Comparing Decision Trees, Logistic Regression, and LASSO Regression for Predicting Heart Disease

Changda Li (T00705321)

Supervisor: Dr. Erfanul Hoque

DASC 5420 Project Report

Abstract

This study aimed to compare the performance of various machine learning algorithms, including decision trees, logistic regression, and LASSO regression, in predicting heart disease using a dataset that included multiple key indicators. Exploratory data analysis techniques were utilized to analyze the dataset and gain insight into the distribution and relationships of the variables. In addition, model performance was evaluated based on several metrics, such as accuracy, AUC, precision, recall, and F1 score, on both the original and undersampled data to account for class imbalance. Furthermore, the results indicated that logistic regression was more effective than the decision tree model in most areas, while the performance of LASSO and logistic regression was similar.

1. Introduction

Heart disease is one of the leading causes of death worldwide, claiming millions of lives yearly. Therefore, detecting and preventing heart disease early is essential to improve patient outcomes and reduce healthcare costs significantly. Moreover, machine learning algorithms can help predict the risk of heart disease to enable early diagnosis and timely interventions.

Thus, I will analyze the Personal Key Indicators of Heart Disease dataset in this report. The data set contains eighteen variables of crucial information, including sex, age, stroke, BMI, sleep time, and more, which are collected from around three hundred and twenty thousand individuals.

I aim to compare and analyze the performance of three widely used machine learning algorithms, which are decision tree, logistic regression, and LASSO regression, to predict the occurrence of heart disease using the given features. In addition, by assessing their effectiveness, I aim to pinpoint the most appropriate algorithm for this data set, which could help enhance the accuracy and efficiency of heart disease diagnostic tools.

Therefore, throughout this report, I will briefly describe the data set, then introduce each machine algorithm and how it works. In addition, this report will also describe pre-processing steps applied to the data set before training the algorithms. Furthermore, I will evaluate the performance of each algorithm using appropriate metrics such as accuracy, precision, recall, and F1 score. I will also compare the performance of each algorithm to identify their strengths and weaknesses for heart disease prediction.

Finally, the report concludes with the most effective algorithm for predicting heart disease and potential improvements or limitations. In other words, this study can provide valuable information on using machine learning algorithms to predict heart disease risk.

2. Background

Heart disease, also known as cardiovascular disease, includes a series of conditions that will impact the heart and related blood vessels. According to the World Health Organization, heart disease is the leading cause of death globally, accounting for nearly 18 million deaths annually, or 32% of all global deaths [1]. Because of its significant impact on public health, extensive research has been conducted to understand better, diagnose, and treat heart disease.

Several risk factors contribute to the development of heart diseases, such as age, sex, genetic predisposition, lifestyle choices, and underlying health conditions [2]. Early identification of these risk factors and timely intervention can significantly improve patient outcomes and reduce

the burden on healthcare systems. In addition, in recent years, machine learning has emerged as a helpful tool for predicting and diagnosing heart disease, leveraging data-driven insights to identify patterns and relationships within complex medical datasets [3].

In this report, I aim to go deeper into the performance of these three machine learning algorithms in the context of heart disease prediction using the Personal Key Indicators of Heart Disease dataset. Furthermore, by comparing their performance, I also want to identify the most suitable algorithm for this task.

3. Data

3.1 Overview

The Personal Key Indicators of Heart Disease is a dataset containing critical information about the patients and whether they have heart disease. It has eighteen variables and about three hundred and twenty thousand rows.

In addition, the HeartDisease column is the target variable, and it is binary to indicate whether the patient has heart disease. The other seventeen variables consist of thirteen categorical variables, including Smoking, Alcohol Drinking, Stroke, Diff Walking, Sex, Age Category, Race, Diabetic, Physical Activity, Gen Health, Asthma, Kidney Disease, and Skin Cancer, as well as four continuous variables including BMI, Physical Health, Mental Health, and Sleep Time.

