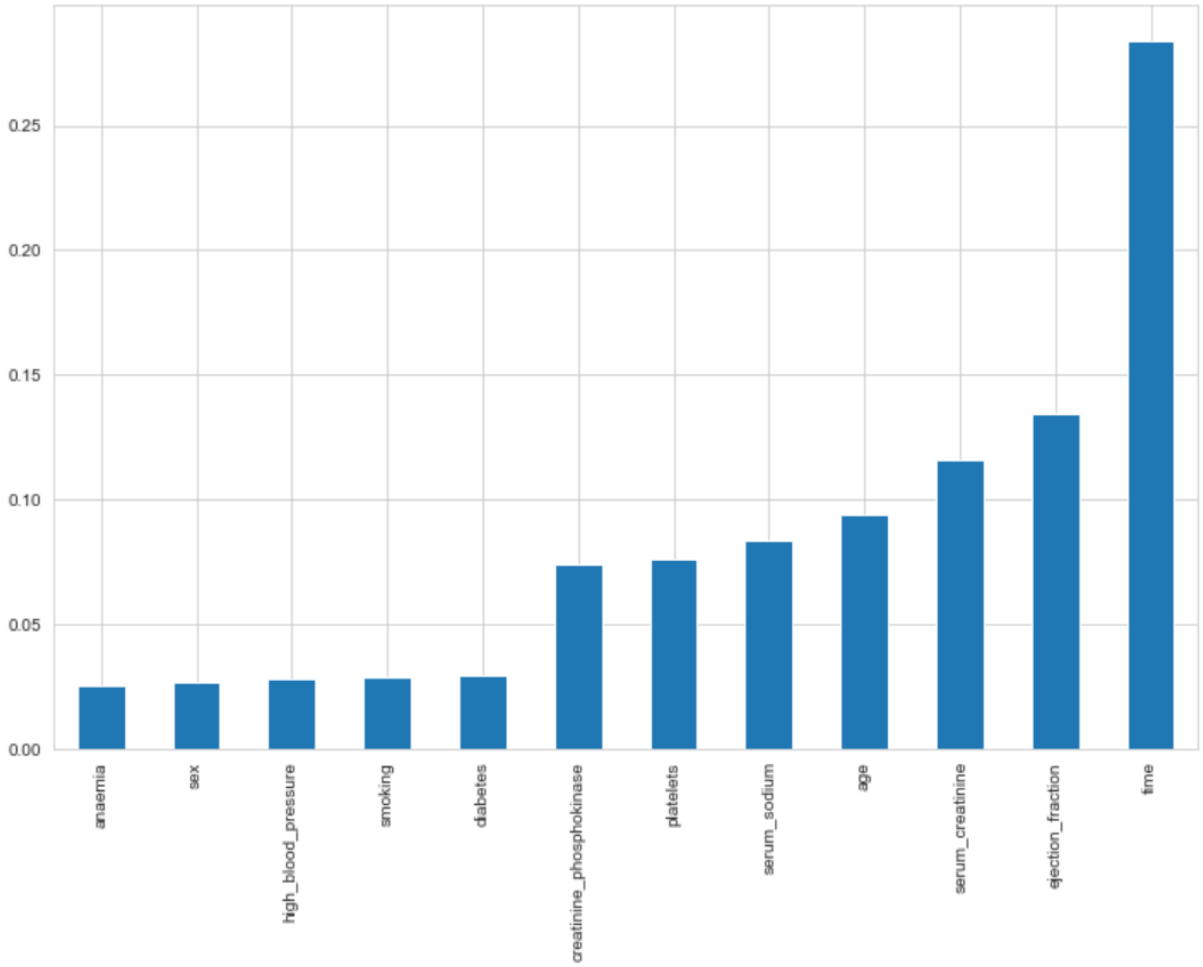
3.1 Exploring data and correlation

Of the 13 features in the dataset, five features are categorized as Boolean with integer representation of 0 and 1. This is the case for anaemia, diabetes, high_blood_pressure, sex and smoking. DEATH_EVENT is the feature which we want to predict. This leaves six fields: age, creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium and time.

In order to find the most import features to predict a patient's heart failure we can create a correlation diagram between all features. This gives a quick overview on what to look at:



Looking at DEATH_EVENT, a few features stand out: serum creatinine and age. Another way to investigate correlation, is to use extra trees classifier which builds decision trees and select the best features. The result is shown in the following graph:

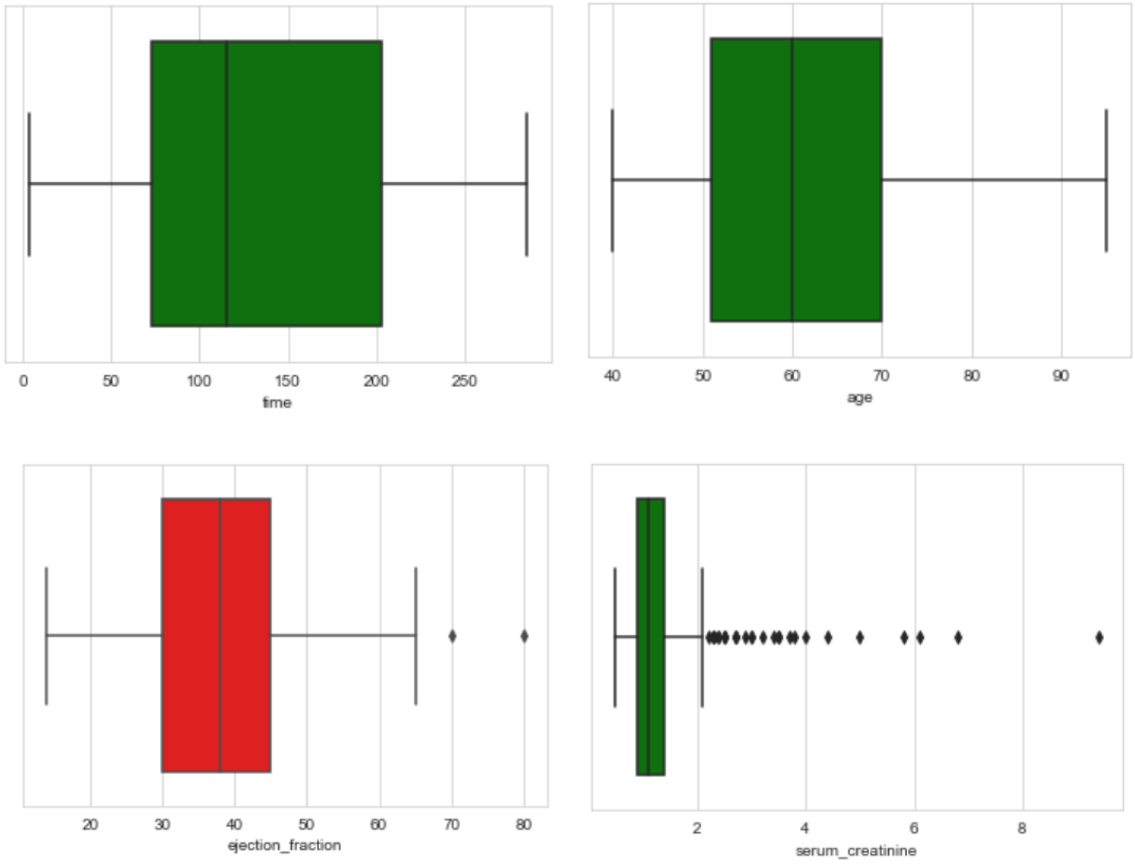


The graph shows time, ejection_fraction, serum_creatinine to be the most important features. Age is below the 0.1 mark and plays a role, too.

