NANYANG TECHNOLOGICAL UNIVERSITY

SCHOOL OF BIOLOGICAL SCIENCES

The hospital mortality prediction for ICU-admitted HF (Heart Failure) patients using MIMIC III dataset

(BS6200 Essential Machine Learning for Biomedical Science Project report)

Han Wenhao

Matriculation Number: G2201054J

26 / 10 / 2022

Content

- ## Introduction

Heart failure (HF) is a complex clinical syndrome that causes a patient's heart to not pump enough blood (ventricular insufficiency) to meet the oxygen needs of vital organs and tissues in the body. The result is the accumulation of fluid in the lungs and/or various parts of the body, resulting in congestion and edema. This also explains some of the common symptoms of the disease, such as dyspnea, fatigue, and poor exercise tolerance [1].

HF is a chronic disease that worsens over time, causing progressive remodeling of the heart. This changes the size and shape of the heart, which in turn impairs ventricular function in two ways:

Systolic HF (HF with reduced ejection fraction): The walls of the ventricles become thinner and weaker, causing the ventricles to dilate and reduce their ability to eject blood.

Diastolic HF (HF with constant ejection fraction): Impairment of diastolic function due to thickening and hardening of the ventricular wall due to tissue hypertrophy.

The image [2] below shows two types of Heart failure:

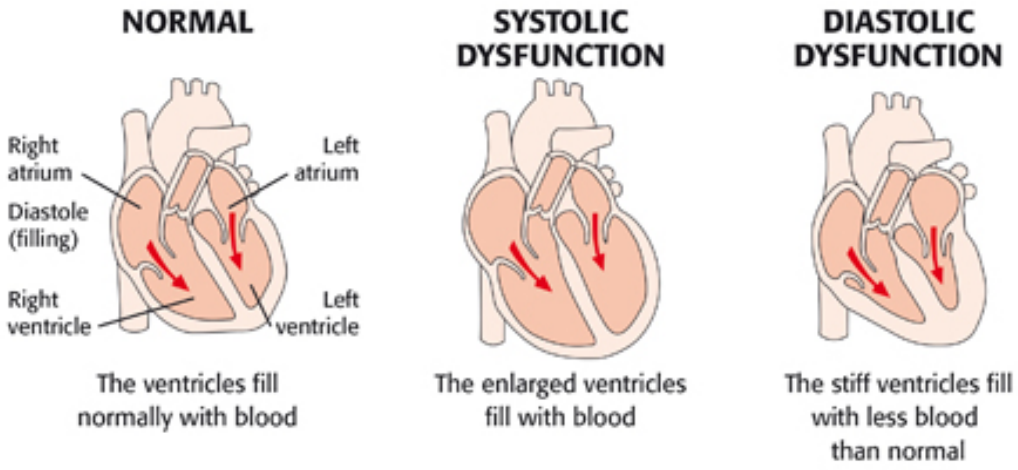


Figure 1. Types of heart failure

- ## Problem statement

MIMIC-III (Medical Information Mart for Intensive Care) comprises information relating to patients admitted to Intensive care units at a large tertiary care hospital. It covers 38,597 distinct adult patients and 49,785 hospital admissions between 2001 and 2012.

The predictors of in-hospital mortality for intensive care units

(ICU)-admitted HF patients remain poorly characterized. This project is aimed to develop and validate a prediction model for all-cause in-hospital mortality among ICU-admitted HF patients. Try to find the features that associate with the hospital mortality of HF.

- **Data preprocessing**

1. Feature selection

The dataset is extracted from the following tables in the MIMIC III dataset: ADMISSION, PATIENTS, ICUSTAYS, DICD DIAGNOSIS, DIAGNOSISICD, LABEVENTS, DLABIEVENTS, CHARTEVENT, DITEMS, NOTEEVENTS, and OUTPUTEVENTS.

The selected features (attributes) can be categorized into the following groups: Demographic characteristics, vital signs, comorbidities, and laboratory variables. I selected 50 features (attributes) in total, the graph below displays the features contained in each group:

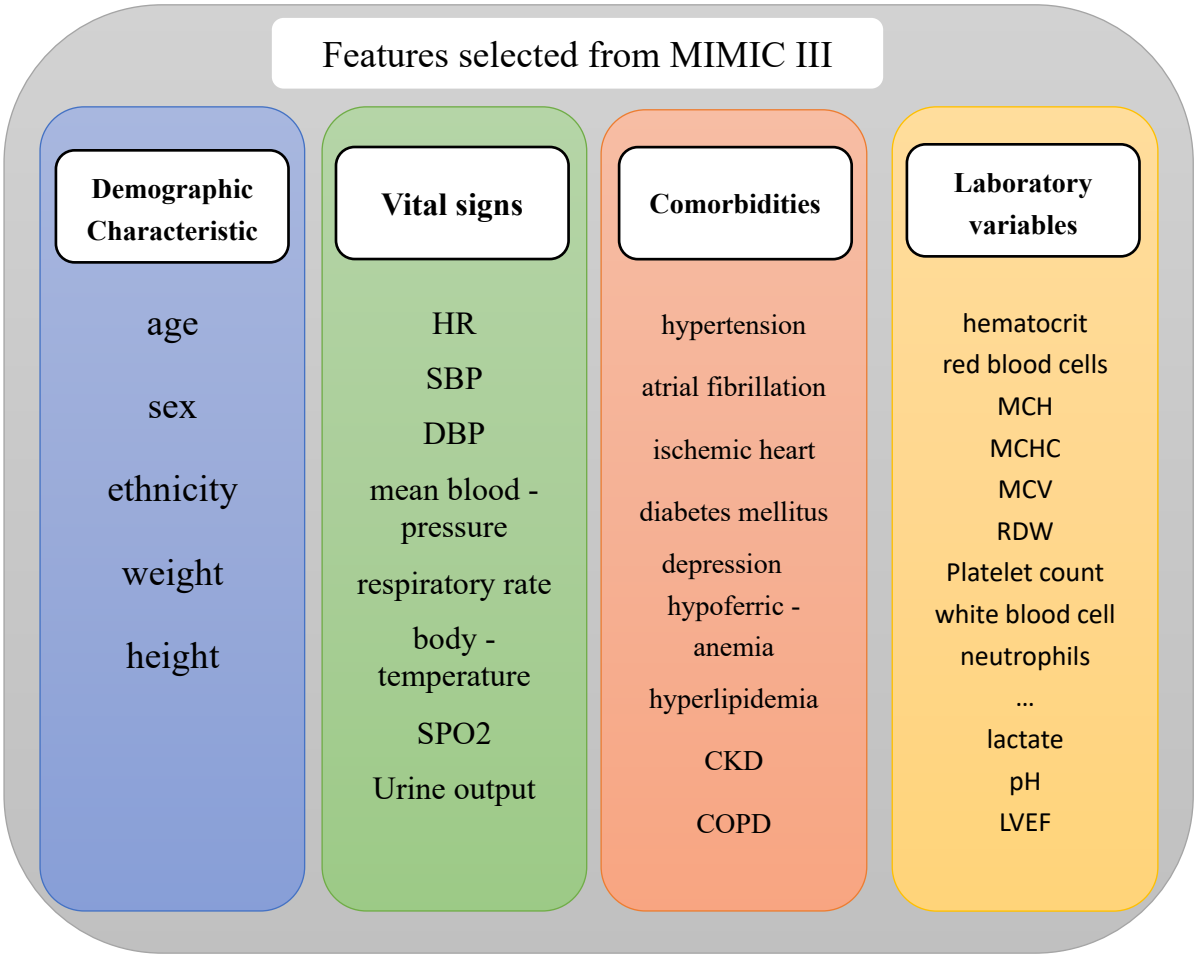


Figure 2. The features contained in four groups

2. Data exploration

In this section, different kinds of plots will be applied to explore the underlying information of the datasets. The analysis is performed based on the plots to explore the association between features and hospital mortality of HF patience admitted to the ICU.

1) age

According to the histogram below, hospital mortality increases with the age. Patients aged 50-55 and over 65 had higher mortality. Heart failure may significantly cause death in patients aged 80 to 90.

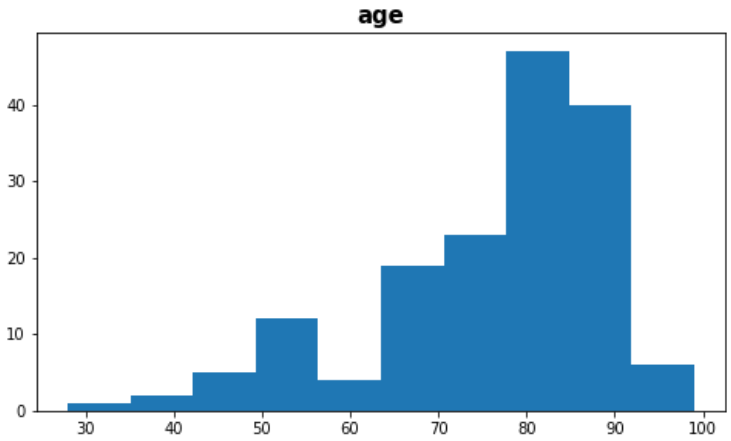


Figure 3. The histogram of age (death)

According to the violin plot below, unlike the distribution of surviving patients, the ages of dead patients were concentrated in the 80s and 90s. The 'outcome' is the final state of patients, '0' represents the patient is alive, and '1' represents the patient is dead.

Figure 4. violin plot (age VS. outcome)

2) gender

According to the pie chart below, the hospital mortality of female HF patients is 14.4%, while that of males is 12.8%. Female patients have a higher risk of death than male patients which is consistent with the results of clinical statistics.

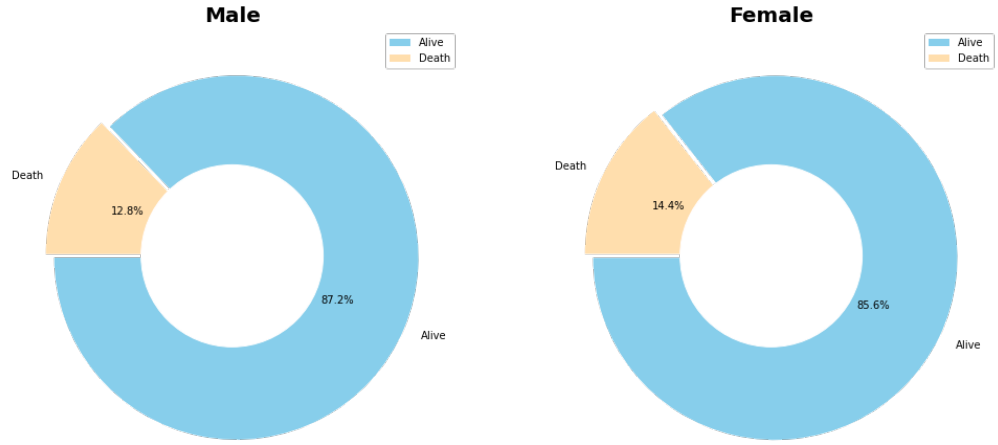
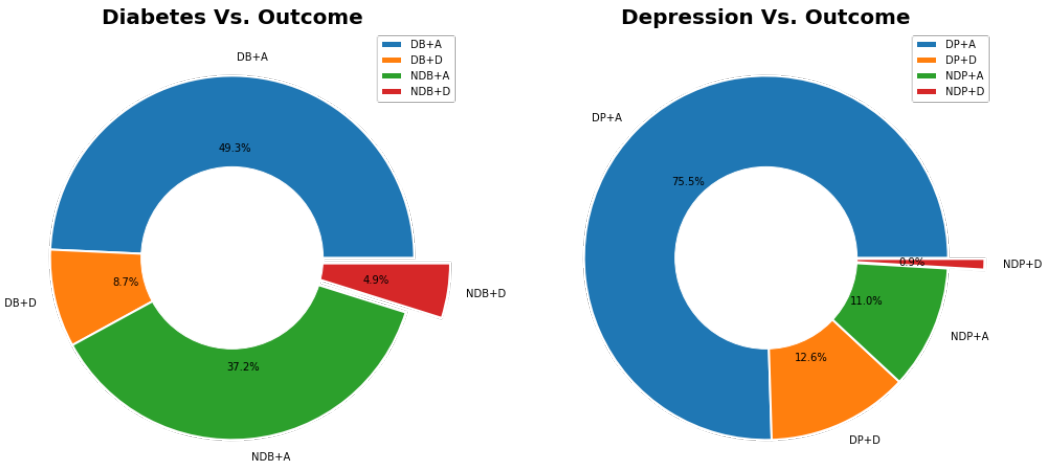


Figure 5. pie chart (gender VS. outcome)

3) complications

Four complications of heart failure (diabetes, depression, hypertensive and renal failure) are selected as features in this study. Four pie charts are plotted to find the correlation between those four complications and heart failure mortality.

