Predicting Heart Disease[[1]](#footnote-2)\*

Julio, Beckett, Adam

CS 270, Spring 2024

Department of Computer Science

Brigham Young University

Abstract

In 2019, heart disease was the number one cause of death in the US and almost half of US adults are affected by cardiovascular disease [American Heart Association, 2019]. With this disease affecting so many, the ability to diagnose it accurately and efficiently is paramount. In this paper, we examine a dataset of standard measurements (taken every time a patient visits a hospital) with heart disease classifications and use machine learning to find which measurements are the best indicators of heart disease. We begin this process by examining data distributions and correlation matrices. After this, we analyze the data through clustering to determine the importance and impact of certain features. We continue by analyzing the data using logistic regression, then attempt to classify data using a decision tree model. Building on this model, we decided to examine several random forest models with different feature sets and comparing their results. Through these processes, we will find that systolic BP, BMI, age, and glucose are among the best indicators for heart disease.

1 Introduction

In general medicine, accurately predicting the likelihood of cardiovascular disease in a patient is becoming increasingly valuable. Nearly half of adults in America have cardiovascular disease. In just the last year, the age-adjusted death rate from cardiovascular disease increased noticeably. Following the horrible trend set by the last several years, heart disease remained the number one cause of death in America [American Heart Association, 2024].

Considering the importance of early diagnosis of heart disease in fighting these trends, this project aims to find factors that indicate a high risk of heart disease. Ultimately, we hope that identifying these factors will lead to earlier diagnosis and better education.

2 Methods and Data

2.1 Data Source

This project uses public data available from the Kaggle data repository. This dataset includes several features measured from patients with and without heart disease. We decided to use this data as it provides a wide variety of features to examine, as well as including specific classification for each instance. Additionally, we found that this dataset was quite noisy and ripe for exploring various analysis techniques. A Juptyer Notebook of our work on the data can be found here:

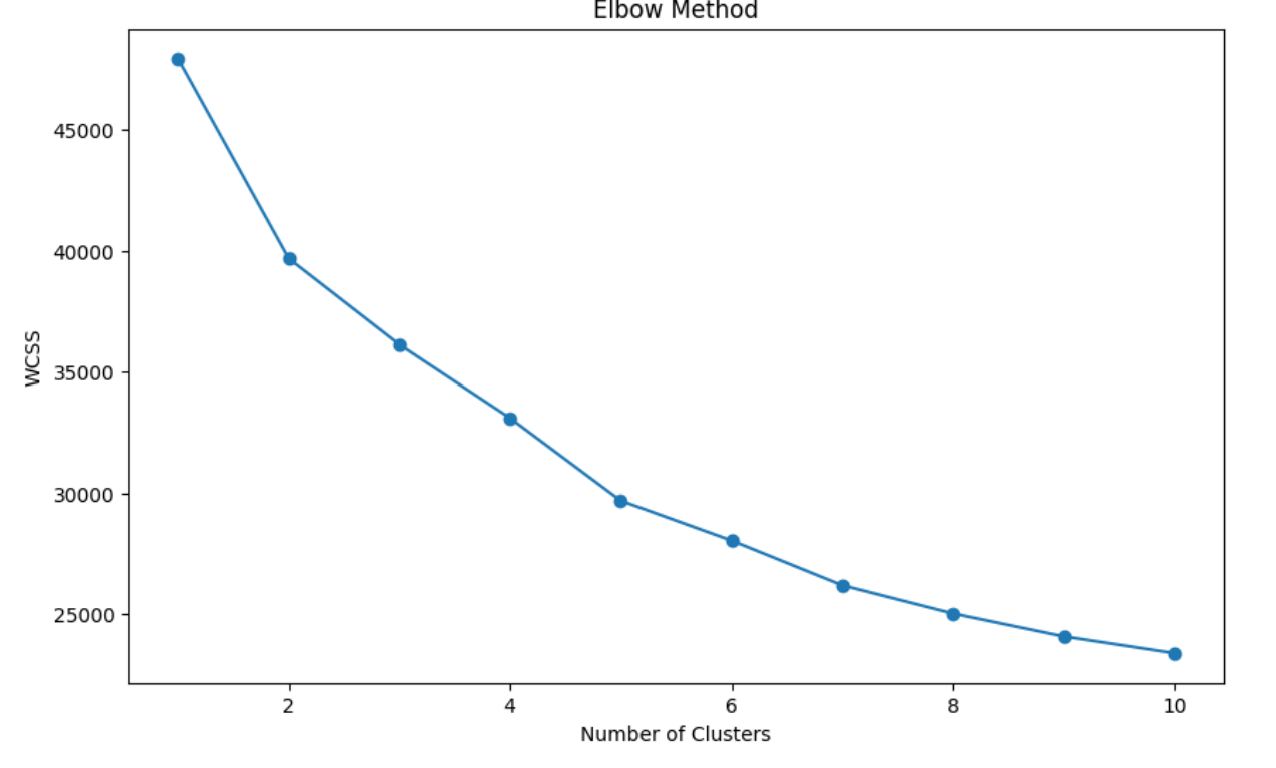
<https://github.com/juliosud/cs270/blob/main/heart_disease_explore.ipynb>

2.2 Dataset

Table 1 shows an example instance the dataset. This includes various features recorded in a standard appointment with a physician. The final feature (CHDRisk) is the instance class, indicating if a patient has an increased risk for heart disease.

|  |  |
| --- | --- |
| sex | male |
| age | 48 |
| education | 1 |
| smokingStatus | yes |
| cigsPerDay | 20 |
| BPMeds | 0 |
| prevalentStroke | 0 |
| prevalentHyp | 0 |
| diabetes | no |
| totChol | 245 |
| sysBP | 127.5 |
| diaBP | 80 |
| BMI | 25.34 |
| heartRate | 75 |
| glucose | 70 |
| CHDRisk | no |

Table 1 – A sample individual in our dataset.

One of the difficulties with this dataset was data normalization. Not every feature in the set was of the same type. For example, sex was a categorical feature, age was an integer, BPMeds was a binary true/false variable, diabetes was a yes/no string value, and so forth. This meant we had to prepare the data before training our machine learning models. The data also included missing values, which we considered in our preprocessing.

To handle these missing values, we used a simple imputer to replace the value with the mean. In categorical features, missing features were filled with ‘missing’. To standardize the data, we used sklearn’s StandardScalar function, since kMeans is sensitive to how the data is scaled. Next, we used OneHotEncoder to encode categorical features. Finally, we used a preprocessing pipeline (the ‘column transformer’ in our code) to prepare subsets of features.

2.3 Selected Models

As our dataset was noisy and large, we decided to try and reduce / cluster our data using k-means and PCA. After this didn’t result as intended, we opted to try decision tree models, logistic regression, and random forest models.

3 Initial Results

We started by analyzing our feature distributions and creating a correlation index, to better understand our features and data. Interestingly, we found that most features weren’t closely correlated, except for the two blood pressure readings with each other and prevalent hypertension.

