Predictive Modeling of Coronary Heart Disease

Kirsten Fung, Angela Hong, Britney Hoang

DS 3001: Machine Learning

Professor Johnson

April 16, 2024

I. Summary

For our project, we explore whether predictive algorithms can accurately predict the likelihood of Coronary Heart Disease (CHD) based on the factors we have selected. Using the Framingham Heart Study (FHS) data set, our central research question is: *Which predictive algorithm has the best accuracy of predicting the chance of an individual contracting CHD?* The findings of this project can be beneficial in identifying risk factors, allowing for preventative measures and early intervention of CHD.

We began our cleaning process by selecting relevant variables from the FHS dataset and dropping the unnecessary columns. In terms of approaching the null values, we opted to replace them with the median values to preserve the overall data distribution. This approach aimed to prevent potential skewing that could compromise data integrity and validity. Furthermore, we renamed any necessary values in order to make them more understandable for both ourselves and other parties. For instance, we converted binary representation of variables, such as gender, into more intuitive labels, facilitating ease of interpretation.

In our data analysis phase, we initially created a heatmap to uncover any notable correlation among our selected variables. This revealed that the strongest correlations were between currentSmoker and cigsPerDay, as well as between diabetes and glucose. While these associations were expected–given the link between diabetes and glucose levels, as well as the behavioral patterns of smokers–they provide valuable insights into the potential risk factors for CHD. Despite not identifying any obvious correlations from the matrix, we proceeded with further graphing to dive deeper into the relationship between other variables and CHD to deepen our understanding of the illness. Overall, we observed a more significant relationship between age and CHD risk, whereas factors like cigarettes per day, BMI (body mass index), and glucose exhibited less predictive power in relation to CHD.

This led us to conduct more of our own analysis and continue with regression testing. We employed four predictive algorithms: decision trees, linear regression, k-nearest neighbor (KNN) regression, and KNN classification. We ultimately chose classification as they are used to predict which values are most likely when the outcome is categorical which aligns best with our model of testing against CHD–a binary variable. The results indicated that the KNN classification model performed the best since it had the highest accuracy value of 0.854 with an optimal k of 22. This means that considering the 22 nearest neighbors provided the best balance between capturing local patterns in the data and avoiding overfitting. However, with respect to regression analysis, the model that yielded the best R^2 value was our linear regression model with an R^2 value of 0.0894 and an RMSE value of 0.3389.

II. Data

After close inspection of the training and testing datasets, we landed on 10 variables to use in our CHD predictive model. Below are the selected variables and brief explanations of the reasoning behind our choices:

sex (the recorded sex of the observations with 1 denoting a participant coded as male): We believed that sex could be an influential factor as CHD could be more likely or prevalent in one sex over another. Gender could be a significant variable due to biological differences.

age (age at the time of medical examination in years): Age can be an influential factor and individuals could be more at risk of developing CHD the older they get.

currentSmoker (Current cigarette smoking at the time of examinations): Smoking is well-known to have negative long-term consequences on the body and so by looking at this variable, we can see if they are more at risk with this unhealthy habit.

cigsPerDay (Number of cigarettes smoked each day): By observing the number of cigarettes per day, we can see the severity of the smoking and if it could have a long-term impact on one's cardiovascular system.

prevalentStroke (Prevalent Stroke (0 = free of disease)): This variable could demonstrate if having a stroke makes an individual more prone to CHD.

prevalentHyp (Prevalent Hypertensive. Subject was defined as hypertensive if treated): Hypertension is a leading one of the leading causes of stroke and this is imperative to include in this study.

diabetes (Diabetic according to criteria of first exam treated): High blood glucose levels can be damaging to the body's organs and can ultimately lead to a stroke (among other major health problems).

BMI (Body Mass Index, weight (kg)/height (m)^2)): BMI is used to calculate the amount of body fat based on one's height and weight. Thus, BMI could be used to demonstrate being overweight, which increases the risk of high blood pressure and cardiovascular disease.

glucose (Blood glucose level (mg/dL)): High blood pressure is another leading cause of strokes and can lead to other health complications.

tenYearCHD (The 10-year risk of coronary heart disease (CHD)): This binary variable can indicate if the individual has a likelihood of developing the disease within a decade. This serves as our dependent variable for this project.

The testing and training sets were cleaned the same way using the methods as follows. We began by reading the provided CSV into a new data frame. The functions ".columns" and ".shape" were called to see an overview of the data. The original columns in the csv were 'Unnamed: 0', 'sex', 'age', 'education', 'currentSmoker', 'cigsPerDay', 'BPMeds', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose', 'TenYearCHD'. This list was reduced to the ten variables mentioned above. The test set has 1060 records and the train set has 3180 records, about three times larger. Having this much data, we expect that conclusive results can be made to find variables that are an accurate predictor of CHD. Looking at the column names, we decided to drop "Unnamed: 0" since we knew it would be unused in our predictive models.

Next, we began going through each of the variables described above. The ".info()" function was called to see the number of non-null values present and their associated data type. The number of nan values was calculated by subtracting the non-null count from the total number of entries. This told us that the 'cigsPerDay', 'BMI', and 'glucose' variables had nan's present. We discussed different approaches to handle these missing values. We briefly considered completely dropping them from the dataset, however, in order to keep the integrity and accuracy of the data, we decided to replace all null values with the median of the data. This ensures that the data does not become skewed due to our cleaning methods.

Lastly, variables such as 'sex', 'currentSmoker', 'prevalentStroke', 'prevalentHyp', 'diabetes', and 'tenYearCHD' produced binary values 0 and 1. However, these numbers represent responses such as 'Male', 'Female, 'Yes', 'No', etc. Thus, we believed that they should return string values as such. We renamed the values using the ".replace()" function to make the data more understandable. This ensures the model that we build will be easier to interpret for a reader with no background knowledge of CHD. In the GitHub repository, there are two sets of cleaned variables. One set contains the binary responses that were used to create the prediction models (cleaned_test_binary and cleaned_train_binary). The other set contains the renamed responses which were used to create the visualizations (cleaned_test and cleaned_train).

III. Results

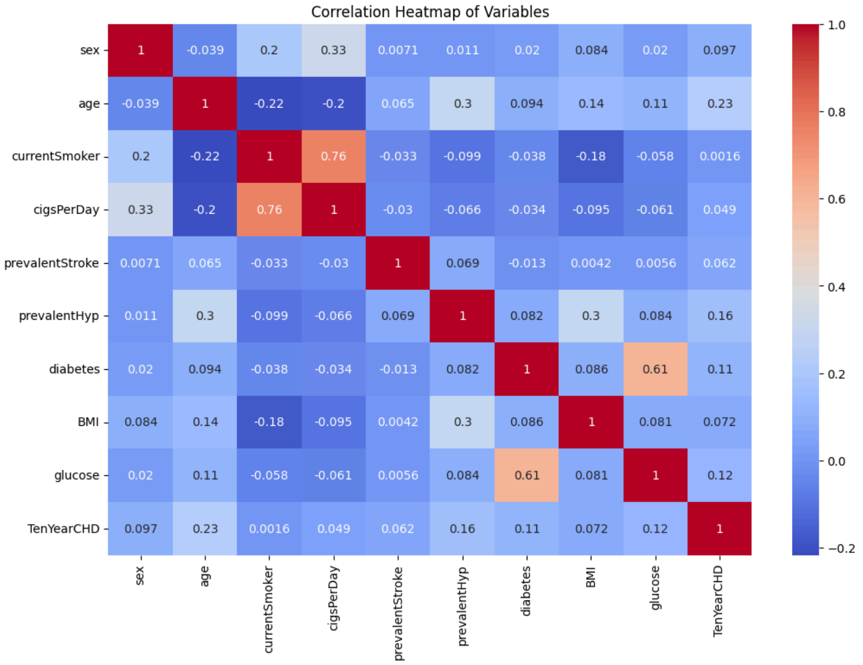

Figure 1: Correlation Heatmap of Variables

As previously noted, the heatmap depicted in Figure 1 did not reveal any significant correlations between the variables. We observed higher correlation coefficients between variables such as cigsPerDay and currentSmoker, as well as between diabetes and glucose. These associations were anticipated and ultimately deemed inconsequential to directly predicting the likelihood of developing CHD. Notably the variable demonstrating the strongest correlation with TenYearCHD was age. Thus, we explored this relationship further in the following graphs.

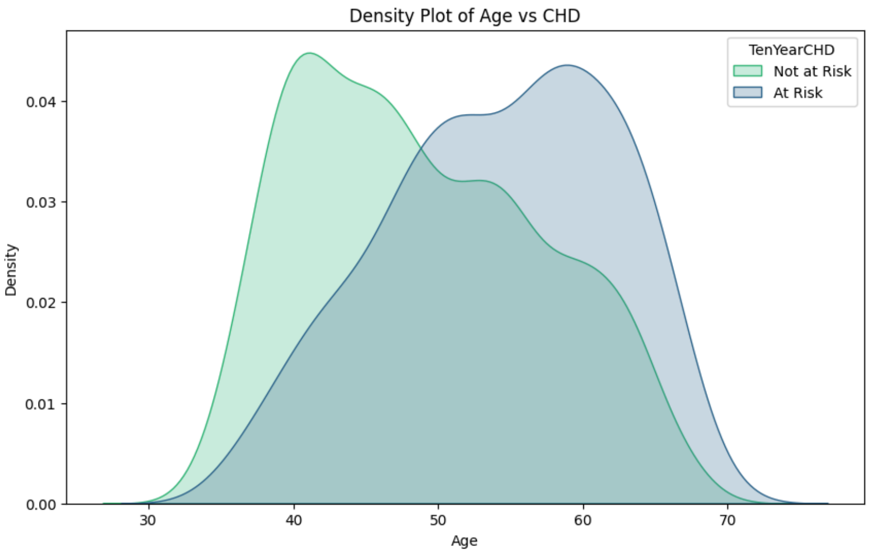

Figure 2: Density Plot of Age vs. CHD

In Figure 2, our goal was to examine the impact of age on the risk of CHD. Within the 'Not at Risk' category, a distinct right-skewed distribution is seen, indicating that individuals aged 30-50 generally exhibit a lower risk of CHD. Conversely, for those 'At Risk' of CHD, we observed an evident left-skewed distribution, with a higher concentration of individuals, typically aged 55 and older, afflicted with CHD. Thus, this graph reveals the trend of reduced CHD risk among younger individuals, while emphasizing the heightened susceptibility among older age groups.

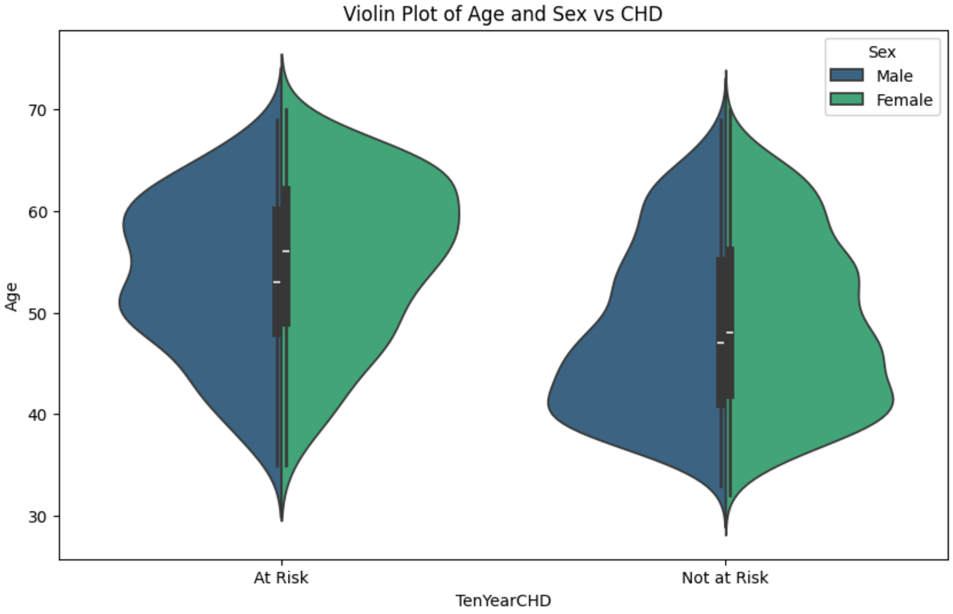

Figure 3: Violin Plot of Age and Sex vs. CHD

Based on the insights from Figure 2, we wanted to investigate further and explore whether sex serves as an influential factor in the development of CHD, potentially being more prominent in males or females. Figure 3 demonstrates similar findings to Figure 2, older individuals demonstrate a heightened susceptibility to CHD compared to their younger or middle-aged counterparts. Our analysis revealed that there is no significant difference between males and females in terms of CHD risk. Thus, sex does not appear to be an influential factor in determining an individual's likelihood of developing CHD.

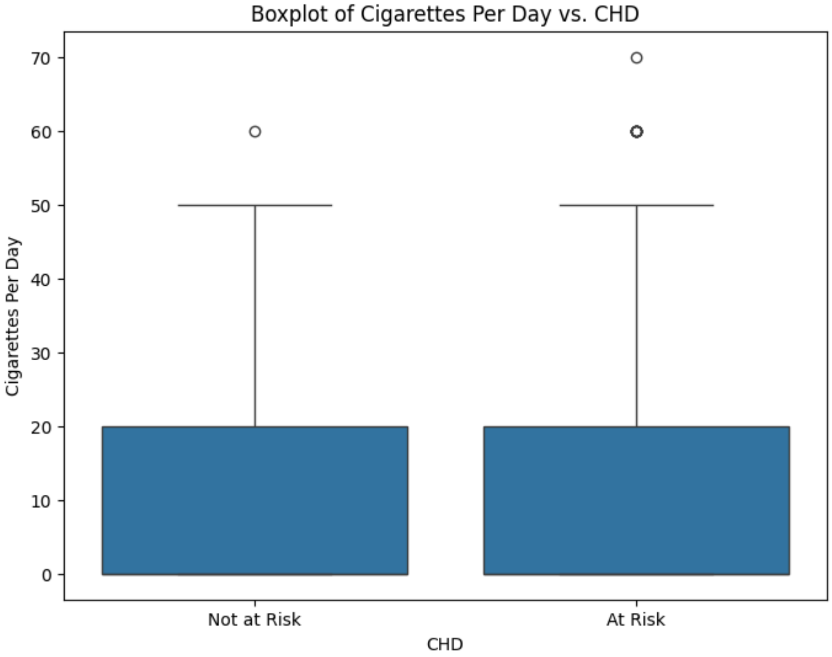

Figure 4: Box of Cigarettes Per Day vs. CHD

Initially, we hypothesized that individuals who engage in regular smoking habits or consume a higher volume of cigarettes would exhibit an elevated risk of CHD. However, Figure 4 contradicts this assumption, showing that there is no substantial correlation between the number of cigarettes smoked per day and the 10-year risk of CHD. Despite a greater number of outliers in the higher range within the "At Risk" boxplot, the overall pattern closely mirrors that of the "Not at Risk" counterpart.

