Project Report

Heart Failure Classification

Ella Wang, Yanfei Wang, Jingyi Wang

Table of Content

1. Project

1.1 Project Description

First, just to shortly describe “Heart Failure”, heart failure happens when the heart cannot pump enough blood and oxygen to support other organs in your body. It is a serious condition, but it does not mean that the heart has stopped beating. In the United States, there are about 6.2 million adults who have heart failure.

This project interests us since the death following heart failure could be caused by various factors. There is a rumor that the death caused by heart failure is most likely caused by unhealthy lifestyle/habits, such as tobacco smoking, which is also included as a feature in the dataset. Testing if the rumor is accurate might reveal to the public the “fact” to individuals to determine if they are at a high risk of having heart failure by themselves, not only currently, but also for a certain age group in the future.

The data are collected from the clinic datasets and the target variable is “Death_Event”, which shows if the patient is diseased during the follow-up period. Since it is a boolean variable, the chosen classification models are built and evaluated.

1.2 Project Deliverables

1. Project Charter (Sep 16th)
2. Ensemble a trained model with preliminary results of the training and testing. (Nov 12th)
3. Flask and Heroku application.
4. Final report of the project in addition to the cloud (Heroku) deployed link.

1.3 Project Steps

1. Exploratory data analysis
2. Binary classification baseline evaluations
3. Ensemble method evaluation
4. Export the model as pickel

5. Develop the web-based app on the flask
6. Deploy it on AWS/Azure

1.4 Project Constraints

1. The dataset does not have a huge amount of data points.
2. The information source of the dataset. The dataset gathering process could be limited to certain areas or institutions. This would limit the variety of the data and would influence the accuracy of classifications on the variables.
3. All the variables are only related to the patient's own health condition; No other environment or other conditions are into consideration.

2. Introduction

Heart problems are always a huge health concern. Many factors have a significant influence on the heart condition which may cause death. UCI has collected data on different health conditions on a sample of patients, and whether their condition leads to the death of heart failure. The purpose of this project is to analyze the dataset and find out which factor has a significant influence on the death of heart failures.

Holy Cross Hospital is good for treating heart diseases, however, they are finding some methods to help potential patients to prevent heart failure. Thus, it is important for them to find out the important factors that would cause death from heart failure. In our application, the hospital can input the history data of patients and the website will output the important features based on our models, which will help them make better decisions and help people prevent heart failure in advance.

3. Analysis of the Dataset and Models

3.1 Preprocessing and Exploratory Data Analysis

Table 1 Summary of the variables

parameters	age	creatinine_ phosphokinase	ejection_ fraction	platelets	serum_ creatinine	serum_ sodium	time
count	299.00	299.00	299.00	299.00	299.00	299.00	299.00
mean	60.83	581.84	38.08	263358.03	1.39	136.63	130.26
std	11.89	970.29	11.83	97804.24	1.03	4.41	77.61
min	40.00	23.00	14.00	25100.00	0.50	113.00	4.00
25%	51.00	116.50	30.00	212500.00	0.90	134.00	73.00
50%	60.00	250.00	38.00	262000.00	1.10	137.00	115.00
75%	70.00	582.00	45.00	303500.00	1.40	140.00	203.00
max	95.00	7861.00	80.00	850000.00	9.40	148.00	285.00

