Coronary Heart Disease Project

Catherine Xu, Jillian Haig, Erin Moulton, Riley Tomek, Ashrita Kodali, and Rushvi Patel

Summary:

In the following report, the team built a predictive algorithm to predict the likelihood of a person developing coronary heart disease (CHD), a disease in which the buildup of plaque in the coronary arteries leads to blockages, reducing blood flow to the heart. The Framingham Heart study data explores multiple variables that may be associated with CHD; utilizing this, two relevant variables were selected: diabetes and prevalentHyp. This longitudinal dataset has tracked patients from 1948, covering three generations of participants and offering insight into a person's ten-year risk of CHD, along with valuable health details such as smoking habits, medication usage, blood pressure levels, and more. Prior to choosing the significant variables, the team began the variable cleaning process. Once the data was concise, the team ran a preliminary exploratory data analysis (EDA) that revealed that diabetes, hypertension, and stroke history were significant influences of the development of CHD. After experimenting with multiple different models (using the training data) including K-nearest neighbors (KNN) regression, multiple linear regression, decision trees, and KNN classification, the team found that the accuracy of the model using KNN classification was much greater than those of the other models; specifically, the R-squared value for the other models were very small while the accuracy obtained in the final model was relatively high. The variables chosen for the main analysis are not continuous variables but instead binary, so this provided further encouragement for the group to use a classification model rather than a regression model. After testing the same

model with the testing data, the accuracy obtained was nearly the exact same as the accuracy of the training data, so performance did not drop.

Data:

The relevant data utilized comes from a subset of the data available from the Framington Heart Study, a study which began in 1948 and samples patients from Framington, Massachusetts. The study is currently on its third generation of patients and studies a variety of different variables related to patients' health. Mostly scatter plots and kernel density plots were used to explore data and pick variables that seemed to have the strongest relation to the TenYearCHD variable. A regression was conducted on several of the variables to determine which variables were the most significant; from this regression, it was determined that prevalentStroke, prevalentHyp, and diabetes had the largest coefficients (0.242, 0.109, and 0.240, respectively) compared to the other variables tested. However, the team decided to not include the prevalentStroke variable as the sample size was only 21. Consequently, the relevant variables chosen for analysis include diabetes, prevalentHyp, and TenYearCHD which are all binary variables. The first variable, diabetes, contains information on whether or not the patient is "diabetic according to criteria of first exam treated." The values are either a 0 if the patient is not diabetic or a 1 if they are. The next variable, prevalentHyp, examines if individuals have been identified as having hypertension at the beginning of the study or also at a different time during the study; the values for the patients are a 0 if no hypertension is prevalent and a 1 if there is. These first two variables are the explanatory variables and will be used in the model to predict the final relevant variable, TenYearCHD. TenYearCHD measures the presence or absence of coronary heart disease over a ten year period; a 0 represents no coronary heart disease where a 1

represents the presence of the condition. The datasets (both training and testing) were already relatively clean. After looking closer into the values in the dataset, it was found that there were some NAs which the team decided to drop or impute.

Results:

The model the team decided would be the best to use for analysis was a model using KNN classification and confusion matrices. To run this model, the KNeighborsClassifier package was imported from sklearn.neighbors, and the train_test_split package was imported from sklearn.model_selection to split the sample. After splitting the sample and solving for optimal k (which was optimal from 4 to 100), the team constructed the following accuracy plot:

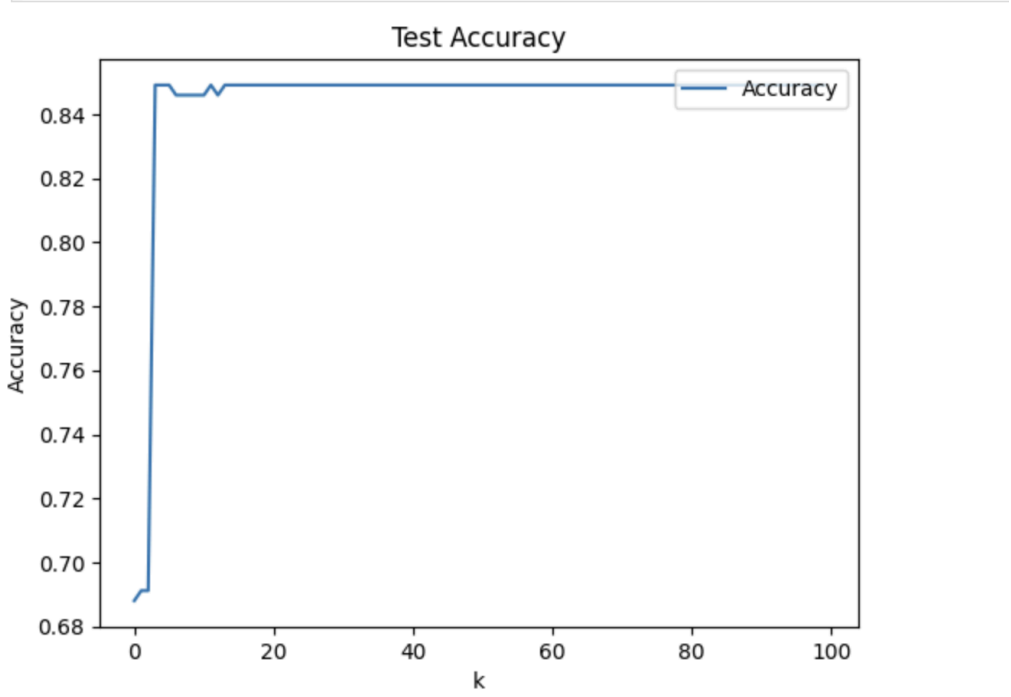

Figure 1: Accuracy plot with optimal K for training dataset

Based on this plot, the accuracy of this model is about 0.86, leading to the conclusion that the model is relatively accurate. A confusion matrix was also constructed from this model and is shown below:

col_0	0
TenYearCHD	
0	1350
1	240

Figure 2: Confusion Matrix using the training dataset

This confusion matrix is essentially predicting that individuals without prevalent hypertension or diabetes (0) are unlikely to develop CHD within the ten-year period, while a smaller portion might potentially develop CHD (1). This observation explains the relatively high accuracy obtained in the previous accuracy plot. The same process was repeated for the testing dataset and a similar plot and confusion matrix was obtained:

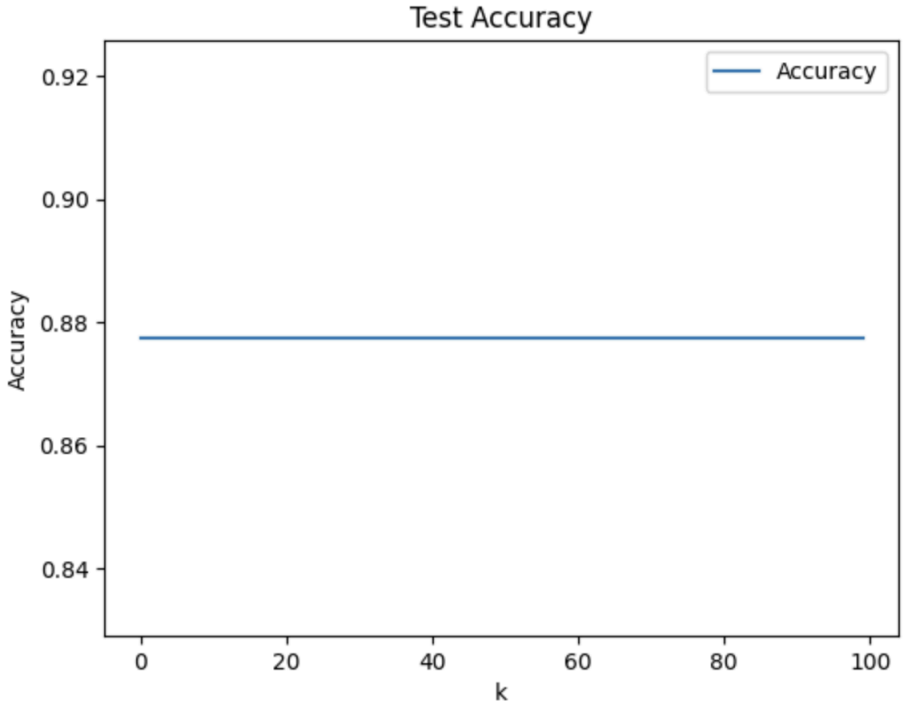

Figure 3: Accuracy plot with optimal K for testing dataset

Based on the results from this accuracy plot, it is evident that our model is likely not over-fitting. Instead, the accuracy for the model increased slightly from 0.86 using the training data to just under 0.88 using the testing data. A confusion matrix was also obtained for this data following the same logic as the first confusion matrix:

col_0	0
TenYearCHD	
0	401
1	56

Figure 4: Confusion matrix using the testing dataset

The accuracy obtained from the classification model was the highest compared to other models the team tested. After importing the KNeighborsRegressor package from sklearn.neighbors, a plot of SSE was constructed which is shown below:

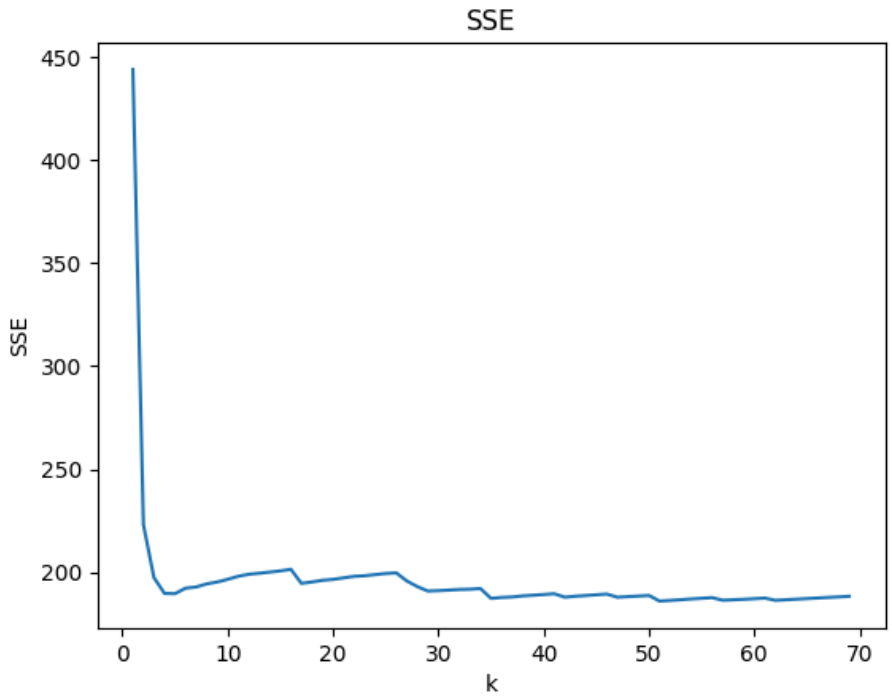

Figure 5: Plot of SSE and k using the training dataset

For the optimal k of 51, the SSE was calculated to be 185.9. The SSE is the lowest at the optimal k, signaling that the model performs particularly "better" at this optimal k than at other levels of k.

One of the other models the team considered was multiple linear regression. After importing the LinearRegression model from sklearn.linearmodel, a regression using two relevant variables (prevalentHyp, and diabetes) was run, and the R-squared was computed to be 0.0368 which seemed extremely low in comparison to the classification model.

A decision tree model was also attempted by the group. After importing the DecisionTreeClassifier and plot_tree packages from sklearn.tree, a model was run using the dummy variables prevalentHyp and diabetes as the explanatory variables and TenYearCHD which was the response variable. The visualization for the results of the decision tree is shown below:

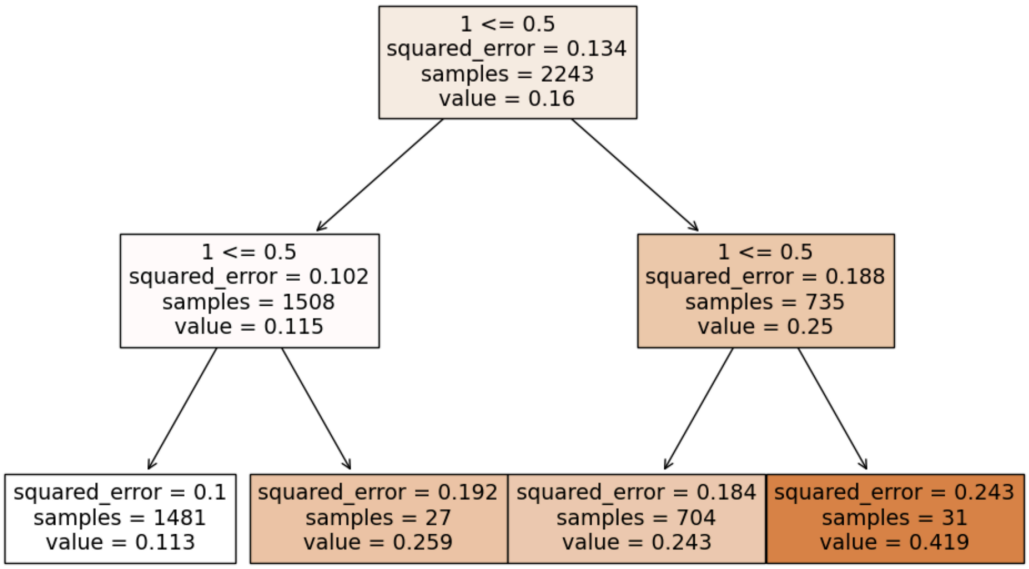

Figure 6: Decision tree visualization for training data

The accuracy obtained for this model was 0.696 which was not too much lower than the accuracy obtained from the classification model. However, the group believed that the classification model would still be more appropriate given the characteristics of the relevant variables and the fact that the accuracy for the classification model was the highest.

Conclusion:

In all, the group worked to generate the most accurate model for predicting the likelihood of future Coronary Heart Disease development. Using the variables from the Framingham Heart study, the team began the variable cleaning process. Numerous NA values were identified within the dataset, prompting removal and imputation before moving on to exploratory data analysis. Diabetes and the prevalence of hypertension (prevalentHyp) were chosen as prominent potential predictors of developing CHD as they yielded high coefficients observed in an initial regression analysis. Once these predictors were identified, multiple models were created using various techniques including K-nearest neighbors, multiple linear regression, decision trees, and

k-nearest neighbors classification. These models were then assessed using parameters such as the accuracy and R-squared value. Of all these models, the KNN classification model appeared to be the best model to continue with as it had the highest accuracy along with the highest R-squared value. In order to assess the accuracy, an accuracy plot was generated that tracked the accuracy as the k value increased. From the plot, it appeared that the accuracy of the model was 0.86, indicating that the model makes accurate predictions. Looking at the SSE plot, it appears that the SSE continues to be minimized as the k-value increases. At a k-value of 51, the SSE is the smallest at approximately 185.9. Finally, looking at the classification models accuracy plot, it appears that the accuracy plot remains constant, indicating that the model was not overfit to the training data set and indicating generalizability of these results. Based on metrics including the accuracy, SSE, and the R-squared value, the best model is KNN classification. While the model is accurate, there are a few drawbacks with this model. The SSE for classification model is inaccurate as KNeighborsRegressor was used in the for loop instead of KNeighborsClassifier as KNeighborsClassifier kept running into errors when trying to calculate the SSE. The inaccurate SSE might not be the best parameter to use in order to compare the classification model against the other models. Furthermore, the data from the Framingham data set might not be representative of the US population. The study was first conducted almost 70 years ago and while it has followed their sample of patients, its applicability to the present-day United States population and worldwide generalizability is questionable. Continuing, further analysis should be conducted to build better models. It might be important to look at transformations and potentially transform some of the variables using a sine or log transformation in order to see how they now relate to the likelihood of developing CHD. Furthermore, it might be helpful to look at a certain

subset of the population instead of trying to generalize to the US (ie: focusing on women in their 20s) to get a better idea of what is happening.