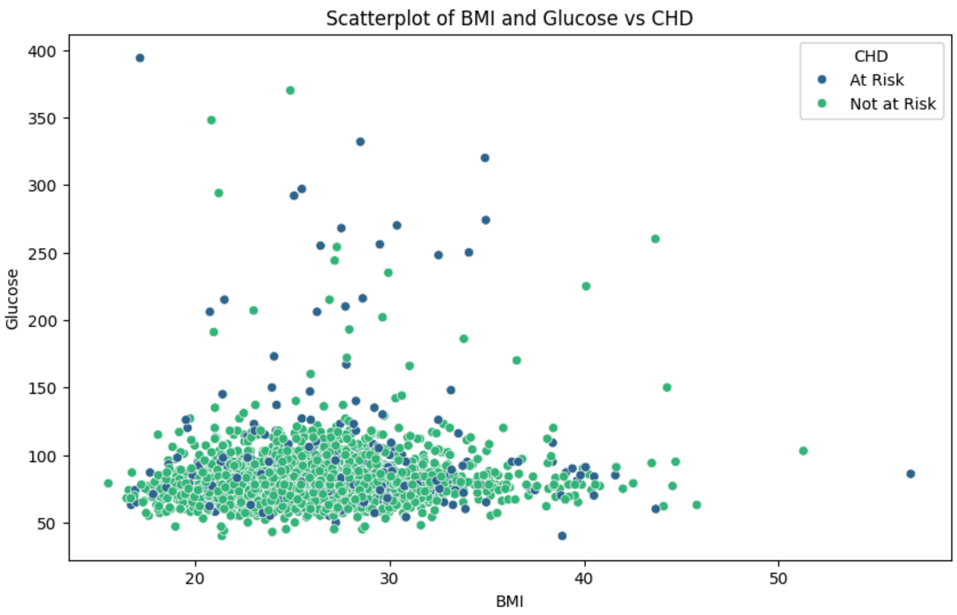

Figure 5: Scatterplot of BMI and Glucose vs. CHD

In Figure 5, we wanted to explore the relationship between a selection of remaining variables and the 10-year risk of CHD. So, we created a scatterplot depicting the relationship between BMI and glucose levels among individuals classified as “At Risk” and “Not at Risk” of CHD. Again, our analysis yielded no discernible conclusions from the visualization, as no evident correlation or patterns emerged between the variables.

To run our regression models, we used built-in functions provided in the Python library and the notes provided in this class to run our models. In the linear regression model, its peak performance yielded an R^2 of 0.0894 and an RMSE of 0.339. Next, we produced a decision tree algorithm which yielded an R^2 of 0.0357 and an RMSE of 0.349. Lastly, we created a KNN regression algorithm which yielded an R^2 of 0.0580 at an optimal K of 137 and an RMSE value of 0.345. Of our regression models, our linear regression performed the best. This could be because linear regression assumes a linear relationship between our independent and dependent variables. However, although our goal was to use regressive predictive modeling algorithms, we ultimately decided to also run a KNN classification algorithm due to the type of data we were predicting. Moreover, by doing so, it yielded the highest accuracy with a score of 85.4% or a value of 0.854 with an optimal k of 22. Overall, every model we ran, produced a relatively low RMSE value. This indicates that our model's predictions are generally close to the actual values in the dataset. Thus, the differences between the predicted values and the observed values are relatively small and the model has a high level of accuracy in its predictions.

IV. Conclusion

The objective of the project was to utilize predictive algorithms to forecast the likelihood of Coronary Heart Disease (CHD). The data used for this project is a subset of the Framingham Heart Study dataset, containing various demographic, lifestyle, and health-related variables. We believe that we selected what we deemed the most pertinent factors for predicting CHD susceptibility. Our analysis revealed the most positive correlation between age and CHD, indicating an increased risk as an individual gets older. However, we did not find many other variables that were high indicators of the disease.

Our findings demonstrated that our best predictive model of CHD was the KNN classification model with the best accuracy value of 0.854 with an optimal k of 22. This was followed by the linear regression model, then KKN regression, and then trees. The reason behind running a classification model was that it better aligned with the dataset's characteristics and the qualities of the test. This model achieved a high accuracy score, indicating that it effectively predicts the likelihood of CHD based on our selected variables. The selected variables were suitable for the classification task, as they contributed significantly to the model's predictive performance. However, the predictive model with the highest regression was the linear regression model which had an R^2 value of 0.0894 and an RMSE value of 0.3389.

Overall, each of the models that we employed ended up with fairly positive results. Through utilizing four different models, we used trial and error to test the efficiency of each model, ultimately rendering our findings more credible. Furthermore, we took measures to address potential criticism by testing both classification and regression algorithms. While some may question our variable selection, we believed that a smaller selection of variables would be more manageable which prompted us to carefully consider our reasoning for choosing them.

In terms of additional work outside the scope of this project, we believe that a larger data set beyond the demographics of Framingham, MA would be beneficial in looking at individuals on a national, regional, and/or global scale. Furthermore, an additional race/ethnicity variable could prove to be very helpful in unveiling potential genetic predispositions for CHD. While we acknowledge that our R^2 values across the methods are quite low, this could rather be a flaw in the data collection process given that data is only collected from a sample of patients from MA.

In conclusion, for optimal regression performance, our linear regression model would be the preferred choice. However, if prioritizing accuracy within the dataset, the KNN classification model stands out due to its compatibility with the dataset's characteristics.