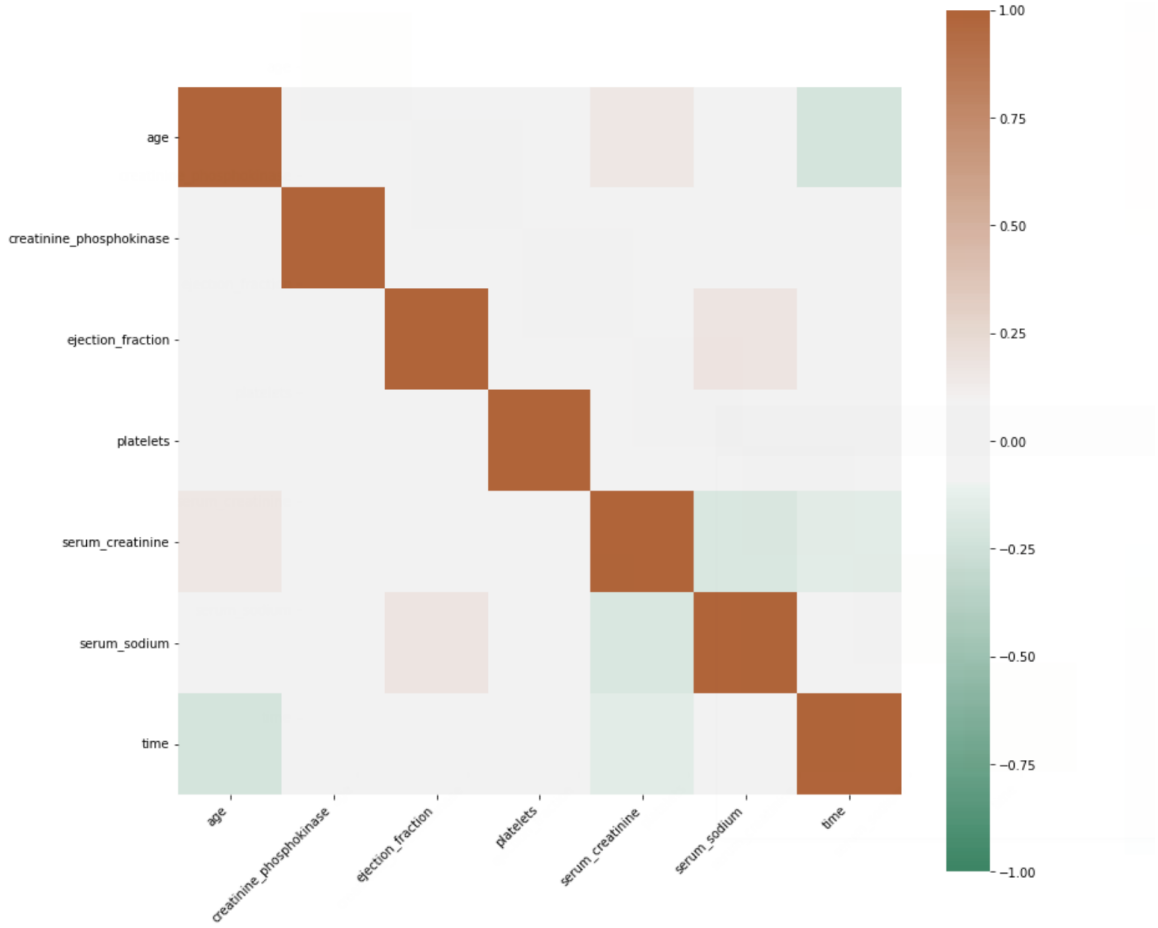


Figure 1 Correlation Plot of numerical variables

Figure 1 doesn't show heavily correlated numeric variables according to the plot. Thus, it is not necessary to delete any variable for further analysis.

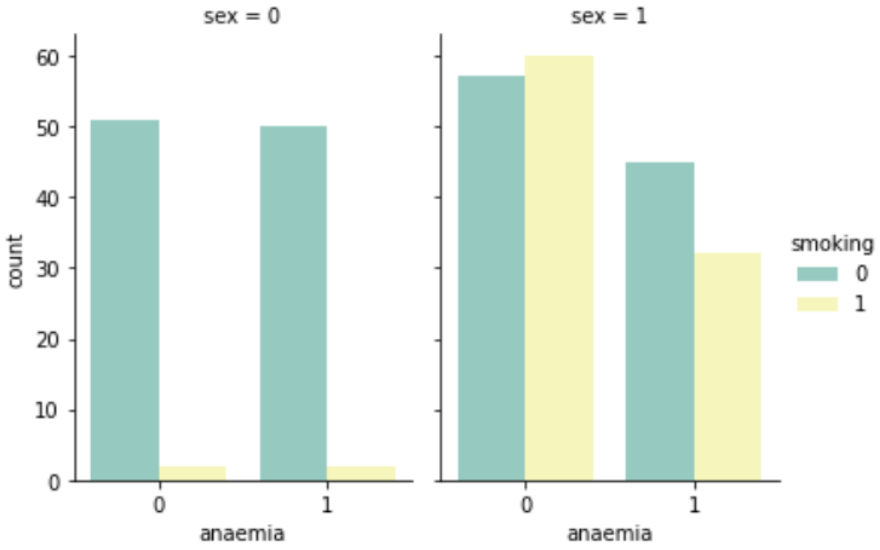


Figure 2 Categorical Plot of anaemia and smoking

Figure 2 shows that there's are significant correlation between sex and anaemia; smoking and anaemia. Smoking will increase the rate of having anaemia and males have a higher chance to have anaemia.

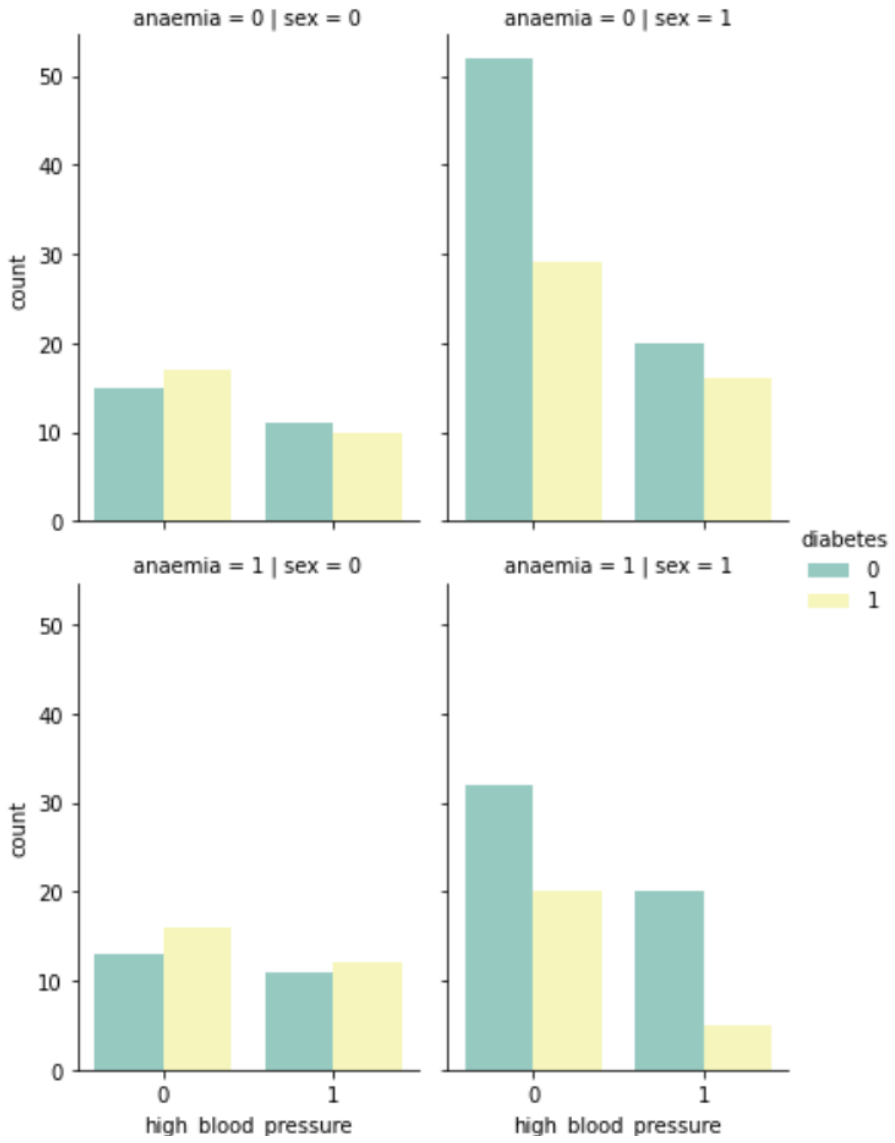


Figure 3 Categorical Plot of anaemia, diabetes, and high blood pressure

Figure 3 shows that there is not a super high correlation between high blood pressure and diabetes and anaemia. However, sex does influence the high blood pressure rate as males are more likely to have high blood pressure and anaemia.

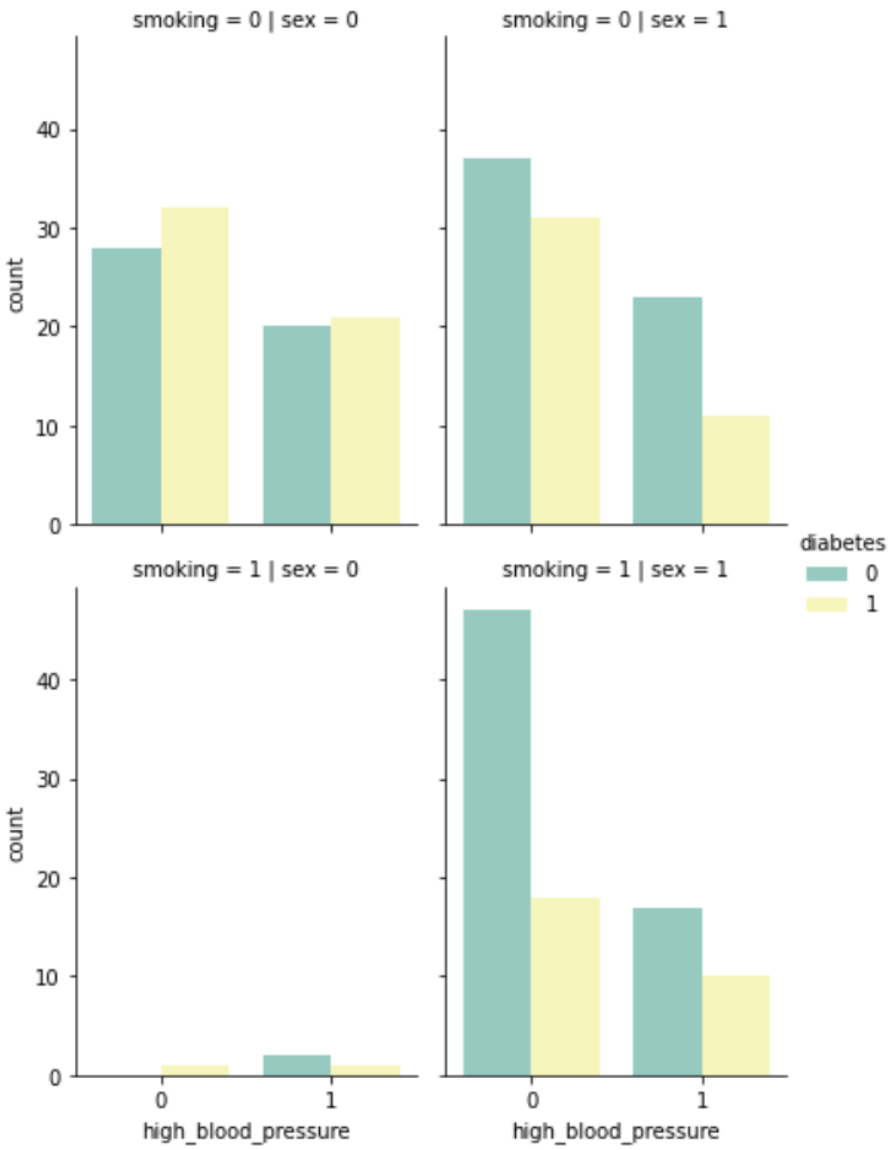


Figure 4 Categorical Plot of smoking, diabetes and high blood pressure

Figure 4 shows there are some correlations. If a person smokes, then it is more likely he/she has high blood pressure. And people who have diabetes are less likely to have high_blood_pressure. Males are usually more likely to have high_blood_pressure.

