A Predictive Study of Heart Attack Triggers Based on Machine Learning

Zijie Feng *

University of California San Diego, La Jolla, 92037, United States

* Corresponding Author Email: zifeng@ucsd.edu

Abstract. With increasing health concerns in society and the enormous surge in digital data generated from technological advancements, the need for efficient and accurate methods of value extraction from data is more evident than ever. Thus, such methods are developed to help medical professionals make thoughtful decisions. This paper explores a piece of health data from the Centers for Disease Control and Prevention (CDC) to find the most influential factors contributing to the occurrences of heart attacks and the best algorithm to help predict such occurrences. To achieve that goal, this paper employed algorithms including Support Vector Machine (SVM), Random Forest, and Logistic Regression, along with a technique called permutation feature importance to identify important features. In the end, Random Forest was found to be the most powerful and consistent algorithm to complete the task of prediction in this case, while also showing that average sleep hours, BMI, and mental health status of individuals can be potential indicators of heart attacks. With this result, this paper marks significant features and algorithms that can be subjects of further and more extensive studies, and eventually be applied to the work of real medical professionals.

Keywords: Heart attack, machine learning, feature importance, random forest.

1. Introduction

Health concerns are on the rise. With the recent pandemic of COVID-19, people started examining their bodies more and more closely just to make sure they can stay healthy. COVID-19 can cause plenty of dramatic changes to a patient's physical health and can invoke many different sorts of diseases. Among all organs, the heart is the most crucial. However, recent research has shown that COVID-19 can deteriorate heart performance in ways including but not limited to inflammation of the heart, arrhythmias, and heart failures [1]. Such complications have influenced both the younger and older generations [1], making everybody in society aware of the significant problems the heart can cause. Besides the impacts of the pandemic, human beings in general have been conducting fewer and fewer physical activities in our daily lives. With the great advancements in technology, jobs that rely entirely on human labor have been fading away, which led to a significantly reduced amount of physical activity performed by people. As a potential consequence, such a situation might have caused many to have health conditions that expose them to cardiac problems. Studies were done to confirm the strong and consistent correlation between continuous daily physical exercises and low risk of heart attack [2]. There are many other studies just like this one, looking for potential reasons why people have heart attacks, or why some people are unlikely to have heart attacks.

Nowadays, the amount of data regarding the health of people is enormous. If data scientists can use proper methods and tools to extract valuable insights from such data, then potential health problems that require immediate attention can be accurately detected. For that purpose, machine learning is exactly needed. With various well-designed models for different situations, such methods can be easily applied to a large amount of health data. Researchers have already tested different models and determined some of the best-performing ones in the case of cardiac health, such as Random Forest, which had a phenomenal heart disease classification accuracy rate of 92.44% measured in one study [3].

This paper would like to explore the significant predictors of a possible heart attack and what kind of models are the fittest for such topics. This paper will go over the details of all parts of the study, including data property introduction, data cleaning and preprocessing, model introduction, and

performance evaluation across models, with data visualization all along to help readers better understand the data and results. Finally, the study will also dive into the limitations of this study and what potential paths researchers can take in the future to advance the obtained results. It is hoped that the research presented in this paper can add more insights to the field of cardiac health analysis.

2. Method and Data

2.1. Data

The data this paper will be exploring originally comes from the Centers for Disease Control and Prevention(CDC) and is a part of the CDC's Behavioral Risk Factor Surveillance System(BRFSS). The BRFSS continues to collect health information by interviewing over 400,000 adults in all U.S. territories every year. The specific dataset this paper deals with contains information collected for the year 2022, with 40 features and 445,132 rows, each row representing a single individual's attributes. The specific qualities of each feature will be discussed in the following section.

2.2. Data preprocessing

As the size of the dataset is quite huge and computation power is limited, it makes sense to drop some rows with null values to reduce the size. After dropping all rows with null values, the dataset's dimension has been reduced to (246022, 40). This should make all the following actions made to the dataset completed quicker.