3.2 Selecting features

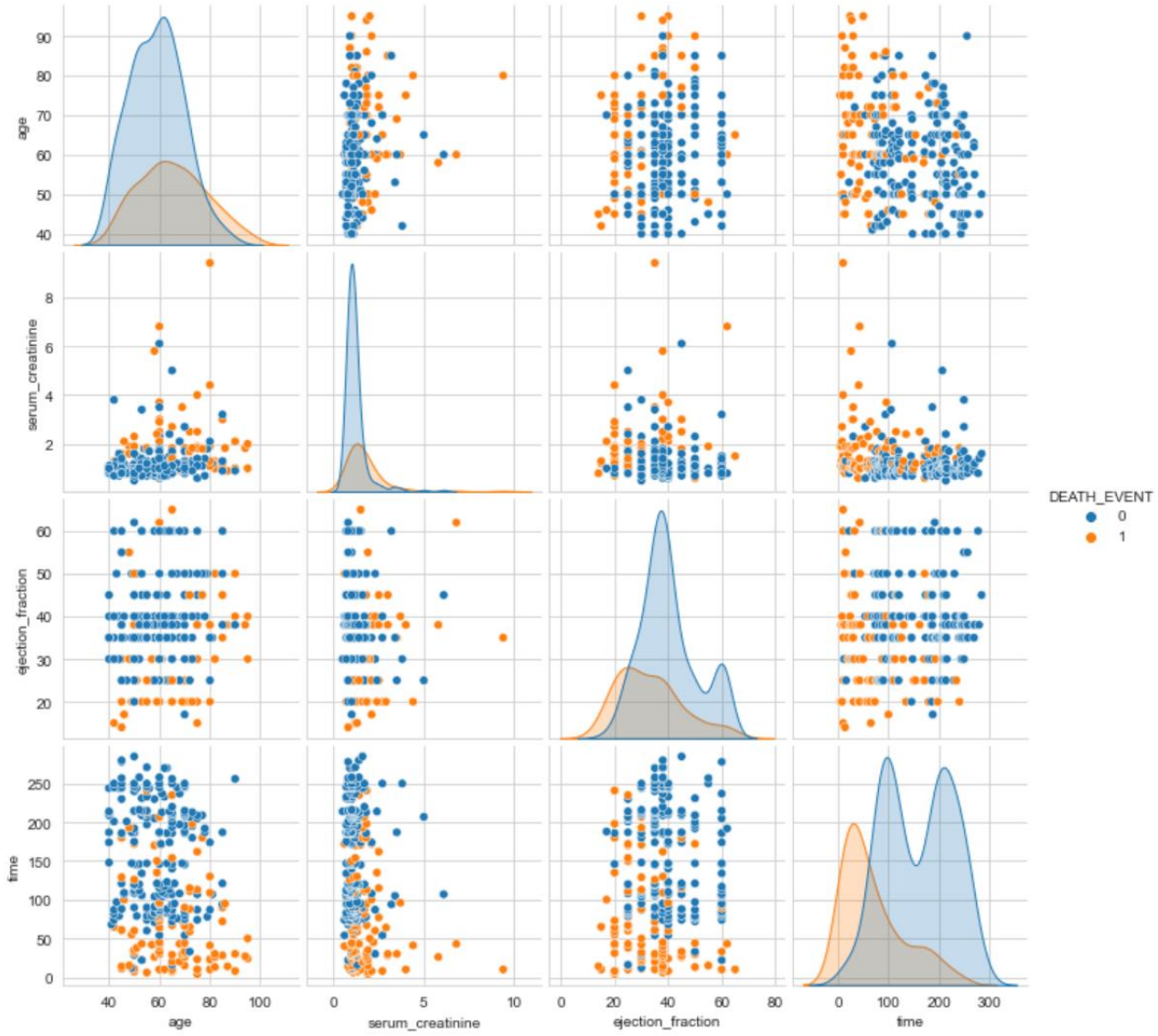
From initial analysis of the dataset, age, time, ejection_fraction and serum_creatinine are the four features to look closer at. As one would imagine features as diabetes, smoking and high blood pressure playing a role these features are investigated to see if they have an impact on predicting death.

