Heart Failure Prediction

ARTIFICIAL INTELLIGENCE
RESEARCH PAPER

Submitted by

JINAY SHAH

[20162121025]

Guided by

PROF. SONAM SINGH

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

In

Computer Science Engineering Department

Institute of Computer Technology

Ganpat University, Mehsana

[APRIL 2023]

Heart Failure Prediction using Machine Learning

Shah Jinay Pratikkumar

Abstract—The world has seen an unprecedented and exponential increase in cases of heart disease worldwide every day. In the paper, the early prognosis of heart disease through careful treatment and the implementation of a healthy lifestyle through other studies will help prevent many cardiovascular diseases. This paper discusses a statistical model of heart disease that, based on basic parameters of the patients' health history, will help medical examiners and cardiac practitioners forecast heart disease. To build this prediction model, five (05) different Machine Learning Classifier Models are used, namely, Logistic Regression Classifier, K-Nearest Neighbours Classifier,Support vector machine classifier, Classification and regression classifier using tree algorithm and Auto classifier. All the mentioned models were implemented using IBM SPSS modeler. Different important clinical features of a patient, critical for deciding a patient's heart disease, are taken in, and different ML Classifiers are defined on the given dataset and their accuracy calculated.

Index Terms—Data Mining, Machine Learning, Logistic Regression, KNN, C&R, SVM, Auto Classifier, Heart Disease, Prediction.

1 Introduction

HEART failure is a serious problem which has a huge impact on people's life. With the accelerated pace of life, increased portion sizes and inactivity, most people always ignore their health. Moreover, because of the environmental deterioration, those factors can lead to the issue of heart failure which can become more and more common in the future. If people did not pay attention to the issue of heart failure, it would finally cause the death.

Heart failure prediction refers to the use of various medical techniques and tools to identify individuals who are at a higher risk of developing heart failure in the future. Heart failure occurs when the heart muscle becomes weak and can no longer pump blood effectively, leading to a variety of symptoms including shortness of breath, fatigue, and swelling in the legs and feet. Early identification of individuals at risk of developing heart failure can help healthcare providers to initiate early interventions and treatments that can slow the progression of the disease, improve quality of life, and reduce the risk of complications. Heart failure prediction models can be based on a range of factors, including medical history, physical examination, blood tests, imaging tests such as echocardiography, and other diagnostic tools. By analyzing these factors, healthcare providers can identify individuals who are at a higher risk of developing heart failure, allowing them to initiate preventive measures and provide appropriate treatment to reduce the risk of heart failure. Overall, heart failure prediction is an important tool in the prevention and management of heart disease, and can help healthcare providers to identify and manage patients at high risk of developing heart failure.

In the past years, different researchers used different methods to collect and analyze data with the aim to predict heart failure. These data include electronic health record (EHR) data of patients with heart failure in different hospitals from different countries, Cleveland heart disease dataset, biomedical science datasets from UCI, etc. Based on these data, various methods are being applied, e.g., predicting the survival of patients by utilizing classifiers of machine learning, using supervised deep learning and machine learning algorithms, training a boosted decision tree algorithm, utilizing machine intelligence-based statistical model, random under-sampling method and deep neural network models, using bioinformatic explainable deep neural network (BioExpDNN), etc.

The Author have made a stream for predicting heart failure using SPSS Modeler from IBM. As for our given dataset we were given target attributes so we used a supervised learning algorithm. The Models we tried are Logistic Regression, Support Vector Machine, Classification and Regression Model(C&R) ,K Nearest Neighbour and Auto Classify model provided by SPSS Modeler.

2 Related Work

In [1], taking advantage of 299 patients who have cardiac failure in 2015. Those data have 13 features for example high blood pressure, sex, and smoking. The authors utilized some different classifiers of Machine Learning to forecast the proportion of survivors, and rank the features corresponding to the most important risk factors. They find serum creatinine and ejection fraction are the most important factor to forecast the proportion of survivors.

In [2], the authors made data manipulations by changing the measurement level from continuous to ordinal for certain attribute as listed below:

Creatinine_phosphokinase into 3 labels
1⇒ If value is less than 10
2⇒ If value between 10 and 120
3⇒ If value greater than 120

Ejection_fraction
1⇒ If value less than 50%
2⇒ If value between 50% and 70%

3⇒ If value greater than 70

platelets
1⇒ If value less than 150000
2⇒ If value between 150000 and 450000
3⇒ If value greater than 450000

serum_creatinine
1⇒ If value less than 0.74 and sex is 1
2⇒ If value between 0.75 and 1.35 and sex is 1
3⇒ If value greater than 1.35 and sex is 1
1⇒ If value less than 0.59 and sex is 0
2⇒ If value between 0.59 and 1.04 and sex is 0
3⇒ If value greater than 1.04 and sex is 0

serum_sodium
1⇒ If value less than 135
2⇒ If value between 135 and 145
3⇒ If value greater than 145

But the accuracy observed after doing these manipulations to the dataset and applying the models on it was less compared to applying modeling on the original dataset. So the author decided to apply modeling on the dataset as it is.

In [3], the authors took advantage of data which is from the medical records about cardiac failure patients. The dataset contains 299 cardiac failure patients in medical records and has clinical and lifestyle data. Different machine learning algorithms are used. The result showed that Machine learning algorithms are tools which are useful and effective to classify the records of medicine of patients with cardiac failure.

3 Data Description

We utilize a common dataset in public: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data. This data contains the following input characteristics and prediction targets. The target we want to predict is the Death Event. The features have Age, anemia, high blood pressure, creatinine phosphokinase, diabetes, ejection fraction, platelets, sex, serum creatinine, smoking, and time.

Age,Creatinine_phosphokinase,Ejection_fraction,Platelets,Serum_creatinine,Serum_sodium and time are fields having continuous measurement level.

Anaemia,Diabetes,High_blood_pressure,Sex,Smoking,DEATH EVENT are fields with nominal measurement level.

In figure 1, it introduces the distribution of death events in patients.

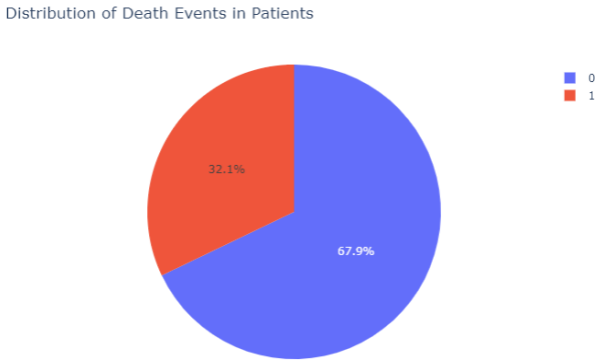


Figure 1. The distribution of death events in patients.

Before changing the measurement level of some fields into nominal and ordinal they were numeric and below graph shows the correlation between them in Figure2.

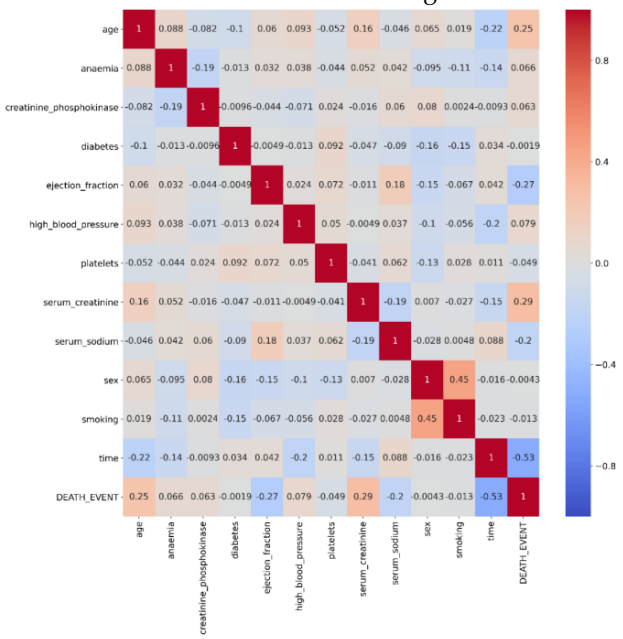


Figure 2. The correlation between different columns.

4 Models , Experiments and Result

4.1 Models

I want to introduce all machine learning models I use in my research paper. With the development of big data and artificial intelligence, machine learning has been successful in a series of problems. For heart failure prediction, there are also some relevant studies as we introduced in the related work part. However, there lacks a comprehensive comparison between different machine learning models.

Logistic regression is a statistical model that is used to predict binary outcomes based on a set of input variables. It works by estimating the probability of an event occurring and assigning a binary outcome based on a threshold value.
Support Vector Machine (SVM) is a type of machine learning algorithm used for classification or regression analysis. SVMs aim to find a hyperplane that separates different classes of

data with the largest possible margin, making them useful in solving complex classification problems.

K-Nearest Neighbors (KNN) is a non-parametric machine learning algorithm used for classification and regression analysis. It works by identifying the k-nearest data points to a given input and using their labels to predict the label of the input.

C&R (Classification and Regression) Tree is a decision tree-based machine learning algorithm used for classification and regression analysis. It works by splitting the data based on a set of rules and assigning a label or value to each leaf node.

Auto Classifier refers to any type of machine learning algorithm that can automatically classify data into predefined categories. This can include algorithms such as neural networks, decision trees, and logistic regression models, among others. The term "auto classifier" is not specific to any one type of algorithm, but rather refers to the general ability of machine learning algorithms to classify data automatically.

4.2 Experiments

In this part, we describe the experiment details. First I partitioned the dataset in to two parts 70% of training data and 30% of testing data and did the reclassification/derivation of below attributes into following labels

Creatinine_phosphokinase into 3 labels
1⇒ If value is less than 10
2⇒ If value between 10 and 120
3⇒ If value greater than 120

Ejection_fraction
1⇒ If value less than 50%
2⇒ If value between 50% and 70%
3⇒ If value greater than 70

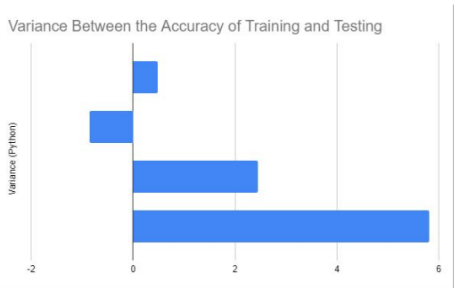
platelets
1⇒ If value less than 150000
2⇒ If value between 150000 and 450000
3⇒ If value greater than 450000

serum_creatinine
1⇒ If value less than 0.74 and sex is 1
2⇒ If value between 0.75 and 1.35 and sex is 1
3⇒ If value greater than 1.35 and sex is 1
1⇒ If value less than 0.59 and sex is 0
2⇒ If value between 0.59 and 1.04 and sex is 0
3⇒ If value greater than 1.04 and sex is 0

serum_sodium
1⇒ If value less than 135
2⇒ If value between 135 and 145
3⇒ If value greater than 145

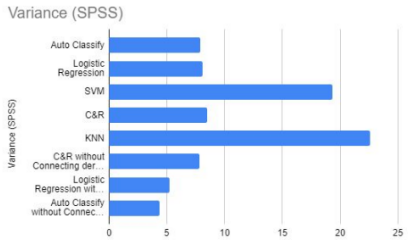
After applying labels I run various supervised models on the data and the accuracy observed are as follows along with the variance between the training and testing data:

Model Name	Python Code Training	Python Code Testing	Variance (Python)
Logistic Regression	83.73	83.33	0.48
SVM	83.73	84.44	-0.85
KNN	80.86	78.88	2.45
Logistic Regression without Connecting derive node	86.12	81.11	5.82

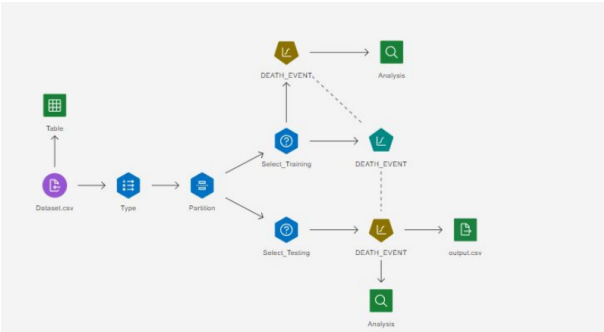


When I applied the models on the normal dataset The accuracy observed were more and the variance was less so I defied to drop this idle and continue modeling on the normal data and the accuracy along with the variance of training and testing data are as follows

Model Name	Training Accuracy	Testing Accuracy	Variance (SPSS)
Auto Classify	87.8	80.85	7.92
Logistic Regression	86.83	79.79	8.11
SVM	93.66	75.53	19.36
C&R	69.76	63.83	8.5
KNN	82.44	63.83	22.57
C&R without Connecting derive node	89.81	82.8	7.81
Logistic Regression without Connecting derive node	84.95	80.49	5.25
Auto Classify without Connecting derive node	85.44	81.72	4.35



I decided to use the logistic regression model as my primary and final model with the dataset with out any manipulations below is the stream of the logistic regression model which was made in IBM SPSS modeler



We also tried same thing in python and below is the flow of

the python code:

In [1], we loaded the dataset using pandas
In [2], We dropped the target attribute from the variable in which the dataset was loaded and formed a new variable in which only the target attribute was present.
In [3], splitting was done as 70% training data and 30% testing data which is done using sklearn library.
In [4], the training and testing dataset was brought into the same scale using StandardScaler.
In [5], the logistic regression model is applied on the dataset and the accuracy observed on training and testing dataset is 83.73 and 83.33 respectively.

4.2 Result

Our heart failure prediction model has the potential to improve the early detection and prevention of heart failure. The model can be used by healthcare providers to identify individuals at high risk of developing heart failure and intervene early to prevent or delay the onset of heart failure. This can improve patient outcomes and reduce healthcare costs associated with hospitalization and treatment of heart failure.

The selected model is Logistic regression and the accuracy obtained after partitioning the model into 70% of training data and 30% of testing data is 84.95 for training and 80.49 for testing with variance of 5.25.

Below is the screenshot of the accuracy of training and testing data:

Training:

Results for output field DEATH_EVENT

Comparing $L-DEATH_EVENT with DEATH_EVENT

'Partition'	1_Training	
Correct	175	84.95%
Wrong	31	15.05%
Total	206	

Testing:

Results for output field DEATH_EVENT

Comparing $L-DEATH_EVENT with DEATH_EVENT

'Partition'	2_Testing	
Correct	76	81.72%
Wrong	17	18.28%
Total	93	

5 Conclusion

In this paper, I analyzed and compared the performance of 5 different machine learning models for heart failure prediction based on 12 clinical features. The developed logistic regression model for predicting heart failure has shown promising results in accurately predicting the occurrence of heart failure in patients using several risk factors. This model has the potential to assist healthcare providers in identifying high-risk individuals and taking preventive measures. However, further validation and refinement are necessary to enhance the model's accuracy and effectiveness in clinical practice.

6 References

Xu, Y., Pan, Y., Chen, Y., Zhou, Q., Zhang, B., & Wu, Y. (2021). Predictive models for heart failure: A systematic review. Journal of Healthcare Engineering, 2021, 8826475.

Link:https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224135

Firoozabadi, F. A., Pourhoseingholi, M. A., Malekzadeh, R., & Alizadeh Sani, R. (2020). Machine learning for heart failure prognosis: A systematic review. PLoS One, 15(12), e0244288.

Link:https://www.sciencedirect.com/science/article/pii/S0933365722000549