Figure 5 Scatter Matrix of heart failure data

3.2 Base Models

Our target variable is Death_Event, and the factors are age, anaemia, creatinine_phosphokinase, diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, sex, smoking, and time. We used these factors to predict the death of heart failures and found out the important factors that cause the death.

Since this is a classification dataset, we used the following classification models to analyze the data: Logistic Regression, KNN, SVC, Decision Tree, Random Forest, Gradient Boosting, Gaussian Naive Bayes, and Bagging.

4. Model Selection

4.1 Best Estimated Model

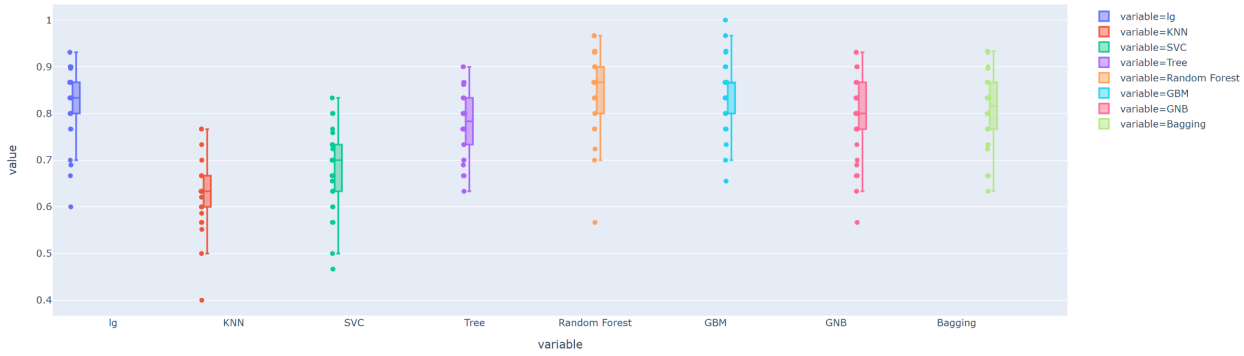


Figure 6 Machine Learning model results

This box plot (Figure 6) shows the model results. We can see that GBM gives the best accuracy value in classification analysis with an average accuracy of 0.84. And Bagging gives the second-best value. So we selected them as the level 0 classifiers and stacked them together to get a stacked model. We used Logistic Regression as the level 1 classifier to aggregate the results of the level 0 models.

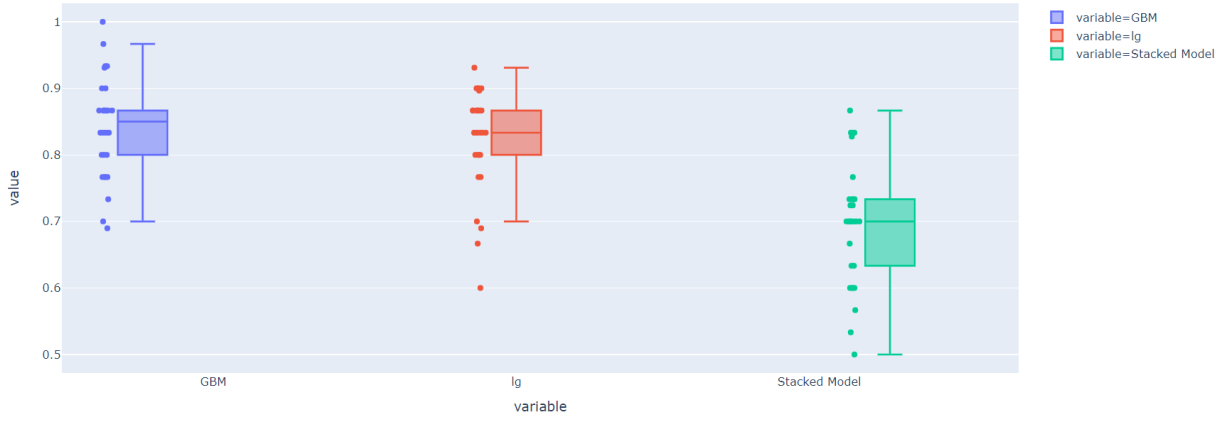


Figure 7 Level 1 model and Stacked model performance

This box plot (Figure 7) shows the performance of GBM, Logistic Regression, and the stacked model. From the plot, the accuracy of the stacked model is lower than the GBM model and the logistic regression model.