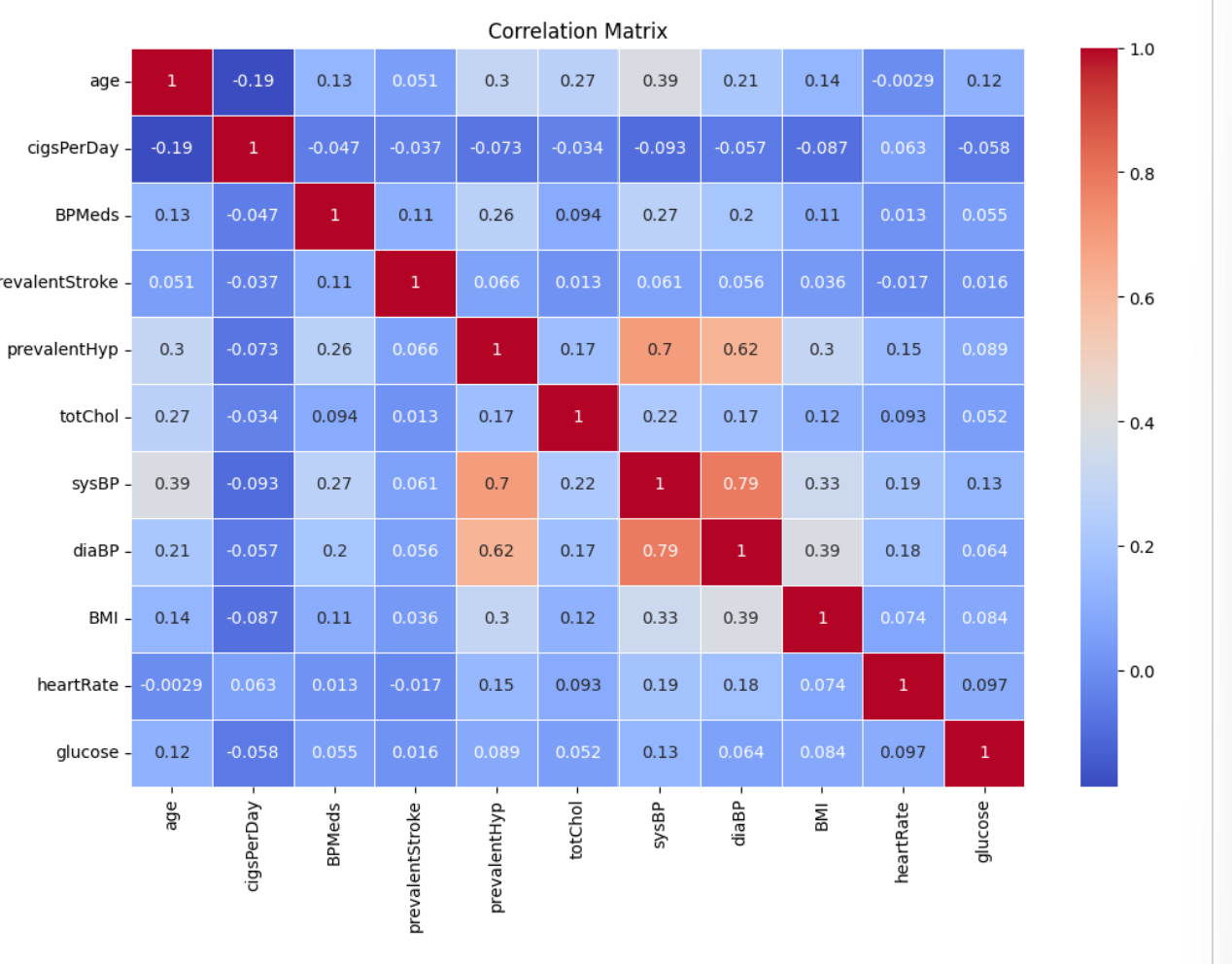
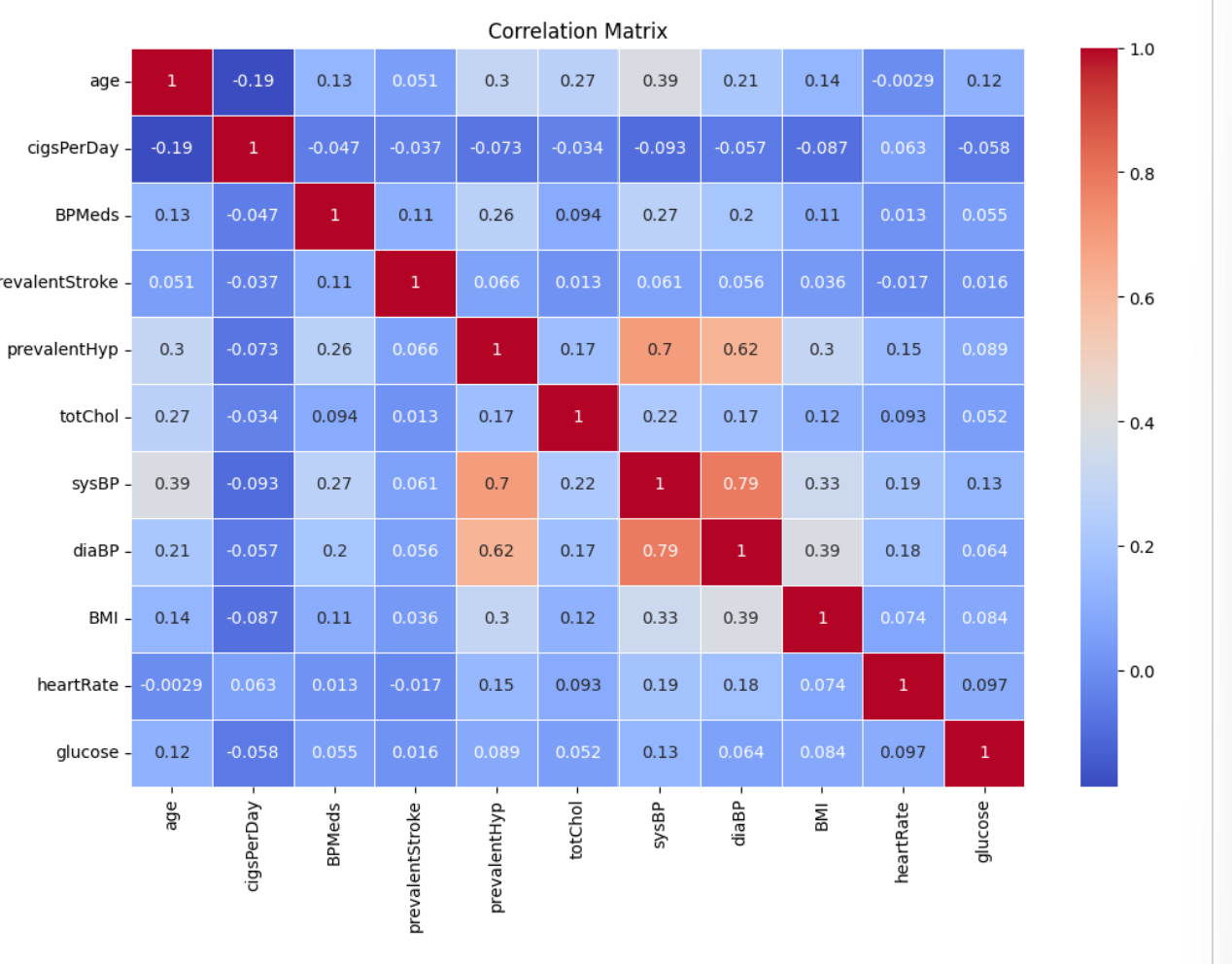


Figure 1 – Initial Clustering

After this, we decided to use a k-means model to run a clustering analysis of our data. We hoped that this would allow us to better understand how our data was packed and see if we could find any general trends before running an analysis. To optimize our k-means algorithm, we ran an initial analysis of different k-values using the elbow method. We eventually found an optimal k-value of 7. The elbow method analysis can be seen in Figure 2.

Figure 2 – The Elbow Method

After finding an optimal k-value for our analysis, we still weren’t finding any trends within our data. As such, we decided to refine our strategy and include other models. We opted to use PCA to reduce the components of our model. With the optimal number of clusters, and using PCA to reduce the dimensionality, our clusters began to like Figure 3.

A chart of a patient cluster

Description automatically generated

Figure 3 – Patient Clusters

This analysis didn’t seem to provide us with many insights. As can be seen in Figure 3, our clusters are widely spread and in proximity together. Some point’s categorized in one cluster are completely separated from their cluster. As we could see that this clustering wasn’t providing much insight, we turned to other methods to classify our data.

4 Improvements

4.1 Logistic Regression and Decision Tree Models

A screenshot of a computer

Description automatically generatedAfter our struggles with k-means analysis with PCA, we decided to try experimenting with other models, hoping to find one that best fit our dataset. We decided to try using a logistic regression model and a decision tree classifier, as these we felt these models could best fit data with lots of features. We were also interested in seeing the feature importance results from the DTC. We used a standard 80-20 training-testing split for our data, then ran the models.

Table 2 – Decision Tree Accuracies

A graph with blue and white bars

Description automatically generatedWe found that the logistic classifier performed much better than we expected, with an accuracy of 0.83. The DTC performed about as we expected, with an accuracy of 0.75. We immediately noted that the models performed similarly on ‘no’ classification (patients without a high risk of heart disease) but the DTC performed abysmally in ‘yes’ classifications (those with an increased risk of heart disease). The full results can be seen in Table 2.

Figure 4 – Feature Importance Bar Graph

We also spent some time analyzing the feature importance results of the DTC. Using these results, were able to figure out which features were the most significant in predicting if a patient is at high risk for heart disease. We found that the features landed in three categories: “highly important” (importance > 0.8), “somewhat important” (0.8 > importance > 0.3), and “less important” (importance < 0.3). The highly important category includes 7 features – sysBP, BMI, age, totChol, diaBP, glucose, and heartRate. The full results can be seen in Figure 4.

While our results from the DTC in regard to feature importance are incredibly promising, we were concerned about making any final conclusions as the ‘yes’ classification accuracy was so low. We wondered if these features were good for knowing if someone didn’t have a high risk of heart disease but would fail to be indicators for a high risk of heart disease. With this in mind, we decided to try implementing a random forest model on the data to see how it would differ from a singular DTC.

4.4 Decision Tree Forest

As we could use our data splits from the DTC with a Random Forest Classifier, we were able to implement one of these models easily. This model was a notable and immediate improvement over the decision tree classifier, which was to be expected. It even made a marginal improvement over the logistical regression model. Our DTC was only able to produce an accuracy of 0.75 with a low ‘yes’ classification precision of 0.27. The RFC model improves these metrics to 0.84 and a whopping 0.79, respectively. The full results of this initial Random Tree Classifier follow in Table 3.

A screenshot of a computer screen

Description automatically generated

Table 3 – Decision Tree Forest Results

We also noted that the feature importance scores produced by the RFC (see Table 6) were nearly identical to those from our DTC. This, combined with a higher ‘yes’ classification score, were promising indicators that these features could be good indicators for heart disease. However, we weren’t entirely sure this was true. To further test this observation, we decided to create a couple more models.

As we were unconvinced that the “highly important” features would serve as good indicators, we decided to create two RFC models for further analysis. The first of these models would classify the data using a combination of the “somewhat important” and “highly important” feature groups. The second would only use the “highly important” group. The results of these models can be seen in the table 4 and table 5, respectively.

A screenshot of a computer screen

Description automatically generated

Table 4 – RFC Accuracy using “Highly Important” and “Somewhat Important” feature groups

A screenshot of a computer screen

Description automatically generated

Table 5 – RFC Accuracy using “Highly Important” features

We found that the overall accuracy of these models remains roughly the same as the RFC using the whole dataset, with precision on ‘no’ classification remaining completely the same. However, we also found a dramatic decrease in the precision of ‘yes’ classification, dropping from 0.79 to 0.67 to 0.56. This shows that the discarded features, despite having less quantifiable importance, impact the model’s ability to classify the data. Further, this likely indicates that the “highly important” features aren’t fully informative indicators for a high risk of heart disease. They only conclusively indicate when a patient is not at a high risk. However, we must note that this is speculation on our end and requires further study. It is also worth noting that the accuracy of these models – though improving on what we had before – is still lower than we’d like. After spending time playing around with various hyperparameters and attempting some feature engineering, we still couldn’t improve our accuracy above the results of the initial random forest classifier.

5 Final Results

5.1 Most Important Features

It must be noted that, while the features we uncovered using out methods can’t fully flag patients with an increased risk of heart disease, they are still informative. Though we have only included the feature importance produced by our DTC and initial RFC, we found that all four of our tree-based models were effective in classifying patients that don’t have an increased risk of heart disease. These models also produced similar results in feature importance; our initial DTC and RFC models ordered the top eleven features the same, with roughly the same weights. The models with reduced dimensionality reflect these results. As such, we can recommend using these feature groupings (especially the “highly important” grouping) to flag individuals who do not have increased risk. The final ordering / weighting of these features can be seen in Table 6.

A screenshot of a computer

Description automatically generated

Table 6 – Final feature importances of initial RFC

6 Conclusions

6.1 Lessons Learned

By using several different models of machine learning, we were able to significantly improve our accuracy and ability to determine the most important features in this data set. Additionally, we were able to find a model in the decision tree forest that would generalize better than our initial attempts, and that had higher accuracy. The “highly important” feature group can be used to classify individuals at a lower risk of heart disease. Age could be expected to be a significant factor leading to heart problems, since older individuals are often the ones that suffer the most from CHD.

6.2 Best Model

The random forest classifier was the best model as far as accuracy was concerned when trained upon this dataset. The regular decision tree was not able to fully capture the results with as high of accuracy as we preferred to see, particularly when classifying those with an increased risk of heart disease. While the random forest was not able to gain above a 90 percent accuracy, it was much more accurate than the decision tree classifier, which had about a 75 percent accuracy.

We learned that it is difficult to improve accuracy for the machine learning models at times, and this was especially true when we experimented with the random forest. We could get it to be 80 percent accurate, but anything above that was difficult to achieve, even with attempts to prepare the data better, move around hyperparameters, etc. This is a part of the process of learning to work with data in machine learning, and we believe it will be valuable to us throughout the future. We do not always achieve optimal results.

7 Future Work

7.1 Additional Features

If we were given the opportunity to collect our own data, we would consider features such as sleep deprivation, and mental health. We know how important both sleep and mental health are to our overall well-being, and it would be interesting to see how, if anything would change for the important features in our data.

7.2 Deep Learning Models

Using deep learning models, along with finding and collecting more data, would help the accuracy of our data. We attempted to do an ensemble forest type of model, but running the data on more deep learning models along with manipulating the data by further narrowing features could be a way to help the project even more.

7.3 More Data

By comparing our data and continuing to learn from other datasets relating to CHD, we would be able to help our model better generalize to unseen data. Additionally, collecting some data of our own, and getting a current dataset would allow us to see how features in predicting CHD have changed over time.

7.4 Post-COVID-19 Study

It would be interesting to do a study on cardiovascular disease since the COVID-19 pandemic.

7.5 Seek a Professional Opinion

While we narrowed down some of the most important indicators for predicting CHD, it would be valuable for us to obtain a professional opinion on the matter. We are not experts in biology, or medicine, and it would be good for us to have another opinion on our results. Doctors see patients each day, and would likely be able to help confirm how reliable our results truly were. Obviously, CHD has many factors that play a role in the causation of the disease, and having an expert to guide us in the process would improve the validity of our study.

Acknowledgments

We would like to acknowledge our thanks to Tyler Trogden for teaching us machine learning, and all the different learning models. Additionally, we would like to point out a special thanks to Julio for helping us find a direction for our project and for picking the dataset.

References

[American Heart Association, 2019] American Heart Association. Cardiovascular diseases affect nearly half of American adults, statistics show. January 2019.

[American Heart Association, 2024] American Heart Association. More than half of U.S. adults don’t know heart disease is leading cause of death, despite 100-year reign. *January 2024*.

[CDC, 2024] Centers for Disease Control and Prevention. Heart disease facts. May 2024.

[Anderson, Keaven M., *et al*, 1991] Anderson, Keaven M., *et al*. Cardiovascular disease risk profiles. *American Heart Journal*, January 1991.

[Bansal, M., 2020] Bansal, Manish. Cardiovascular disease and COVID-19. *Clinical Research & Reviews*, 14(3) 247-250, May 2020.

1. \* These match the formatting instructions of IJCAI-07. The support of IJCAI, Inc. is acknowledged. [↑](#footnote-ref-2)