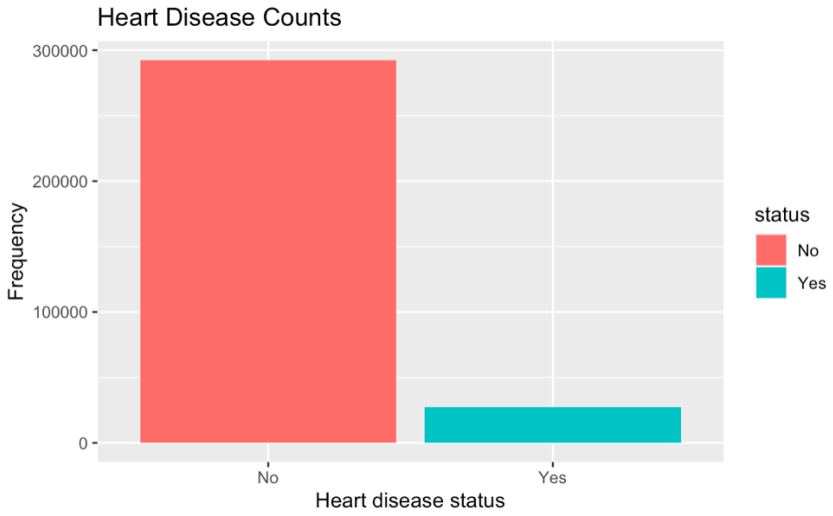

Figure 1. The distribution of heart disease status

3.2 Exploratory data analysis

Moreover, Figure 1 shows that the ratio between the positive (i.e., Yes) and the negative class (i.e., No) is roughly 1 to 11. In other words, the dataset is severely imbalanced. Thus, I might need to undersample the data in the later section for the minority class to avoid the machine learning models bias toward the majority class.

Figure 2 shows that heart disease status has a positive relationship with some categorical variables. For instance, more heart disease patients smoke since the proportion of patients with heart disease among smokers is higher than among non-smokers. The figure also shows that more and more patients have heart disease as age increases. Patients who have strokes are more

likely to have heart disease. Furthermore, people with poor general health are also more likely to be heart disease patients. The figure gives a general idea of the relationship between heart disease status (Yes or No) and other features.

Overall, the aim of analyzing this dataset is to predict the risk of heart disease based on several personal key indicators. Understanding the connections between the independent variables and the target variable can gain valuable insights into risk factors for heart disease. Statistical analysis techniques can be applied to this dataset to detect patterns and relationships that can help us to predict the likelihood of heart disease.

3. Methods

Figure 2. The proportion of heart disease status for each categorical variable.

3.1 Data Preprocessing

Scaling of data is a crucial preprocessing step that involves standardizing the continuous feature values to a standard scale. This is particularly important for machine learning models because they are sensitive to the scale of input features. If the features are on different scales, it can lead to an imbalance in the models since some features might have more weight than others. Therefore, I will scale the value of the continuous variables of the dataset to ensure the models are as accurate as possible for the study.

I will also perform a train-test split for the dataset, which means the dataset will be divided into training and testing sets. Doing so is typical for evaluating the performance of decision trees, logistic regression, and LASSO regression. In my study, 70% of the dataset will be a training set, and the rest 30% will be a testing set.

Moreover, as mentioned above, the dataset needs to be balanced; thus, I will undersample the dataset. Undersample is a helpful technique to balance uneven datasets by keeping all the data in the minority class and reducing the data size of the majority class to match the size of the minority. In addition, I will fit the models on both the original and balanced datasets to see whether balancing the data will improve the performance of those models.

3.2 Decision trees

Because the task of our study is to classify whether the patient is with or without heart disease, a decision tree is a good choice for the job. Decision trees are a type of supervised learning algorithm used for both classification and regression tasks. Moreover, the algorithm works by recursively partitioning the data into smaller subsets based on the values of different features and creating a tree-like model that can be used for predictions.

The math behind this algorithm involves selecting the best split at each tree node based on a criterion that measures how well the split separates the data into different classes or groups. There are several popular splitting criteria, including Gini impurity and entropy. For this study, I will use the "train" function in the "caret" library, and the function uses the default splitting criteria of the "rpart" function, which is the Gini index for classification. The Gini impurity measures the probability of misclassifying a randomly chosen element from the set. Furthermore, the algorithm selects the split that minimizes the weighted sum of the Gini impurities of the resulting subsets.

Once the best split is selected, the algorithm recursively applies the same process to the subsets until a stopping criterion is met, such as reaching a maximum depth or having a minimum number of samples at each leaf node. As a result, the final result is a tree-like model that can be used to make predictions on new data by traversing the tree based on the values of the input features and assigning the class label of the leaf node reached by the traversal.

Moreover, I will fit the decision tree model to the training set with 10-fold cross-validation to tune the hyperparameter cp. The complexity parameter (cp) controls the decision tree's size and selects the optimal tree size. The larger the cp value, the smaller the tree and the simpler the model, which reduces the risk of overfitting. On the other hand, smaller cp values allow for more complex models, which may increase the risk of overfitting. Thus, it is essential to choose the optimal cp, so I use cross-validation to find the optimal cp. Then, I will use the optimal cp to fit the model and then use the model on the testing set for evaluating the model performance.

3.3 Logistic regression

Logistic regression is another good choice for classification tasks since it predicts the probability of binary outcomes based on other independent variables. The logistic regression model is based on the logistic function, also known as the sigmoid function, which is given by:

$$\pi_i = p(y_i = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}$$

Where the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are regression coefficients, x's are the independent variable values, and e is the Euler's number.

Furthermore, the power of e is the linear combination of the 17 independent variables of the dataset and their regression coefficients. The sigmoid function transfers the linear combination to the probability of the target variable HeartDisease taking 1. Moreover, the essential part of the

logistic regression is to estimate the values of coefficients that maximize the likelihood function, which is given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

In order to get the maximum likelihood estimate (MLE) of the coefficients, we need to take log of the likelihood function, so we have the log-likelihood function as:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i log \frac{\pi_i}{1 - \pi_i} + log(1 - \pi_i) \right]$$

Thus, we can get the MLE by taking the derivative of the log-likelihood function with respect to the coefficients and setting them equal to zero:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi_i] = 0$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[(y_i - \pi_i) x_{ij} \right] = 0, \text{ where } j \in [1, p]$$

After having the MLE of the coefficients, we can plug them into the sigmoid function to predict the probability of the HeartDisease taking value 1 (i.e., Yes). Moreover, the default threshold of 0.5 will be used to classify new observations into either 0 or 1 based on the predicted probabilities for this study.

In addition, 10-fold cross-validation will be applied for training the logistic model to get the validation accuracy. Then the model will be tested to get the test accuracy to evaluate the model performance.

3.4 LASSO regression

Since there are 17 independent variables in the dataset, feature selection might help improve the model performance. LASSO regression is a linear regression that adds a penalty term to the least square equation that could be solved to get appropriate coefficients. Moreover, because the penalty term is the L1 norm of the coefficients, it will shrink them to zero; in other words, it will help the model eliminate unrelated variables.

The L1 norm of the coefficients is defined as the sum of the absolute values of the coefficients. Therefore, the LASSO regression objective function is defined as the sum of squared errors plus the product of a constant λ and the L1 norm of the coefficients:

$$\mathrm{E} = \mathrm{RSS} + \lambda \Sigma |\beta|$$

Where E is the objective function, RSS is the sum of squared residuals, and λ is a tuning parameter that controls the value of the penalty.

For this study, a logistic regression model with LASSO will fit the training set with 10-fold cross-validation to find the optimal λ. Once the optimal value of λ is calculated, it will be used to fit the model to the testing set for evaluating the performance.

3.5 Evaluation Metrics

It is common to use evaluation metrics to measure the performance of machine learning models. Since the task of this study is classification, accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC) will be used to evaluate the models' performance.

Accuracy is the proportion of correct predictions the model makes, which is equal to the number of correct predictions divided by the total number of predictions.

Precision is the proportion of true positives out of all the positive predictions, whereas recall is the proportion of true positives out of all the actual positives. Furthermore, the F1 score combines both precision and recall into a single value, given as:

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Therefore, based on the above formula, the F1 score measures the balance between precision and recall. In other words, a high F1 score means both precision and recall are high, indicating a good model performance. However, a low F1 score means either precision or recall is low, telling the model needs improvements.

AUC measures the model's ability to distinguish between positive and negative examples. Thus, it is calculated by calculating the area under the curve plotted by the true positive rate (sensitivity) against the false positive rate (1-specificity) at various thresholds.

In addition, the better the models perform, the higher the evaluation metrics used in this study.

The GitHub link for the code: https://github.com/changdali1207/DASC-5420-Final-Project

4. Results and Discussion

Table 1. Models' performance for the original and undersampled dataset.

Case	Model	Dataset	Validation Accuracy	Test Accuracy	AUC	Positive (YES)			Negative (No)		
						Precision	Recall	F1 Score	Precision	Recall	F1 Score
1	Decision Tree	Original	0.9145	0.915	0.645	0.553	0.039	0.073	0.917	0.997	0.955
2		Undersampled	0.6807	0.710	0.747	0.735	0.657	0.694	0.690	0.763	0.725
3	Logistic Regression	Original	0.9159	0.916	0.841	0.548	0.105	0.176	0.922	0.992	0.956
4		Undersampled	0.7671	0.763	0.841	0.756	0.778	0.767	0.771	0.749	0.760
5	LASSO	Original	0.9160	0.916	0.839	0.569	0.094	0.161	0.921	0.993	0.956
6		Undersampled	0.7669	0.763	0.841	0.756	0.776	0.766	0.770	0.750	0.760

As mentioned in the dataset section, the dataset is highly imbalanced. Therefore, all the models are fitted to original and balanced datasets to see how data imbalance impacts the models' performance. Since three models are used in this study, there are six different cases. Table 1 shows all the values of evaluation metrics of three models for two different scenarios.