4.2 Model Performance Evaluation

4.2.1 GBM Model

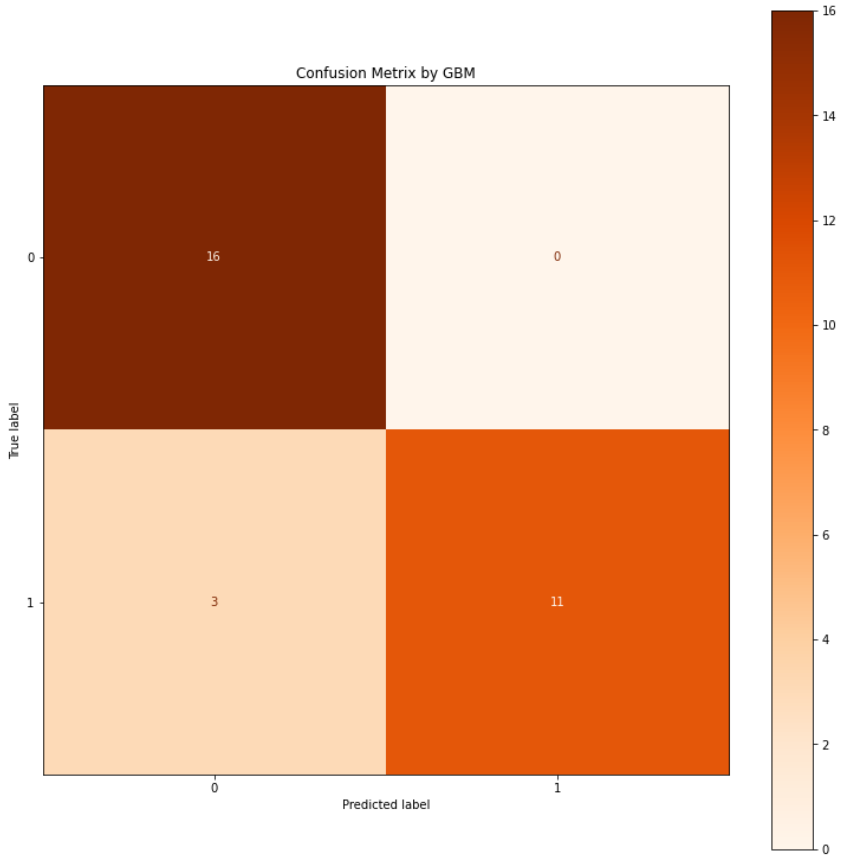


Figure 8 Confusion Matrix by GBM

Figure 8 shows relatively high accuracy with only 3 data points mispredicted. The accuracy of the GBM model is 90%.

4.2.2 Logistic Regression Model

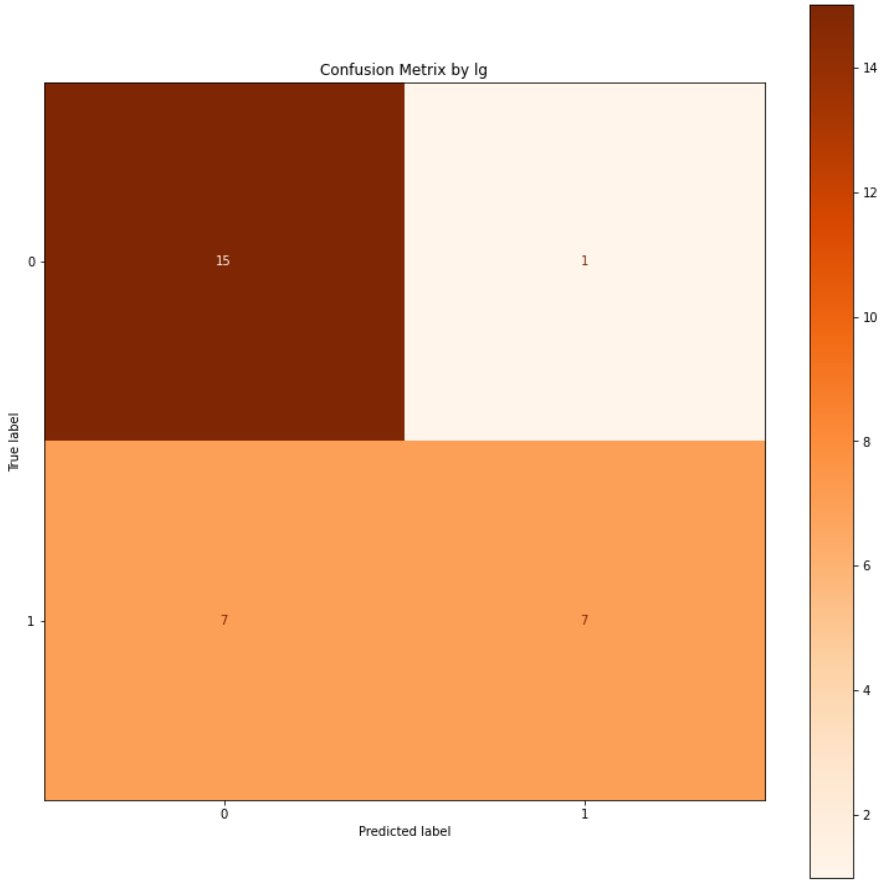


Figure 9 Confusion Matrix by Logistic Regression

This confusion matrix (Figure 9) shows that there are 8 data points mispredicted by the logistic regression model, so the accuracy is 73.33%, much lower than the GBM model.

4.2.3 Stacked Model

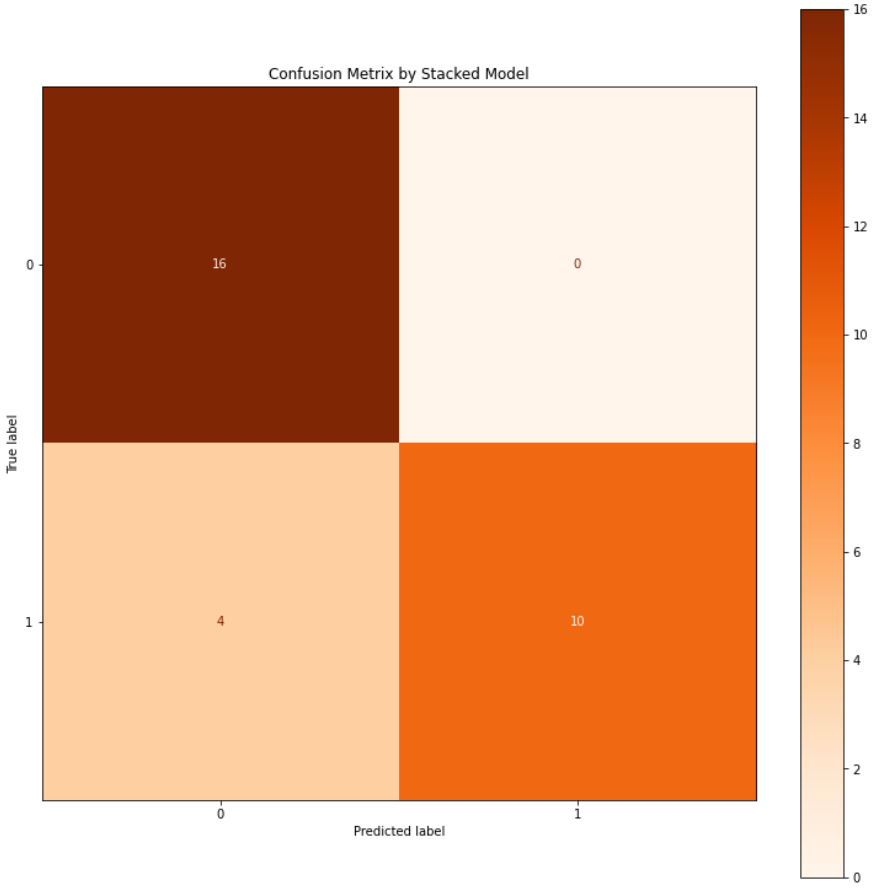


Figure 10 Confusion Matrix by Stacked model

Figure 10 shows the accuracy of 86.67% for the stacked model. There are 4 data points mispredicted by the model. The stacked model gives a better prediction than the logistic regression model but worse than the GBM model.

After careful comparison, it shows the GBM model has the highest accuracy which would be the best fit model for this dataset.