A simple box plot is made of the selected key features to show outliers and distribution of values:



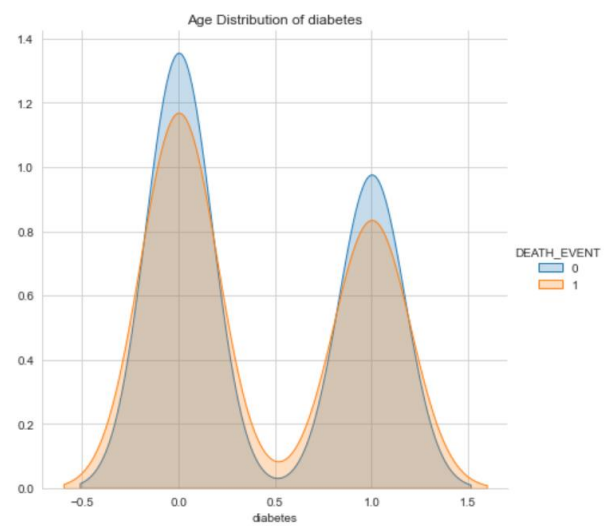
Neither time nor age have any outliers. Ejection fraction show outliers from 70 to 80, which are two records that are removed from the dataset. Serum creatinine looks as if there are many outliers, but the deviation is also narrow. There are 34 outliers that wasn't removed from the dataset.

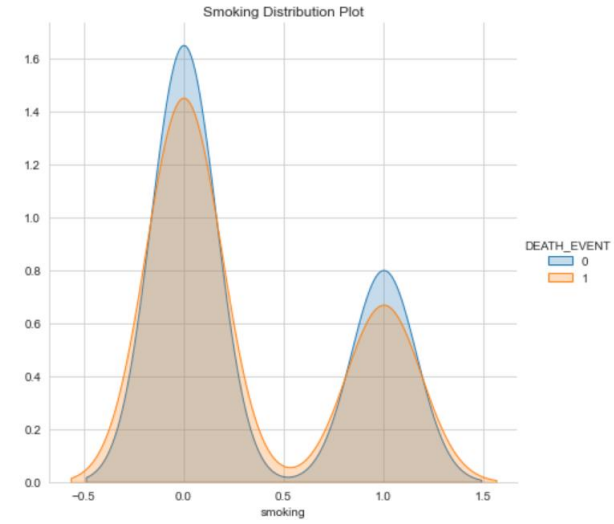
A closer look at the features (age, time, ejection_fraction and serum_creatinine) reveals more about which of the features play a role:

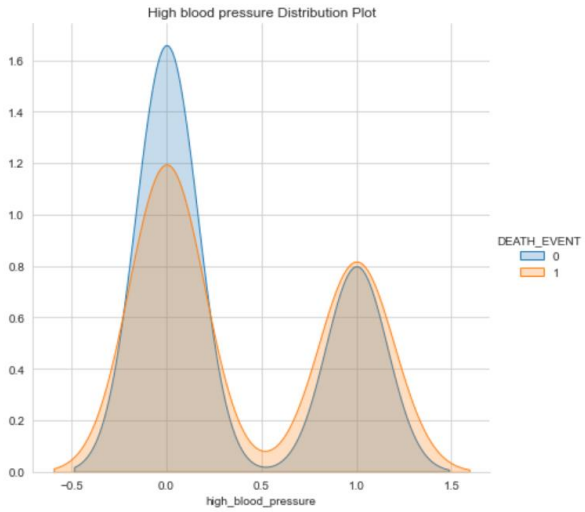


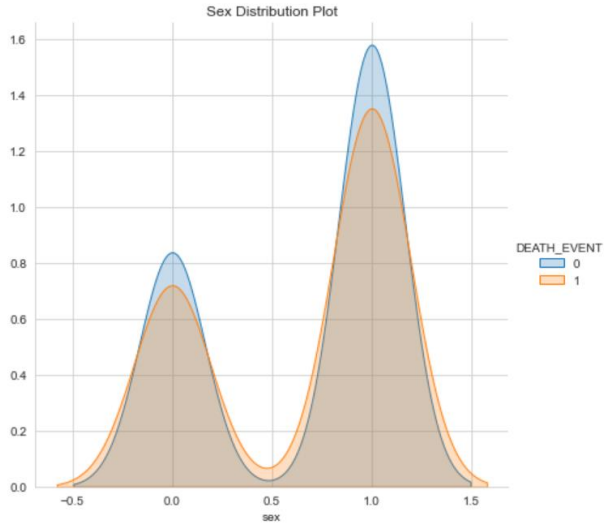
This confirms the features ejaction_fraction and serum_creatine as the key features. Age and time does play a role, but the main action are on the two.

This still leaves the question if diabetes, smoking, sex and high blood pressure are important factors influencing prediction of death by heart failure. Let us look at how close these features correlate with DEATH_EVENT in the dataset:









As we can see, all these features do not interfere with prediction of DEATH_EVENT.

3.3 Testing various models

As the dataset is only 297 records (after removing 2 outliers mentioned in the last section), a question was how much should go into the test set. One problem to avoid was overfitting and another having enough test data for the models to work. The final solution was 75% into training and 25% into testing.

The following five models have been trained and evaluated:

- Decision Tree

- K Nearest Neighbor

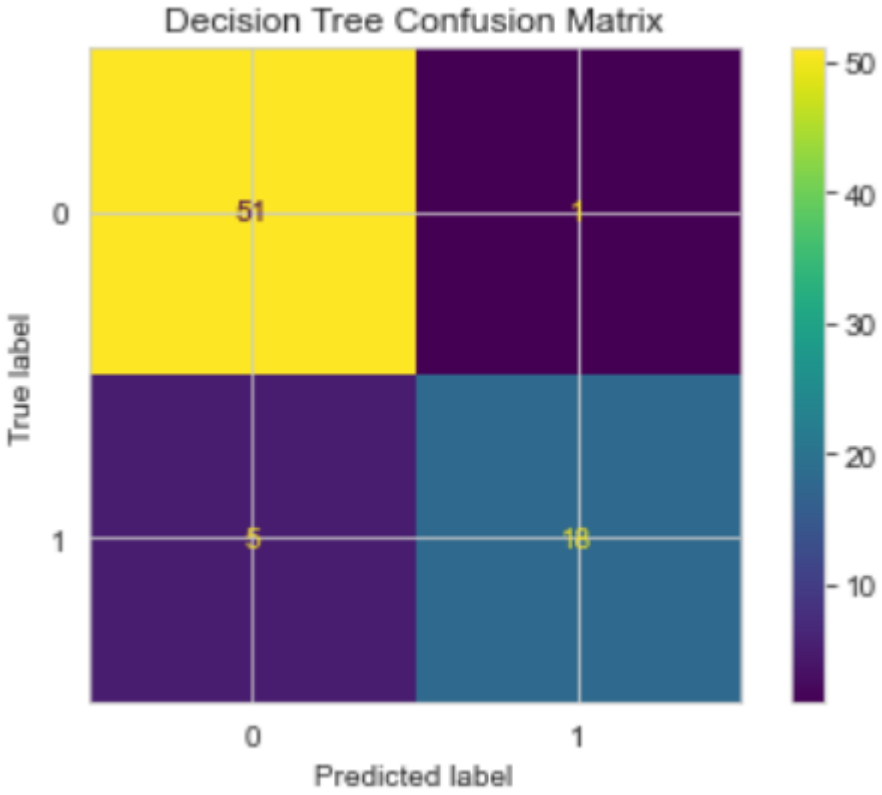
- Support Vector Machines

- Logistic Regression

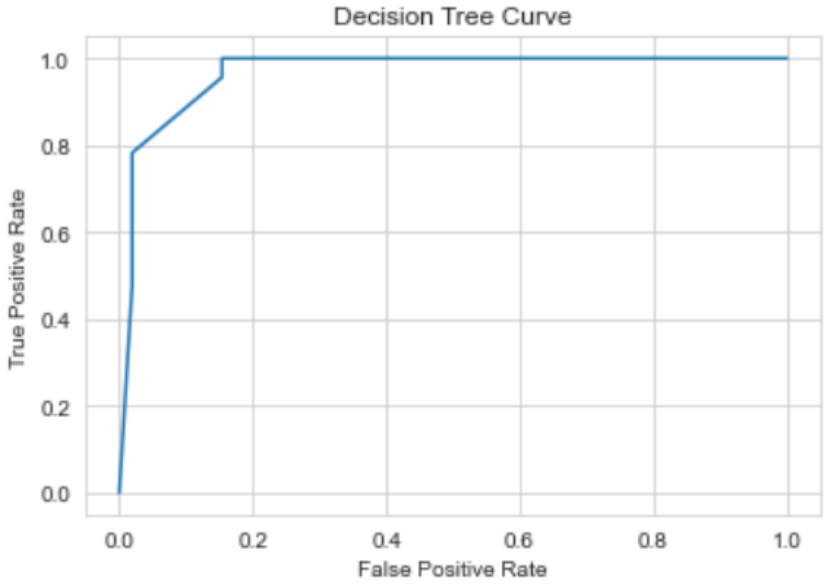
- Random Forest

3.3.1 Decision Tree

Decision Tree showed an accuracy of 0.92 and a F1-score of 0.917

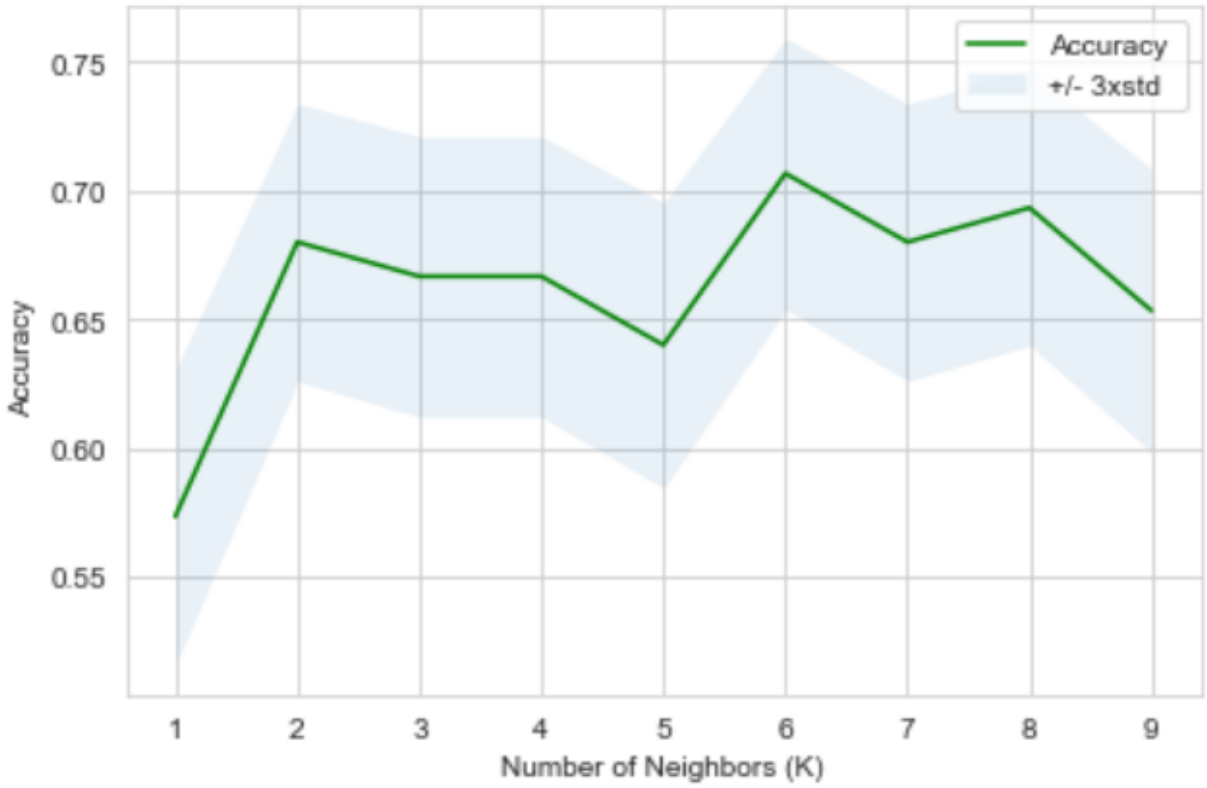


Decision Tree has a very high accuracy, but could have signs of overfitting. The following graph show the relation between true positive and false positive predictions.



