Prediction of Heart Failure using Classification Model

Chun Hin Attic Lee - 001183418

Abstract

Cardiovascular disease is a real-world problem that is a leading death of cause globally. Diagnose the disease at early stage is always the best for the patients, but it is always a challenge. The aim of this paper explore different machine learning algorithm and to evaluate and compare their performance on heart failure prediction. Various algorithm were used including Logistic regression, L1 and L2 regularization, KNN, SVC, Decision tree, and Random forest. Performance was measured by accuracy score, F1 Score and AUC Score. The best model of this study is using L2 regularization algorithm, with 87.68% accuracy and 93.36% AUC score.

1. Introduction

Cardiovascular disease is a term that describing the various disease that involve of heart and blood vessels which can lead to serious symptom such as angina, heart failure and heart attacks. According to WHO (2021), Cardiovascular disease is the leading cause of death globally: In 2019, an estimated of 17.9 million people died because of Cardiovascular disease, which equals to 32% of all global deaths.Heart attack and stroke causes 85% of these death cases.

In this research, a machine learning solution will be designed to solve a real-world topic and problem which can be solved by machine learning effectively. Cardiovascular disease is a serious problem, however, it can be difficult to diagnose at the early stage. This can be beneficial from machine learning by providing prediction, potential patients can get proper advice and treatment that can prevent their health goes worse before the situation become irreversible and more difficult to cure.

1.1. Objective

Early detection and management is essential for people with cardiovascular disease or at high cardiovascular risk, which means with one or more risk factor such as high blood pressure, diabetes. The objective of this work is to predict if the target is at a high risk of being diagnosed as a cardiovascular disease patient. The machine learning model can act as an early screening of cardiovascular disease. When someone has a risk factors that is determined by the machine learning model, they can seek for treatment from their doctor, hence reducing the chance to become advanced heart failure and facilitating their recovery process, and reducing the rate of death causing by Cardiovascular disease.

2. Methods

Classification algorithm were be used for achieving the desired outcome that is a supervised machine learning method that the model based on the data given to produce hypothesis that predict future events(Kotsiantis, Zaharakis, Pintelas, et al. 2007).Binary classification problems are common with health-related data, such as 'infected' and 'not infected','disease' and ' no disease', those are the classes label that require to be classified and predicted. Classification algorithms will be used for solving the problem as it can utilize the previous data from the diagnosis of the patients, and find out the correlation between different factor and heart disease, and predict the result of whether new data is at high risk of heart failure.

Classification is the most effective approach to solve this problem. In classification, the dataset splits into three different set:Train,Validation, and test set. Training set, is a set of data that used to train an model to learn the pattern of the data and find the optimal parameter of model. Validation set, is a set of data to validate model performance and prevent overfitting in the training phrase. Test set, is a set that test the model after training, determining the accuracy of the model.(Suthaharan 2016)

In this study, multiple supervised algorithms were implemented and compared to find out the best model for this classification problem. Logistic Regression, L1 regularization, L2 regularization, Random forest, Support vector machine,K-Nearest Neighbors algorithm, and Decision Tree were used. The dataset used in this study was 'Heart Failure Prediction Dataset.' from fedesoriano. The dataset is the largest heart disease dataset available nowadays that combines 5 dataset with over 11 common features for research purpose. The algorithms were compared to dummy baseline model.

2.1. Logistic Regression

Logistic regression is one of the most popular model for classification. It is use binary classifications, which suits the case of the heart disease dataset, which the goal of the classification model is to predict whether the input is 'Heart disease' or 'No heart disease'. n denoted as the number of predictor variables,e is the base of the natural logarithm, α and β are the parameter of the model, P is probability of 1, X is the independent variable that is the rolling mean of the dependent variable. The equation of logistic regression is shown below

$$P = \frac{e^{(\beta_0+\beta_1 X_1 \cdots + \beta_n X_n)}}{1 + e^{(\beta_0+\beta_1 X_1 + \cdots \beta_n X_n)}} \quad (1)$$

2.2. Regularization

Regularization is a method that prevent the model from overfitting by penalizing high-valued regression coefficients. L1 regularization will tend to shrink coefficients to zero by estimating the median of the data while the L2 regularization tends to shrink coefficients evenly by estimating the mean of the data to avoid overfitting.

