Heart Failure Analysis

02405051

May 2024

1 Introduction

Heart failure is one of the Cardiovascular diseases, which kills millions of people every year. Heart failure occurs when the heart is unable to pump sufficient blood to the body. Lots of diseases can lead to heart failure, including high blood pressure, diabetes, or other heart conditions (Chicco and Jurman [2020]). Predicting the survival of each patient with heart failure symptoms has become an important topic for medical doctors and physicians. However, there are still some problems in this area. For example, most of the models do not have high predicted accuracy (Smith et al. [2011]). In addition, although scientists have identified a wide range of predictors and indicators, there is no consensus on their relative impact on predicting survival (Levy et al. [2006]). More can be done in this area to solve the above problems and provide more reliable results.

In this study, we focus on patients who had heart failure and want to predict whether they will survive or not by considering their clinical features. The classification models we use include Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), Naive Bayes, Logistic regression and Random forest. We mainly focus on answering the following questions:

- Which classification model performs best when predicting patients' survival?
- How do clinical features affect patients' survival?

2 Data

The whole dataset comes from the Kaggle website. There are 5000 medical records in the dataset, and all of these patients had heart failure during their follow-up period. There are also 12 clinical features in the dataset and they are listed in table 1. In the following classification analysis, we use 'DEATH_EVENT' as our response variable and try to use the remaining features to predict the response variable.

3 Exploratory Analysis

Before doing the classification modelling, it is necessary to have an initial look at the data. Figure 1 includes the boxplots of continuous variables across different survival groups. As can be seen, all of the features show different distributions across 'DEATH_EVENT' groups, which means it is reasonable to include them in the model to predict the survival of patients. However, 'Creatinine phosphokinase' and 'Serum creatinine' features have heavy-tailed distributions. To increase the classification accuracy, we need to make transformations to those features, since some classification methods are built based on the distance between samples.

After trying some transformation methods, we choose to apply log transformation on the 'Creatinine phosphokinase' variable and reciprocal transformation on the 'Serum creatinine'

Feature	Description
age	Age of the patient (years)
anaemia	Decrease of red blood cells or hemoglobin (0:No, 1:Yes)
creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood (mcg/L)
diabetes	If the patient has diabetes (0:No, 1:Yes)
ejection fraction	Percentage of blood leaving the heart at each contraction
high blood pressure	If the patient has hypertension (0:No, 1:Yes)
platelets	Platelets in the blood (kiloplatelets/mL)
sex	0: woman, 1: man
serum creatinine	Level of serum creatinine in the blood (mg/dL)
serum sodium	Level of serum sodium in the blood (mEq/L)
smoking	If the patient smokes or not (0:No, 1:Yes)
DEATH_EVENT	If the patient died during the follow-up period (0:No, 1:Yes)

Table 1: Clinical features of Patients

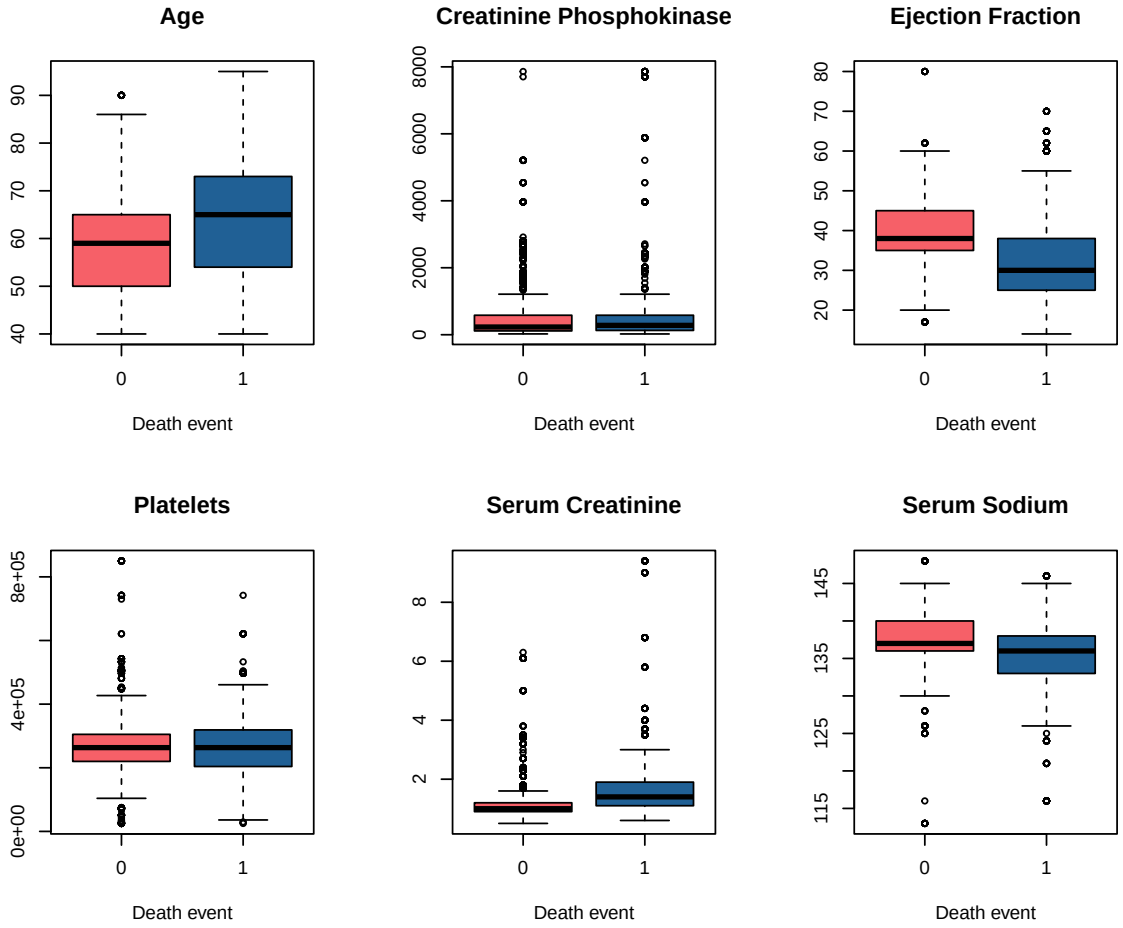


Figure 1: Boxplots of continuous variables.

variable. The boxplots of these two variables after transformations can be seen in figure 2. Now their distributions look better and there are fewer outliers.

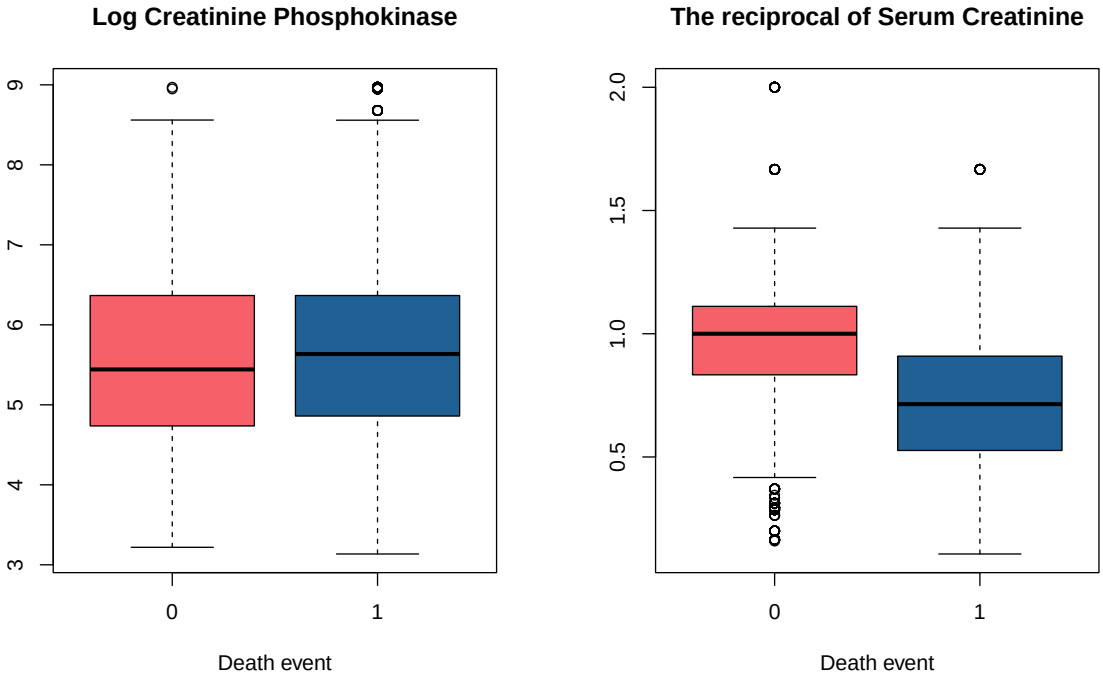


Figure 2: Boxplots of continuous variables after transformation.

Then, we use figure 3 to demonstrate the correlation between features. Positive correlation is shown in blue and negative correlation is shown in red. To be specific, our response variable 'DEATH_EVENT' has a positive correlation with age and the related coefficient is 0.25. Also, 'ejection fraction', 'the reciprocal of serum creatinine' and 'serum sodium' have negative correlations with the survival of patients, with the corresponding coefficients of -0.29, -0.38 and -0.23 respectively. Apart from these, the remaining features have minimal correlations with the response variable.

Figure 4 shows the distribution of the 'DEATH_EVENT'. There were 1568 patients dead during the follow-up period and 3432 patients survived.

In terms of the distributions for the variables, figure 5 contains the density plots for those continuous variables across survival groups. To illustrate, the density plots for 'log creatinine phosphokinase' and 'platelets' features are similar across the two survival groups. The density plot of age indicates that older patients might have a higher risk of death. However, patients with lower levels of ejection fraction, the reciprocal of serum creatinine and serum sodium seem to have a higher risk of death.

Figure 6 provides some barplots depicting the distributions of various binary categorical variables across two groups. Those plots show that the percentage of patients with anaemia, diabetes or high blood pressure who experienced a death event is substantially higher than those without these diseases. Moreover, a higher count of males experienced a death event compared to females, suggesting a possible gender difference in mortality risk in the dataset. In contrast, the percentage of death events seems to be the same across smoking groups and non-smoking groups, indicating smoking might not be the major feature affecting the patients' deaths.

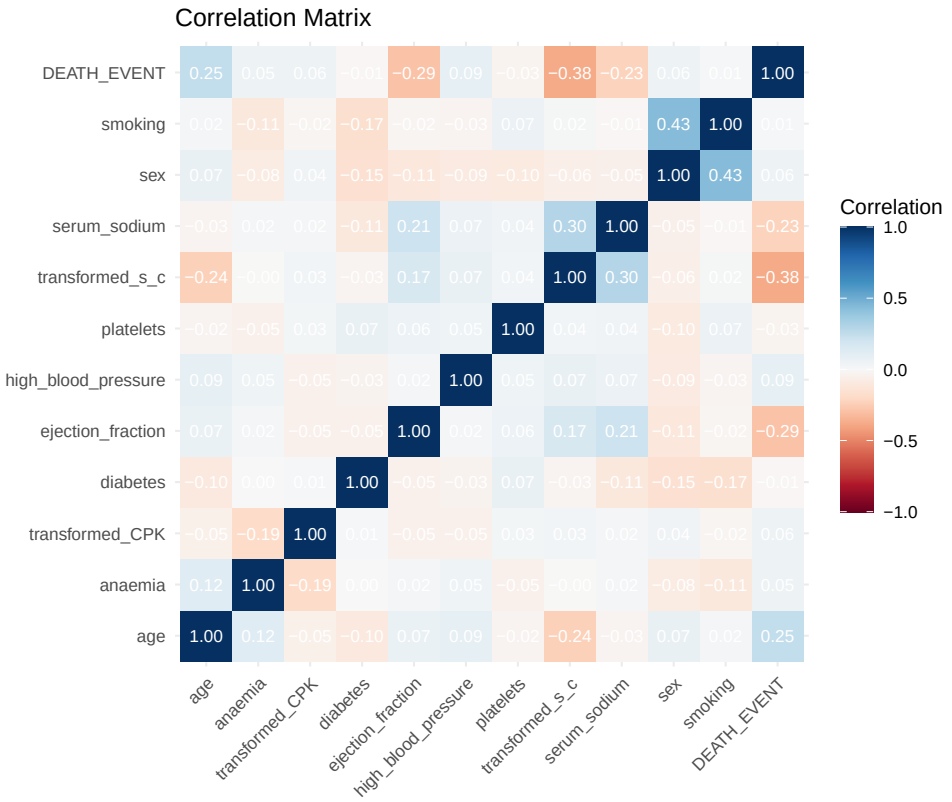


Figure 3: Correlation matrix for variables.

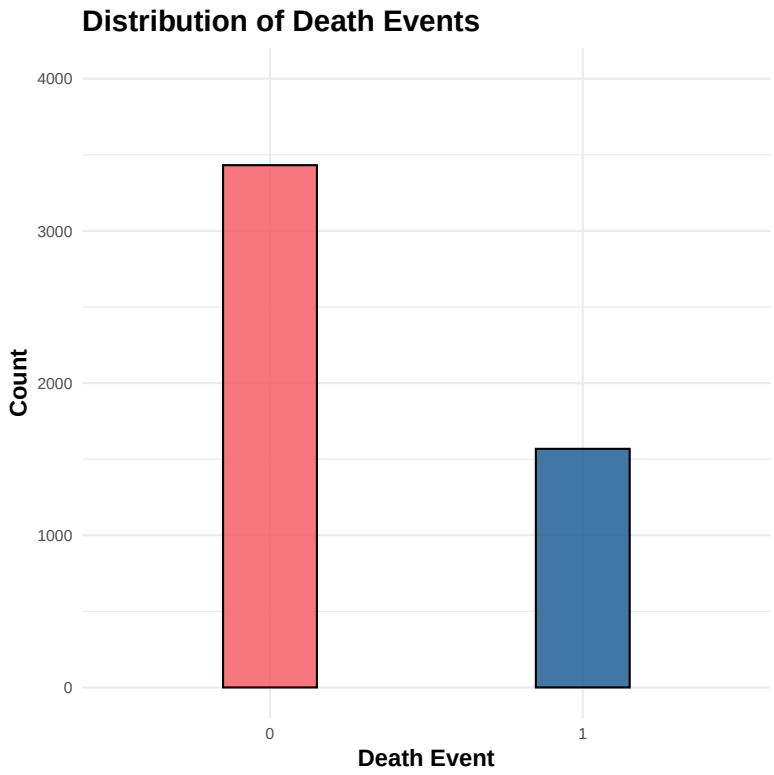


Figure 4: Distribution of response variable.

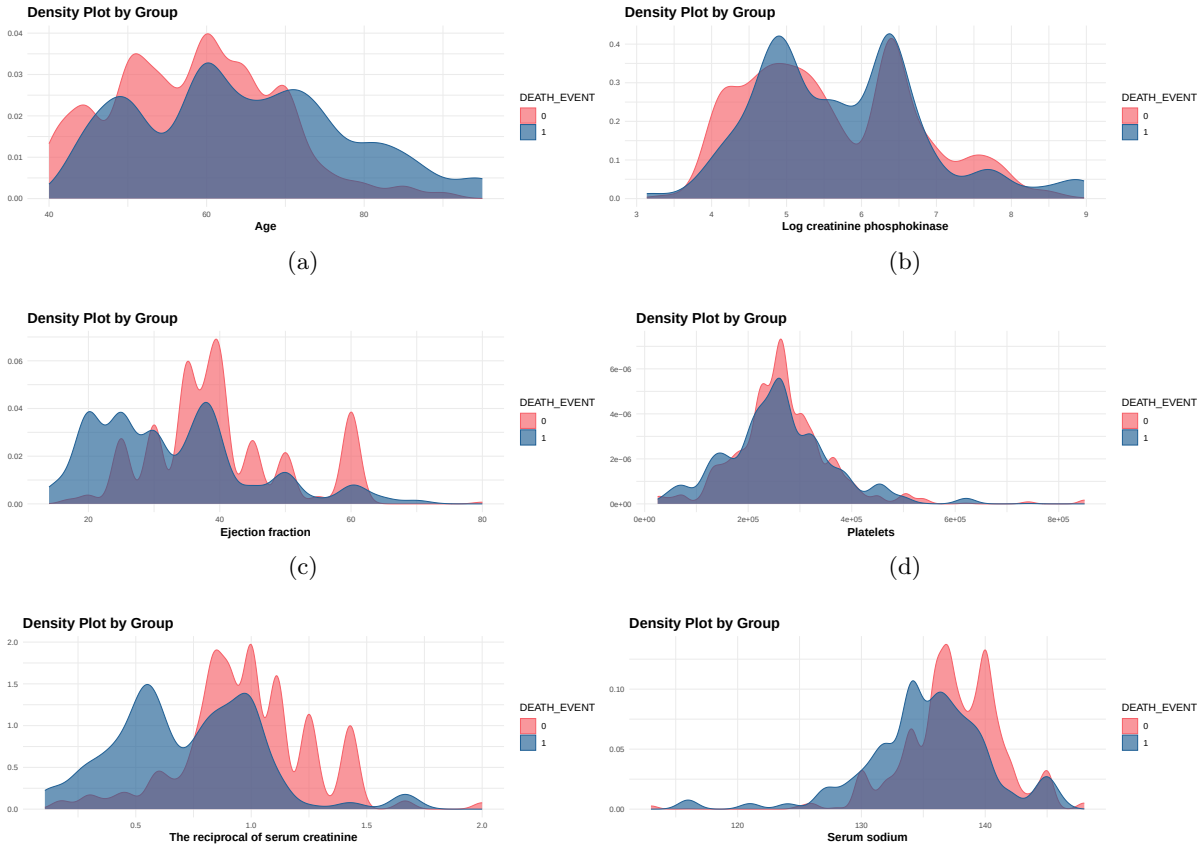
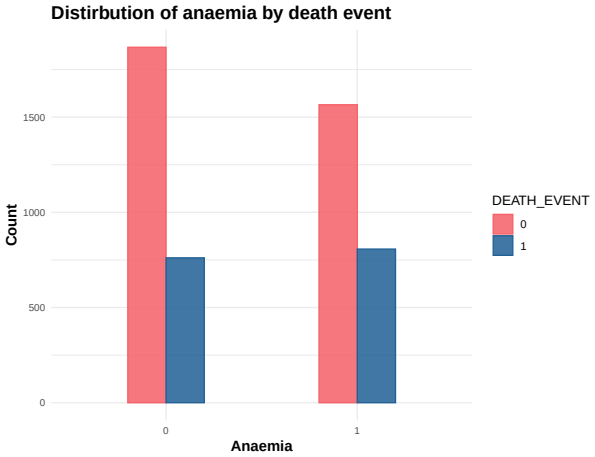
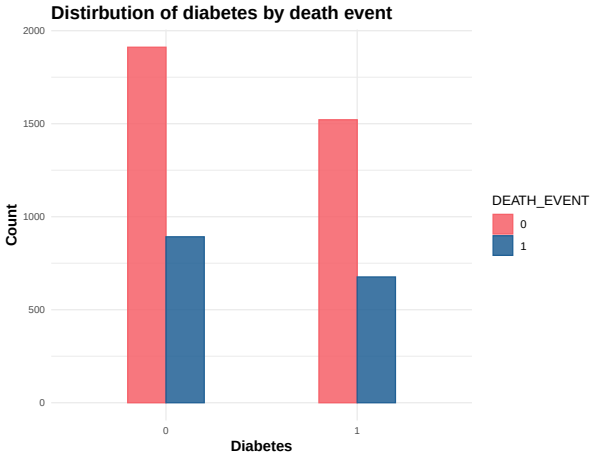


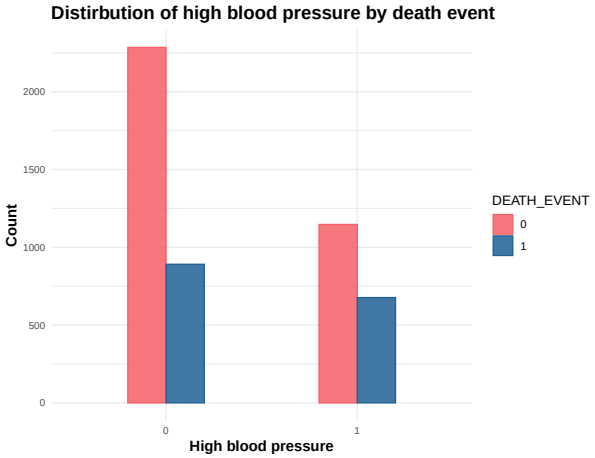
Figure 5: Density plot of different continuous variables across groups.



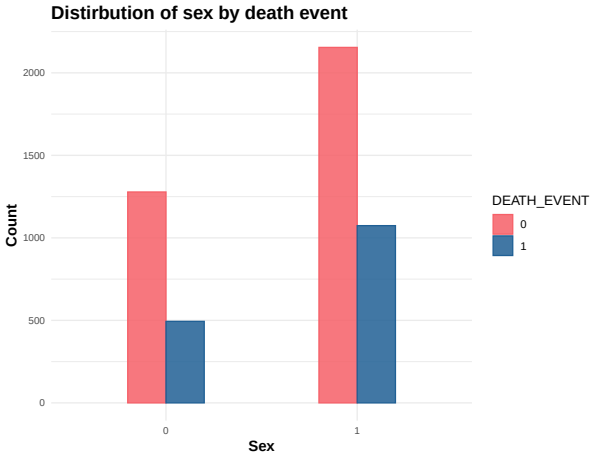
(a)



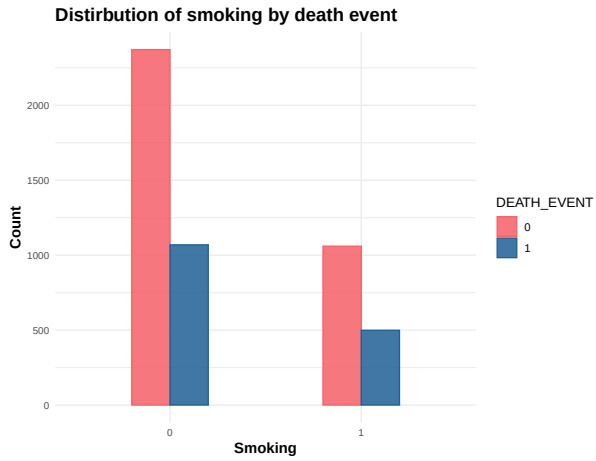
(b)



(c)



(d)



(e)

Figure 6: Barplot of different categorical variables across groups.

4 Analysis and results

Before doing the modelling, we first do the standardisation for all the continuous variables, since lots of classification methods account for the distance between samples, just as I said in the transformation section. Large differences in the scale of features might decrease the prediction accuracy. We do the standardisation by abstracting the variables' mean and then dividing by their standard deviation.

We also divide the whole dataset into the training set and test set. The training set is used to train the model and we compare the prediction performance of different models on the test set. After the splitting, there are 3750 samples in the training set and 1250 samples in the test set.

4.1 LDA and QDA

We first consider the Linear discriminant analysis (LDA) and Quadratic discriminant analysis (QDA). The basic idea behind the LDA is to project the high-dimensional data to a line, by decreasing the distance between samples from the same group and increasing the distance between samples from different groups. After fitting the model on the training set, we find a linear combination of features that best separates two survival groups and the coefficients of the line can be seen in table 2. We can find that the absolute values of the coefficients 'age', 'ejection fraction', 'high blood pressure' and 'the reciprocal of serum creatinine' are about 0.5. Among them, 'Age' and 'high blood pressure' have positive coefficients, while the remaining features have negative coefficients. These results coincide with what we have found in the correlation matrix in figure 3.

Then we apply the trained model on the test set to make predictions. The results can be seen in table 3.

Variable	**Coefficient**
Age	0.428
Anaemia	0.247
Log creatinine phosphokinase	0.173
Diabetes	-0.089
Ejection_fraction	-0.515
High_blood_pressure	0.429
Platelets	-0.0097
The reciprocal of serum creatinine	-0.672
Serum_sodium	-0.194
Sex	-0.035
Smoking	0.041

Table 2: Coefficients of Linear Discriminants

Predicted \ Actual	0	1
0	765	198
1	93	194

Table 3: Confusion Matrix of LDA

The QDA is the extension to the LDA and the decision boundary for QDA is non-linear. After training the QDA on the training set, the results of the prediction on the test set can be

seen in table 4. The predictions of QDA are more accurate than those of the LDA since the diagonal of the confusion matrix has more value for QDA.

Predicted \ Actual	0	1
0	769	157
1	89	235

Table 4: Confusion Matrix of QDA

4.2 Naive Bayes

Naive Bayes is a probabilistic classification model based on Bayes' Theorem. Naive Bayes assumes all the features are independent of each other given the class label. This assumption simplifies the computation of the posterior probability for each class, and the Naive Bayes classifier assigns a new sample to the class with the highest posterior probability. After training the model, the confusion matrix of the test set can be seen in table 5. This classifier does not perform better than the previous two methods. One of the possible reasons might be that some variables have moderate correlations with each other, which are ignored by the assumption of the Naive Bayes.

Predicted \ Actual	0	1
0	755	183
1	103	209

Table 5: Confusion Matrix of Naive Bayes

4.3 Logistic Regression

Logistic regression is a generalised linear model with the binary response variable, which is used for classification. The probability that the sample belongs to class 1 can be defined as follows:

$$P(Y = 1|X) = \sigma(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{11} X_{11}),$$

where $\beta_0, \beta_1, ..., \beta_{11}$ are the parameters of the model and $X_1, X_2, ..., X_{11}$ are the variables. Also $\sigma(x)$ can be defined as

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

After fitting the model on the training set, the estimated coefficients can be seen in table 6. The results show that 'age', 'anaemia', 'log creatinine phosphokinase', 'ejection fraction', 'high blood pressure', 'the reciprocal of serum creatinine' and 'serum sodium' have statistically significant effects on the response variable since their p-values are extremely small. The estimated coefficient represents the change in the log odds of the dependent variable for a one-unit change in the predictor variable. For example, for the age variable increase of one year, the odds of the patients being dead will get multiplied by exp(0.56), assuming other variables are held constant.

Then the trained model is applied to the test set. To make a classification decision, the predicted probability $\sigma(x)$ is compared to 0.5. If $\sigma(x) > 0.5$, the sample is assigned to class 1; otherwise, it is classified as class 0. The confusion matrix can be seen in table 7.

Coefficient	**Estimate**	**P-value**
(Intercept)	-1.448	$< 2e-16$***
Age	0.560	$< 2e-16$***
Anaemia	0.417	2.14e-06***
Log creatinine phosphokinase	0.294	2.21e-11***
Diabetes	-0.071	0.418
Ejection_fraction	-0.693	$< 2e-16$***
High_blood_pressure	0.601	1.40e-11***
Platelets	-0.053	0.216
The reciprocal of serum creatinine	-0.899	$< 2e-16$***
Serum_sodium	-0.173	5.2e-05***
Sex	-0.153	0.128
Smoking	0.119	0.242

Table 6: Logistic Regression Coefficients

Predicted \ Actual	0	1
0	771	197
1	87	195

Table 7: Confusion Matrix of Logistic regression

4.4 Random forest

We also consider the Random forest model, which is an ensemble method for classification and regression tasks. Random forest is the extension of the decision tree by building many trees on different sub-samples of the whole dataset during the training process and assigning the sample to the class that is the majority class predictions by individual trees.

The random forest can also help us to understand which features contribute most to the model's prediction accuracy. The variable importance can be seen in figure 7. The left plot shows the variable importance based on the mean decrease in accuracy. A higher value of the mean decrease in accuracy indicates that the variable is more important for the model accuracy. As can be seen, 'the reciprocal of serum creatinine ', 'ejection fraction', 'age', 'log creatinine phosphokinase ' and 'platelets' are the top 5 important variables for the model. The right plot shows the variable importance based on the mean decrease in Gini impurity, which is the measurement of the purity of node splits. Although some of the variables have different levels of importance compared with the plot on the left, most variables are ranked in roughly the same order of importance. Especially, the top 4 important variables are the same in these two plots.

Then the trained model is applied to the test set, and we get the confusion matrix as table 8. The random forest has good prediction accuracy on the test set.

Predicted \ Actual	0	1
0	852	11
1	6	381

Table 8: Confusion Matrix of Random forest

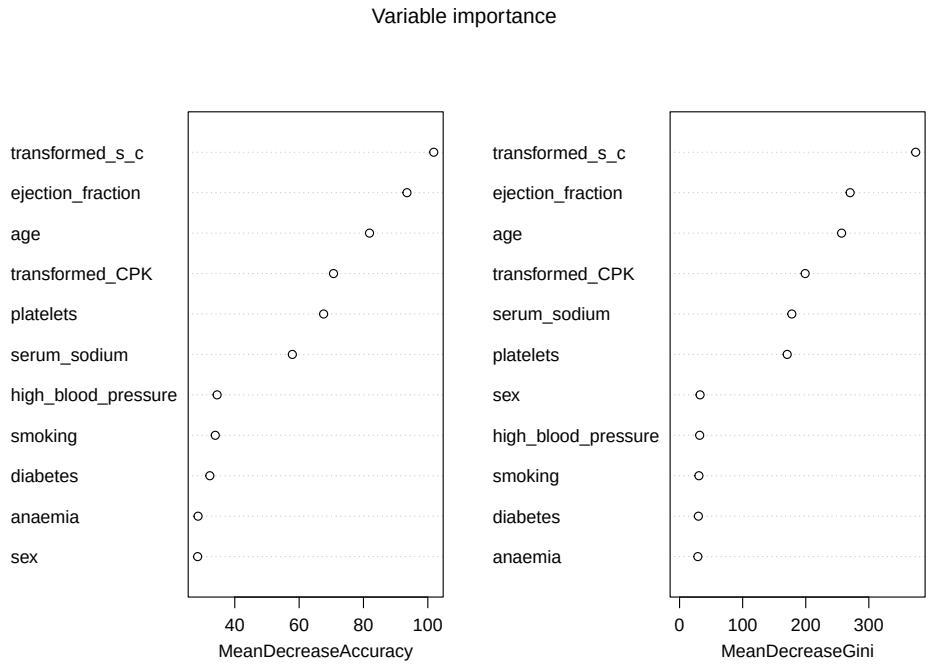


Figure 7: Variable importance of Random forest model.

5 Conclusion

5.1 Prediction performance

The prediction accuracy on the test set for different models can be computed from the confusion matrix and the results can be seen in table 9. Overall, the Random forest has the highest prediction accuracy, which is 0.9864. This value is far higher than the accuracy of other methods. QDA model is the second best model. There is no such difference in prediction accuracy between the remaining three models.

Model	**Prediction accuracy**
LDA	0.7672
QDA	0.8032
Naive Bayes	0.7712
Logistic Regression	0.7728
Random forest	0.9864

Table 9: Prediction accuracy on the test set for different models

5.2 Clinical features effect

Although different models suggest slightly different important features, there are still some common variables between them. The common important features include 'age', 'ejection fraction' and 'the reciprocal of serum creatinine'. These variables play an important role in model accuracy and have effects on the survival of patients with heart failure.

References

D. Chicco and G. 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, 2020. URL https://api.semanticscholar.org/CorpusID:211018036.

W. C. Levy, D. Mozaffarian, D. T. Linker, S. C. Sutradhar, S. D. Anker, A. B. Cropp, I. Anand, A. Maggioni, P. Burton, M. D. Sullivan, B. Pitt, P. A. Poole-Wilson, D. L. Mann, and M. Packer. The seattle heart failure model. *Circulation*, 113(11):1424–1433, 2006. doi: 10.1161/CIRCULATIONAHA.105.584102. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.105.584102.

D. H. Smith, E. S. Johnson, M. L. Thorp, X. Yang, A. Petrik, R. W. Platt, and K. Crispell. Predicting poor outcomes in heart failure. *The Permanente Journal*, 15(4):4–11, 2011. doi: 10.7812/TPP/11-100. URL https://www.thepermanentejournal.org/doi/abs/10.7812/TPP/11-100.