3.3.2 K nearest neighbor

Selecting the K value for this model was done in a loop from 1 to 10 where the K with the highest value was selected. The graph shows the results from each k tested:



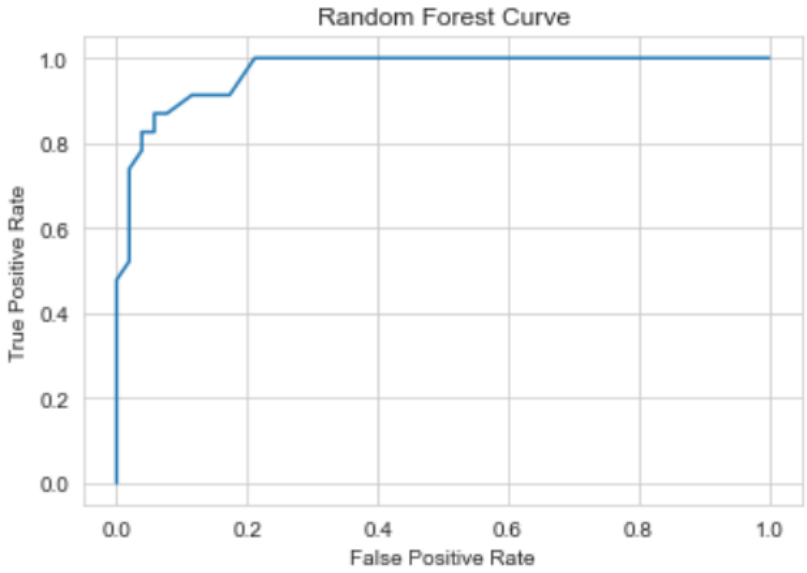
As can be seen, k = 6 delivered the highest accuracy with a value of 0.706 and a F1-score of 0.66

3.3.3 Support vector machines (SVM) and Logistic Regression

Due to the distribution of data, SVM and Logistic regression was only made to make sure there were no surprises. The accuracy of SVM was 0.69 and a F1-score of 0.609. Logistic regression showed an accuracy of 0.69 and a F1-score of 0.568.

3.3.4 Random Forest Classifier

Random forest classifier is a model that works with a number of decision tree classifiers and works on various subsets of the data which is averaged. The reason to take this model is that decision tree in 3.3.1. showed a very high accuracy, but with fear of overfitting. Random forest classifier tries to control overfitting. The accuracy of the model is 0.92 with an F1-score of 0.568



3.4 Selecting model

The results for the various models are:

Model	Accuracy	F1 score
Decision Tree	0.92	0.917
KNN	0.706	0.66
SVM	0.609	0.609
Logistic regression	0.609	0.568
Random forest tree	**0.92**	**0.568**

From the results, decision tree models have the best prediction. Logistic regression is the worst model to use. A concern for decision tree is overfitting. The recommended model is random forest tree that gives a very high accuracy while still being useful to unseen data.

4. Results

The results show that ejection fraction and serum creatinine is sufficient to predict heart failure. Follow up time is an important factor, nevertheless the analysis shows that these two features are more accurate.

5. Discussion

Although diabetes, smoking and high blood pressure have an impact on patients, these features do not change the predictability of using ejection fraction and serum creatinine. It could be interesting to collect more data on diabetes, smoking and high blood pressure to see exactly how these features play a role. From a medical point of view, diabetes and smoking play a secondary role when dealing with heart failure. The dataset does not show any effect large enough for these markers to be important.

It must be mentioned that the dataset has only 299 records from two hospitals in Pakistan. So, training and test set were small.

6. Conclusion

Using ejection fraction and serum creatinine from patients' medical records is an important marker for doctors to indicate if a patient will die from heart failure. Doctors and patients will benefit from using this tool in daily practice.

References

1. WHO Cardiovascular diseases (CVDs). (https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))

2. Heart.org: Ejection Fraction Heart Failure Measurement. (https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement#:~:text=What%20is%20%E2%80%9Cejection%20fraction%E2%80%9D%3F%20Ejection%20fraction%20%28EF%29%20is,left%20ventricle%20is%20pushed%20out%20with%20each%20heartbeat.)

3. Why are blood creatinine levels checked? (https://www.medicinenet.com/creatinine_blood_test/article.htm)

4. Learn why doctors test creatinine levels after heart attack. (https://www.healthguideinfo.com/heart-attack/p65166/#heart-attack--creatinine)

5. : Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020) (https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5#citeas)

6. BMC Dataset. (https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5#Sec2)