4.2.4 Feature Importance

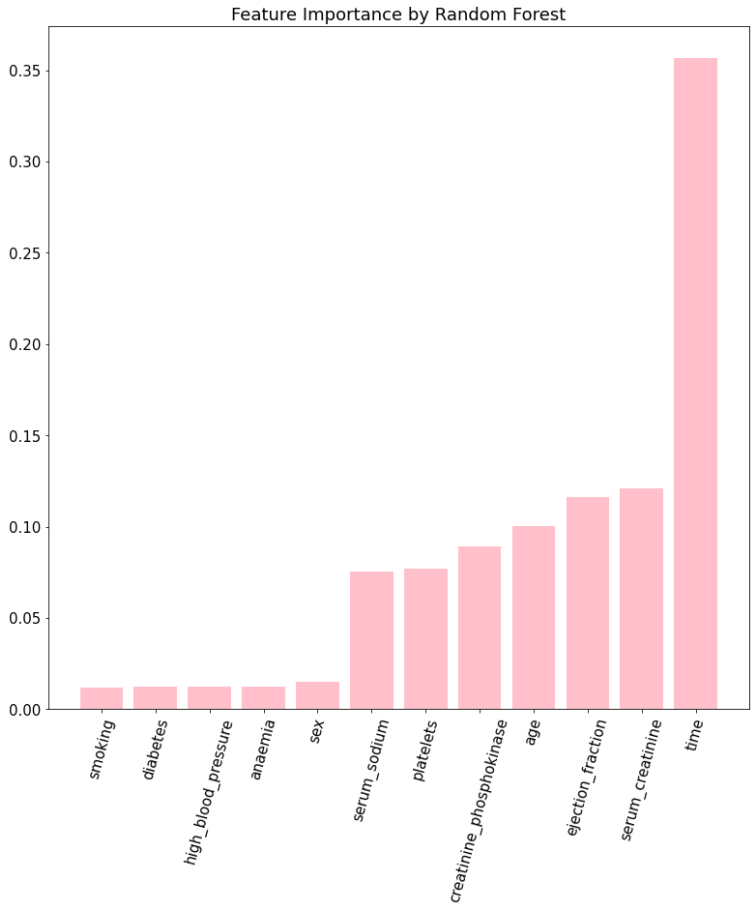


Figure 11 Feature Importance by Random Forest

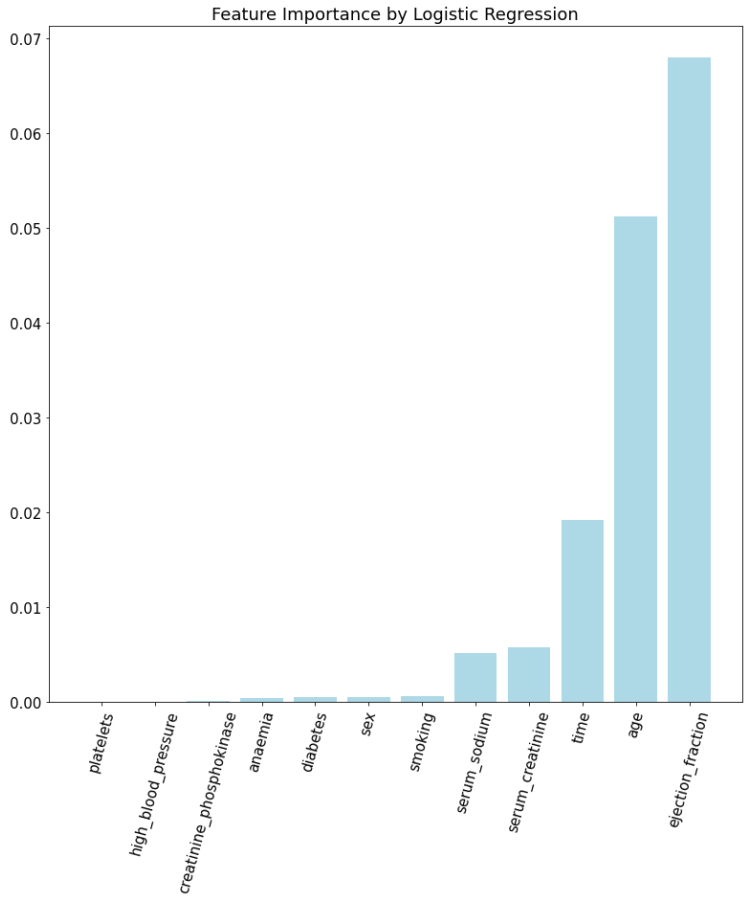


Figure 12 Feature Importance by Logistic Regression

Among all the models, only the Tree model and logistic regression model will give the feature importance rank. By those models, the two plots presented above give the importance of all the features. By random forest model, “time”, “serum_creatinine” and “ejection_fraction” are the top 3 significant features, while by the logistic regression model, “ejection_fraction”, “age” and “time” are the top 3.

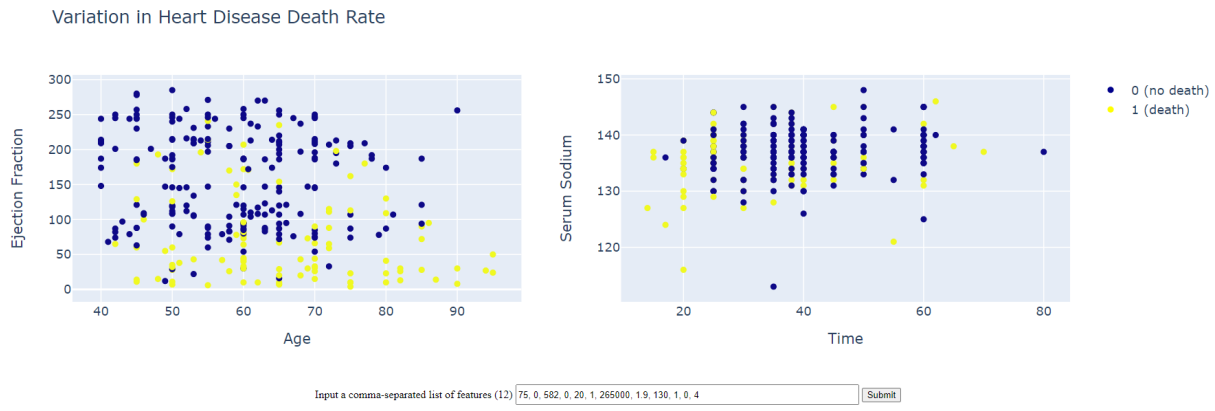
Among which, ejection fraction measures how much blood the left ventricle pumps out with each contraction, serum creatinine, which is the product of human muscle metabolism, reacts closely with non-enzymatic dehydration. Therefore, blood creatinine is closely related to the total amount of muscle in the body and is not easily affected by diet.

5. Heroku Application

We used Heroku's deployment feature to connect to a Github repository and used it to deploy the model. Users can go to https://anly605projectgroup2.herokuapp.com/ to build the model. They can utilize the following five feature vectors to test the model. A table showing the variables and example values for each variable is included below the input field. Users may also choose to create any feature vectors to fit and test the model.

Here are five feature vectors that users can use for testing:

1. 75, 0, 582, 0, 20, 1, 265000, 1.9, 130, 1, 0, 4
2. 65, 1, 52, 0, 25, 1, 276000, 1.3, 137, 0, 0, 16
3. 65, 0, 146, 0, 20, 0, 162000, 1.3, 129, 1, 1, 7
4. 53, 0, 63, 1, 60, 0, 368000, 0.8, 135, 1, 0, 22
5. 65, 1, 160, 1, 20, 0, 327000, 2.7, 116, 0, 0, 8



Variable Descriptions for Heart Disease Data set

Variable Name	Variable Description	Variable Type	Variable Range	Example Value
Age	*Number of age between 0 to 100*	Continuous	40-95	75
Anaemia	*A deficiency in the number or quality of red blood cells in your body (1 if Deficiency)*	Boolen	0 or 1	0
Creatinine Phosphokinase	*An enzyme in the body. It is found mainly in the heart, brain, and skeletal muscle.*	Continuous	23-7861	582
Diabetes	*A chronic (long-lasting) health condition that affects how your body turns food into energy. (1 if this condition is TRUE)*	Boolen	0 or 1	0
Ejection Fraction	*A measurement, expressed as a percentage, of how much blood the left ventricle pumps out with each contraction.*	Continuous	14-80	20
High Blood Pressure	*A common condition in which the long-term force of the blood against your artery walls is high enough that it may eventually cause health problems, such as heart disease. (1 if High Blood Pressure)*	Boolen	0 or 1	1
Platelets	*Are colorless blood cells that help blood clot.*	Continuous	25100-850000	265000
Serum Creatinine	*A blood test done to determine the amount of creatinine present in the blood.*	Continuous	0.5-9.4	1.9
Serum Sodium	*The amount of sodium relative to the volume of water in the blood*	Continuous	113-148	130
Sex	*Gender (1 if female)*	Boolen	0 or 1	1
Smoking	*1 if cigeratte*	Boolen	0 or 1	0
Time	*Time*	Continuous	4-285	4

Figure 13 Screenshot of the Heroku application

The above graph displays the scatter plot of four most important features of the dataset, that is, age, ejection fraction, time and serum sodium. The color of the points represents the category of the target variable, that is, Death_Event. Users can provide input to the text box below the graph. After inputting the feature vector, the website will display the updated graph with predictions from the feature vector given from the user's input.

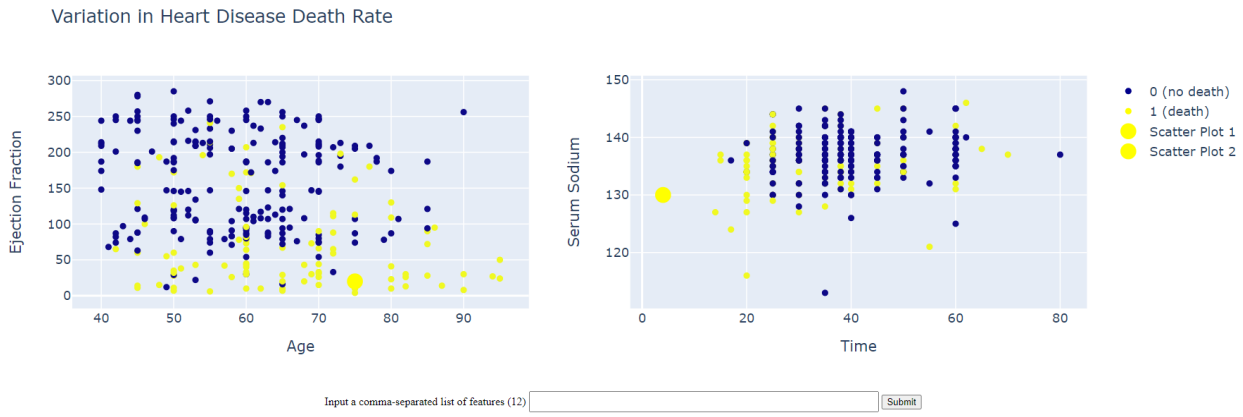


Figure 14 Predicted vector using 75, 0, 582, 0, 20, 1, 265000, 1.9, 130, 1, 0, 4

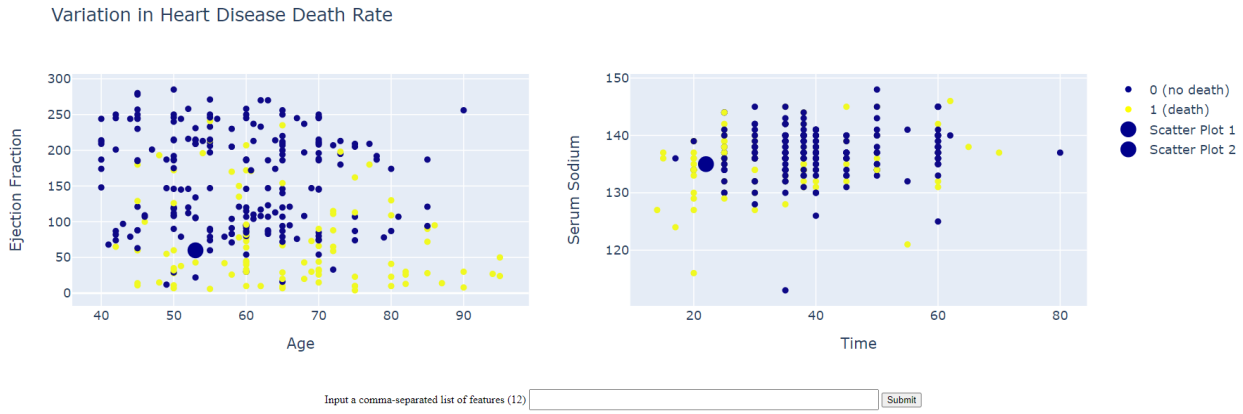


Figure 15 Predicted vector using 53, 0, 63, 1, 60, 0, 368000, 0.8, 135, 1, 0, 22

Figure 14 and Figure 15 show the updated graphs with the predicted Death_Event in yellow or darkblue on the plots.

6. Conclusion

In conclusion, many health habits that are closely related to heart failure, such as high blood pressure, and smoking, are not as significant as intuition based on the study. However, having a healthy lifestyle is still impactful to preventing heart failure. Here comes a critical point: how should people prevent Heart Failure from happening to the maximal level based on this study? The answer would be: in order to reduce the chance of heart failure, it would be important to regulate health conditions and control the self health habit. Especially, the study shows exercising is more useful in heart failure prevention than diet since the significant feature serum creatinine is closely related to the total amount of muscle in the body, which will be gained by working out, and is not easily affected by diet.

If moving a few steps away from the current features, there are also indirect features that are hard to catch and record. For example, in the studies of Heart Failure in 2020, the pandemic is considered one of the big factors in Heart Failure. And it is believed that more studies will focus on this and the related fields.

In the future, our team would like to optimize and finalize this model on more clinical data about Heart Failure.