Model Comparison for Cardiovascular Disease Prediction

## Daniel Theng MATH 191T

## December 2, 2020

**Contents**

1. [Introduction 3](#_TOC_250003)
2. [Methods and Results 4](#_TOC_250002)
3. [Discussion 5](#_TOC_250001)
4. [References 6](#_TOC_250000)

# Introduction

Globally, cardiovascular diseases and the number 1 cause of death which means more people die from cardiovascular disease than any other singular cause [4]. Cardiovascular diseases refer to any disorders that pertain to the heart or the blood vessels, which include various diseases where the blood vessels constrict limiting blood to certain areas such as the heart, appendages, or brain (coronary heart disease, cerebrovascular disease, and peripheral artery disease) or could manifest in a malfunction in any other part of the cardiovascular system [4]. Acute effects from cardiovascular diseases are heart attacks and strokes where there is a blockage of blood flow to the heart or brain [4]. The common mechanism is fatty deposits that narrow the vessels restricting blood flow to the vital organ [4]. Another cause is for a fatty build to occur elsewhere, but becomes dislodged and moves to the heart or brain where it causes the blockage. It is agreed that there isn’t a singular cause of cardiovascular diseases rather it presents itself with the occurrence of a wide range of risk factors which include tobacco use, diet, physical activity, obesity, excessive alcohol use, diabetes, hypertension, and hyperlipidaemia (increased fat concentration in the blood stream) [4].

Because of the severity and global nature of cardiovascular disease naturally physicians want better ways to predict cardiovascular disease. This fact compounded with the overwhelming complexity and volume of risk factors the usage of computers to predict cardiovascular disease is promising and many papers have used several programs and models to predict cardiovascular disease and new predictors are constantly being discovered. This project implements 3 very common multipurpose models, Logistic Regression, Random Forest Classifier, and Support Vector Machine. The data used contains 11 features: age, gender, height, weight, systolic and diastolic blood pressure, cholesterol levels, blood glucose level, smoking habits, alcohol consumption, and physical activity. All of these features are the simplest risk factors. The objective of this project is to accomplish the following:

* Find any general, easily observable trends between the predictors and the presentation of cardiovascular disease.
* Determine if different classification models can accurately predict the presence of cardiovascular disease given the data original predictors and engineered predictors.
* If this is the case then of 3 models, Logistic Regression Classifier, Random Forest Classifier, and Support Vector Machine, which one performs the best.
* Lastly, which predictive features are the most important to the models.

The entire project was done in a Jupyter notebook where the data was loaded. A pairplot was used to explore the data and take a quick look at what the data offers. Once all the models are created, the data will be trained and scored. The time to fit and score the models as well as the scores will be recorded to help compare the models to each other. This process is then repeated after engineering features and then one final time after removal of low contributing predictors.

While other studies use more complicated models and cutting edge predictors, this project plans to use accessible classifiers in tandem with common risk factors as predictors, to conclude which of these commonplace models and predictors are the best.

# Methods and Results

The first to take a quick look at the data and identify any clear trends. The data contains 12 features: age, gender, weight, height, systolic(ap\_hi) and diastolic blood pressure(ap\_lo), cholesterol levels, glucose levels, smoking habits, alcohol consumption habits, physical activity habits, and the presence of cardio vascular disease. The latter being the target variable. The data contains 70000 patients, with 34979 patients having cardiovascular disease, which is 49.71% of the data.

This was done using the seaborn package’s pairplot function. Moreover because of the number of features, a pairplot with all the variables would have made each individual plot too small to be usable, therefore each of the variables were split into 3 group, of which a pairplot of each group was made. The variables were split by their feature type, objective, examination, and subjective. Age, gender, height and weight are the objective variables. Systolic and diastolic blood pressure, cholesterol and glucose levels are examination variables. Smoking, alcohol consumption, and physical activity are the subjective variables. Each pairplot is shown below.

Chart, calendar

Description automatically generated

**Figure 1. Pairplot of Objective Variables; age, gender, height, weight**

A picture containing text, white, bath

Description automatically generated

**Figure 2. Pairplot of Examination Variables: systolic and diastolic blood pressure, and cholesterol and glucose levels.**

A picture containing calendar

Description automatically generated

**Figure 3. Pairplot of Subjective Variables: smoking, alcohol consumption, and physical activity.**

The part was to create the 3 models and fit them using features given in the data. Using scikit-learn’s train\_test\_split function, the data was split into training and test data. Then the models were fit and scored. The length of time to fit and score each model was recorded as well as the scores. Time to train, time to score, and score for each model can be found in Table 1 listed as Fit Time 1, Score Time 1, and Score 1 respectively.

The data allowed for some new features. Specifically, obesity and hypertension are often correlated with cardiovascular disease [2][3]. BMI has shown to be a useful tool in gauging an individuals body weight [2]. BMI could be calculated from the height and weights of each patient by the equation:

From the BMI feature, weights classes could easily be assigned to each patient. BMI from under 18.5 considered underweight, 18.5 to 25 was considered normal, 25 to 30 considered overweight, 30 to 35 considered first level obesity, 35 to 40 considered second level obesity, 40 to 45 considered the final level of obesity

Hypertension or high blood pressure is usually diagnosed if blood pressure is above 140/90, so if a patient had a systolic pressure at least 140 or a diastolic pressure of at least 90 then they were assigned to the high blood pressure group [3]. If a patient had a systolic pressure of at least 180 or a diastolic pressure of at least 120, then they were assigned to the severe hypertension group.

The models were refit with the new data and scored. Again the length of time to fit and score each model was recorded with the scores. Time to train, time to score, and score for each model this second time can be found in Table 1 listed as Fit Time 2, Score Time 2, and Score 2 respectively.

Lastly, feature importance for the Logistic Regression and Random Forest classifiers were graphed. Although the Support Vector Machine has a feature coefficient as well, they can only be calculated if the model is linearly separable, if the data can be separated with a single line which this model is not. Bar graphs of the feature importance for Logistic Regression and Random Forest are shown below.

Chart, waterfall chart

Description automatically generated

**Figure 4. Shows the feature coefficients for each feature.**

Chart, histogram

Description automatically generated

**Figure 5. Shows the feature importances for each feature.**

Finally, removing the least contributing features; glucose levels, smoking, alco, and weight status the models were fitted scored one last time. Again the length of time to fit and score with the scores were recorded. Time to train, time to score, and score for each model this third time can be found in Table 1 listed as Fit Time 3, Score Time 3, and Score 3 respectively.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic Regression** | **Random Forest** | **SVM** |
| **Fit Time 1 (sec)** | 0.929679871 | 7.91764617 | 119.530364 |
| **Score Time 1 (sec)** | 0.057277918 | 1.279073954 | 22.80775595 |
| **Score 1** | 0.691142857 | 0.714904762 | 0.604190476 |
| **Fit Time 2 (sec)** | 1.448005199 | 8.550495863 | 154.075804 |
| **Score Time 2 (sec)** | 0.038278818 | 1.265856028 | 24.91919398 |
| **Score 2** | 0.705333333 | 0.712333333 | 0.603571429 |
| **Fit Time 3 (sec)** | 1.3947389125823975 | 8.577803134918213 | NA |
| **Score Time 3 (sec)** | 0.027863025665283203 | 0.7556169033050537 | NA |
| **Score 3** | 0.7053333333333334 | 0.705047619047619 | NA |

**Table 1: Shows the times and scores for each model, each trial. Note the NA values for SVM for the third trial because feature importance could not be calculated.**

There are some notable values in this table. For example, from a simply eye test removing 4 features didn’t significantly decrease or increase the fit time, score time or score. On the other hand, the addition of 3 features (BMI, Weight Status, and hypertension) did seem to increase the fit time by quite a bit. The percent increase between the Fit Time 1 and Fit Time 2 for the Logistic Regression and Random Forest models are quite large (roughly 56% and 8% respectively). While the fit times between 1 and 2 the score times seemed to stay fairly consistent and the same can be said for the accuracy. For the Logistic Regression model, it benefitted from the extra features and didn’t lose accuracy from the removal of the least significant features. Conversely, the extra features were a slight detriment to the Random Forest model and the removal of the least significant features did impact its accuracy, decreasing it by about 1%.

# Discussion

Examining the seaborn pairplots not a lot of trends are easily observed. Nearly every single one of the plots the points overlap each other and there are no clear separations. The only exception is the age distribution, where patients with cardiovascular disease tend to be a little older than healthy individuals. This makes sense, because as established the relationship between cardiovascular disease and its risk factors is very complex so some scatter plots won’t very useful.

Looking at Table 1, it is clear that the Support Vector Machine is the worst classifier to use, not only did it take the longest to fit, but was the least accurate by 10% and didn’t improve with the created features. The scores between the Logistic Regression classifier and Random Forest were very close, but the Random Forest model didn’t improve with the addition of the created features. Overall the accuracy of these two models is about 70%, which is quite low for diagnostic tests for cardiovascular disease.

Inspecting at Figure 4 the feature importances are quite confusing. BMI was the most influential feature but the BMI coefficient thus the feature predicted the patient was health, which seems counterintuitive. There is other coefficient that don’t seem to add up. For instance height and weight are positively correlated, but their coefficients point in the opposite directions. Systolic blood pressure contributes to a positive result, but diastolic blood pressure barely contributes to any direction.

Feature importance for the Random forest model is much clearer. Here it seems that the most basic and general predictors are the most important such as age, height, weight, blood pressure, and BMI. Much of the categorical variables were not as important to the models. It seems that Random Forest places much more important for features that have a wide range of possible values. All of the binary categorical variables did worse than the tertiary categorical variables. This reflects a weakness in this model, because smokers, alcohol consumption, and physical activity often are strongly correlated with the other variables. For example, a very active individual is more likely to have a higher BMI because the index doesn’t differentiate between muscle mass or fat mass, but this model would take that high BMI value as very important without the context of the physical activity.

There are a couple reasons why my model specifically will be not be found in every hospital in the country. Other than the fact that it just isn’t accurate enough, the models can’t tell what form of cardiovascular disease is present. Treatment options depend on how the cardiovascular disease presents itself, but these models don’t discriminate between them which decreases their usefulness. Most diagnostic tools focus on diagnosing issues with specific parts of the cardiovascular system, because that information is much more useful for physicians.

# References

[1] Downey, A. (2016). *Think Python*. O'Reilly Media.

[2] Kannal, W. B., D'Agostino, R. B., & Cobb, J. L. (1996). Effect of weight on cardiovascular disease. *The American Journal of Clinical Nutrition*, *63*(3), 419–422. https://doi.org/https://doi.org/10.1093/ajcn/63.3.419

[3] Sowers, J. R., Epstein, M., & Frohlich, E. D. (2001). Diabetes, Hypertension, and Cardiovascular Disease. *Hypertension*, *37*(4), 1053–1059. https://doi.org/https://doi.org/10.1161/01.HYP.37.4.1053

[4] World Health Organization. (2017, May 17). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).