An Investigation Into Cardiovascular Diseases Using Classification Machine Learning Models

ISE/DSA 5103-001

Nasri Binsaleh, nasri.binsaleh-1@ou.edu

Mick Christensen, mickbc@ou.edu

Rithsek Ngem, rithsek.ngem-1@ou.edu

Onintsoa Ramananandroniaina, ony@ou.edu

EXECUTIVE SUMMARY

The leading cause of death worldwide is cardiovascular (heart) diseases. In this project, the probability of heart failure was modeled using 918 observations of 11 features linked to heart failure. Thus, the dataset was limited in number of observations and features. The amount of information was limited for a disease that affects millions of people annually, and the dataset was reduced by 20% to create a test set. Thus, the reliability of the models may be less than desired to be applied in practice.

A logistic regression model performed the best with an accuracy of 0.87 and a kappa of 0.735, closely followed by a support vector machine model and a random forest. Two MARS and AdaBoost models performed the worst with accuracies of 0.85 and kappas of 0.689. Hence, the recommendation would be to go forward with logistic regression models in the future.

BACKGROUND

Introduction

The leading cause of death worldwide is cardiovascular diseases. Each year, around 17.9 million lives are lost due to this type of disease (Khan, 2021). Heart failure is one of the most common events that occurs because of cardiovascular disease. The dataset in this project contains 11 features that could be used to predict the probability of heart failure that could lead to death. It is greatly beneficial to be able to detect the disease early on to prevent life threatening events related to cardiovascular disease.

The purpose of this paper is to use a heart failure dataset to assess the likelihood of an event that can be attributed to cardiovascular disease. The dataset is a combination of five datasets related to heart disease (Fedesoriano, 2021). In creating the dataset, all common features were extracted, which resulted in 11 features. These 11 features can be used to predict heart disease output class. Output class of 1 means a person has heart disease and output class of 0 means a person does not have heart disease.

Data description

The dataset contains 11 predictor variables, 6 numeric variables and 5 factor variables. The response variable is whether the patient has a heart disease (1) or not (0). The description of each variable can be seen below.

1. Age: age of the patient [years].
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

Initial exploratory data analysis

Figure 1 and Figure 2 below show the amount of the elements in the features to visualize how the data looks like in our dataset. One important observation is that in a variable called “Cholesterol”, the value of zero had been recorded the greatest number of times. So, the variable was explored more in depth. Additionally, the “FastingBS” variable was changed from a yes/no variable to a categorical variable.

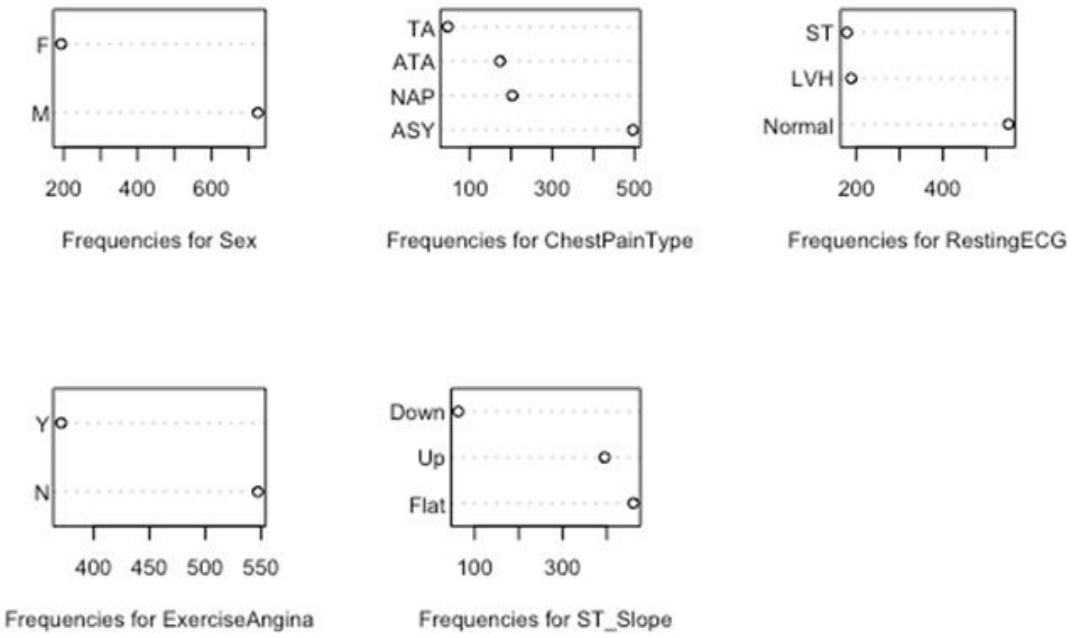

Figure 1: frequency plots of factors

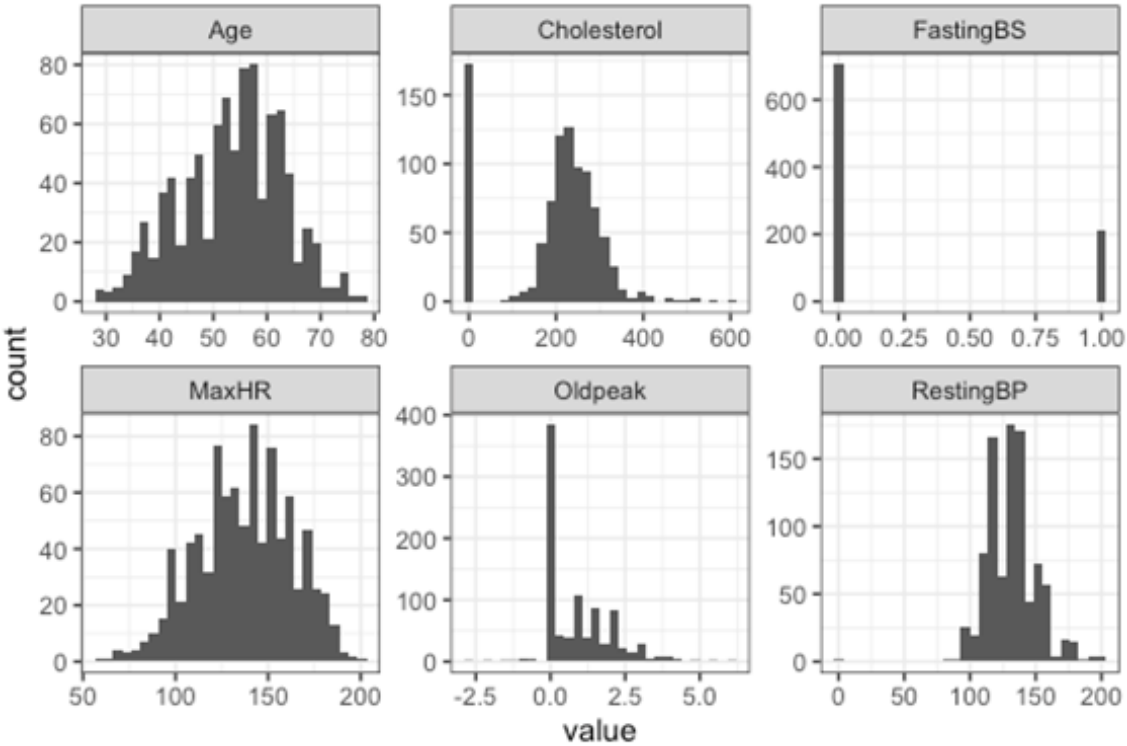

Figure 2: histograms of numeric variables

Figure 3 below shows a histogram of the “cholesterol” variable and Figure 4 shows the boxplot of “Age”, “Resting BP”, “Cholesterol”, and “Max HR” in observation with and without heart disease. The number of zeroes in the “Cholesterol” variable influenced the box plot greatly. It was concluded that it is not reasonable that humans can have zero cholesterol, and so number zero was treated as a missing value.

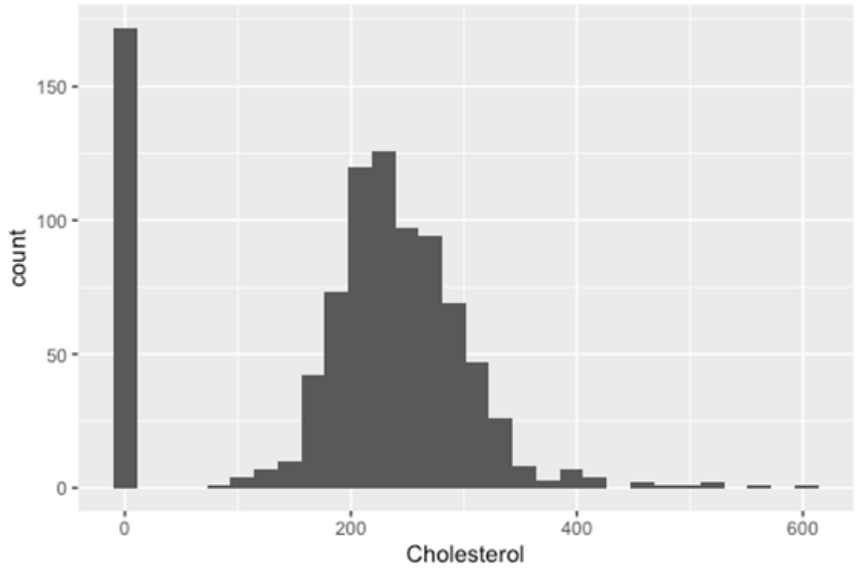

Figure 3: Plots for cholesterol VS its count

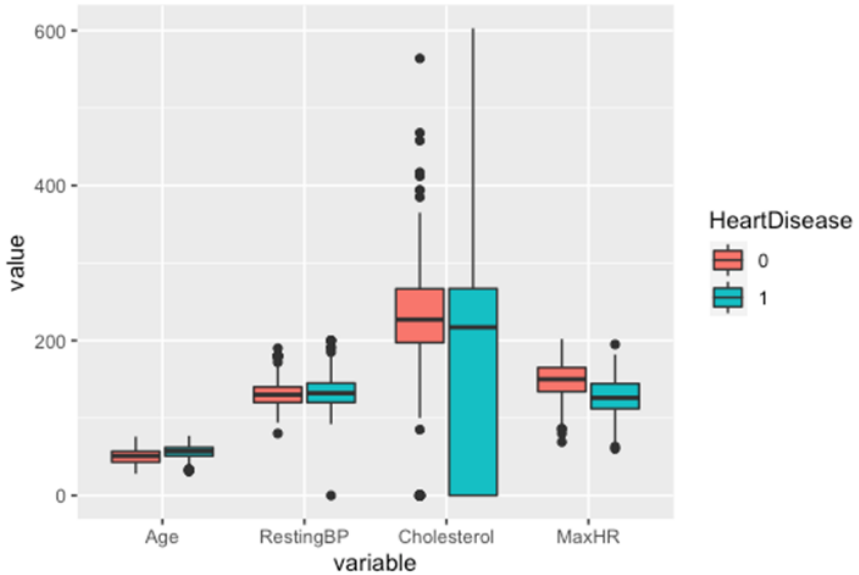

Figure 4: Box plots showing Age, Resting BP, Cholesterol, and Max HR in observation with and without heart disease.

Figure 5 below visualizes the number of observations with and without heart disease. It shows that about 400 observations are from people without heart disease, and about 500 observations are from people with heart disease. The amount between the two does not vary greatly, which allowed us to use accuracy as a measure of the performance of the model. Additionally, the kappa importance measure was used to complement accuracy.

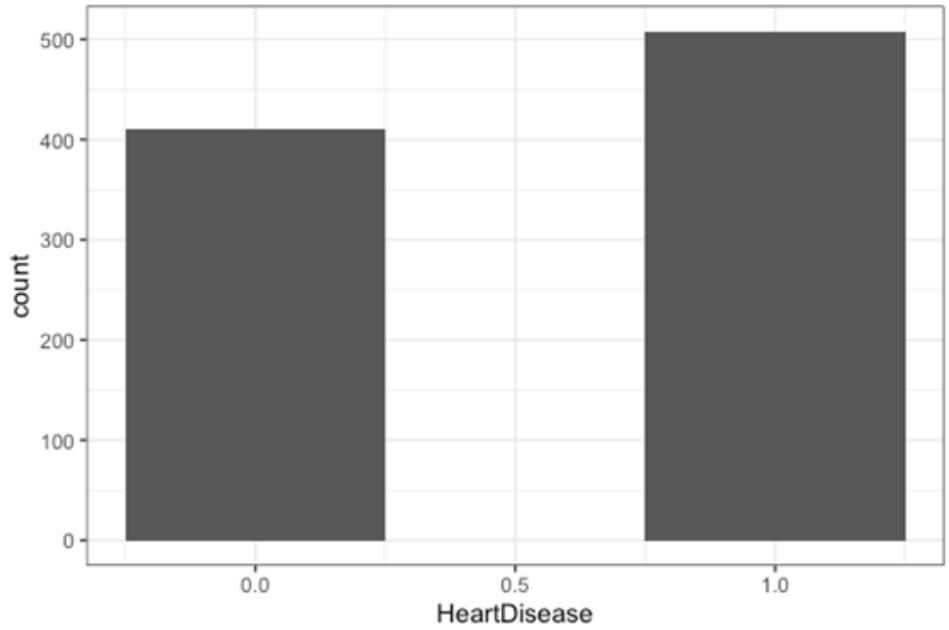

Figure 5: A bar chart comparing the amount of observation with and without heart disease.

METHODOLOGY

Data preparation

From the data quality report in Appendix A, it was found that there were no missing values in the data frame. However, some values were stored as zeroes for some variables such as "cholesterol" when they are likely to be missing. A patient had 0 resting blood pressure, which is impossible for a living human. Because there was just the one missing value, mode imputation was applied. For cholesterol, logistic regression was used for imputing the missing

values using the MICE package. Additionally, after testing for skewness, it did not appear to be significant and to warrant further investigation.

The dataset only had factors with a maximum of 4 levels, therefore, factor levels were not collapsed to a lower number of levels.

Feature Engineering

The recommended maximum heart rate for a person of a certain age is:

$$\max\ \_heart\_rate = \ 220 - patient\ age$$

Based on this, another feature was engineered that indicates whether the person has a maximum measured heart rate exceeding 90% of the recommended maximum heart rate.

Description of modeling approach

Our proposed solution was to use classification methods to optimize the models that give an accurate prediction of cardiovascular disease. Five classification modeling approaches were used in this project and the one that gave the most accurate result was chosen. Those models are Logistic Regression, Random Forest, Support Vector Machines (SVM), Multivariate Adaptive Regression Splines (MARS), and an AdaBoost Tree Ensemble. The classification model is a great choice because the main purpose of the project is to predict whether a person has cardiovascular disease or not.

To reduce the overfitting of the models, 5-fold cross validation was used to tune the models. The author of the dataset did not provide another test set to test the accuracy of the created models. Therefore, 20% of the data was randomly selected to be a test set onto which the cross-validated models that were built on the remaining 80% were applied. After attempting to

use 70% of the data to train and 30% to test, it was decided that the models using 20% test and 80% train performed better.

RESULTS

					Train/Test CV Performance	
Model	**Method**	**Package**	**Hyper parameter**	**Selection**	**Accuracy**	**Kappa**
Logistic Regression	glm	caret	N/A	N/A	0.851/ 0.870	0.696/ 0.735
Random Forest	rf	caret	mtry	2	0.861/ 0.859	0.717/ 0.712
SVM	svmPoly	caret	sigma, C	0.041, 4	0.861/ 0.864	0.717/ 0.723
Mars	earth	caret	nprune degree	7, 1	0.862/ 0.848	0.719/ 0.689
AdaBoost	adaboos	fastAdaboost	nIter, method	15, Adaboost.M1	0.835/ 0.848	0.663/ 0.689

Table 1: Results of five models

The logistic regression model performed the best based on the accuracy and kappa measures, closely followed by the support vector machine model. The AdaBoost model proved to perform the worst alongside the MARS model on the test data.

CONCLUSION

The purpose of this paper was to assess the likelihood of a cardiovascular event by applying classification models. Logistic regression, random forests, support vector machines, Mars, and AdaBoost models were created, and the highest performers based on accuracy and kappa values were selected. The logistic regression model ended up being the most accurate and had

the greatest kappa value of all the models, so the recommendation would be to go forward with that model in the future. However, the dataset was limited in number of observations and features. With only 918 observations of 11 features, the amount of information was limited for a disease that affects millions of people annually. Additionally, the dataset was reduced by 20% to create a test set, which made the training set that much smaller. Thus, the reliability of the models may be less than desired to be applied in practice.

REFERENCES

Fedesoriano. (2021, September 10). *Heart Failure Prediction Dataset*. Kaggle. Retrieved October 22, 2021, from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Khan, T. (n.d.). *Cardiovascular diseases*. World Health Organization. Retrieved December 16, 2021, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Nicholson, C. (2021). Intelligent Data Analysis [R code]. Retrieved from The University of Oklahoma DSA/ISE 5103 Canvas.

APPENDICES

Appendix A

Numeric variables:

variable	N	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
Age	918	0	0	50	5.447	53.511	28.0	47	54.0	60.0	77.0	9.433
RestingBP	918	0	0	67	7.298	132.397	0.0	120	130.0	140.0	200.0	18.514
Cholesterol	918	0	0	222	24.183	198.800	0.0	173	223.0	267.0	603.0	109.384
FastingBS	918	0	0	2	0.218	0.233	0.0	0	0.0	0.0	1.0	0.423
MaxHR	918	0	0	119	12.963	136.809	60.0	120	138.0	156.0	202.0	25.460
Oldpeak	918	0	0	53	5.773	0.887	-2.6	0	0.6	1.5	6.2	1.067
HeartDisease	918	0	0	2	0.218	0.553	0.0	0	1.0	1.0	1.0	0.497

Factor variables:

Variable	N	missing	missing_pct	unique	unique_pct	freqRatio	1st mode	1st mode freq	2nd mode	2nd mode freq	least common	least common freq
Sex	918	0	0	2	0.218	3.76	M	725	F	193	F	193
ChestPainType	918	0	0	4	0.436	2.44	ASY	496	NAP	203	TA	46
RestingECG	918	0	0	3	0.327	2.94	Normal	552	LVH	188	ST	178
ExerciseAngina	918	0	0	2	0.218	1.47	N	547	Y	371	Y	371
ST_Slope	918	0	0	3	0.327	1.16	Flat	460	Up	395	Down	63