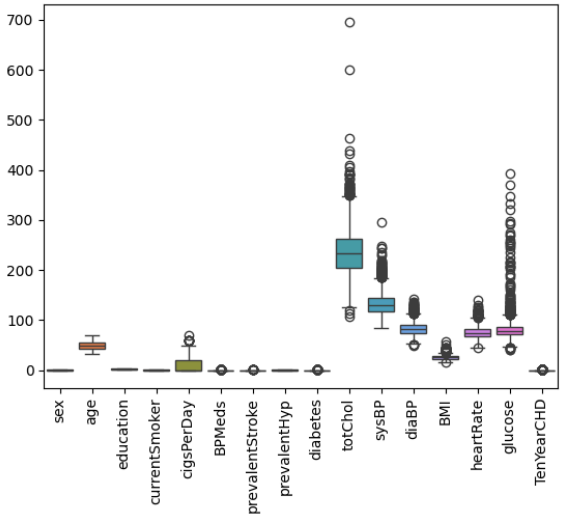
Anisha Poudel, Jinwen Mu

Summary

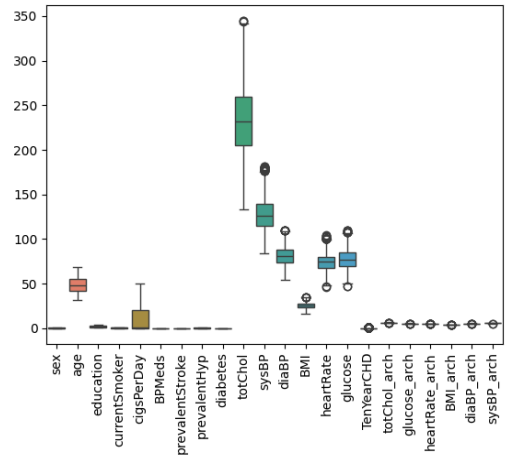
The purpose of this study is to build predictive algorithms that predict the likelihood of a person developing coronary heart disease (CHD). The data used in this study is a subset of the data available from the Framingham Heart Study, which began in 1948 and is now in its third generation. The data includes 16 variables that include sex, age, education, and other variables specifically related to health. We found this project important as it is helpful for medical professionals to know the correlation between numerous variables and the likelihood of developing coronary heart disease. Attaining this insight might help with preventive actions for patients who have signs of early coronary heart disease developing. Our primary methodology included linear regression with a stepwise selection of significant predictors based on the R-aquares improvements. Our best performing model has an R- square of 0.108. This indicates modest predictability.

Data

The dataset for this project originated from the Framingham Heart Study, encompassing a range of variables. The first challenge encountered while cleaning the data was treating missing data, especially in the ‘education’ category. 2.67% of data entries had such missing data.



The boxplot showcased the data without any cleaning. You can see that there are numerous outliers for multiple variables. Outliers can skew and have significant effects on your models. Thus, we removed outliers removing data points that fell outside the bounds of the upper and lower bound.



This was after each variable was cleaned and checked for outliers.

Results

'TenYearCHD' variable was used for training and testing and target variable for training and testing. A linear regression was performed in each feature variable and single features were extracted for training and testing purposes. After, the R-squared and MSE values are calculated for each variable. Next, variables were chosen that had an R-square greater than 0.01 to make a multiple linear regression. Higher the R-square value, the more strongly related they are to the target variable. Additionally, the negative R-squared were not included as they are likely to decrease the fitness of the model. The final multi linear model chose included these features: sex, age, education, currentSmoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, sysBP, diaBP, BMI, glucose, sysBP_arch, glucose_arch , BMI_arch , and diaBP_arch. This was the final multi-linear model we chose with r-square 0.108 which is better than the first model which includes all features, and better than all the models with a couple selected features.

[we should provide visualization of final multi-linear model]

Conclusion

This project aimed to understand the precursors of coronary heart disease (CHD) using predictive algorithms and historical medical data. Our analysis was rooted in linear regression models, creating a model with an R-squared value of 0.108. This indicates that our predictors have a modest influence on CHD risk, yet a considerable portion of variability remains unknown.

Some criticism of our project may include the modest predictive power of our final model and the decisions to exclude outlines. We recognize that the complex nature of CHD's precursors inherently constraint the degree of which a finite selection of variables can predict its likelihood for patients. Furthermore, the choice to remove outliers was intentionally done to enhance the model's representativeness for the general population. These criticisms are valid and acknowledged however, we affirm that our methodology was deliberate and done to reflect real life scenarios.

Looking forward, future research could explore the integration of machine learning algorithms that can encompass nonlinear relationships and interactions between predictors more accurately. Furthermore, additional variables could be incorporated, such as genetic markers and lifestyle data, to exchange predictability capabilities.