2.3. Decision tree

Decision tree is a popular model for classification and regression problem. The algorithm is like a flow-chart and a tree-base model, each nodes is a if/else question, leads to different branches, hence to make the 'best' possible decision.(Moret 1982) However, this model is prone to overfitting when the tree become complex, they tend to have a good fit on the dataset, which also means it do not generalize the data well and result in overfitting.

2.4. Random forest

Random forest is another tree-based model that uses ensemble methods, by using an collection of several weaker decision trees as sub-models to build a stronger model and the prediction are produced by the summarizing the prediction of each of the sub-models.(Biau and Scornet 2016)

2.5. K-Nearest Neighbor algorithm

K-Nearest Neighbor algorithm is a simple algorithm. It classifies data points that are nearby into a same class.When there is a new data point, the model makes predictions by finding the a similar instance in the training set.

2.6. Support vector machine

Support vector machine is a algorithm that assign the data into a multiple dimension features environment. A linear decision boundary is formed in this environment, the best boundary created is called hyperplane. The closest data point to the hyperplane is called support vector point. The goal of the algorithm is to maximize the margin between the hyperplane and Support vector point.(Cortes and Vapnik 1995)

3. Experiments

In the experiment of this research, Dataset used in this study was obtained from Kaggle. Refer to Table 1, the dataset has 12 attributes and 11 features and it is combines from 5 different heart datasets from the UCI Machine Learning Repository. The initial size of the data is 918.

	Attributes
1	Age
2	Sex
3	ChestPainType
4	RestingBP
5	Cholesterol
6	FastingBS
7	RestingECG
8	MaxHR
9	ExerciseAngina
10	Oldpeak
11	ST_Slope
12	HeartDisease

Table 1. Dataset Attributes (fedesoriano 2021)

3.1. Experimental settings

In this study, the task is to build models for prediction of the class based on selected attributes. And the objectives is to build a model that can successfully predict the heart failure. Dummy classifier was used in this study as a baseline model for comparisons with other models in this study. The baseline model gives a baseline performance of a model, which is the probability of success without any learning.

Data preprocessing is essential in machine learning as it standardize the data and transform it understandable format for the algorithms. During this stage, duplicate and missing value must be dropped. Then, removing the 0 value of 'RestingBP' as there is impossible to have 0 blood pressure, it an abnormal value. The 'Cholesterol' also having 0 value that is abnormal, replace them with the median of the column. Finally, using scaling features to normalize the independent variables of the features. Once finishing data preprocessing, the dataset is ready to be processed. Process the data with one-hot encoding for better prediction as the algorithm would not treat a higher value of variable as more important than the others. Then, The dataset was split into 70% of training set and 30% of test set and trained with dif-

ferent models. Evaluate metrics can be obtained and tuning hyperparameter until getting a model that is ideal.

3.2. Evaluation criteria

In this study, accuracy score was used as a metric to evaluate the trained model. Accuracy score is a common classification evaluation metric. The equation of accuracy score is shown below,

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives. In simpler word, Accuracy is the ratio of correct prediction to the total prediction, which means the rate of successful prediction. Accuracy score is not suitable for imbalance data as it will give us the accuracy on the majority class, the accuracy is misleading. In this case, further metrics are required such as F1 score and ROC Curve.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3}$$

The F1 score calculated from harmonic mean of precision and recall and can handle imbalance classes better.

ROC is a probability curve that predicting the probability of a binary output of a model, and AUC is a score that the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. It tells how well is the model's ability to classify samples (Narkhede 2018).

In this study, accuracy score is suitable as the dataset is fairly balanced. However, F1 score and AUC score were also considered for evaluating the model.

3.3. Results

Models	Accuracy Score	F1 Score	AUC Score
L2 regularization	0.8768	0.8957	0.9336
Logistic Regression	0.8732	0.8930	0.9334
L1 regularization	0.8696	0.8896	0.9340
Random forest	0.8659	0.8889	0.9299
svm	0.8514	0.8754	0.9281
KNN	0.8261	0.8519	0.8288
Decision Tree	0.7790	0.8063	0.7870
Dummy	0.6123	0.7596	0.5000

Table 2. Models Performance

According to Table 2, the L2 regularization model has the best performance among all of the classification models, which has a 0.8768 accuracy score, 0.8957 F1 score, and 0.9336 AUC score. Figure 1 shows the confusion matrix of the different models. Figure 2 shows the importance feature level of different models.

