A Multi-Layer Perceptron Model with Random Forest Checking Feature Importance for Determining Cardiovascular Disease Risk in Patients

Luke Innarelli
Mountain Ridge High School
22800 N 67th Avenue Glendale, AZ, 85310

Abstract:

Cardiovascular disease is the leading cause of death in first-world countries. It's identifiable and oftentimes preventable using basic medical data on each patient. This machine learning study attempts to improve easy access to knowledge about the risk of cardiovascular disease to help patients be prepared for possible future options in their life plans. The model used, based on a multi-layer perceptron classifier, would use a publicly available dataset of 303 patients with basic medical information such as age, sex, chest pain, cholesterol levels, and others, as well as already reported risks of cardiovascular disease. Based on its findings it would then determine the patients' risk of cardiovascular disease using a 1 as high risk and 0 as low risk with an accuracy rate of 86%. This approach can be applied to help improve not only the efficiency inside of a hospital but also provide patients with greater access to vital information.

Keywords:

Random Forest, multi-layer perceptron classifier, Cardiovascular disease, Permutation importance

Introduction:

Cardiovascular disease has been the leading cause of death globally, and every year, there are around 17.9 million deaths due to the many different forms of this disease [1]. However, a recent study has shown that over half of the US adult population doesn't recognize cardiovascular disease as the leading cause of death [2]. If the general populace were able to gain an automated process for receiving vital data about their risk of heart disease, they could shape future life choices.

This research utilized the health data of 303 patients categorized into different risks of heart disease in order to build a model that could determine this risk on new patients. Using either a multi-layer perceptron classifier or a random forest classifier variations in hidden layer sizes and learning_rate_init was made in order to build a more accurate model in order to do so.

Methodology:

This work used the dataset of 303 patients with known high-risk or low-risk of heart disease [3]. This dataset included things like the patient's age, sex, resting ECG, and chest pain, among others. All of this data was already classified into a numerical system with numbers meaning different things for each piece of data. For example, chest pain was classified into 1-4, with 1 being typical angina, 2 being atypical angina, 3 being non-anginal pain, and 4 being asymptomatic. The dataset also included a split of 165 high-risk patients to 139 low-risk patients. We split this data into 90% to 10% for training and testing, respectively.

Results:

We ran this model multiple times using Python 3.12 with PyCharm with differing rates of learning_rate_init and changing hidden layer sizes, as well as the number of sizes in order to try

to find a more optimized version of the model. Afterward, we ran the same process but instead used a random forest classifier to determine the importance of each data point using feature importance to determine the relative importance of each feature. Figure 1 demonstrates the relative importance of each of the features, with caa (number of major vessels) being the most important and fbs (fasting blood sugar levels) being the least important.

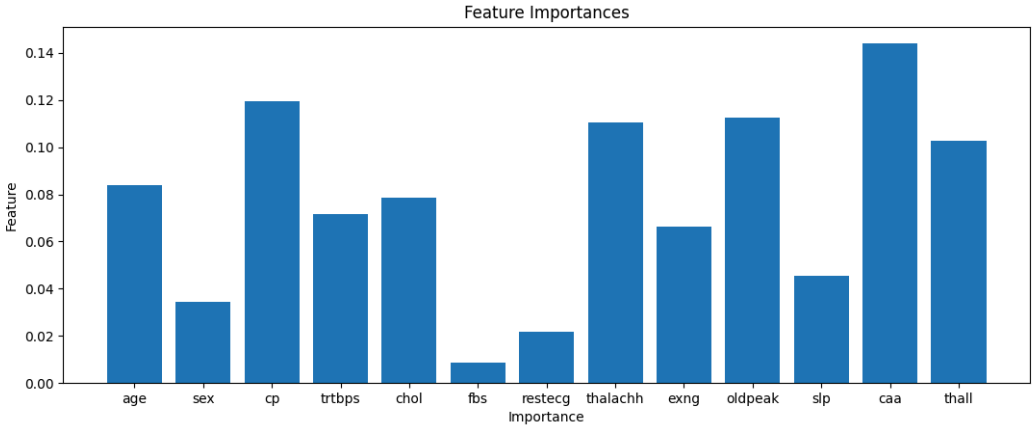

Figure 1. An overview of the different importances of features based on the random forest classifier

We used permutation importance on the original model built by the multi-layer perceptron instead of using the random forest built-in to gain a more accurate view of the importance of every feature, as shown in Figure 2 below. This Figure, in comparison with the previous one, agrees that fbs is the least important but also believes that resting ECG and slp (slope) are equally unimportant. While these two figures agree that caa is most important, however, Figure 2 shows that it has a wide range of importance depending on the patient, unlike Figure 1 showing it is most important on average in every case.

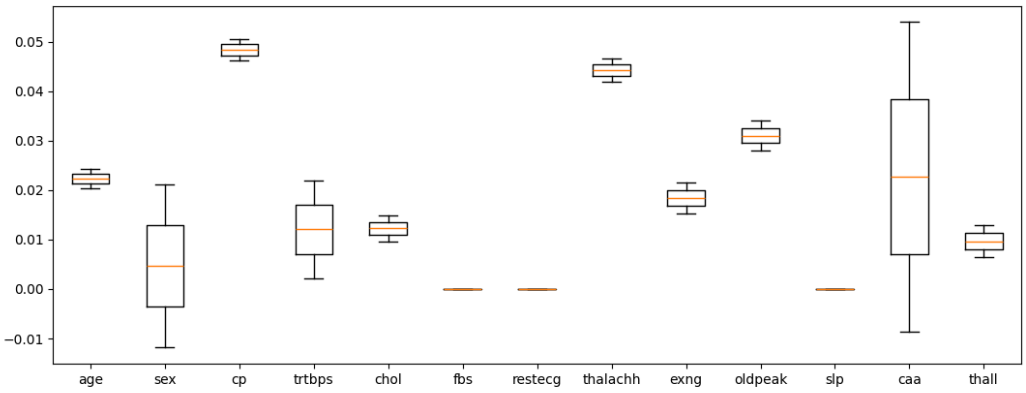

Figure 2. An overview of the different importance of features using permutation importance.