Now, all features in the dataset should be inspected so it can be determined whether they are a good fit in the possible models that are going to be deployed. The first feature that needs to be examined would be the target variable, which would be whether the individual has had a heart attack before, with a column name "HadHeartAttack" and values of "Yes" and "No". In a classification problem like this, it is important to check the ratio between different labels in the dataset.

As shown in Figure 1, it is clear that this data set has a very unbalanced distribution: "No" is about 17 times the "Yes". There are many different ways to deal with an unbalanced dataset. Common methods include Data Level Methods, Over-Sampling Techniques, SMOTE, Cluster-Based Over-Sampling, and Under-Sampling Techniques [4]. For this dataset, it is appropriate to use the Under-Sampling Technique, which will sample a portion of the majority class and keep the minority class unchanged. To further reduce the size of the dataset, the technique was adjusted in this study: instead of keeping all 13435 minority labels and having the same number of majority labels, 5000 data points from each label are randomly selected to make the new dataset. Now the dataset only has 10,000 rows and an equal number of "Yes" and "No" labels, as shown in Figure 2.

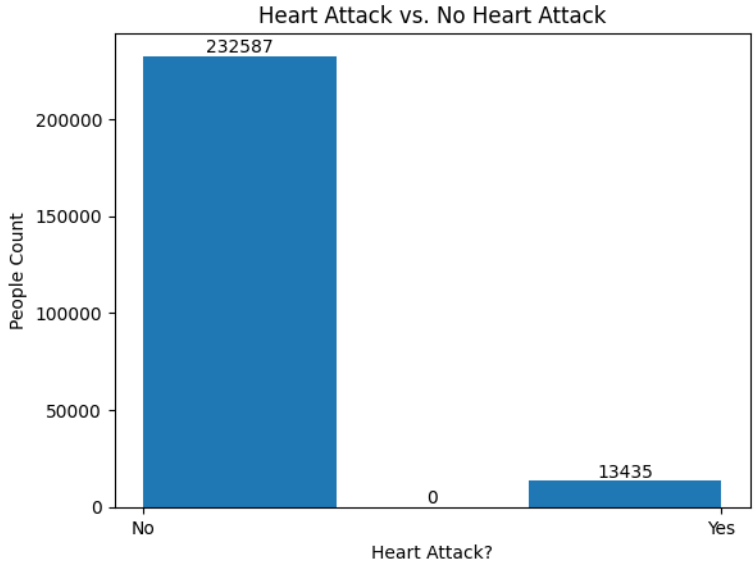

Figure 1. Heart Attack Count vs. Non-Heart-Attack Count in Dataset

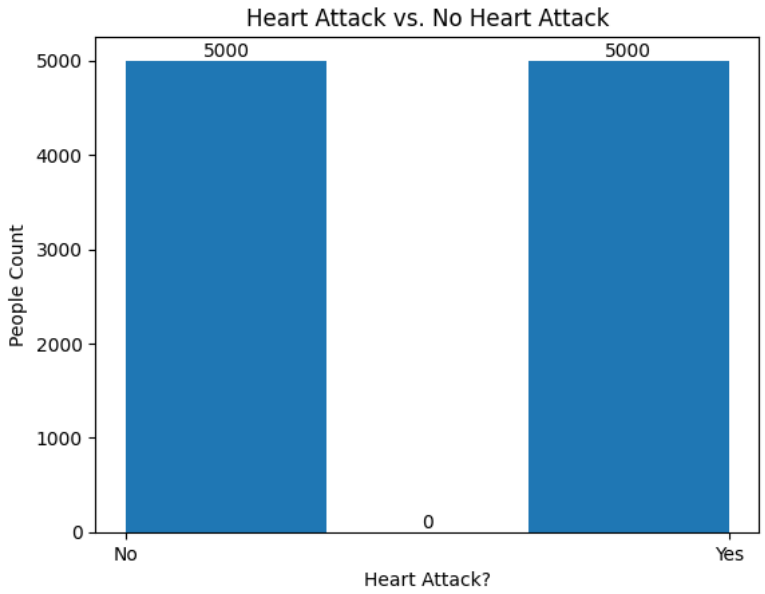

Figure 2. Heart Attack Count vs. Non-Heart-Attack Count in Modified Dataset

Moving on to the rest of the features, there are 3 kinds of features in the dataset: numeric, ordinal, and nominal, and they all need to be transformed into a better format that's easier to work with.