Different colors represent patients with different states. Blue represents the patients with complication that are alive, orange represents patients with complication that are dead, green represents patients without complication that are alive, and red represents patients with complication that are dead.



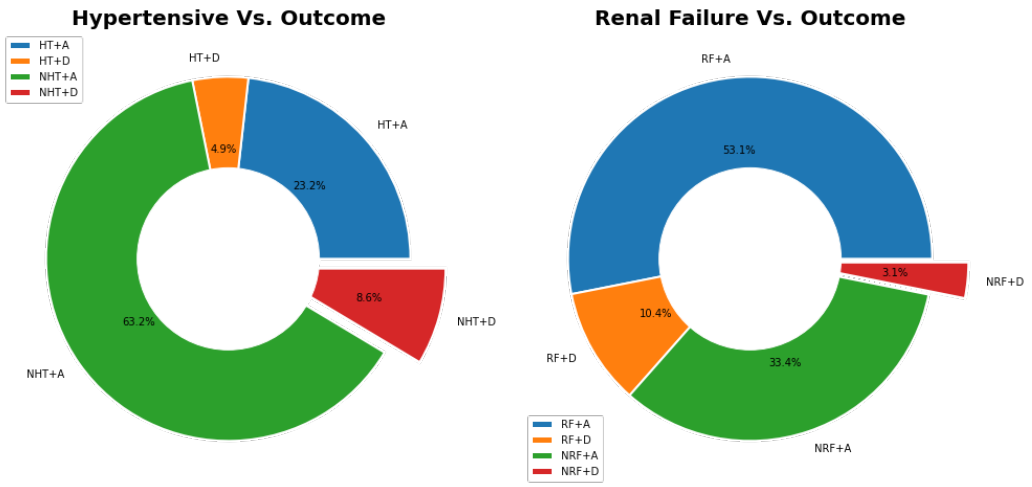


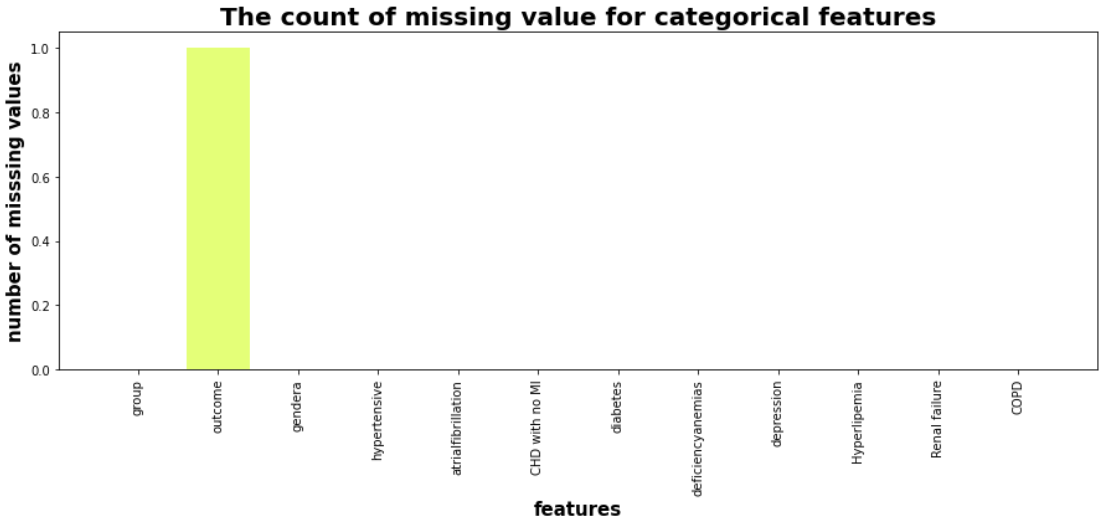
Figure 6. pie chart (four complications VS. outcome)

According to the pie charts above, most patients with heart failure also have diabetes, depression, and renal failure. The patients with these four diseases have higher mortality than those without, which indicates that those four complications may improve heart failure mortality. It should be notice that most patients with heart failure are very depressive. So psychological therapy should be paid more attention to in the treatment of heart failure.

3. Data cleaning

Not all features contain useful information, the feature 'group' and 'ID' are eliminated before model training.

The bar charts below show the missing values in continuous and categorical features. According to the bar charts, the missing values mainly exist in continuous features, and only one sample missing the class label 'outcome'.



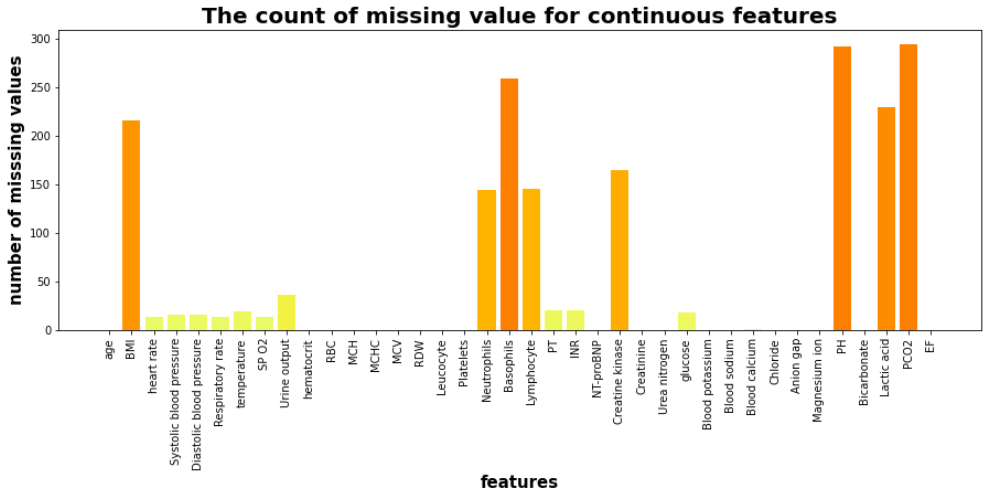


Figure 7. The missing values in continuous and categorical features

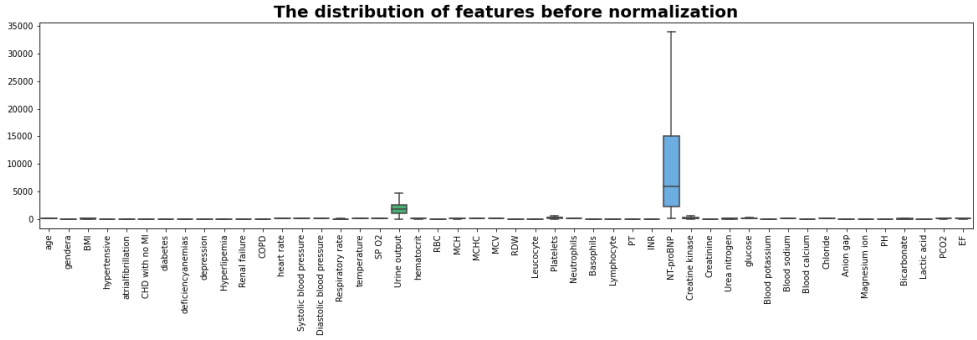
The missing values in continuous features are filled with the mean of the available data in each column. Then, eliminate the samples missing class labels.

Filling the missing values with means may minimize the information loss but could result in batch effect. Only one sample misses the class label, so eliminating it barely causes information loss.

4. Data normalization

Some models and algorithms applied in this study require normalization on the dataset. PCA is applied for dimension reduction and visualization. Logistic regression, KNN, and SVM are applied for prediction.

The z-score normalization is applied in this study to ensure the dataset is normally distributed. There are only three possible values for categorical features which are 1, 0, and -1. So the normalization is only applied on continuous features. The box plots below show the distribution of feature before and after normalization.



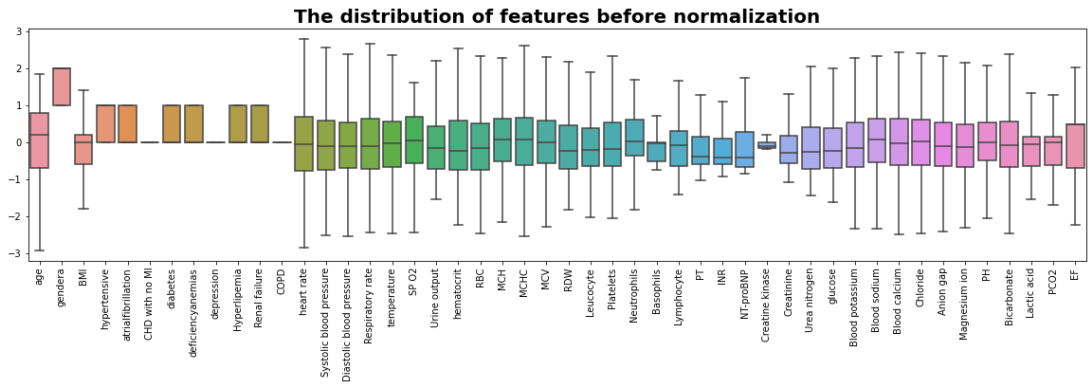


Figure 8. The distribution of features before and after normalization

- **Methodology**

1. Training and testing strategy

Before performing model training, the distribution of the class label is checked using a pie chart. According to the pie chart, the sample with the class label '1' (death) only occupies 13.5 % while the sample with the class label '0' (alive) occupies 86.5%. It is obvious that the class label of the dataset is imbalanced.

To handle the imbalance of class labels, a few strategies are applied in the training and test procedures. Firstly, stratified splitting is applied to make sure the class label in the training and testing set has the same distribution as the original dataset. In addition, the class weight is added based on the distribution of the class label. Apart from accuracy, the AUC is applied in model evaluation because it is not sensitive to imbalance.