In the end, our final model gained an 86.397% accuracy and showed promise to be able to be used in the future. Figure 3 represents the different models produced over time and their accuracies.

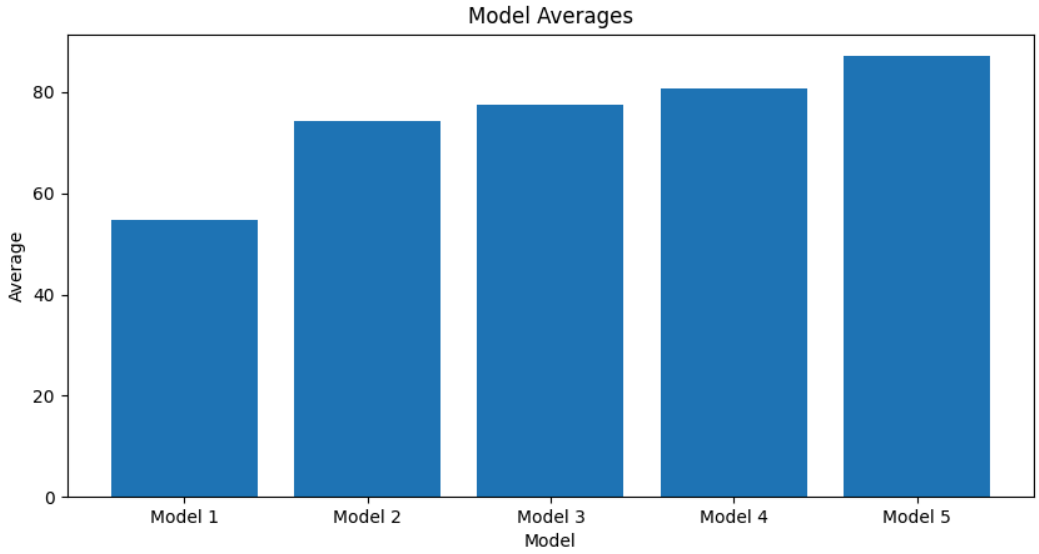

Figure 3. The averages of the accuracies of the models in consecutive order

Discussion:

This research successfully developed a high-accuracy model for determining the risk of cardiovascular disease within patients. It also determined the relative importance of the features used inside of these models and presents a possible chance at improving said model by using these importances to either focus on or cut out certain features that may limit or improve patient accessibility.

In a similar study [4] performed by Madhumita Pal, Smita Parija, Ganapati Panda, Kuldeep Dhama, and Ranjan K Mohapatra, where they trained a model to detect high and low cardiovascular disease risk they discovered similar findings. They were able to gain a similar accuracy of 86.9% for their MLP model, and while some features helped determine cardiovascular disease risk better than others, most of them remained important in discovering high or low risk.

The main limitation of this research is the low amount of data to work with, only having a sample size of roughly 300 patients to train a model for potentially millions of patients. This data is often hard to obtain and may affect patient privacy, so available data is very limited. While the output could also have been a limiting factor as it only had a high or low option to risk, there was no medium risk and no percentages to this risk, meaning the model could never tell the middle ground for patients.

If we were to further this research, a greater data size with a greater diversity could potentially help increase the accuracy, while testing for more ideas on permutation importance could help pinpoint the specific features that we should focus on to train the models.

Conclusion:

Using the multi-layer perceptron model, we were able to produce a model that would be able to determine a patient's risk of heart disease by looking at their medical data with an 86.397% accuracy for the multi-layer perceptron model and 83.871% for the random forest model. This model shows promise for the future, and using permutation importance, we were able to find out that the fbs, or fasting blood sugar levels, are relatively unimportant when

determining the risk of heart disease. With this new knowledge and hopefully improvements in the future, we can reduce the amount of information needed for the model to produce an accurate result, increasing accessibility even possibly to the common household, barring some simple equipment.

Acknowledgments:

Efthimios Gianitsos

References:

[1] “Cardiovascular Diseases.” *World Health Organization*, World Health Organization, www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. Accessed 3 Oct. 2024.

[2] “More than Half of U.S. Adults Don’t Know Heart Disease Is Leading Cause of Death, despite 100-Year Reign.” *American Heart Association*, newsroom.heart.org/news/more-than-half-of-u-s-adults-dont-know-heart-disease-is-leading-cause-of-death-despite-100-year-reign. Accessed 3 Oct. 2024.

[3] Heart Attack Analysis & Prediction Dataset (2020) Rashik Rahman, https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/discussion?sort=hotness

[4] Risk prediction of cardiovascular disease using machine learning classifiers (2022) Madhumita Pal, Smita Parija, Ganapati Panda, Kuldeep Dhama, and Ranjan K Mohapatra https://pmc.ncbi.nlm.nih.gov/articles/PMC9206502/