Regarding the numeric features, a way of standardization is needed because different numeric features might have different ranges of numbers and thus their importance in the model can vary greatly and unintentionally, causing misleading final results [5]. Among different methods for data standardization, a standard scaler is a very common and useful tool. One study found the standard scaler to be the best standardization method for the auto-detection of epileptic seizure events, which is also a health problem that this paper deals with [6]. Thus, a standard scaler was used to transform numeric values into how far away in terms of standard deviation units from the mean of each feature.

Regarding ordinal features, they can be simply ordered into different ranks within each feature, ranging from 0 to however many more levels they need.

Regarding nominal features, there is no reasonable order that the values could be put into so One-Hot encoding will be needed to expand these features into more columns, with each column representing a certain value in the original feature. One thing to note here is that some of the nominal features are binary features meaning they only have 2 possible values: "Yes" and "No". Converting these features into pairs might lead to issues such as collinearity. This arises from the direct prediction of a 1 in one column when there is a corresponding 0 in another. Such a transformation can introduce problematic correlations between the features, thereby affecting the integrity of the data analysis. However, the program can be instructed to delete one of the columns in each pair to avoid such a problem. Therefore, all nominal features have now been turned into columns of 1's and 0's.

Now all features are ready to be calculated and selected using feature selection methods. Before using algorithms to determine which features to retain, some features need to be removed manually for predetermined reasons. Features like "State", "Sex", and "RaceEthnicityCategory" are removed because they might cast unwanted bias to the model and lead to false foundations of predictions since these attributes should not affect one's health. Other features like "LastCheckupTime", "ChestScan", and "HighRiskLastYear" are removed because they might be directly related to patients who had heart attacks, thus causing collinearity. Also, other features that indicate whether patients have had other kinds of diseases are removed because this paper intends to focus on the indicators themselves for heart attacks; adding these features might obscure the indicators. "HeightInMeters" and "WeightInKilograms" are removed because "BMI" is already in the dataset. "ECigaretteUsage" is removed because it only shows presence in the young generation, not applicable to the entire population in the dataset.

Finally, Scikit-learn's selection kBest method was used to determine which features to retain after the previous manual deletion. Now comes the step where we can use Scikit-learn's SelectKBest

method to determine which features to keep, after the previous manual removals. At the moment, there are 16 features left in the dataset, 4 numeric and 12 categorical. It is reasonable to keep all the numeric features and do feature selection on the categorical ones since there are far more categorical features present. With the categorical features, SelectKBest can calculate the chi-squared statistic between each feature and the class, measure dependence between stochastic variables, and finally select the desired number of features to retain. The original intention was to retain 6 categorical variables; however, looking at the scores SelectKBest calculated for each feature, "AgeCategory" has an abnormally high score compared to others, which became the reason this feature was dropped, too.

2.3. Modeling

Support Vector Machine (SVM) is a good model to try first. The essential idea of SVM would be to classify data that are linearly separable from each other. However, in many cases, including the situation this paper deals with, data could be more complicated than just having linear relationships. Thus, SVM also developed a technique called the kernel trick to deal with non-linear relationships but still keeping the original linear structure of how an SVM works. There are many different kernels and different parameters associated with an SVM. With a grid search, a C value of 100, a gamma value of 0.001, and a radial basis function kernel were found to give the best performance of SVM in our case.

Another powerful and popular model would be a Random Forest. Many studies regarding health issues have implemented Random Forests as the solution; one paper mentioned an accuracy of 94% for the prediction of COVID-19 cases, pointing out the strength of Random Forests in classification problems like this [7]. A Random Forest is created by a set of Decision Trees. Each tree in the Random Forest will be different based on the bootstrapped dataset they get and the randomly assigned features to consider. The Random Forest utilizes all these different trees to make a decision collectively. This solves the problem of being too sensitive to the training dataset when using a single decision tree. The randomized process of building a Random Forest helps the model to confidently deal with various situations. The grid search indicates that a max depth of 10, a minimum sample split number of 10, and an estimator number of 100 would give us the best performance in this case.

Finally, a machine learning algorithm that is older than both SVM and Random Forest, Logistic Regression can always be utilized especially in a binary classification problem. Logistic Regression deals with the probability of an individual being in one class or the other, depending on all the other attributes of such individual. It's a rather simple model, giving us the likelihood of a certain class being true; it's normally decided that an individual belongs to a class if that probability is above 50%. Logistic Regression might not work with complicated problems very well but it's still worth getting it involved just to see how it performs compared to the others. A grid search shows a C value of 0.1, a max iteration number of 100, and a l2 penalty can give us the best performance.

3. Results

This section mainly compares the performance of the above three models. In Figure 3, the purpose of each matrix is to display how many mistakes and successes the models made with their predictions. As shown in Figure 3, each matrix has 4 elements indicating 4 categories: True/False Positive/Negative. True and False represent whether the prediction made was correct and Positive and Negative refer to the decision itself made by the models: 1 for Positive and 0 for Negative, as shown on the x-axis, "Predicted label". On the other side, the y-axis represents the "True label", also distributed to 2 values: 0 and 1. Now that the structure of the matrices is clear, it's necessary to focus on the most important pieces of information Figure 3 provides. In a severe health problem, as this paper deals with, people who are labeled positive (patients) would be a top priority. Starting with Matrix 1, the actual patients are denoted by True Label 1, which leads us to the 2 elements in the bottom row of the matrix: in total we have 279 + 689 = 968 patients, and 689 of them were accurately identified as positive which means they can now start receiving treatments as if this is in the real

world. On the other hand, 279 False Negatives were made meaning 279 patients were missed, and this is the number that needs to be reduced as much as possible. This is the same as saying the number of True Positives needs to be as large as possible. The same logic applies to all 3 matrices.

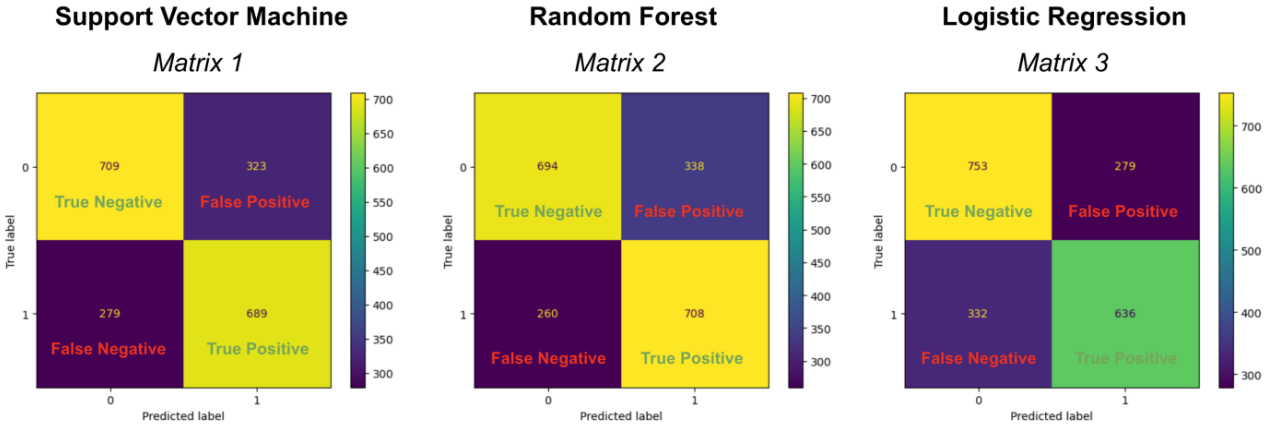

Figure 3. Confusion Matrices for SVM, Random Forest, and Logistic Regression

It is important to keep in mind the necessity of making the ratio of patients being detected as large as possible, which helps determine the better metric of the two shown in Table 1. Here is how precision is calculated:

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{1}$$

As shown in the formula above, the nominator, True Positive, is the number that needs to be increased and so this precision score seems to be appropriate for the situation here. However, it is also important to note what the denominator stands for. Here, the precision's denominator represents all positive predictions made by the models. Is this what matters the most? It's necessary to examine another metric, recall:

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

The nominator remains the same as precision but the denominators have changed to represent all actual patients, and this is indeed the most important matter. Out of all the actual patients, the ratio of people being detected needs to be as large as possible. Thus, recall is chosen to be the metric for model performance evaluation.

Table 1. Recall and Precision Scores for All 3 Models

Model	Recall	Precision
Support Vector Machine	0.71	0.68
Random Forest	0.73	0.68
Logistic Regression	0.66	0.70

According to Table 1, Random Forest has the highest recall score. Thus, this algorithm will be chosen to do further analysis to discover which features matter the most to provide some useful information for health advice.

With the tuned Random Forest model, a method called permutation feature importance can be employed. It works by measuring the model performance with the original test dataset, setting the baseline performance, and then randomly shuffling values within one of the features, just to measure the model performance on the current shuffled dataset again to compare this performance to the baseline performance. If performance drops from the baseline, this means the randomly shuffled feature has a positive effect on the original model performance, and vice versa. As shown in Figure

4, each feature has an associated importance score calculated by subtracting shuffled performance from baseline performance, which is a subtraction of recall scores in this case; the more positive and larger the importance score, the more beneficial impact a feature has on model performance. As Figure 4 displays, "SleepHours" has a very large and positive importance value compared to other features, along with "BMI" and "MentalHealthDays" showing potential for being worth considering as influencers for heart attacks.

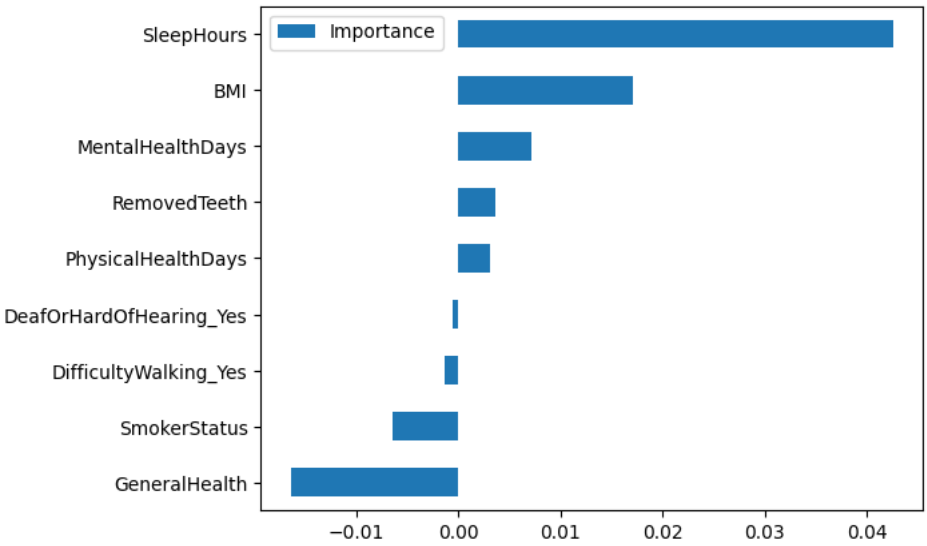

Figure 4. Average Importance of Each Feature.

4. Discussion

In any study that involves data analysis, often, what matters the most is the data itself, not the advanced techniques scientists could employ. So, it is necessary to reevaluate the data used in this study here. First, the data was collected only in the U.S. territory so any conclusions or discoveries should be restrained, potentially not applicable to the population of the entire world. As one study points out, individual health levels could be associated with the health levels of one's nation. The study shows that even in Europe, where there are many countries geographically close to each other, there seems to be a trend for Eastern European countries to have individuals report the poorest health conditions [8]. Thus, it's reasonable to keep conclusions from this paper to applications in the U.S. only. Second, this dataset mostly contains self-reported data which can be subject to inaccuracy. While this kind of data does reveal a different side of people's health conditions as the provided self-reported data can determine the subjective status of individuals, it may be more effective to take exact health data from individuals by having them take physical examinations. The biggest problem with self-reported health data is that not all people have the same level of awareness and knowledge about their health. One study pointed out that the ability to understand health information is inversely related to physical inactivity, and this ability also has various relationships with other factors, making it evident that people's capability to understand and convey health information themselves can vary dramatically [9]. For future studies, it could be optimal to have a mix of self-reported data and data from physical examinations including blood sugar level, blood pressure, and more. This way, the prediction model could be more accurate with the presence of examination data but could also offer insights about what self-reported features could be influential, which is great information to provide to the public because these are features that people could monitor daily with ease. Besides the quality of data, this paper finds Random Forest to be a competitive algorithm and this discovery aligns with other studies in the field, including one study that indicated Random Forest outperformed their Convolution Neural Network (CNN)-based model in the prediction of heart diseases, while CNN-based models are one of the most popular and powerful models in machine learning nowadays [10].

Thus, motivation is given for future studies regarding the extensive application and variation of Random Forests in this field.

5. Conclusion

This paper, through the data from the CDC, showcased that average sleep hours, BMI, and mental health can all be worthy candidates when considering the causes of heart attacks. Although the accuracy of the dataset is not guaranteed to be perfect, this can still provide evidence and motivation for further research and investigation regarding these factors. In terms of modeling, this paper along with the aforementioned paper that showed Random Forest outperforming the CNN-based model, selected Random Forest to be the fittest. Although that study differs from this paper, mainly in terms of data – the mentioned study uses examined physical data only, for instance, resting blood pressure – it still helps to confirm the capability of Random Forest in predictions of health problems. In the end, the diagnosis of diseases is still in the hands of doctors and other medical professionals. Machine Learning algorithms can provide insight for them to make better decisions but it is not appropriate to rely entirely on the algorithms to make judgments and assumptions about people's health conditions. As this paper pointed out, factors including average sleep hours, BMI, and mental health can all be considered by medical professionals and Random Forest can also be utilized by them to aid their decision-making. Future research and practices will make the field more knowledgeable and capable.

References

[1] Eric J, Topol. COVID-19 can affect the heart. Science, 2020, 370: 408 - 409.

[2] Ralph S. Physical activity as an index of heart attack risk in college alumni, American Journal of Epidemiology, 1978, 108 (3): 161 – 175.

[3] Obasi. T and Omair Shafiq M. Towards comparing and using machine learning techniques for detecting and predicting heart attack and diseases, IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, 2393 - 2402.

[4] Hasib K. M. A survey of methods for managing the classification and solution of data imbalance problem,2020, arXiv preprint arXiv: 2012. 11870.

[5] Gregory P. Dietl, Mary E. On the measurement of repair frequency: how important is data standardization? Palaios, 2013, 28 (6): 394 – 402.

[6] Thara D.K., Prema Sudha B.G, Fan X. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques, Pattern Recognition Letters, 2019, 128, 544 - 550.

[7] Iwendi C. COVID-19 patient health prediction using boosted random forest algorithm, Frontiers in Public Health, 2020,8.

[8] Karen M. Health differences between European countries, Social Science & Medicine, 2007, 64 (8): 1665 - 1678.

[9] Aaby A, Friis K, Christensen B, Rowlands G, Maindal HT. Health literacy is associated with health behaviour and self-reported health: A large population-based study in individuals with cardiovascular disease. European Journal of Preventive Cardiology, 2017, 24 (17): 1880 - 1888.

[10] Ram Kumar, R.P., Polepaka, S. Performance comparison of random forest classifier and convolution neural network in predicting heart diseases. In: Raju, K., Govardhan, A., Rani, B., Sridevi, R., Murty, M. (eds) Proceedings of the Third International Conference on Computational Intelligence and Informatics. Advances in Intelligent Systems and Computing, Springer, Singapore. 2020, 1090.