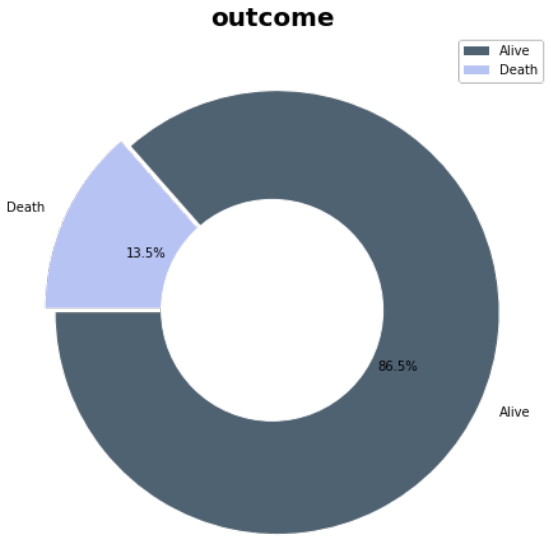


Figure 9. The imbalance of the class label

2. The workflow of experiment

The flow chart below displays the main procedure of the experiment. The k-means clustering and hierarchical clustering are applied to group the sample with high similarity in the sample clusters. By checking the features of samples in the same clusters, the internal association of different features. Then, the PCA and t-SNE are applied for dimension reduction and visualization of the result of clustering.

Cross-validation is applied in the evaluation of predictors. The stratified splitting is performed before normalization to avoid data leakage. Logistic regression is applied for binary class prediction in this study. Apart from prediction accuracy, the MAE (Mean Absolute Error) and MSE (Mean Squared error) are applied to check the performance of regression. The decision tree is a basic model applied in prediction. To improve the performance of the decision tree, the ensemble tree models (random forest and XGBoost) are also applied in prediction and feature selection. The feature selection is based on the feature importance in the tree-based model.

KNN, Naïve Bayes, and SVM classifiers are also applied in prediction. The voting is performed on the three optimized classifiers (with the best parameter) to achieve the best predictor.

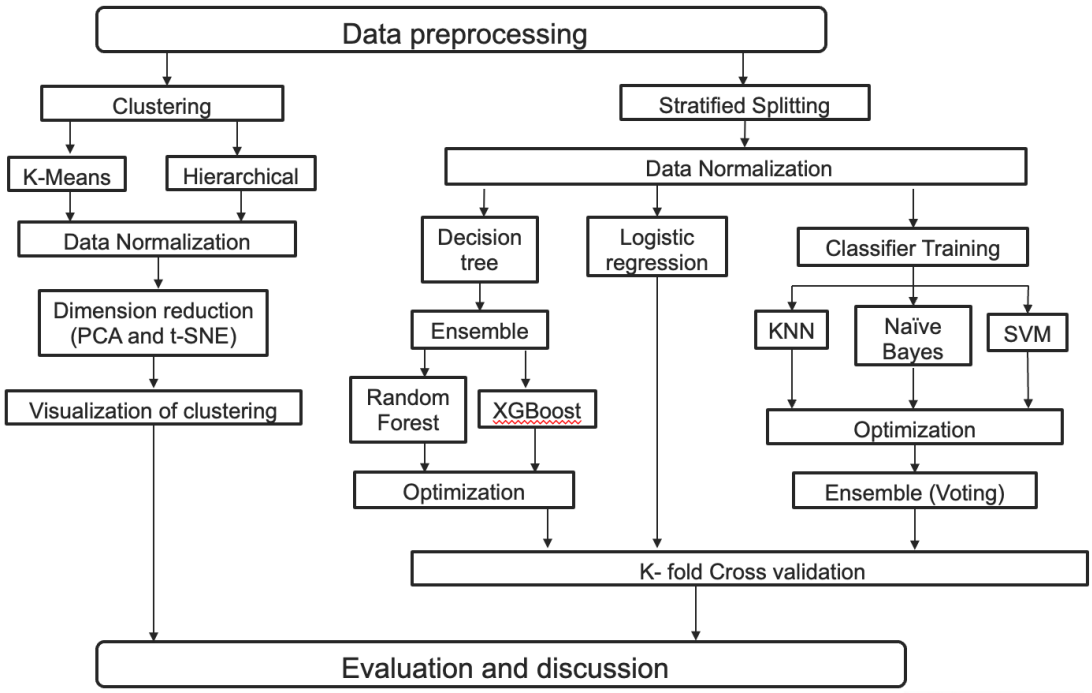


Figure 10. The workflow of the experiment

3. Experimental study and analysis

In this section, the model selection and optimization, and result analysis will be explained in detail.

1) clustering

The best K of K-means clustering is selected based on the inertia (SSE) curve. According to the curve, the inflection point occurs when K is equal to 3. So the K-Means clustering is performed with K equal to 3.

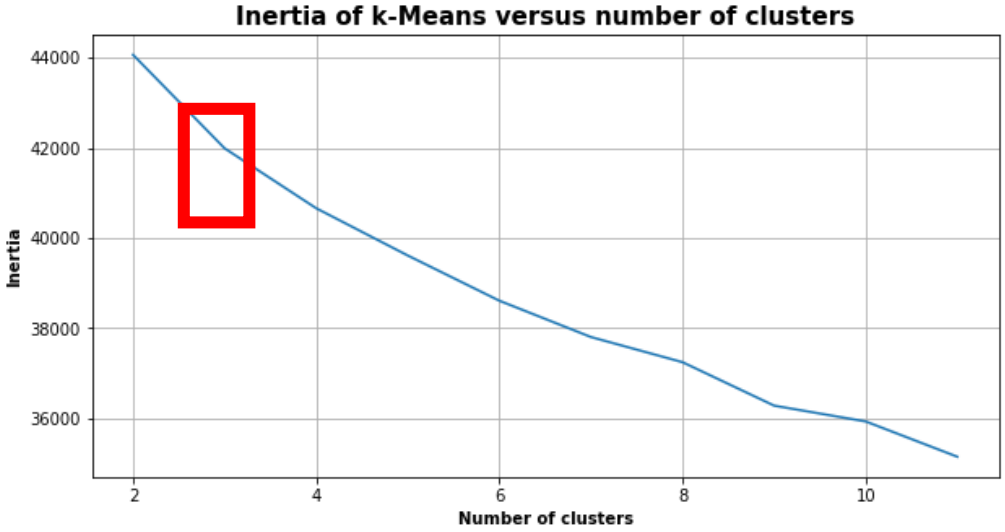
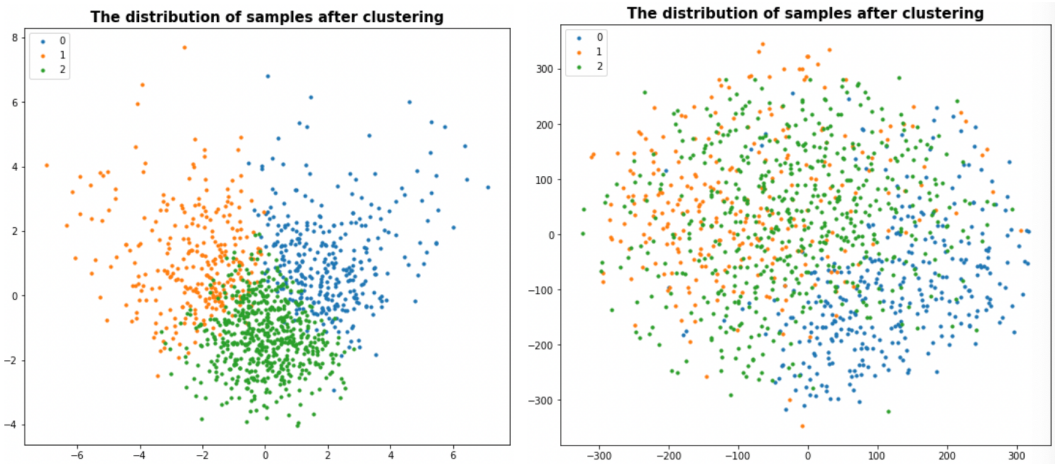


Figure 11. The inertia curve of K-Means clustering

The PCA and t-SNE are applied for the visualization of clustering. The scatter plot below shows the result of K-Means clustering, the samples are grouped into three clusters.



(a) PCA (b) t-SNE

Figure 12. The result of K-Means clustering

The result of hierarchical clustering is displayed using the dendrogram. According to the plot, the samples are categorized into five clusters.

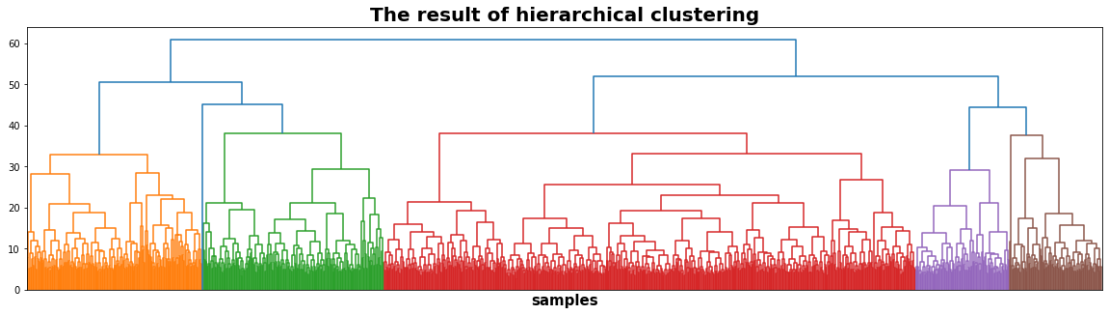


Figure 13. The result of hierarchical clustering

The different results of the two clustering algorithms are probably due to the randomness of the initial centroid location selection. In the next chapter, the features that play a decisive role in the two clustering methods will be compared and analyzed.

2) logistic regression

Logistic regression is usually applied in binary classification even if it is a regression model. The essence of logistic regression is to assume that the data follow logistic distribution, and then use maximum likelihood estimation to estimate parameters.

In this project, logistic regression is applied to fit the distribution of the data and predict patient survival. The table below displays the performance of the logistic regression.

According to the table, the MAE is the same as the MSE both of which are relatively low. This may indicate that the distribution of the data is well fit for logistic distribution. The classification accuracy on the test set is 85% while this score for the train set is over 90% which is higher than the test set. The results indicate the logistic regression model performs well in the prediction of patient survival and the overfitting didn't occur in model training.

Table 1. The evaluation scores of logistic regression

Metric	MAE	MSE	RMSE	Accuracy (train)	Accuracy (test)
Scores	0.1489	0.1489	0.3859	0.8511	0.9075

3) tree-based models

The decision tree classifier applied in this project is ID3 which builds the tree based on entropy and information gain. The max depth of the tree is set as 4 and the minimum sample leaf is set as 5. The parameters of the random forest and XGBoost model are consistent with the decision tree.

The number of estimators is the hyperparameter of random forest and XGBoost. To line graph below shows the classification accuracy of two models with different numbers of estimators. According to the plot, the inflection point occurs when the number of estimators is 80. So the optimal number of estimators is 80.

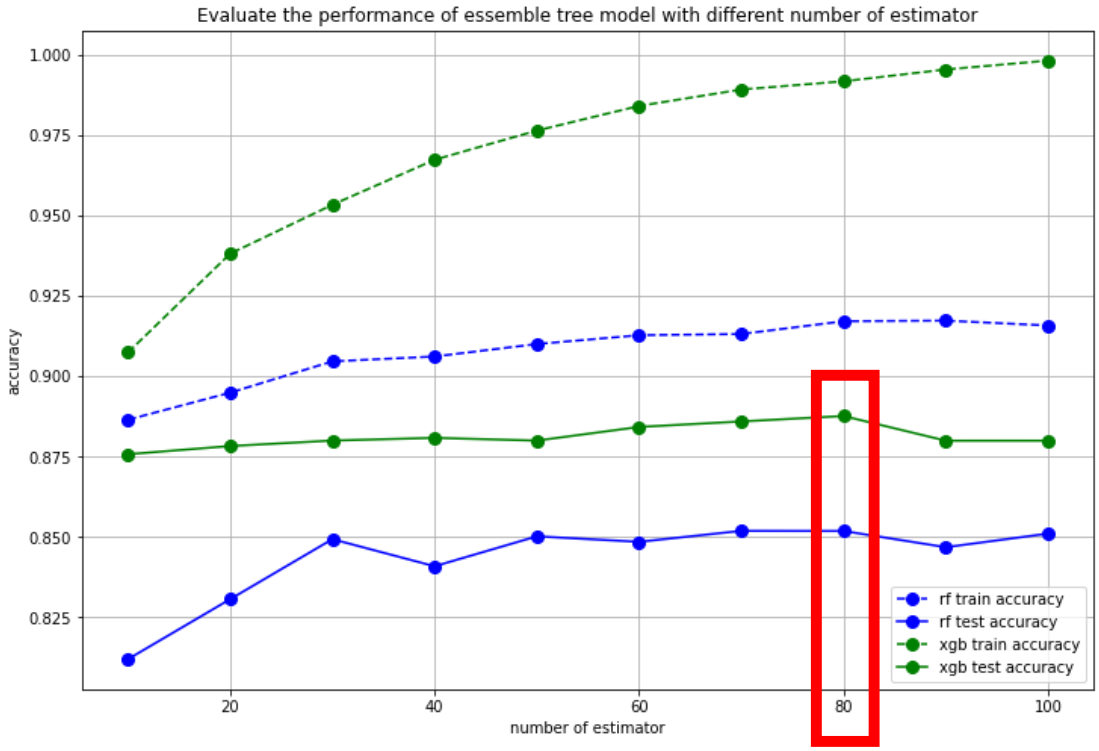


Figure 14. Evaluation of the number of estimators

The table below shows the evaluation scores of three tree-based models. According to the table, the XGBoost has the highest classification accuracy and precision, while the recall of this model is the lowest among the three models, which indicates that the highest accuracy result from the accurate prediction of surviving patients. But it can barely predict dead patients. The random forest has the highest F1-score and AUC and the second-highest accuracy and recall, which indicates that it is the most robust one among the three tree-based models.

Table 2. The evaluation scores of tree-based models

	Accuracy	Precision	Recall	F1	AUC
DT	0.70	0.24	**0.53**	0.33	0.67
RF	0.84	0.43	0.47	**0.44**	**0.81**
XGBoost	**0.88**	**0.67**	0.26	0.38	0.80

4) classification models

a) KNN classifier

The K is the hyperparameter of the KNN classifier. The line graph below shows the classification of the KNN classifier with different K. Also, a function is defined to select the optimal value of K based on the accuracy of cross-validation. According to the plot, the classification accuracy of the training set and testing set gradually approaches with the increase of K. The classification accuracy on the test set peaks at K of 14.

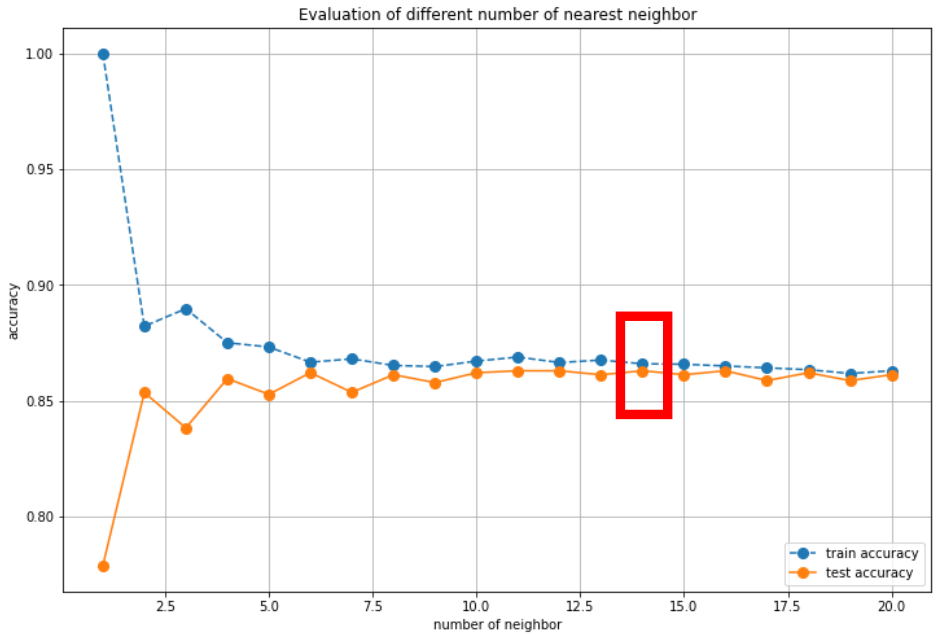


Figure 15. The evaluation of the number of estimators

b) Naïve Bayes classifier

Five Naïve Bayes models are evaluated in the project, including Gaussian, Bernoulli, categorical, compliment, and Multinomial Naïve Bayes (NB) classifiers. Theoretically, models other than the Gaussian NB classifier are more suitable for categorical or discrete data classification.

According to the table below, the precisions and recalls of Bernoulli and categorical NB classifiers are 0 which indicates that those two models cannot predict the death of the patients. The Gaussian NB classifier has the highest precision, F1-score, and AUC, while the complement and multinomial NB classifiers have the highest recall (50%). Overall, the performance of the Gaussian NB classifier is more robust than the other two NB classifiers even though the other two have better performance in death prediction.

Table 3. The evaluation scores of five Naïve Bayes classifiers

	Accuracy	Precision	Recall	F1-score	AUC
Gaussian	0.85	**0.45**	0.35	**0.38**	**0.78**
Bernoulli	**0.86**	0	0	0	0.64
Categorical	**0.86**	0	0	0	0.64
Complement	0.68	0.22	**0.50**	0.30	0.64
Multinomial	0.68	0.22	**0.50**	0.30	0.64

c) SVM classifier

The SVM classifier with the 'rbf' kernel is applied in this study. C and gamma are two hyperparameters of this classifier. The C is the penalty coefficient, which balances the classification interval margin and the misclassification of samples. The gamma is the kernel parameter that controls the 'spread' of the kernel (how broad the decision region is).

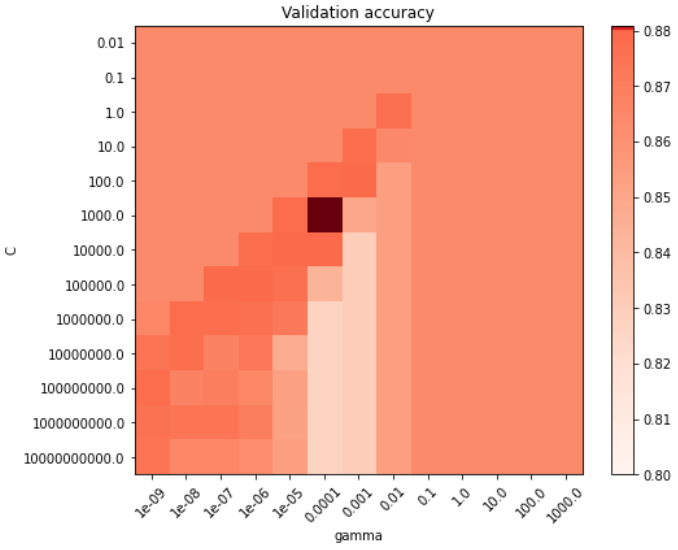


Figure 16. Selection of the optimal C and gamma for SVC

The heatmap above compares the classification accuracy of SVC with different C and gamma. The higher saturation indicates better performance in classification. According to the heatmap, the SVC has the best performance when C is equal to 1000 and gamma is equal to 0.0001.

5) ensemble learning

Soft voting is performed on the optimal KNN, Naïve Bayes, and SVM classifiers. The plot below shows how soft voting works when the three models are assigned the same weight in voting.

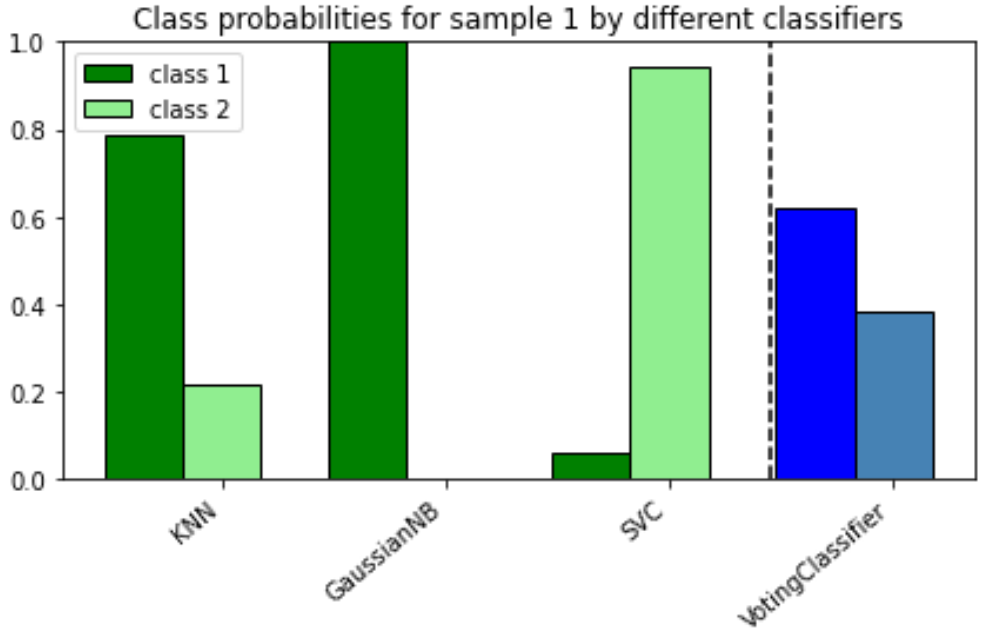


Figure 17. The procedure of soft voting in one prediction

The table below compares the performance of different classifiers and soft voting (the weight assigned to three classifiers are 1, 2, 3).

Table 4. The evaluation scores of four classifiers

	Accuracy	**Precision**	**Recall**	**F1-score**	**AUC**
KNN	0.87	0.30	0.03	0.06	0.77
NB	0.85	0.47	**0.39**	**0.41**	0.78
SVC	**0.88**	**0.69**	0.20	0.31	0.80
Voting	**0.88**	0.67	0.26	0.37	**0.81**

- **Achievement and discussion**

1. Summary of the results

The table below compares all classification methods applied in this study (Confusion matrix and ROC curves see appendix). According to the table, the XGboost, SVC, and Voting have the highest classification accuracy. The SVC has the highest precision and the lowest recall among the three models, which indicates that it has a high false positive rate. Both XGBoost and voting are ensemble learning models, their AUCs are relatively high which indicates the ensemble improves the robustness of the prediction.

The decision tree model has the highest recall, even higher than the random forest and XGBoost, which indicates that the ensemble of the tree model may reduce the prediction accuracy of death.

The recalls of all models are relatively low in all models, this situation is probably caused by the imbalanced class label. Besides, the selected features may not contain enough information about the dead patient which could be another cause of the low recalls.

In conclusion, the project achieves an optimal predictor (voting classifier) with 88% accuracy.

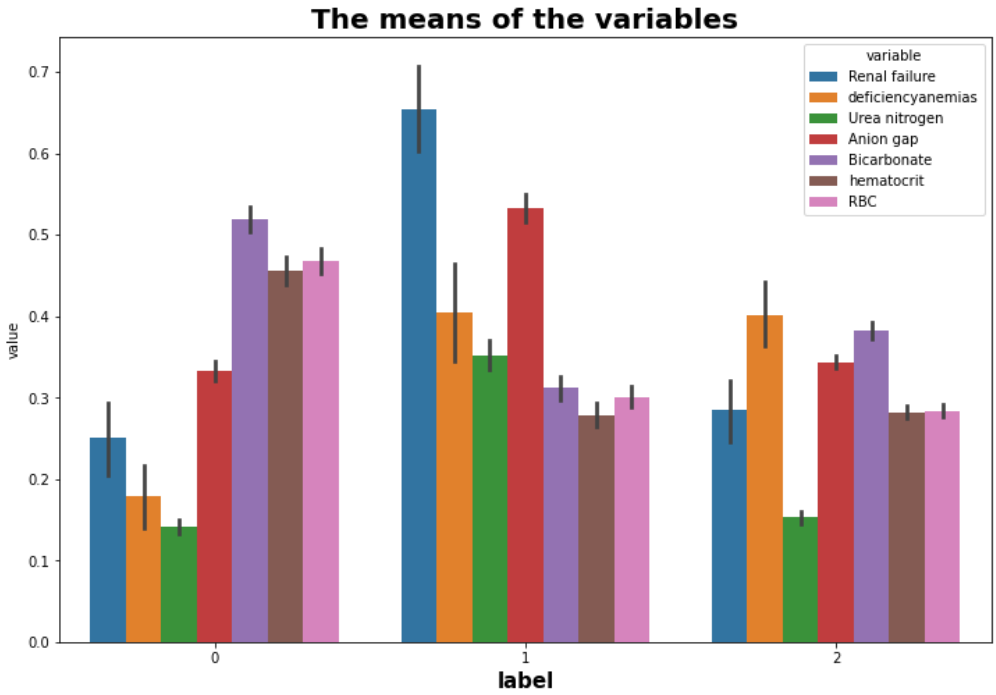
Table 5. The evaluation scores of all models applied

	Accuracy	**Precision**	**Recall**	**F1-score**	**AUC**
LR	0.87	0.63	0.12	0.19	0.74
DT	0.70	0.24	**0.53**	0.33	0.67
RF	0.84	0.43	0.47	**0.44**	**0.81**
XGBoost	**0.88**	0.67	0.26	0.38	0.80
KNN	0.87	0.30	0.03	0.06	0.77
NB	0.85	0.47	0.39	0.41	0.78
SVC	**0.88**	**0.69**	0.20	0.31	0.80
Voting	**0.88**	0.67	0.26	0.37	**0.81**

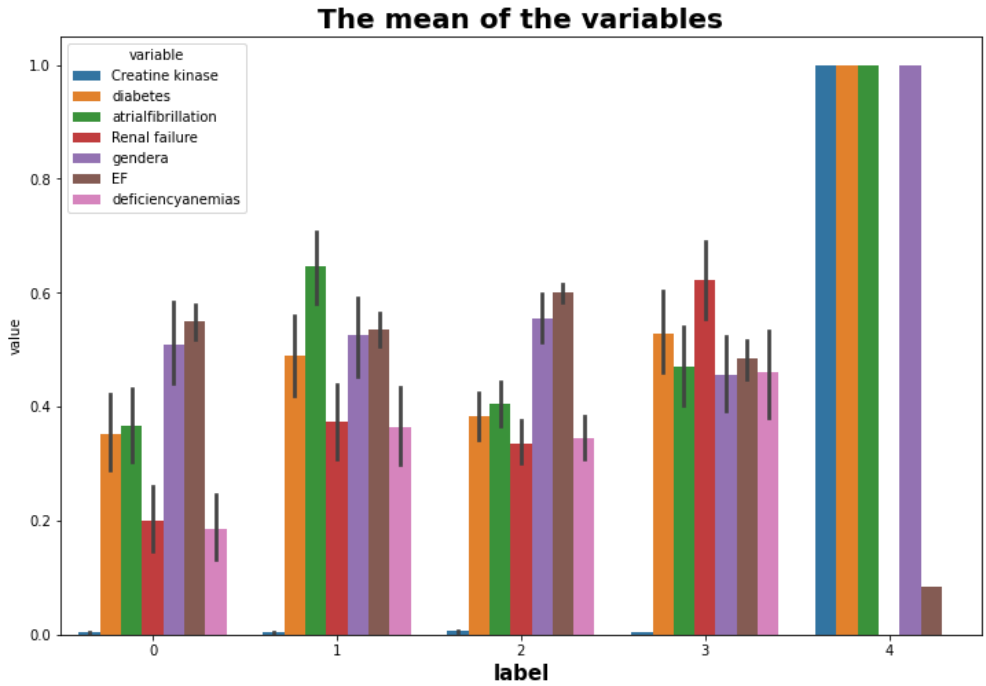
2. Discussion of the important features

1) clustering

The bar charts below show the decisive features for K-Means and hierarchical clustering.



(a) K-Means clustering



(b) Hierarchical clustering

Figure 18. The decisive features of clustering algorithms

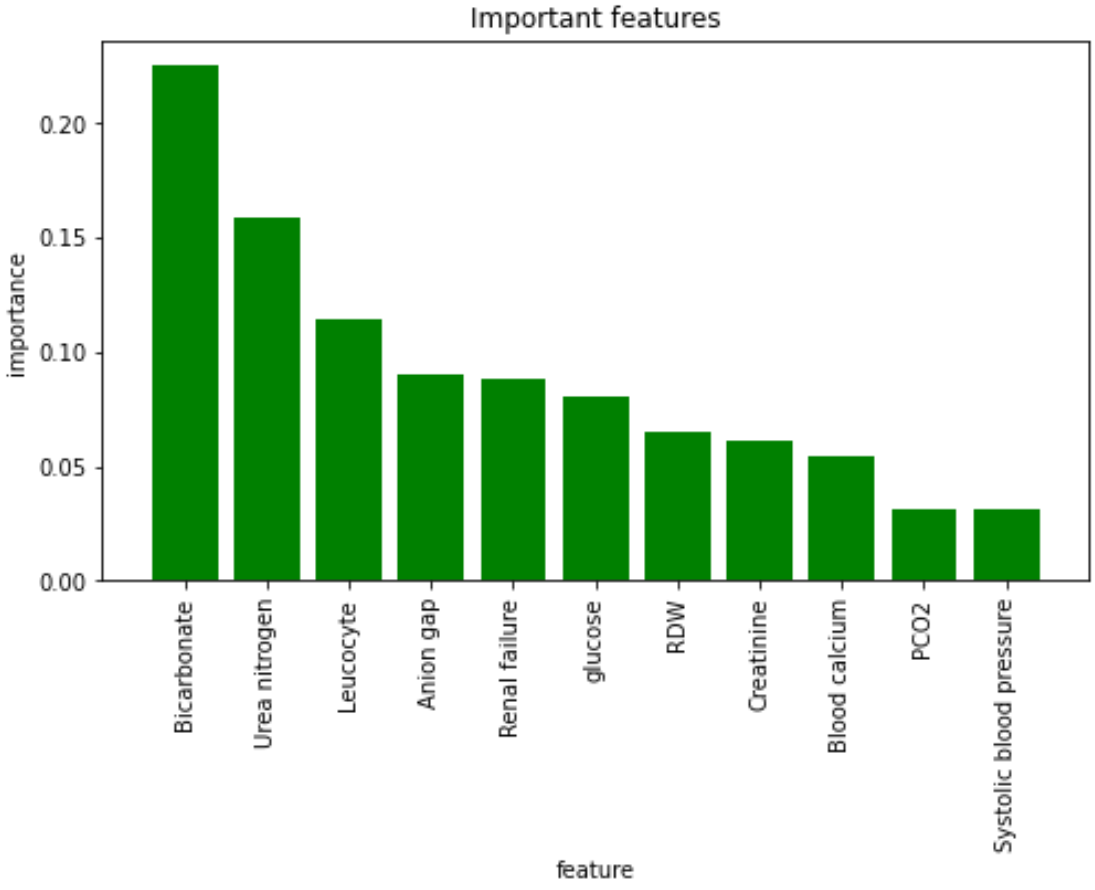
According to the Figure 18. The renal failure and deficiency anemias are important in both K-Means and hierarchical clustering. Renal failure is a complication of heart failure, it can improve hospital mortality.

Deficiency anemia – a condition in which blood lacks adequate healthy red blood cells. The blood cells carry oxygen to the body's tissues. So people with deficiency anemia usually feel fatigued. Patients with failure also get this symptom.

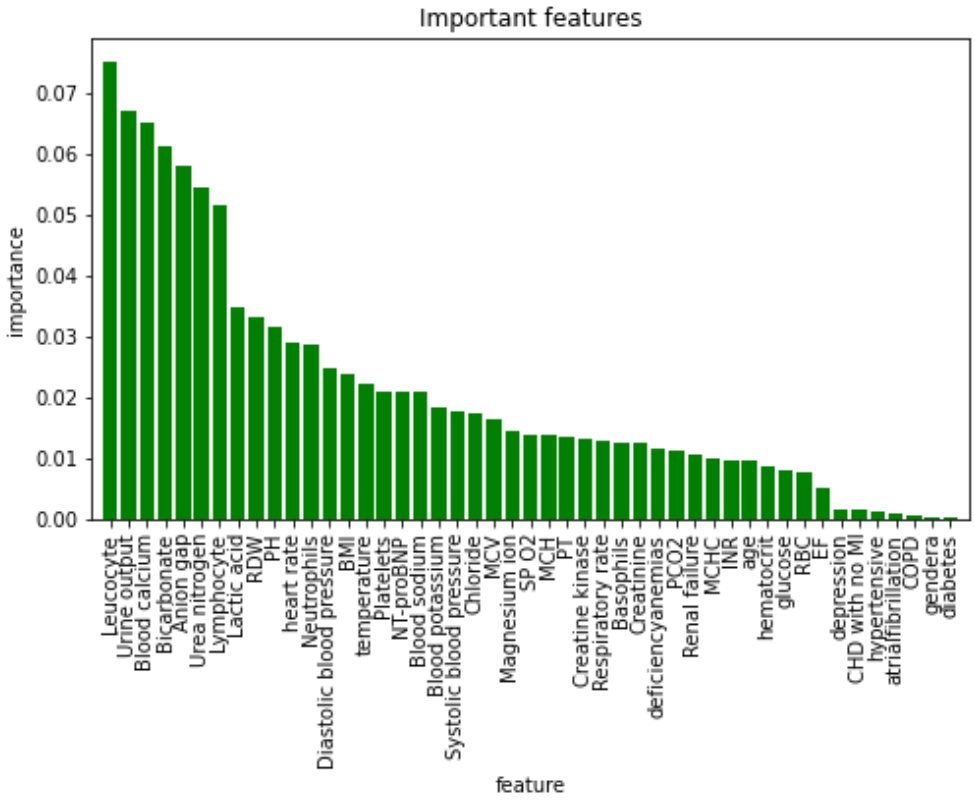
According to the former research [3], Iron Deficiency anemia (ID) is widely present in patients with heart failure with an estimated prevalence of over 50% in ambulatory patients. ID in those with heart failure appears to worsen symptoms, reduce the quality of life, and increases mortality risk.

2) tree-based models

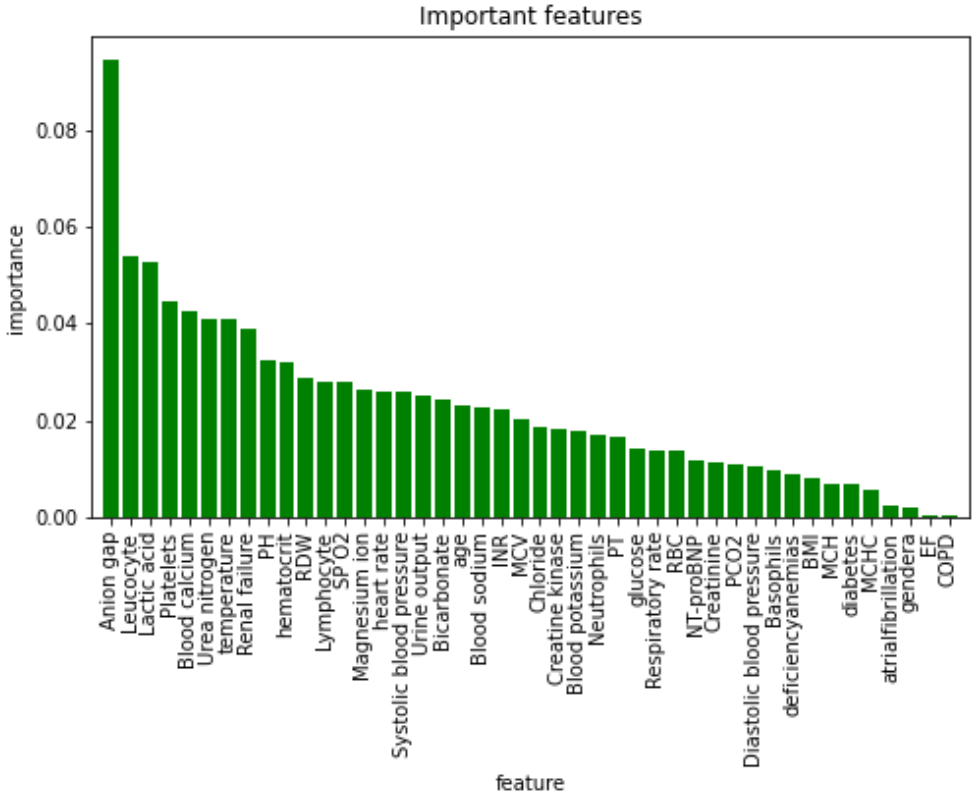
The bar charts below show the important features selected by tree-based models.



(a) Important features selected by the decision tree



(b) Important features selected by random forest



(c) Important features selected by XGBoost

Figure 19. Important features selected by tree-based models

The intersection of features selected by the tree-based models is 'Urea nitrogen' and 'Blood calcium'.

The blood urea nitrogen (BUN) test reveals important information about how well the patient's kidneys are working. A high BUN level means the kidneys aren't working well and is associated with mortality in patients with renal dysfunction and HF [4]. This feature also reveals the high association between renal failure and hospital mortality of heart failure.

Calcium plays a key role in cardiac muscle contraction and metabolism. Calcium particles enter the heart muscle cells during each heartbeat and contribute to the electrical signal that coordinates the heart's function. Calcium particles also bind to machinery within the cell that helps the cell to squeeze together ("contract"), which makes the heart pump blood [5]. It is recommended that physicians check blood calcium levels in the elderly, as hypocalcemia is a reversible cause of heart failure.

3. Future direction for further improvement

According to the evaluation scores of all models, the accuracy is relatively high, but the recalls of most models are lower than 50%. This problem may be caused by the insufficient dataset, rather than the limitation of the models.

In this project, the missing values in continuous variables are filled with the mean of the rest available data in this column. This operation could cause the batch effect in the experiment. Instead, filling in the value with the KNN-predicted values might retain most information of the dataset.

The imbalance of the class label could be another reason for the death prediction. So, one solution is generating the new samples with the class label '1' (death) using SMOTE or GAN to balance the dataset. The other method is performing oversampling and undersampling on the dataset before model training.

Another direction to go further is to apply different ensemble learning models on the optimal classifiers like bagging, boosting, and Adaboost to improve the performance of prediction.

- # Reference

[1] Authors/Task Force Members, et al. "ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC." *European heart journal* 33.14 (2012): 1787-1847.

[2] Heart Failure (HF) (Congestive Heart Failure) By Nowell M. Fine , MD, SM, Libin Cardiovascular Institute, Cumming School of Medicine, University of Calgary, MERCK MANUAL Consumer Version, 2022. Link: https://www.merckmanuals.com/home/heart-and-blood-vessel-disorders/heart-failure/heart-failure-hf

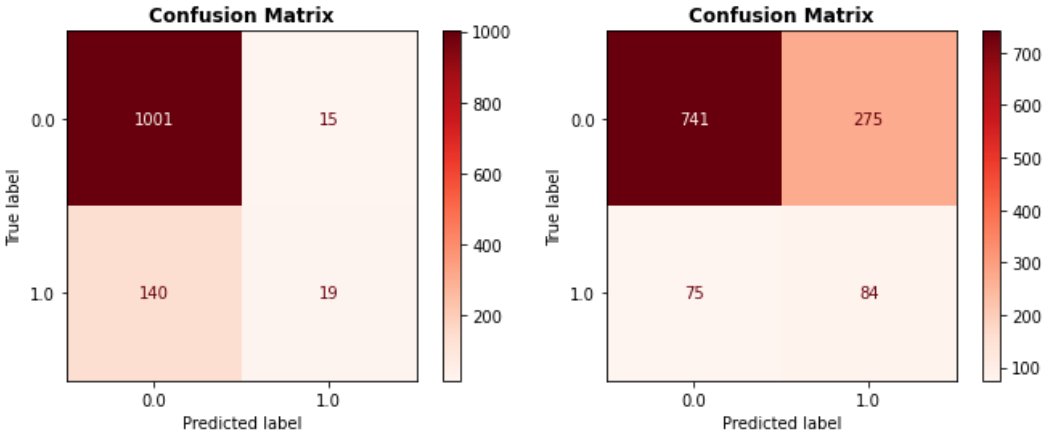
[3] von Haehling, Stephan, et al. "Iron deficiency in heart failure: an overview." JACC: Heart Failure 7.1 (2019): 36-46.

[4] Testani JM, Coca SG, Shannon RP, Kimmel SE, Cappola TP. Influence of renal dysfunction phenotype on mortality in the setting of cardiac dysfunction: analysis of three randomized controlled trials. Eur J Heart Fail 2011; 13: 1224–1230.

[5] Sutanto, Henry, and Jordi Heijman. "The role of calcium in the human heart: with great power comes great responsibility." Front Young Minds 7.65 (2019): 10-3389.

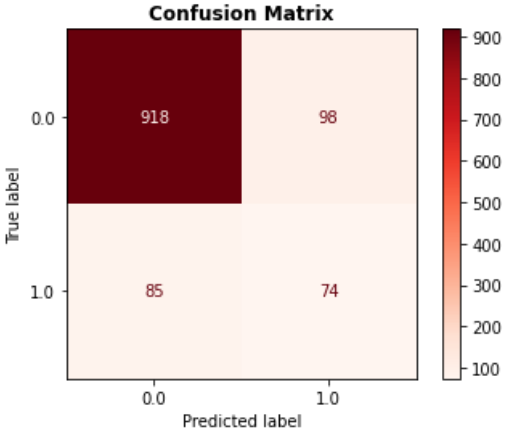
- # Appendix

1. The confusion matrix of all models

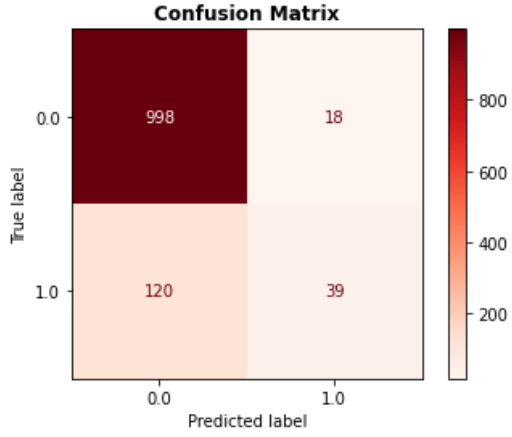


(a) Logistic regression

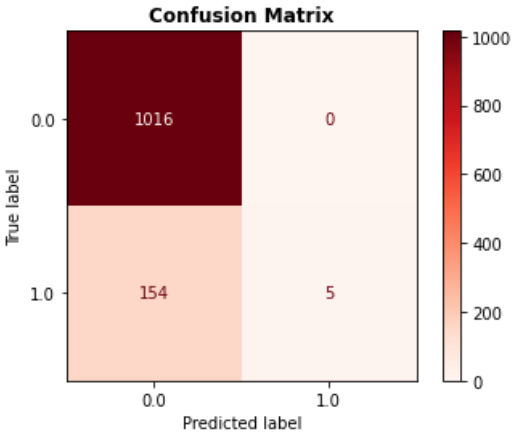
(b) Decision tree



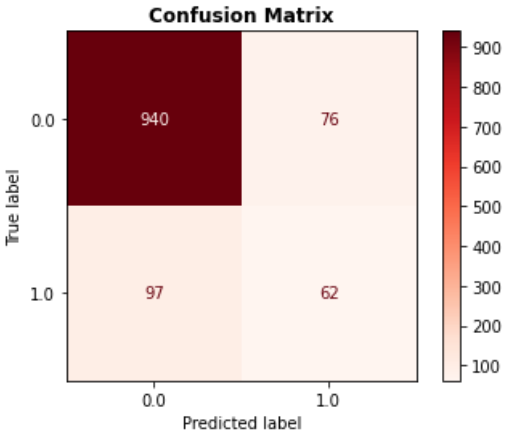
(c) Random Forest



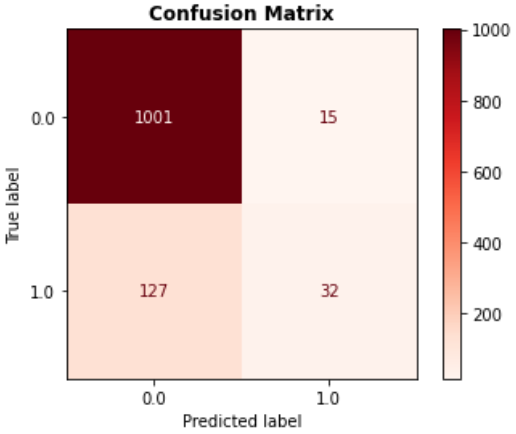
(d) XGBoost



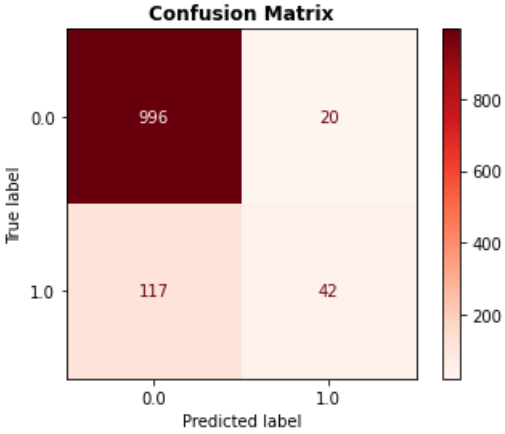
(e) KNN



(f) Naïve Bayes

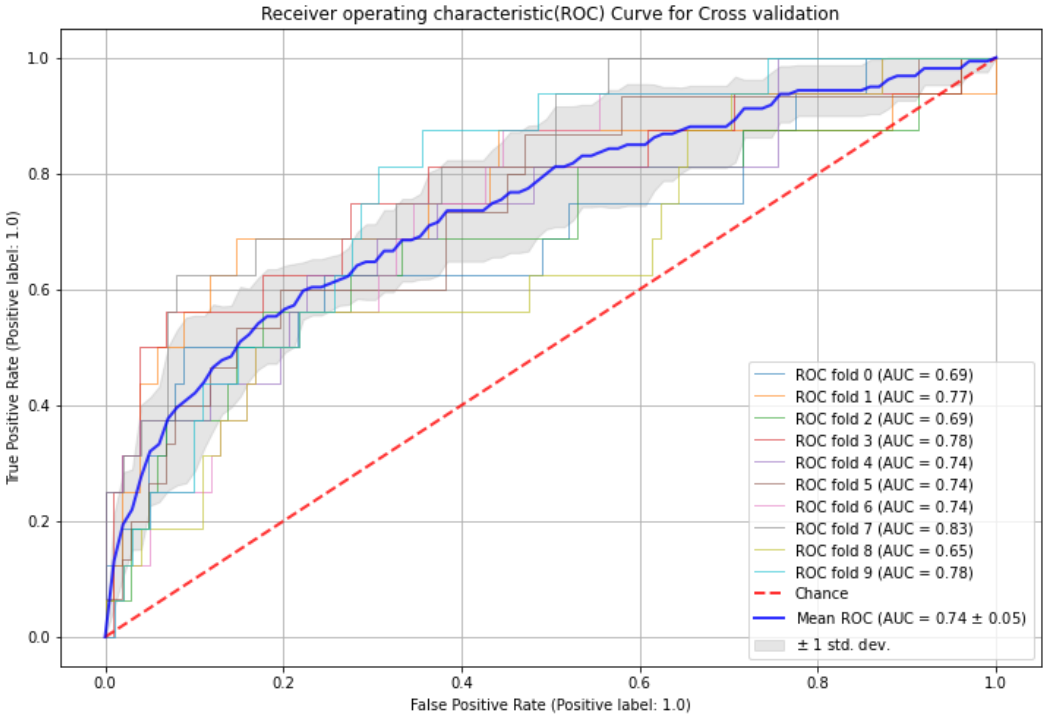


(g) SVC

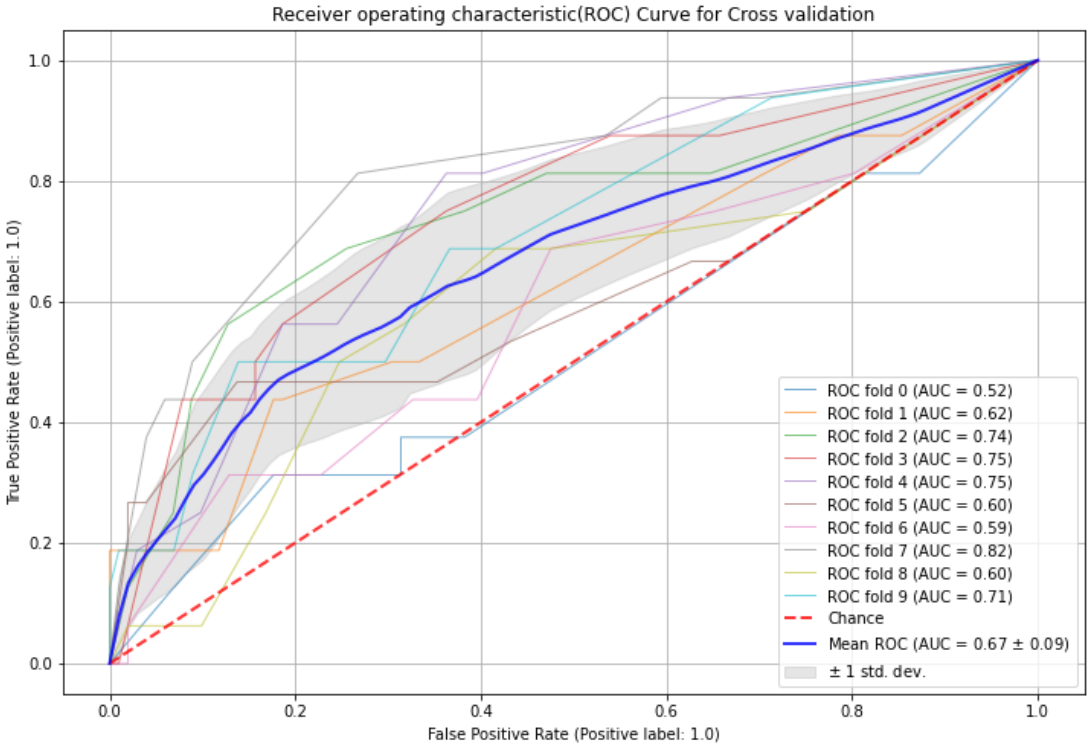


(h) Voting

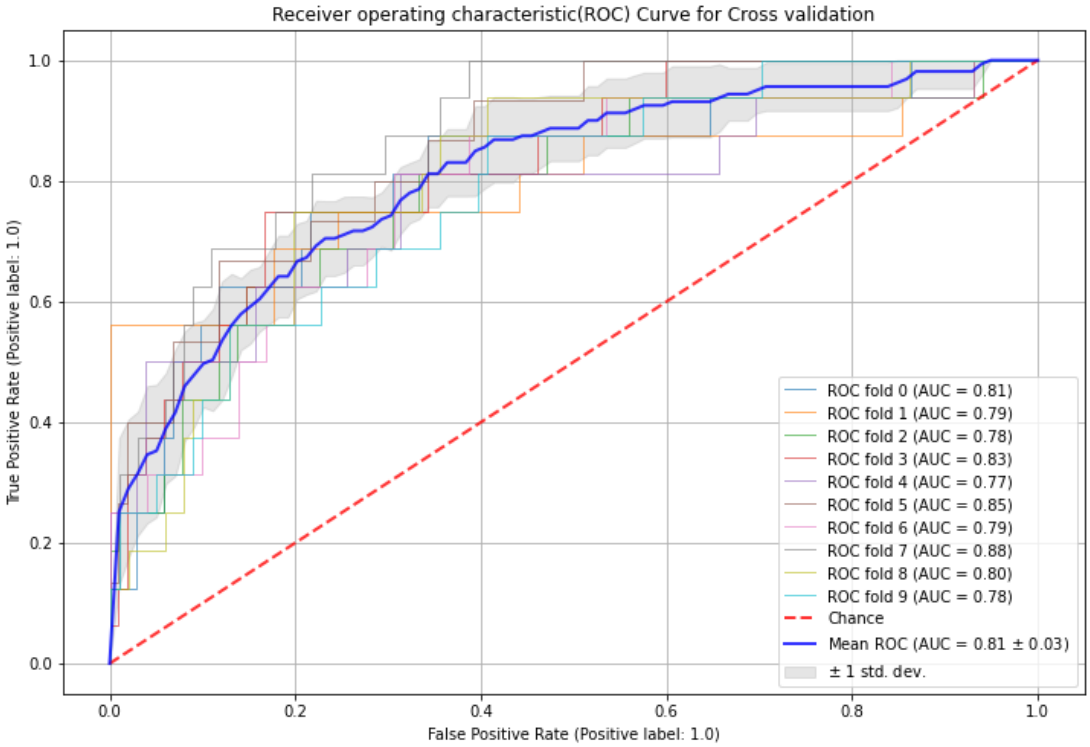
2. ROC curve of all models



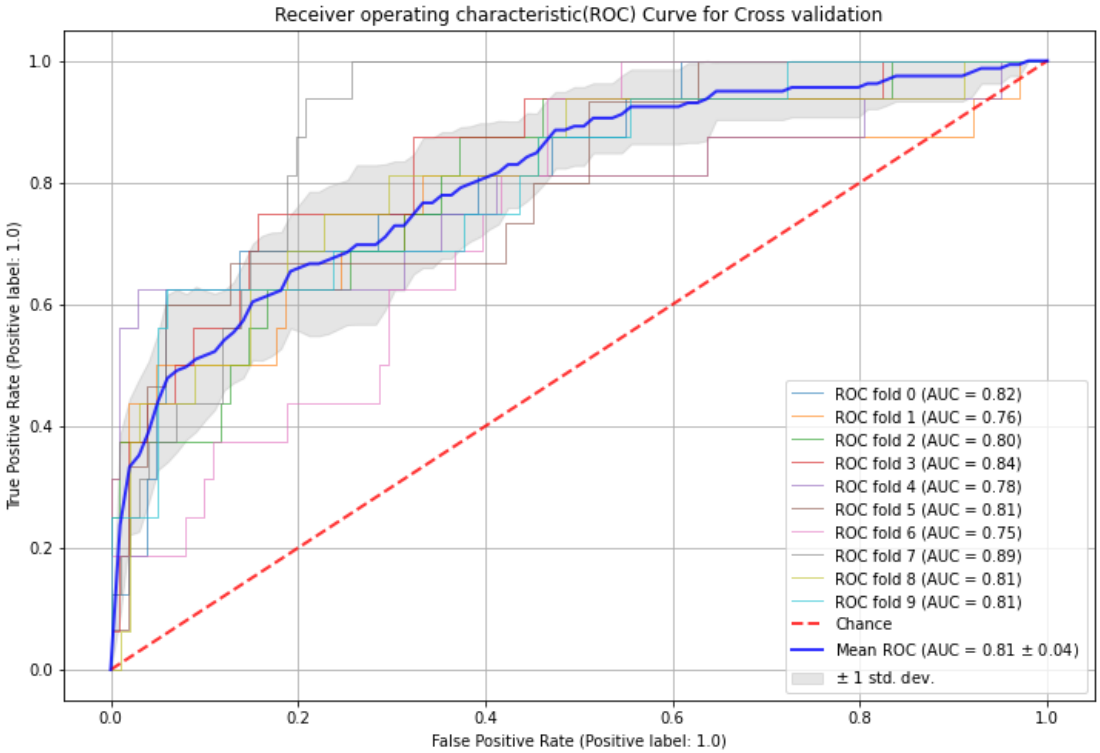
(a) Logistic regression



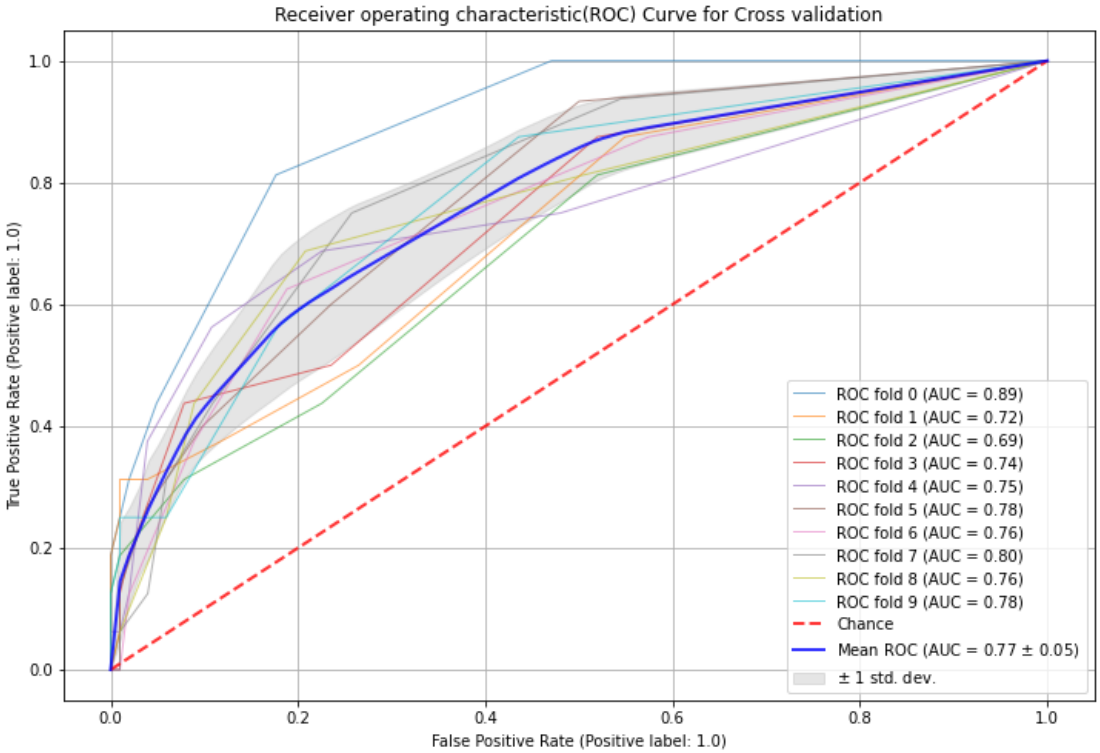
(b) Decision tree



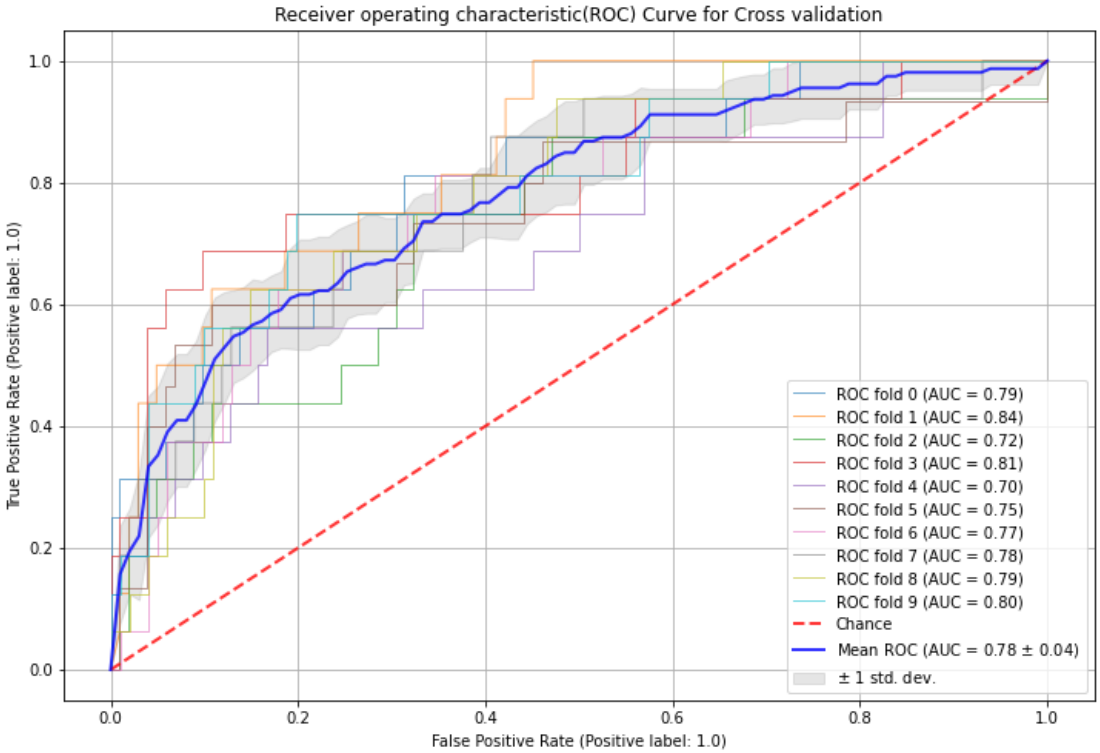
(c) Random forest



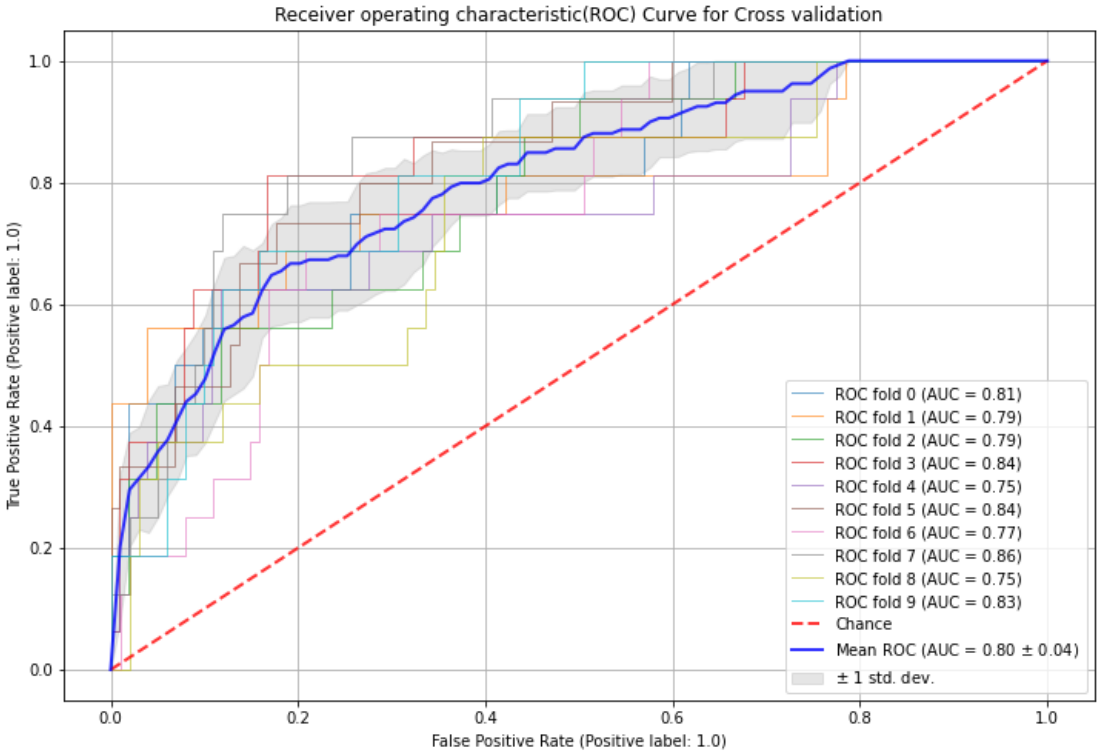
(d) XGBoost



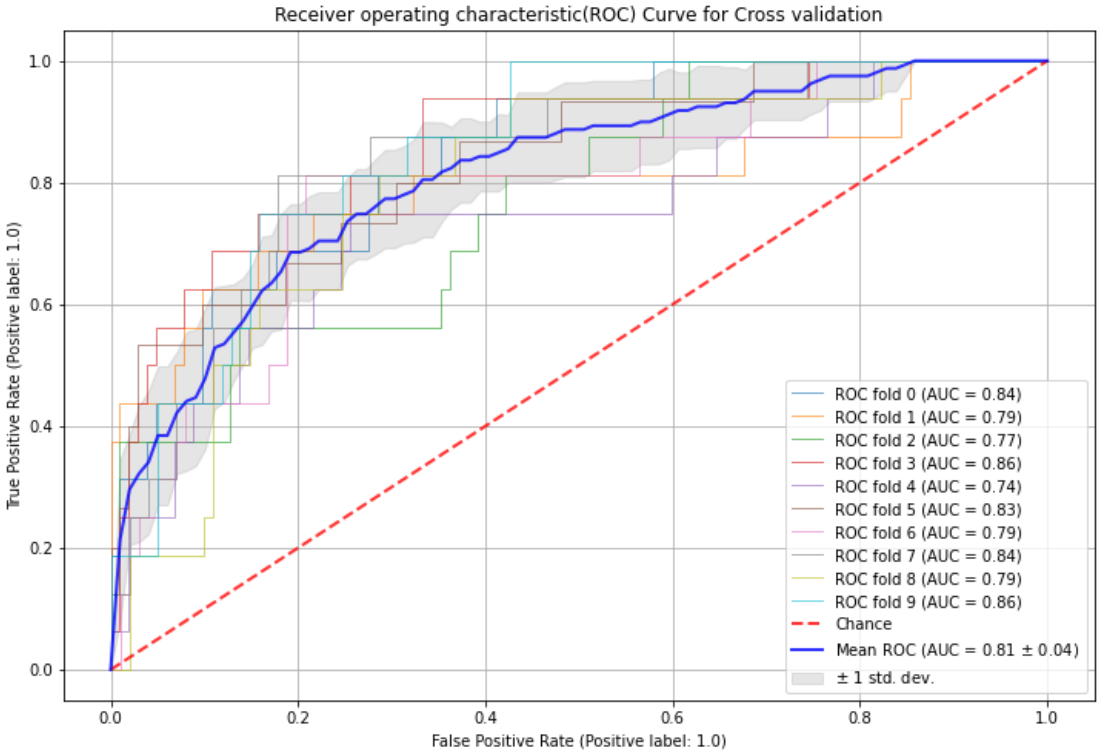
(e) KNN



(f) Naïve Bayes



(g) SVC


Receiver operating characteristic(ROC) Curve for Cross validation
True Positive Rate (Positive label: 1.0)
False Positive Rate (Positive label: 1.0)
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.84)
ROC fold 1 (AUC = 0.79)
ROC fold 2 (AUC = 0.77)
ROC fold 3 (AUC = 0.86)
ROC fold 4 (AUC = 0.74)
ROC fold 5 (AUC = 0.83)
ROC fold 6 (AUC = 0.79)
ROC fold 7 (AUC = 0.84)
ROC fold 8 (AUC = 0.79)
ROC fold 9 (AUC = 0.86)
Chance
Mean ROC (AUC = 0.81 ± 0.04)
± 1 std. dev.

(h) Voting