# Project 2 – CHD

**Author(s):**

**Gabriel Jackson**

**Naad Kundu**

**Thao Nguyen**

**Kayla Nguyen**

**Halbert Nguyen**

**Ben Willoughby**

**Omar Zeineddine**

**DS 3001: Foundations of Machine Learning**

**April 16, 2024**

**Summary**

Our main focus for this project was to build predictive algorithms to determine the likelihood of a person developing coronary heart disease. We could utilize many variables such as sex, age, education, current smoker status, cigarettes per day, etc. We used provided data to build our training model and to test our model. First, we cleaned our data by updating any NaN values. We needed to clean several variables, including education, glucose, BPMeds, totChol, BMI, cigsPerDay, and heart rate, in terms of NaNs. We then saved these cleaned datasets into updated CSV files. After cleaning the data, we used linear regression and decision trees to predict the likelihood of developing coronary heart disease. For our linear models, we used the sklearn library. We first read in the CSV files, including the test data. Our linear regression model results were interesting, particularly due to our dummy education variable. We decided to compare two linear models: one with all variables and one excluding the education variable. We found that the adjusted R^2 values for the model with all variables were higher than those for the model without the education variable, indicating that the model with all variables had more explanatory power for predicting CHD. Our decision tree model also yielded interesting results—it had an 85% accuracy rating. However, it did end up having a lot of false positives, incorrectly predicting CHD in 143 cases where the person did not have CHD. The importance of education in predicting CHD suggests potential socio-economic factors influencing heart health, warranting further investigation into how these predictors interact. Our next steps will focus on refining the models by integrating additional variables and employing more advanced machine learning techniques to enhance predictive accuracy and reduce false positives. This continued refinement will aim to minimize overfitting while exploring the impact of unexamined variables.


**Data**

In terms of the data, two csv files were provided- one for training our model and one to test the model. The variable we are trying to predict is 10YearCHD, which represents the 10-year risk of coronary heart disease. Initially, we had to clean the training dataset. For our education variable, we replaced all NAs with 0s, in order to keep its value numerical. For our glucose variable, we filled all NAs with 85, which is the human average. This was done to keep the data from being skewed for our models. For our BPMeds variable, we replaced all NAs with 0- to indicate that they are not being taken. For our totChol (total cholesterol), BMI (body mass index), cigsPerDay, and heartRate variables, we decided to drop all the NANs. We then saved these cleaned datasets into their respective updated csv files to prepare them to be used by botb our linear regression model and our decision tree. To begin our linear regression models, we first updated our education variable. One big challenge we encountered was with our education variable. The education variable was categorical, but it appeared numerical. We proceeded to change the entries into text, so that we could easily create dummy variables. We replaced its values from 0-5 to 'Unknown education', 'Some high school', 'High school/GED', 'Some college', 'College', respectively. Overall, our dataset contains multiple dummy variables. For many of the variables, the representation is intuitive, as 1 is true and 0 is false. But for our sex variable, 0 represents female and 1 represents a male. We proceeded to construct a linear model with no intercept, using the variables sex, age, our education dummy variable, and currentSmoker, cigsPerDay, BPMeds, prevalentStroke, and prevalentHyp. Calculating R^2 and adjusted R^2, we got 0.0998 and 0.0943 respectively. We proceeded to make another linear model, this time without the education variable, and our R^2 was 0.0832 and our adjusted R^2 was 0.0790. It didn't seem like education played much of a factor in coronary heart disease, which is why we decided to create two models- one with the education variable and one without it. Looking at the R^2 values, it seems like the model with all available variables, including education, had more explanatory value due to the higher adjusted R^2 value. We encountered some challenges when trying to make our predictive model. The variable we were trying to predict, 10YearCHD, can only take a binary value of 0 or 1, meaning we have a linear probability model. This implies that the predicted values are understood as probabilities of the event occurring. Also, under linear probability models, the regression line will never fit the data perfectly if the dependent variable is binary and the regressors are continuous. This means the R^2 value, our primary tool for measuring the explanatory power or 'usefulness' of a model, loses its interpretation. Since linear models predict probabilities, we decided to try and use a decision tree as well, to compare our results.

To make a decision tree, writing the code was fairly simple. We also used the sklearn library for the decision trees. We used the same cleaned data files as we did for our linear models. A couple of challenges did arise though- the biggest one was attempting to limit the amount of false positives the decision tree kept giving us. The decision tree incorrectly predicted yes for 143 cases when they had no coronary heart disease. The accuracy was 85%.

**Results**

In order to predict the likelihood of a person developing Coronary Heart Disease, three predictive models were considered using training and test data from the Framingham Heart Study. This data includes the following variables: sex, age, currently a smoker, cigarettes per day, use of blood pressure medication, stroke prevalence, prevalent hypertension, diabetic, total cholesterol, systolic and diastolic blood pressure, body mass index, heart rate, glucose, and the 10 year risk of coronary heart disease. Overall, the three predictive models used were k-nearest neighbors, decision trees, and (ordinary, least-squares) linear regression. The first predictive model used was decision trees, which presented a series of decisions in the form of a tree of nodes that predict if a person is likely to develop Coronary Heart Disease ('Yes CHD') or not ('No CHD'). The training and test data were used to create a 5-layer Decision Tree Classifier from SKLearn, which can be seen in Figure 1:

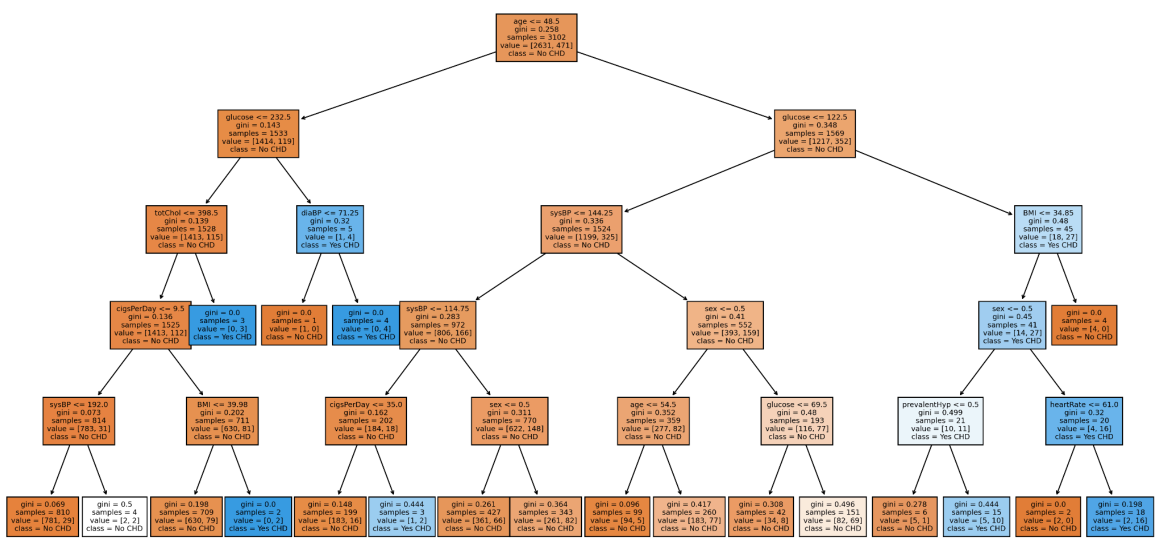


**Figure 1: The Decision Classification Tree**

The decision tree will make a decision at each node, which after making a series of decisions in the tree leads to a terminal node that will make a prediction on whether or not a person is likely to develop Coronary Heart Disease. The terminal nodes that end in a ‘Yes CHD’ prediction are colored from a range of white to blue depending on how many samples were used to make this terminal decision. Additionally, the terminal nodes that end in a ‘No CHD’ prediction are colored from a range of white to orange, depending on how many samples were used to make this terminal decision.

Along with creating this decision tree diagram, the decision tree was tested with the testing data to produce a confusion matrix that determined accuracy, specificity, sensitivity, and MCC score of the tree. The confusion matrix showed 85% accuracy, 0.90% specificity, 98.76% sensitivity, and a MCC score of 10.68%. This means that with a specificity rate of 0.90% there was a high false positive rate and that with a sensitivity of 98.76% that it was good at detecting actual Coronary Heart Disease cases. Therefore, the low specificity and high sensitivity leads to the model favoring ‘Yes CHD’ strongly. This can be seen in its low false negative rate and its high false positive right. Additionally, the MCC score of 10.68% showed that it was not good at balancing errors made within the tree. However, it did have an accuracy of roughly 85% in the testing data, which demonstrates its ability to correctly classify a person as likely to develop CHD. While this decision classification tree was correct 85% of the time, it can only account for up to 5 nodes (5 decisions in the tree). Using a higher number of layers is possible, but makes the diagram (Figure 1) harder to interpret. Therefore, the researchers chose to use 5 layers to balance readability with performance. This comes at a cost, where roughly 5 decisions can be made to determine if a person is likely or not to develop Coronary Heart Disease (despite there being more than 5 variables). While the tree does not consider all variables, it does indicate which variables are most likely to determine developing CHD, such as a higher age, glucose level, etc. It’s also fairly straight-forward to read these variables as they are presented in mostly-human readable form presented as conditions.

The last predictive algorithm used was the K-Nearest Neighbors Classification algorithm. This algorithm was used to form each datapoint in the training set into clusters that would be labeled as either having Coronary Heart Disease or not. To determine the best number of clusters (k) to choose, the optimal cluster was chosen from 1 to 150 clusters. For each cluster, the number of clusters with the highest accuracy was found to be 6 clusters with an accuracy of roughly 85.48%.

The third predictive model used was linear regression, which creates a predictive equation using the variables of the data as input to determine whether or not a person is likely to develop Coronary Heart Disease. In the equation, each variable is associated with some weight, which will either increase or decrease the odds of developing Coronary Heart Disease. To determine whether or not

Discussion of methodology:

Linear regression: Talk about linear regression

- How we tested on education excluded and education included
- Cite Appendix for tables
- Talk about how there was a higher r-squared for the education included table and that there was a slightly lower r-squared for the education table
- Talk about how linear regression was a poor choice due to how the data is discrete 0 or 1.

Decision trees: Talk about decision trees

- How the decision tree was created
- The diagram made for the decision tree
- The results of the decision tree

Analysis of how the

## Conclusion

Overall, our project embarked on an ambitious task to predict the likelihood of developing coronary heart disease (CHD) using a subset of data from the Framingham Heart Study. Through the methods of linear regression models and decision trees, we explored various factors that could increase risk factors in developing CHD, such as sex, age, education, smoking habits, and medical history. Our approach required us to focus on the fundamentals: meticulous data cleaning, thoughtful variable transformation, and strategic model selection in order to adhere with the project's objectives of experimenting with machine learning models, and having fun.

One of the project's significant findings was the varied impact of education on the risk of CHD, revealed through comparing models with and without the education variable. This result highlights the interesting relationship between socio economic factors and health, suggesting that education level can serve as a precursor to activities that impact heart health. Additionally, the decision tree model further complemented our analysis with an 85% accuracy rate; however, with a good amount of false positives. This goes to show the complications of predicting binary values (yes/no) in data regarding health. It is important to note that while our decision tree model did generate a higher number of false positives than ideal, this is often a challenge in medical prediction models, particularly when prioritizing sensitivity. In future iterations, adjusting the decision threshold and exploring cost-sensitive learning could better balance the trade-off between sensitivity and specificity, reducing false positives without significantly compromising the ability to detect actual cases of CHD.

As for further exploration, we could incorporate other machine learning techniques such as k nearest neighbor, in hopes of enhancing accuracy and reducing false positives. Further, exploring the relationship between other variables in the dataset can reveal other potentially unexpected predictors of CHD. Another way we can further elaborate on this data is to account the model for change in risk factors over time. This would provide us a different and dynamic point of view on CHD prediction. Moreover, doing further analysis to include more information such as genetic information, dietary habits, or even psychological markers can add depth to our models. These multidimensional analyses allow for an intricate platter of factors influencing CHD, presenting deeper insights and more detailed predictions.

Wrapping everything up, although our project may be not perfect, it gives us a glimpse into the world of health data science. Going from the raw data to cleaned_data to predictive models is complicated and tedious, but the experience gained from this project is what makes it possible for us to adopt a thirst for knowledge, hopefully understanding more of how we can decrease the risk factors of CHD. As technology progresses and machine learning models become even smarter, we are getting closer to a future where predictive models can provide personalized insights on improving health, increasing human lifespan.

## Appendix