For the features importance level, as it shown in Figure 2, the 'OldPeak', which is the decrease of the ST segment during exercise shown in the electrocardiogram. A ST depression can indicates a heart disease and associate with angina (Medicine, Populations, and Criteria 2010). 'OldPeak' is the highest correlation in linear models, and the 'Sex_M', which indicates male, is the second highest correlation in linear models. In tree-based model, 'ST_slope_up', which means the ascending slope of heart of ST segment during exercise, and it can associate with heart disease(Finkelhor et al. 1986). Hence, it has significant correlation to heart disease in tree-based algorithms.

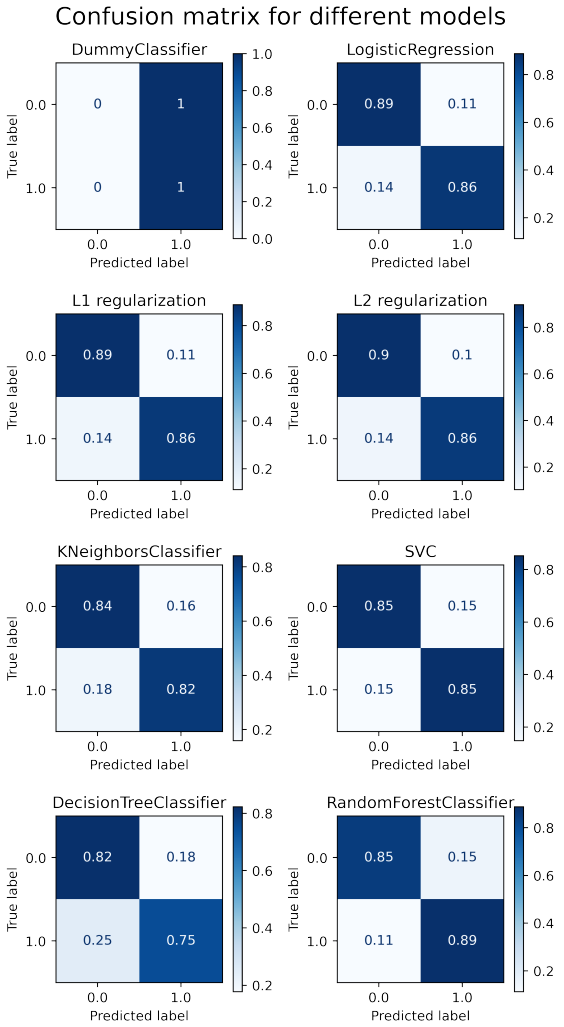

Figure 1. Confusion Matrix for different models

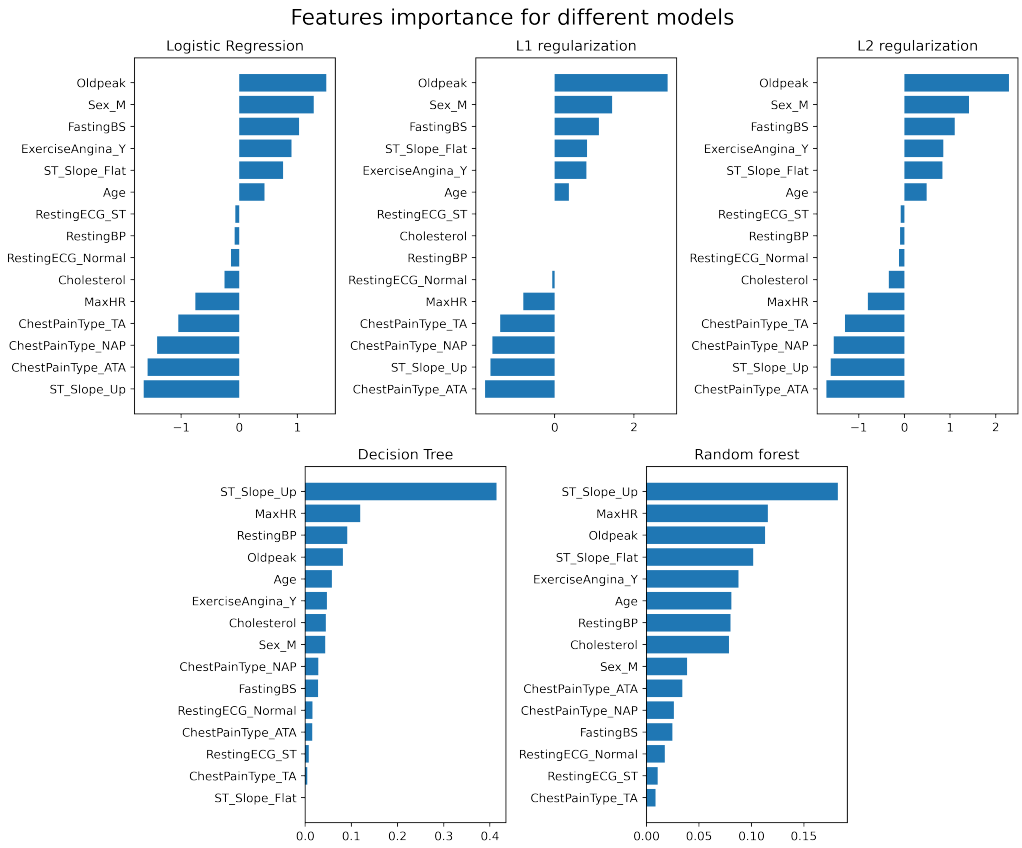

Figure 2. Features importance for different models

3.4. Discussion

Referring to Table 2, the L2 regularization model has the best performance among the others, with highest accuracy score and F1 score. Overall, linear model performs better than tree-based models. This is because the dataset has fewer noise variable than explanatory variable, while true and false positive result is appear more frequent in random forest model when the explanatory variable is high (Kirasich, Smith, and Sadler 2018). The result of the feature importance level shows that people with ST depression and male would more likely to be predicted as 'Heart disease'. In this study, it is difficult to conclude on whether the predictor variable of this model is correctly representing the cause of heart disease as it need to be proved by medical professional. While for this study, it has enough evident to prove that this model is success.

4. Conclusion

Detecting a heart disease in early stage is challenging in real-world. However, machine learning can solve this problem handily. In this study, a successful machine learning model was designed to predict the heart disease of a person from their body testing result. The result of the experiment had achieved 87.68% accuracy with L2 regularization algorithm. The model can assists doctor's work of screening people at high cardiovascular risk. Nonetheless, it is also possible to extend it for detecting other diseases.

References

Biau, Gérard and Erwan Scornet (2016). "A random forest guided tour". In: *Test* 25.2, pp. 197–227.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

fedesoriano (Sept. 2021). "Heart Failure Prediction Dataset". In: URL: https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Finkelhor, Robert S. et al. (Aug. 1986). "The St Segment/Heart Rate Slope as a predictor of coronary artery disease: Comparison with quantitative thallium imaging and conventional ST segment criteria". In: *American Heart Journal* 112.2, pp. 296–304. DOI: 10.1016/0002-8703(86)90265-6.

Kirasich, Kaitlin, Trace Smith, and Bivin Sadler (2018). "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets". In: *SMU Data Science Review* 1.3.

Kotsiantis, Sotiris B, I Zaharakis, P Pintelas, et al. (2007). "Supervised machine learning: A review of classification techniques". In: *Emerging artificial intelligence applications in computer engineering* 160.1, pp. 3–24.

Medicine, Institute, Board Populations, and Committee Criteria (Dec. 2010). *Cardiovascular disability: Updating the social security listings*, pp. 1–278. DOI: 10.17226/12940.

Moret, Bernard M. E. (Dec. 1982). "Decision Trees and Diagrams". In: *ACM Comput. Surv.* 14.4, pp. 593–623. ISSN: 0360-0300. DOI: 10.1145/356893.356898. URL: https://doi.org/10.1145/356893.356898.

Narkhede, Sarang (2018). "Understanding auc-roc curve". In: *Towards Data Science* 26, pp. 220–227.

Suthaharan, Shan (2016). "Machine learning models and algorithms for big data classification". In: *Integr. Ser. Inf. Syst* 36, pp. 1–12.

WHO (June 2021). *Cardiovascular diseases (cvds)*. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).