

Heart Failure prediction

Ke-Chen Lin

Abstract

This project uses the dataset that contains health condition information of patient who have cardiovascular diseases. Factor analysis and a prediction model is built by decision tree method, which are used to find out any factor at high risk and if the mortality of patient can be predicted.

Motivation

Cardiovascular diseases (CVDs) are the top 1 cause of death globally, with approximately 17 million deaths each year, which account for 31% total deaths worldwide.

Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality from heart failure.

If heart failure can be predicted by a machine learn model that it can be used by hospitals to do the treatment with early detection and management.

Dataset(s)

This dataset I found in Kaggle (<https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>) which contains 12 features such as age, sex, diabetes, smoking, follow-up period, hypertension, anaemia, and levels of chemical substance. This dataset also recorded mortality in 300 samples.

Data Preparation and Cleaning

Data preparation will use filter to select high correlation coefficient factor for further analysis.

Research Question(s)

Which factors have a high risk to cause heart failure?

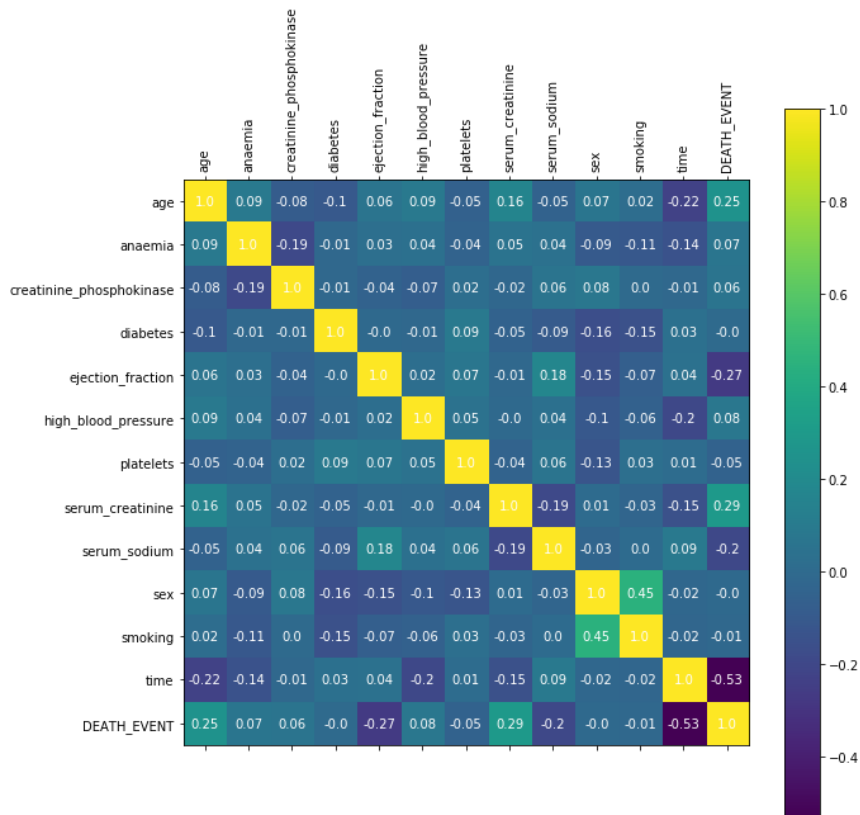
Is heart failure can be predicted with machine learning mode?

Methods

I firstly use correlation coefficient analysis to choose features which have higher correlation.

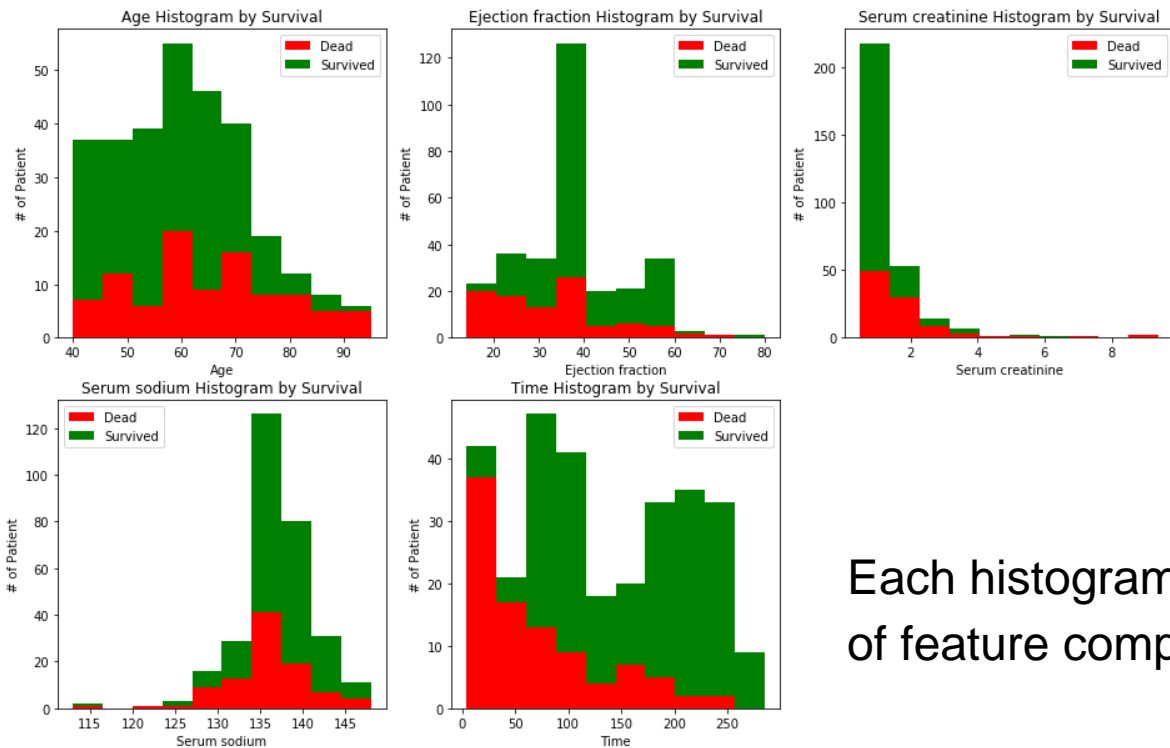
Next, the selected features dataset and death result are separated with train/test method, then do the prediction with a decision tree model.

Findings



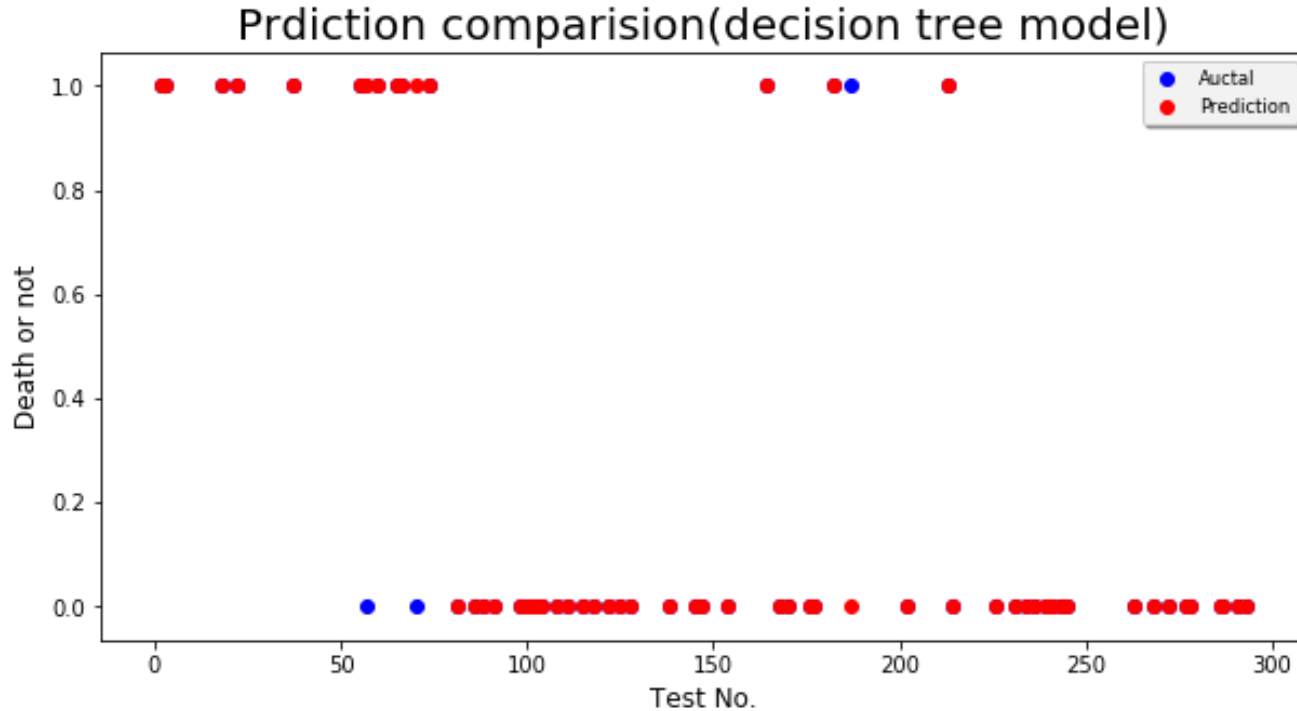
The features are selected for correlation coefficient with 'DEATH_EVENT' > 0.1 , 'age', 'ejection_fraction', 'serum_creatinine', 'serum_sodium', and 'time'

Findings



Each histogram plot represents the distribution of feature compare with dead or survived.

Findings



Accuracy score is 0.95 created by the decision tree model

Limitations

Although the accuracy score is high enough, there may still have other factors that we are not considered. In this prediction model, only decision tree model is tested may also not enough for practical usage.

Conclusions

The coefficient results show mortality by heart failure, which can be influenced by the factor of patient's age, ejection fraction by heart, serum creatinine and sodium level in the blood, and follow-up period.

From the decision tree model, death event can be predicted with accuracy score 0.95, but still needs further evaluation if any more factor be joined.

Acknowledgements

Davide Chicco, Giuseppe Jurman: 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).

References

All the Cardiovascular diseases(CVDs) information is from the website of World Health Organization

https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1