4.1 Overview

As for the performance of the decision tree, the model has a higher test accuracy is 91.5% for the original dataset, indicating that the model correctly classifies most of the observations of the testing set. However, the AUC is 0.645, so the model is not good at distinguishing the two classes. Moreover, looking at the positive (Yes) class, fitting the model to the undersampled dataset could significantly improve the model's performance in terms of precision (0.735 vs. 0.553), recall (0.657 vs. 0.039), and F1 score (0.694 vs. 0.039). As a result, the model that is trained on the undersample dataset is better at identifying the positive instances.

Furthermore, as for the negative class, the model trained on the original dataset performs better in terms of precision (0.917 vs. 0.69), recall (0.997 vs. 0.763), and F1 score (0.955 vs. 0.725). Similarly, the result suggests that the model fitted to the original dataset is better at identifying negative instances.

Overall, the choice between the two models is depended on our specific goals. In other words, if our primary goal is correctly identifying the positive (Yes) class, we should use the model trained on the undersampled dataset. On the other hand, if our goal is to identify the negative (No) class, we might need to use the model trained on the original dataset. In addition, the model trained on the undersampled dataset has a higher AUC and roughly equal F1 scores for both classes; in other words, this model is equally effective at identifying both classes. Thus, this model is ideal although it has relatively lower test accuracy.

The performance of the logistic regression and LASSO is similar to that of the decision tree. For the original data, they have higher test accuracy (91.6%), indicating they also correctly classify most of the test cases. Moreover, one notable thing is that the models have almost the same AUC for both the original and undersampled datasets, which means the models are equally effective at distinguishing between negative and positive classes for both scenarios.

In addition, similarly, the models trained on the original dataset are better at identifying and classifying negative instances. In contrast, the models trained on the undersample dataset better identify and classify positive samples. However, table 1 also shows that the models trained on the undersampled data have a balanced performance for both classes. In addition, since they have a similar AUC, the models trained undersampled dataset should be more desirable than the model trained on the original dataset.

4.2 Comparison

4.2.1 Logistic Regression vs. LASSO

Comparing the results of cases 3 and 4 with that of cases 5 and 6, I did not see a significant improvement. In other words, for this dataset, using LASSO to remove unrelated variables does not help improve the models' performance for both original and undersampled datasets. As a result, I will compare the performance of the decision tree with logistic regression only.

4.2.2 Logistic Regression vs. Decision Tree

As for the original data, the logistic regression has a slightly better test accuracy (0.916 vs. 0.915). Furthermore, the logistics regression has a significantly higher AUC (0.841 vs. 0.645), indicating that it can better identify positive and negative classes. Speaking of the precision for the positive class, the logistic regression has a slightly lower value (0.548 vs. 0.553); on the other hand, it has a much better recall (0.105 vs. 0.039) and F1 score (0.176 vs. 0.073), indicating it

performs better in classifying positive instances. In addition, both methods result in comparable performance for the negative class.

As for the undersampled data, the logistic regression has a better test accuracy (0.763 vs. 0.710); moreover, its AUC is also higher (0.841 vs. 0.747) because of its better ability to distinguish the two classes. Furthermore, it has higher precision, recall, and F1 scores for both positive and negative classes.

Overall, the logistic regression outperforms the decision tree in most aspects; in other words, the logistic regression is more suitable for making predictions of this dataset.

5. Conclusion and Limitations

In conclusion, this study aimed to compare the performance of three machine learning algorithms, including decision trees, logistic regression, and LASSO regression, in predicting heart disease using a dataset containing various essential features. By comparing the performance of the models through evaluation metrics such as accuracy, AUC, precision, recall, and F1 score, we saw that logistic regression outperformed decision trees in most aspects. Moreover, the performance of LASSO and logistic regression were similar, which can be attributed to some data characteristics.

However, there are some limitations of the study. Although some efforts were made to tune the models' hyperparameters, there could still be room for improvement in this area, particularly for the decision tree model, which might have performed better with optimal tuning. In addition, this study only focused on comparing decision trees, logistic regression, and LASSO. Exploring other machine learning algorithms, such as random forests, support vector machines, or neural networks, could give us a deeper understanding of the best models for this dataset.

Therefore, for further research, I could deal with these limitations by optimizing model tuning and evaluating the performance of other machine learning algorithms.

Reference

1. World Health Organization. (2021). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Jordan LC, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, O'Flaherty M, Pandey A, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Spartano NL, Stokes A, Tirschwell DL, Tsao CW, Turakhia MP, VanWagner LB, Wilkins JT, Wong SS, Virani SS; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. Circulation. 2019 Mar 5;139(10):e56-e528.

3. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, Min JK, Tang WHW, Halperin JL, Narayan SM. Deep learning for cardiovascular medicine: a practical primer. Eur Heart J. 2019 Jul 1;40(25):2058-2073. doi: 10.1093/eurheartj/ehz056. PMID: 30815669; PMCID: PMC6600129.

4. Dataset link: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease