Project 2 CHD: Coronary Heart Disease Training Models

Group 7: Anna Brown, Adaire Burnsed, Tamera Fang, Diana Nguyen, Elle Park, Carol Wu

School of Data Science, University of Virginia

DS 3001: Foundations of Machine Learning

Professor Johnson

April 2, 2024

Abstract

The Framingham Heart Study (FHS) is a long-term research endeavor that investigates the risk factors for cardiovascular disease by following generations of participants in Framingham, Massachusetts, which began in 1948. Using a subset of this data, we developed three predictive algorithms — a decision tree, linear regression, and k nearest neighbors (KNN) — to predict the likelihood of a person developing coronary heart disease (CHD), given the risk factors measured by the data frame. The risk factors included in the FHS data subset are as follows: sex, age, education, current smokers, cigarettes per day, blood pressure medicines, previous strokes, diabetes, hypertension, total cholesterol, body mass index, 10-year risk of CHD, heart rate, and different blood pressures. All three models were built on training data, and then tested on testing data. After cleaning the data and training the models, we will compare the testing data R-squared outputs of the three models to examine the effectiveness of the models' predictability both individually and comparatively.

Our group was divided into two subgroups. The first group cleaned the data, and the second group developed the visualizations for analysis. Cleaning included handling missing values, converting variable types, mapping categorical variables, and creating new columns to organize the dataset. First, we created and printed out a decision tree for analysis, assessing the accuracy of the trained model. Then, we implemented a k-nearest neighbor model to determine its accuracy. Finally, we applied a linear regression model and evaluated its performance using metrics such as RMSE and R-squared values. The k nearest neighbor and decision tree models' accuracy scored at 0.833 and 0.752 respectively; however, with the R-squared values were 0.093 and 0.077, the linear regression model needed further training. Overall, this exercise provided an

interesting comparison of three predictive models and displayed the differences in accuracy between them. From this project, we concluded that using k nearest neighbor model training is almost 0.10 more accurate than decision trees in classifying and predicting a patient's risk of developing coronary heart disease.

Data

Our training and testing datasets are extracted from a subset of the Framingham Heart Study data frame. This dataset includes health metrics, hypothesized to be correlated with the development of coronary heart disease, in volunteer patients from 1949. The specific variables included in our dataset are listed in the table below:

Name	**Description (Unit of Measurement)**	**Quantity of Missing Values**
sex	Binary variable of the observed patient's recorded sex: 0 = female, 1 = male	0
age	age at the time of medical examination (years)	0
education	Categorical Variable of the participants education: 1 = some high school, 2 = high school/GED, 3 = some college/vocational school, 4 = college	85
currentSmoker	Binary variable recording if the observed patient is a current cigarette smoker at the time of examinations: 0 = no, 1 = yes	0
cigsPerDay	number of cigarettes smoked each day (cigarettes)	24
BPMeds	Binary variable recording if the observed patient uses anti-hypertensive medication at exam: 0 = no, 1 = yes	37
prevalentStroke	Binary variable recording if the observed patient has a prevalent history of strokes: 0 = no, 1 = yes	0
prevalentHyp	Binary variable recording if the observed patient has a prevalent history of hypertension: 0 = not previously treated for hypertension, 1 = previously treated for hypertension	0
diabetes	Binary variable recording if the observed patient has diabetes: 0 = no, 1 = yes	0
totChol	Total cholesterol (mg/DL)	39

sysBP	Systolic blood pressure (mmHg)	0
diaBP	Diastolic blood pressure (mmHg)	0
BMI	Body Mass Index, weight (kg/height in meters^2)	15
heartRate	Heart rate (beats/minute)	0
glucose	Blood glucose level (mg/DL)	285
TenYearCHD *(predicted variable)*	Ten year risk of coronary heart disease (CHD)	0

In total, there were 485 NAN values throughout our data frame, with over half of these values coming from the glucose variable. We initially created a dummy variable for each column with missing values, where 1 indicates the value is missing. We used these dummy columns to determine how to handle the missing values. While we ultimately decided to drop them from our predictor models, these columns could be used in future models to analyze whether there is a correlation between missing values and the ten-year risk of CHD.

Results

The first model we used to predict the likelihood of coronary heart disease in 10 years is decision trees. This was done by importing ski-kit learning packages to train and test using the categorical variables. Dummies were created and after that step, we split and trained the model with the random state set at 42 and the test size set at 0.2. The accuracy was calculated using the y-test and the y-hat (predicted).

The second model we used to predict the likelihood of coronary heart disease in 10 years (TenYearCHD) is the K-Nearest Neighbors (KNN) algorithm. We used KNN because it provides relatively accurate predictions from a simple model which is a good tradeoff and provides a lot of information. When preprocessing the data used in our KNN model, all missing values were dropped from the raw data frame. According to the correlation matrix on the data frame, ‘age’ (0.23), ‘sysBP’ (0.21), ‘diaBP’ (0.14), and ‘glucose’ (0.12) are the most strongly correlated variables with TenYearCHD. Additional dummy variables were created for all the boolean variables: 'sex', 'currentSmoker', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'BPMeds'. For all dummy variables except ‘currentSmoker’, the mean for 0 is much lower than 1, indicating a higher proportion of the population dies when these boolean variables take the value of 1 rather than 0. Based on the results of the correlation matrix and the dummy variables summary table, we selected 'sex', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'BPMeds', 'age', 'sysBP', 'diaBP', 'glucose' to build a matrix of X of the variables most predictive of TenYearCHD, and a variable y equal to TenYearCHD. Min-max normalization was done on all the variables in X. The samples were next split into 80% for training and 20% for testing. The optimal number of neighbors for a KNN regression for the variables selected was k=41 with a minimum sum of squares error (SSE)

of 160.69 (Figure 5). Using k=41, the accuracy of the KNN model is 0.833.

Lastly, we utilized a variety of kernel density plots to encompass demographic and clinical variables to assess the 10-year risk of coronary heart disease. Additionally, using the describe() feature in our code, we found information for variables we did not create graphs for. For instance, we found the demographic factor of sex showed a stark contrast in CHD incidence, with 19.45% of males exhibiting CHD compared to only 12.16% of females. This finding is in line with existing research that suggests males are at a higher risk for CHD, potentially due to a combination of genetic, hormonal, and lifestyle factors. From a clinical perspective, the boolean variables painted a vivid picture of CHD risk factors. A few graphs not shown but were also investigated using the describe() function include the following: Patients on anti-hypertensive medication ('BPMeds') demonstrated a 34.94% prevalence of CHD, which was considerably higher than those not on medication. This may reflect the severity of underlying hypertension or the presence of other comorbid conditions that may require medication. A history of stroke ('prevalentStroke') was another significant boolean variable, with a 35.29% CHD rate among individuals with a previous stroke, dramatically higher than the 15.32% in those without, indicating prior stroke as a powerful predictor of future CHD. Hypertension ('prevalentHyp') and diabetes also stood out as boolean variables with strong CHD associations. Hypertensive individuals had a 24.29% CHD incidence, more than double the 11.24% in non-hypertensive individuals. Diabetes showed an even more pronounced effect, with diabetic individuals exhibiting a 36.84% CHD incidence compared to 14.84% in non-diabetics, showing a link between diabetes and cardiovascular disease which makes sense as diabetes is known to damage blood vessels which puts patients at a higher risk for CHD. The numeric variables provided

additional granularity. The age distribution (Figure 2A) revealed a clear age-related increase in CHD risk, with a density peak in the late 50s to early 60s, reflecting a rise in cardiovascular risk with advancing age. Smoking habits, represented by the number of cigarettes smoked per day (Figure 2B), showcased two peaks: one at the lower end, representing non-smokers to light smokers, and another peak for heavy smokers, indicating an increased CHD risk at both ends of the spectrum. This bimodal distribution underscores the heightened CHD risk for smokers, irrespective of the number of cigarettes consumed. Total cholesterol ('totChol') levels (Figure 2C) did not seem to have much of a visible impact on our graphs for peaks for CHD and non-CHD are around the same levels. Systolic ('sysBP') (Figure 2D) and diastolic blood pressures ('diaBP') (Figure 2E) displayed similar trends. Body Mass Index ('BMI') (Figure 2F) also displayed a similar trend in which the peaks are around the same for CHD and non-CHD. This was surprising because outside research usually shows a density shift towards higher BMI values for CHD cases, emphasizing the role of obesity as a major modifiable risk factor for heart disease. The heart rate plots (Figure 2G) did not present much variation either, with peaks once again being around the same range. For our last kernel density plot, we notice the same trend again. Glucose levels (Figure 2H) did not show much distinction, with glucose concentrations being markedly around the same.

Now, for some logistics of our model creations, our linear regression model centered on these numeric predictors yielded a training RMSE of 0.337 and a testing RMSE of 0.372, with respective R-squared values of 0.093 and 0.077. Although these values reflect only a modest portion of the variance in CHD risk, they emphasize the multifactorial nature of CHD, where a single predictor is not sufficient for accurate risk assessment. Enhancing the model with

polynomial features, we observed an improvement in the training data fit, with an RMSE of 0.332. However, the testing RMSE slightly increased to 0.378, suggesting the model might be overfitting the training data and highlighting the delicate balance required in model complexity. The scatter plot of the polynomial regression model's true versus predicted values (Figure 3) illustrated the challenge of capturing the full spectrum of CHD risk, with the distribution of residuals (Figure 4) revealing underprediction for higher-risk cases.

Conclusion

Our project used predictive algorithms (decision trees, linear regression, and KNN) to predict the likelihood that a 1948 Framingham, Massachusetts adult patient would develop coronary heart disease. The results from the KNN model showed that 'age', 'sysBP', 'diaBP', and 'glucose' are the most strongly correlated variables with a 10-year risk of CHD. Based on the correlation matrix and summary table, when patients have the dummy variables (except for 'currentSmoker') indicated as a 1, a higher proportion of the patient population dies with those variables. To summarize from a clinical and demographic perspective, our kernel density plots showed males have a higher risk for CHD than females. Patients on anti-hypertensive medication, have a history of a stroke, hypertension, or diabetes, and those in their late 50s and early 60s reflected a significant risk for CHD. Our linear regression with numeric predictors yielded a training RMSE of 0.337/ r-squared of 0.093 and a testing RMSE of 0.372/ r-squared of 0.077.

Our linear regression has a weak correlation between our predictors which does not suggest that the model has good predictive strength. Figure 3 shows a scatterplot in our polynomial regression model where the relationship between predicted and true values is weak. Figure 4 shows a distribution of residuals in a polynomial regression model that is not around zero and is positively skewed with a big kurtosis tail indicating a cluster of outliers depicting variance. Despite these results and graphs are not perfect, their nature creates an opportunity to train our data even more to develop a better predictive model that could potentially be applied in a clinical setting in the future. It also creates an opportunity for discussion surrounding future research to study the flaws in our research or to either confirm or deny the results of our project.

Additional work could include a new variable or variables which provide for example race/ethnicity demographics so that we can see whether these or potentially other cultural factors may contribute to the likelihood of CHD. An example of how this idea of a cultural factor could be a predictor is if one cultural group traditionally ate a diet that helped protect them from developing CHD. Although these future research steps require additional diverse surveying which may take more time and money, the research would have more generalizability so that our findings can be applied beyond our class and be more representative of the Massachusetts population. This would have beneficial implications because historically, many research projects and clinical trials have been done with predominantly white patient populations can limit generalizability beyond the research setting.

Our project can increase awareness of the risk of coronary heart disease and all the possible risk factor variables that contribute to CHD. Despite some faults, our project still contributed to our understanding of the disease and our results could be helpful for others as well. Our results would be more powerful if they were verified by additional supporting research, especially with a new data set that is composed of a diverse group of participants from around the country, or even the world. This project can help understand the ongoing research on this prevalent disease and help distinguish which variables such as age, diabetes, hypertension, and more influence the likelihood that a patient will develop CHD.

Appendix

Figure 1: KNN regression determining the optimal number of neighbors for the variables selected k = 41. The accuracy of the KNN model is 0.833.

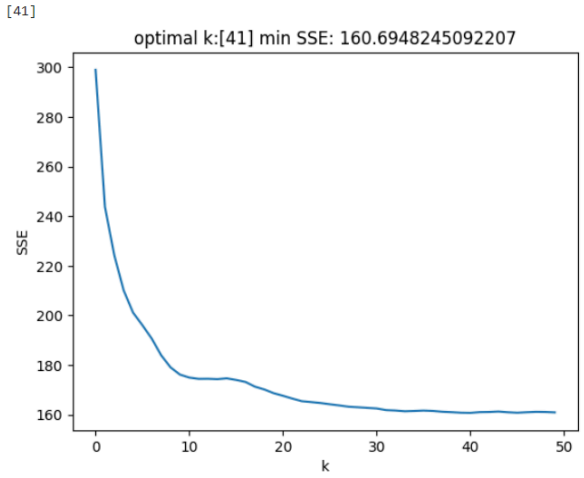

Figure 2A: Kernel density plot of the numeric variable age (years) by CHD status.

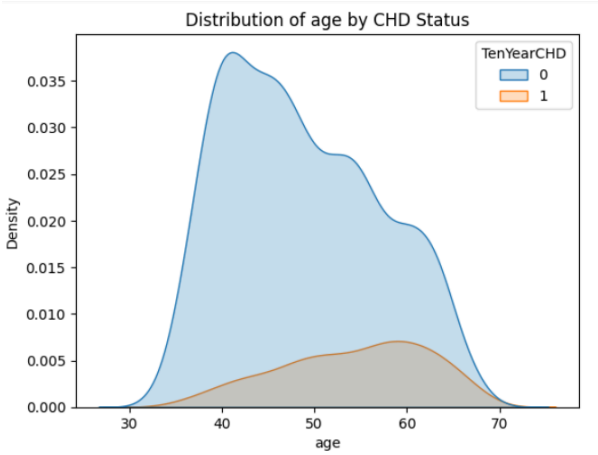

Figure 2B: Kernel density plot of the numeric variable cigarettes per day by CHD status.

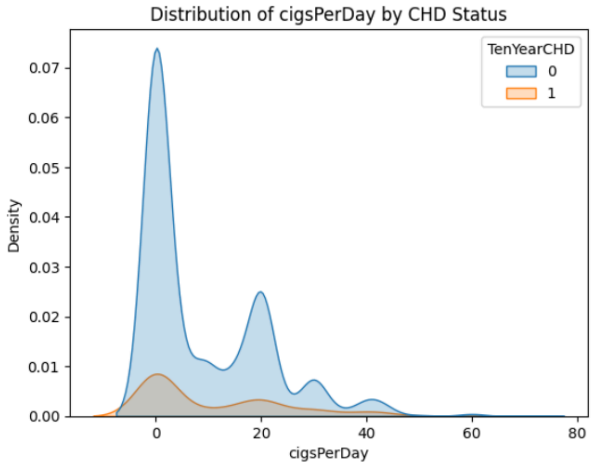

Figure 2C: Kernel density plot of the numeric variable total cholesterol (mg/dL) by CHD status.

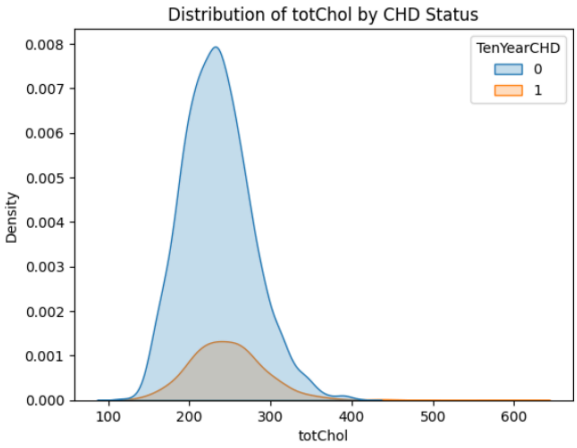

Figure 2D: Kernel density plot of the numeric variable systolic blood pressure (mmHg) by CHD status.

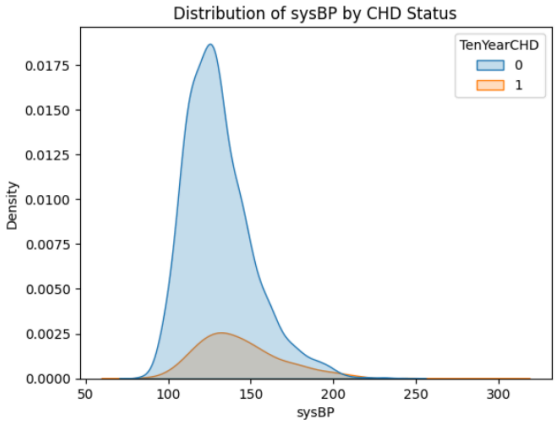

Figure 2E: Kernel density plot of the numeric variable of diastolic blood pressure (mmHg)by CHD status.

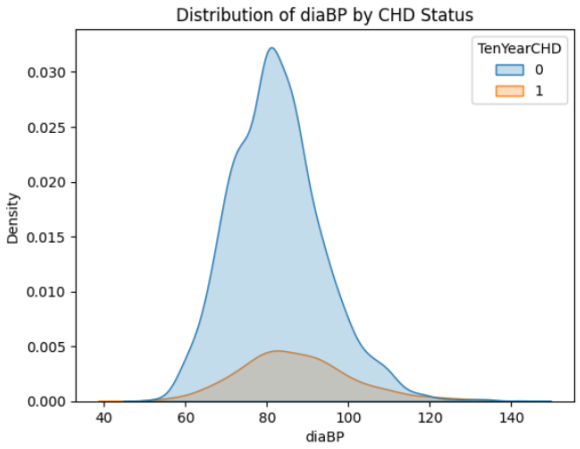

Figure 2F: Kernel density plot of the numeric variable body mass index (kg/m^2) by CHD status.

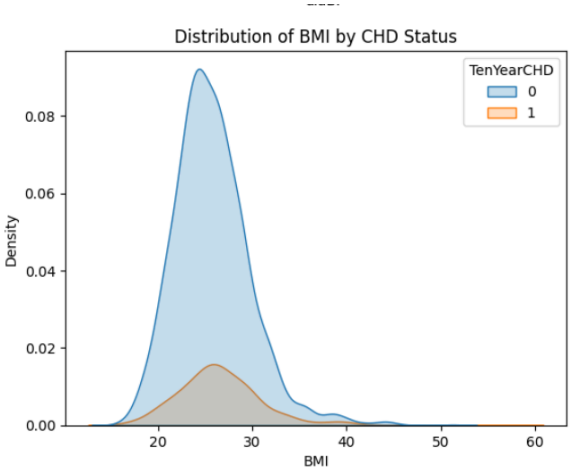

Figure 2G: Kernel density plots of the numeric heart rate (beats/minute) by CHD status.

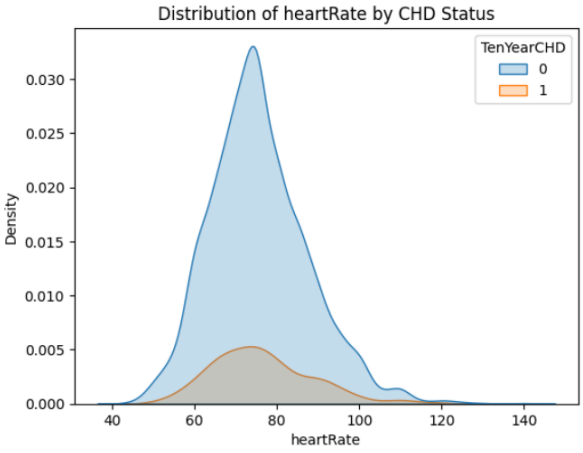

Figure 2H: Kernel density plot of the numeric blood glucose level (mg/dL) by CHD status.

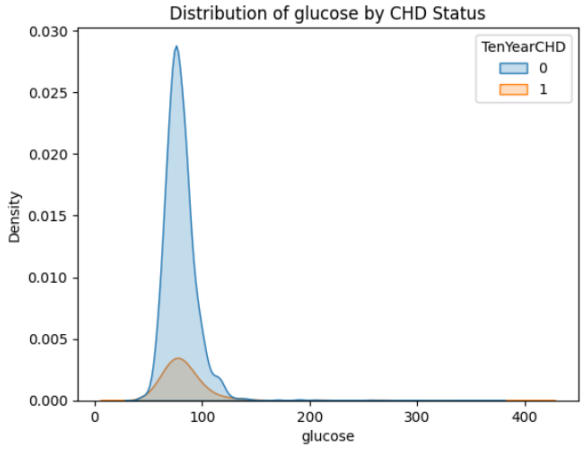

Figure 3:

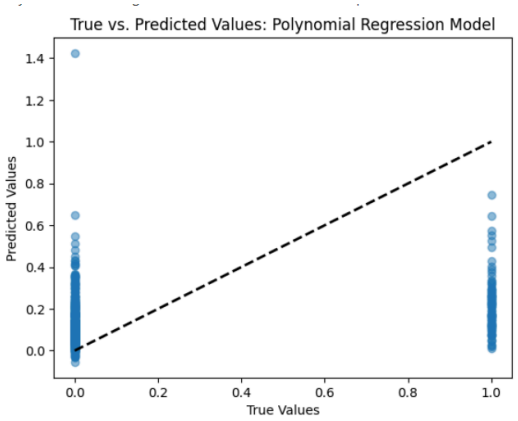

Figure 4: