Developing a Logistic Regression Model to Predict the Risk of Heart Disease: An Analysis of Key Risk Factors

Bhagyashree Thombare & Jasmine Tran

Introduction

According to WHO, coronary heart disease (CHD) is a major cause of morbidity and mortality worldwide, with an estimated 17.9 million deaths each year[1]. CHD is a complex disease that is influenced by several risk factors, including age, gender, family history, smoking, high blood pressure, high cholesterol, and obesity. These risk factors can interact with each other to increase the risk of developing CHD. The body's cholesterol level, smoking behavior, obesity, family history of illnesses, blood pressure, and work environment are all factors that affect heart disease. According to the literature review named "Cholesterol, coronary heart disease, and stroke in the Asia Pacific region", in both Asian and non-Asian populations in the Asia-Pacific region, total cholesterol is strongly associated with the risk of CHD and ischemic, but not hemorrhagic, stroke[2].

Obesity has been identified to be a significant CHD risk factor. People with a high body mass index (BMI) are more likely to have high blood pressure, high cholesterol, and diabetes. In addition, hypertension poses a significant risk for CHD due to the additional strain it exerts on the heart and blood vessels, which over time causes damage and dysfunction.Blood pressure medication can aid in managing hypertension, but they also involve a risk of CHD-related side effects. Elevated blood glucose levels, a symptom of diabetes, are also associated with an increased risk of CHD. We wanted to research how these lifestyle choices are important to our health because in our daily life we frequently discuss how they affect our physical and mental well-being.

It is crucial to investigate how elements like behavioral patterns and other aspects of nutrition and lifestyle may increase the risk of heart disease given the considerable impact it has on public health and healthcare systems. Medical experts use complex statistical research to determine the likelihood that a patient may develop heart disease. In the field of medicine, data analysis helps with disease prognosis, improved diagnosis, symptom analysis, provision of appropriate medications, improvement of treatment quality, cost reduction, and decrease in the mortality rate of cardiac patients. Predicting and diagnosing cardiac disease, which is based on the patient's symptoms and physical examination, is the primary challenge in the medical sector. The primary challenge in the medical field is predicting and diagnosing heart disease, which is based on the patient's symptoms and physical examination.

In this statistical project, we aim to investigate the effects of specific health behaviors, such as medication adherence, and lifestyle choices on heart disease outcomes. Heart disease is a complicated condition that can be brought on by a variety of variables, such as hereditary, dietary, and environmental ones. The goal of this study is to create a thorough model that pinpoints the main causes of heart disease. Our objective is to construct a logistic regression model that can accurately predict the probability of heart disease diagnosis based on a number of risk variables. We predict that individuals with High BMI, high cholesterol, hypertension, and blood pressure medication use are more likely to develop the risk of coronary heart disease (CHD).

[1] World Health Organization. (2021, June 11). Cardiovascular diseases (CVDs). Who.int; World Health Organization: WHO. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Asia Pacific Cohort Studies Collaboration, Cholesterol, coronary heart disease, and stroke in the Asia Pacific region, International Journal of Epidemiology, Volume 32, Issue 4, August 2003, Pages 563–572

Materials and Methods

The data set used for this project purpose is publicly available on the Kaggle website[3]. It is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The data set provides information about 4238 patients and 16 attributes. Each attribute is a potential risk factor including demographic, behavioral, and medical variables. After some data wrangling and cleaning we removed all rows containing missing values. The column education was eliminated as it did not provide enough information for this study purpose. There were few usual points in the numeric variables, however, they are important for the study and analysis and thus were still included in the study. We mutated sum new variables to be in the factor format (Yes,NO) for analysis purpose and BMI as a categorical groups.

Each participant's demographic characteristics, such as age, gender (male/female), and educational level, are provided. The participant's smoking behaviors, whether or not they smoke, and the average number of cigarettes they consume each day are among the behavioral characteristics. Persons who smoked regularly during the previous 12 months were classified as smokers. Additional information on the patient's medical history is provided by the variables, such as BPMeds, a binary variable that indicates whether or not the patient was taking blood pressure medication, prevalence of stroke, a variable that indicates whether or not the patient had previously experienced a stroke, and prevalence of hypertension and diabetes, both of which are also binary. Additional characteristics including total cholesterol level, systolic and diastolic blood pressure, body mass index (BMI), heart rate, and blood glucose level were taken into account when determining patient's present medical status. Height and weight were measured, and body mass index (kg/mˆ2) was calculated. Two blood pressure determinations were made after the participant had been sitting for at least 5 minutes, and the average was used for analyses. Hypertension was categorized according to blood pressure readings. Diabetes was considered present if the participant was under treatment with insulin or oral hypoglycemic agents if casual blood glucose determinations exceeded 150 mg/dL. These factors were included because they can be measured in a clinical environment and because they have been linked to heart disease in the past.

To determine how each predictor variable relates to the response variable, an exploratory data analysis was conducted. A conditional density plot, a box plot, and summary statistics by TenYearCHD risk were used for the quantitative explanatory variables. A segmented bar chart and the relevant table of proportions demonstrating the connection with TenYearCHD were prepared for categorical explanatory variables. Additionally, the distribution of continuous quantitative variables was observed using a histogram.

A mix of simple and multiple logistic regression analysis was employed to construct a model that could accurately predict the probability of risk of heart disease based on the above risk variables. When a set of predictor variables is used to predict the result of a categorical dependent variable, the statistical method used is known as logistic regression. The dependent variable in logistic regression is always binary. Logistic regression is mainly used for prediction and also for calculating the probability of success. The outcome variable, which is based on a combination of medical and behavioral factors, is the 10-year risk of CHD (binary variable, yes =1, no = 0). A likelihood ratio test will be performed to compare our final model to an intercept-only model to determine the effectiveness of the model we created. A stepwise variable selection and LASSO variable selection method were used, adding or removing variables based on their statistical significance and contribution to the model's predictive ability. Several measurements, including the chi-squared, and drop-in deviance test, were used to evaluate the model's goodness of fit.

Results

According to our general exploratory data analysis, the risk of developing coronary heart disease for those who take blood pressure medication is 33.3% compare to those who do not take BP medication with 14.7%. The same pattern applies to those who previously had a stroke. 38.1% of patients with the history of stroke are likely to develop CHD compared to 15.1% of individuals without previous stroke occurrence. The

[3] Dileep. (2019, June 7). Logistic regression to predict heart disease. Kaggle. Retrieved April 6, 2023

proportion of smokers and non-smokers is approximately the same in both ten year CHD risk and no risk group. The number of cigarettes smoked per day have a slight linear trend according to the conditional density plot. The box plots and summary statistics were used to compare quantitative variables those with and without ten year CHD risk. The patient group with heart disease risk had a higher mean age than those without it, and conditional density plot shows stronger relationship between age and ten year CHD risk. There are many observations in the data set whose cholesterol levels are extremely high. On the interval from 100 to 400 mg/dL, the total level of cholesterol is normally distributed while on the wider internal of (100,600), the distribution is highly right-skewed. Age also highly correlates with the chance of developing coronary heart disease. As age increases, the chance increases, with a relatively positive linear relationship. Additionally, the linearity conditions were checked using conditional density plot and holds true for all our models. (Refer to figure 3-4 in the Appendix)

The correlation matrix between the predictor variables provides information to deduce variables that are highly correlated with each other and may need to be removed from the model to avoid multicollinearity. Systolic BP and diastolic BP have strong positive correlation of 0.787, which explains their absence from our analysis. The BMI (0.331) and age (0.389) seem to have moderate positive correlation with systolic blood pressure. We predict this correlation might impact the contribution of BMI and age as predictor variables to determine the risk of ten year CHD. (Refer figure 3 in the Appendix)

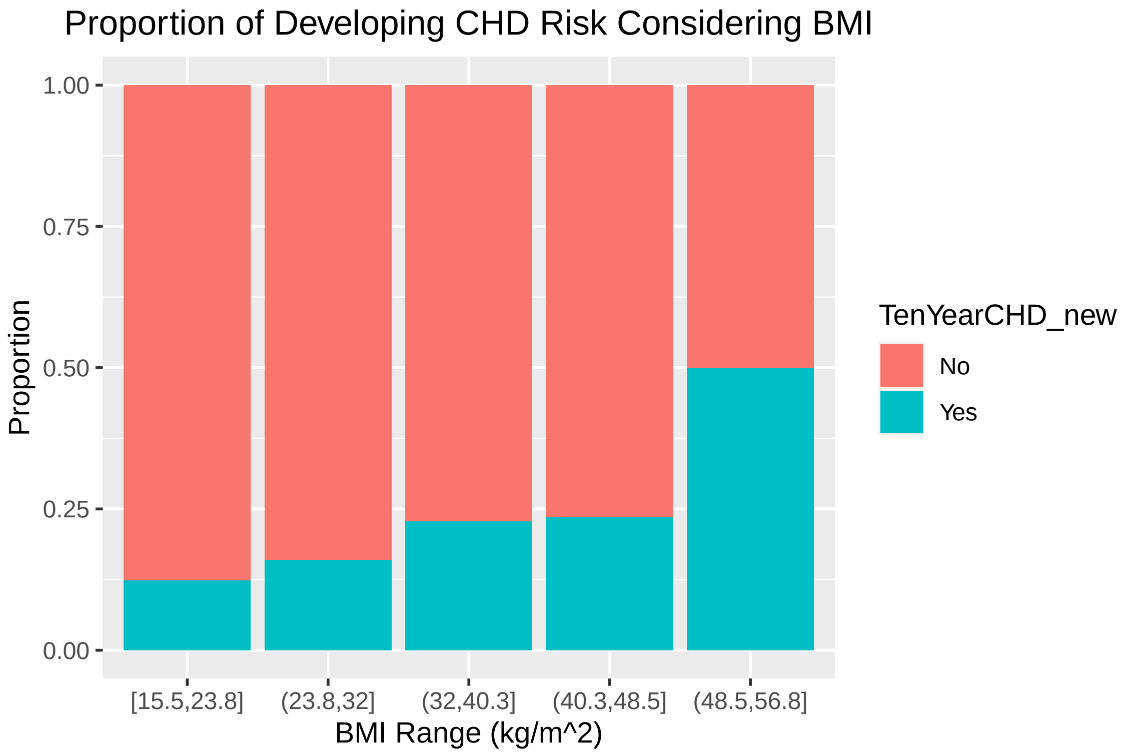


Figure 1: Proportion of developing coronary heart disease by BMI range.

Figure 1 evaluates the likelihood of developing CHD risk while taking BMI into account. There are five categories for BMI, and it shows that the chance of having CHD rises with increasing BMI ranges. The patient's in BMI range of 48.5-56.8(kg/m^2) have 50% chance of developing CHD risk. These BMI values far exceed the normal range and fall under the obese category.

Table 1: Depicts the unadjusted and adjusted odd ratios for chosen predictor variables. The *** indicates the significance at 0.05 level.

	Unadjusted odds ratio	Adjusted odds ratio
BMI	1.054***	1.015
BPMeds	2.908***	1.327
prevalentHyp	2.730***	1.834***
age	1.081***	1.069***

The simple logistic regression analysis carried out for BMI, BPMeds, prevalentHyp, and age variables shows that all of them are significant contributors at 0.05 level. According to the simple logistic regression analysis, blood pressure medication (BPMeds) users have a 2.91 time increased chance of developing CHD over the course of ten years than non-users. When compared to people without a history of hypertension, those with a prevalence of hypertension have 2.73 times higher chance of receiving a ten-year CHD diagnosis than a person without it. The CI-based profile likelihood test provides 95% confidence that each one kg/m2 increase in BMI is associated with between a 3.19% and a 7.61% increase in the odds of being diagnosed for ten-year CHD. A one year increase in age is associated 1.08 times increase chance of being diagnosed with ten-year CHD.

Furthermore, multiple logistic regression carried out using the same variables showed that BMI and blood pressure medication (BPMeds) are not significant anymore. This might be because additional variables in the model such as hypertension prevalence and age is accounting for the effect of them. The multiple regression analysis showed that hypertension prevalence is associated with 1.83 times increase in chance of being diagnosed for ten-year CHD after accounting for BMI, BPMeds and age. After accounting for age, BMI, and BP medications, hypertension appears to be a strong significant predictive variable at the 0.05 significance level.

Furthermore, the correlation matrix indicated that some variables are associated and might interact with each other. To further observe if these correlation impact the model predictions, we created a model adding interaction between BMI and systolic BP. The interaction model gave significant p-values for the BMI, prevalentHyp, age and the interaction term at 0.05 significance level(Appendix). A drop in deviance test provides statistically significant evidence with G value of 16.97 and p value 3.70e-05 that adding the interaction between BMI and systolic BP improves our ability to explain variability in patients' probability of being diagnosed for ten-year CHD risk. Hence, we consider systolic BP as one of our important predictor variable in determining the risk of being dignosed for ten-year CHD.

Additionally, to build an appropriate model, the LASSO variable selection method was employed. It helped to identify the most important predictors for ten year CHD risk. This is particularly useful since this study deals with high-dimensional data sets with a large number of potential predictors.

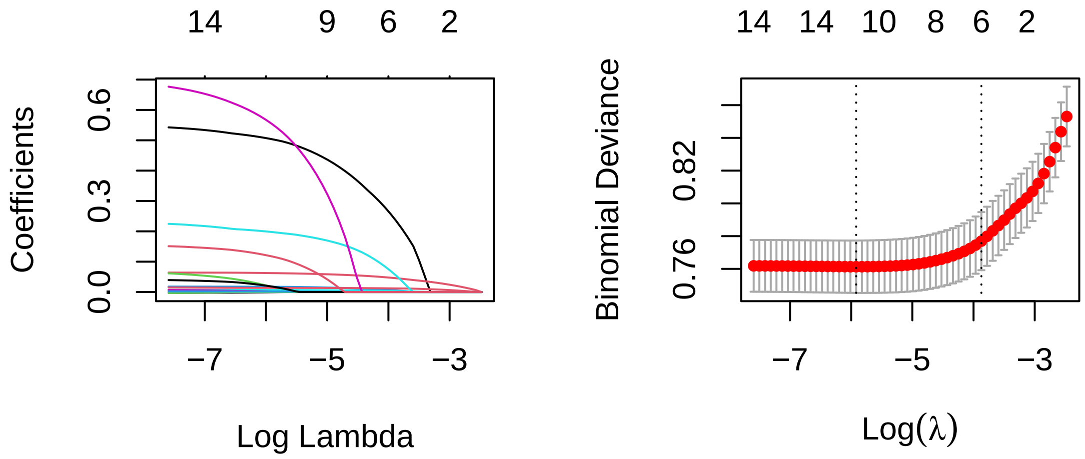


Figure 2. LASSO Variable Selection Model

Our LASSO model prediction shows 14 predictor variables at the start and then the number of predictor variables decrease. According to the LASSO model gender, age, cigarettes per day, glucose level, systolic blood pressure, and prevalence of hypertension explains the variability in ten year CHD risk diagnosis the most. Stepwise selection method also provides similar results and include predictor variables approximately same as the LASSO method. However, the stepwise model includes total cholesterol and prevalent of stroke which are not part of LASSO model, instead it consists of systolic blood pressure.(refer to table 3 in the Appendix)

The difference in results is due to the varied degree of penalties for different methods. However, in our data set there are few variables that overlap such as current smoker and cigarettes number, glucose and diabetes, systolic and diastolic blood pressure. All these variables are somewhat correlated and this relationship affect the effectivness of the given predicor variable in the model.

Discussion

In the present study, we developed a simple logistic regression model to predict the probability of risk of heart disease based on several categorical risk factors, including high BMI, high cholesterol, hypertension, and BPmeds use. Our findings indicate that these risk factors are significant predictors of heart disease risk, with high BMI and hypertension showing the strongest association with the outcome variable.

According to our LASSO variable selection approach, age, gender, cigarettes per day, glucose level, systolic blood pressure, and hypertension prevalence appear to be significant factors in the prediction of risk of coronary heart disease. Males have a larger risk of developing heart disease than females do, and coronary heart disease is more common in men. Current literature suggests that this is because men have worse coping mechanisms (physiologically, behaviorally, emotionally) that lead to reduced adaptability to stressful situations as compared to females, increasing their risk for CHD[4]. It's interesting to note how the LASSO model does not include some of the factors that we believe to be important predictors, such as BMI and BPMeds. It is likely that additional variables in the model are accounting for the effect of these removed variables on the result. For instance, there is frequently a strong correlation between BMI and other indicators including age, gender, and cholesterol levels. The effect of BMI can be redundant and so removed by the

[4] Weidner.(2000).Why Do Men Get More Heart Disease Than Women? An International Perspective. Journal of American College Health, 48(6), 291–294. https://doi.org/10.1080/07448480009596270

LASSO because the other variables are already present in the model and are highly predictive of the risk of ten year CHD.

The results of our logistic regression model suggest that individuals with high BMI and hypertension are at a significantly higher risk of developing heart disease. This is consistent with previous research, which has shown that obesity and hypertension are major risk factors for cardiovascular disease (CHD) and contribute to the development of other chronic conditions such as type 2 diabetes, stroke, and heart attack[5]. In addition, our study found that individuals with high cholesterol and BPmeds use were also at a higher risk of heart disease, although the effect sizes were smaller than those for BMI and hypertension.

Based on these important risk factors, our logistic regression model offers a useful tool for medical professionals to evaluate a person's risk of heart disease. Healthcare professionals can create targeted therapies to lower the risk of CHD and stop negative consequences by identifying people who are at high risk of heart disease.

Despite the fact that our study has several advantages, such as a large number of samples and extensive statistical analysis, there are some drawbacks to take into account. First off, the demographics in this study's data were not diverse in terms of ethnic and racial groups because they came from a specific region This limits our ability to generalize the data from the Massachusetts town of Framingham to a larger population. Furthermore, as this study is an observational one, we are unable to draw any conclusions about causality from it. Furthermore, other potential confounding variables including family history, physical activity, and diet that may affect the risk of heart disease were not taken into account in our study. The ongoing study of Framingham town now started considering these factors and have larger diverse population which potentially will increase the credibility of the study.The ongoing study of Framingham Town has now begun to take these issues into account and has a larger, more diversified population, which may enhance the study's credibility.

[5] Cercato, & Fonseca, F. A. (2019). Cardiovascular risk and obesity. Diabetology and Metabolic Syndrome, 11(1), 74–74. https://doi.org/10.1186/s13098-019-0468-0

Reference

1. Asia Pacific Cohort Studies Collaboration, Cholesterol, coronary heart disease, and stroke in the Asia Pacific region, International Journal of Epidemiology, Volume 32, Issue 4, August 2003, Pages 563–572.

2. Dileep. (2019, June 7). Logistic regression to predict heart disease. Kaggle. Retrieved April 6, 2023.

3. Dinas, Koutedakis, Y., & Flouris, A. D. (2013). Effects of active and passive tobacco cigarette smoking on heart rate variability. International Journal of Cardiology, 163(2), 109–115.

4. Karason, Mølgaard, H., Wikstrand, J., & Sjöström, L. (1999). Heart rate variability in obesity and the effect of weight loss. The American Journal of Cardiology, 83(8), 1242–1247.

5. Participants | Framingham Heart Study. (n.d.)Www.framinghamheartstudy.org. https://www.framinghamheartstudy.org/participants/

6. Weidner. (2000). Why Do Men Get More Heart Disease Than Women? An International Perspective. Journal of American College Health, 48(6), 291–294. https://doi.org/10.1080/07448480009596270

7. World Health Organization. (2021, June 11). Cardiovascular diseases (CVDs). Who.int; World Health Organization: WHO. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

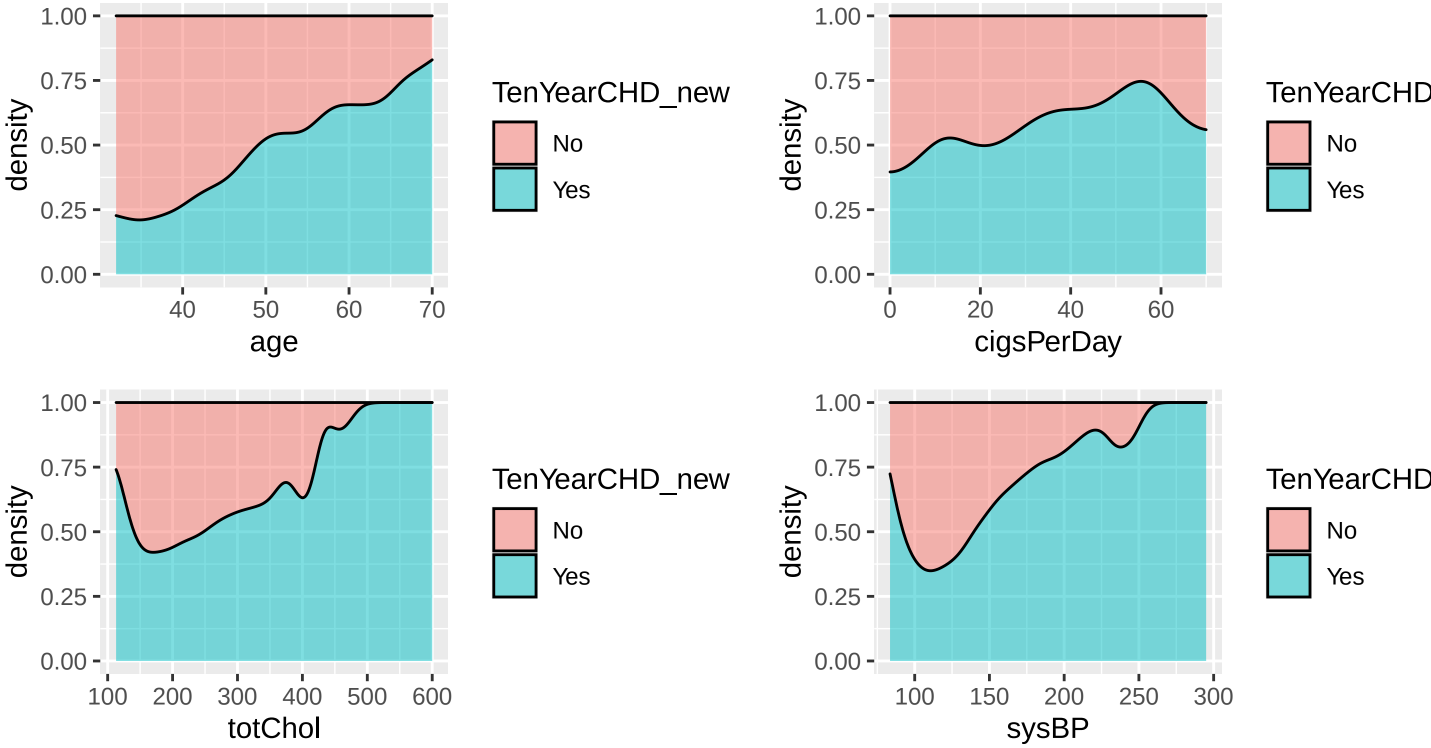
Appendix

Table 2: Data set and variable overview: Explanations of variables

Name	Variable Role	Type	Values	Units
male	Potential confounder	Categorical	0 or 1	NA
age	predictor	Quantitative	30-70	years
education	Potential confounder	categorical	1-4 or NA	NA
currentSmoker	Predictor	Categorical	0 or 1	NA
cigsPerDay	Predictor	Quantitative	0 - 70	NA
BPMeds	Predictor	Categorical	0 or 1	NA
prevalentStroke	Predictor	Categorical	0 or 1	NA
prevalentHyp	Predictor	Categorical	0 or 1	NA
diabetes	confounder	Categorical	0 or 1	NA
totChol	confounder	Quantitative	107-693	mg/dL
sysBP	predictor	Quantitative	83.5-295	mmHg
diaBP	Confounder	Quantitative	48-142.5	mmHg
BMI	Predictor	Quantitative	15.54-56.8	kg/m^2
heartRate	Response	Quantitative	44.143	BPM
glucose	Predictor	Quantitative	40-349	mg/dL
TenYearCHD	Response	Categorical	0 or 1	NA

Figure 3. Correlation Matrix (EDA-linearity)

Figure 4. Conditional Density Plots For Quantitative Variables (EDA-linearity)



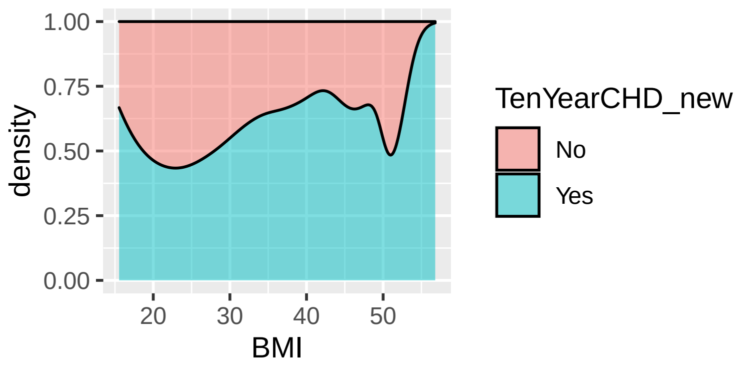


Figure 4. Segment Bar Charts For Categorical Predictors in EDA

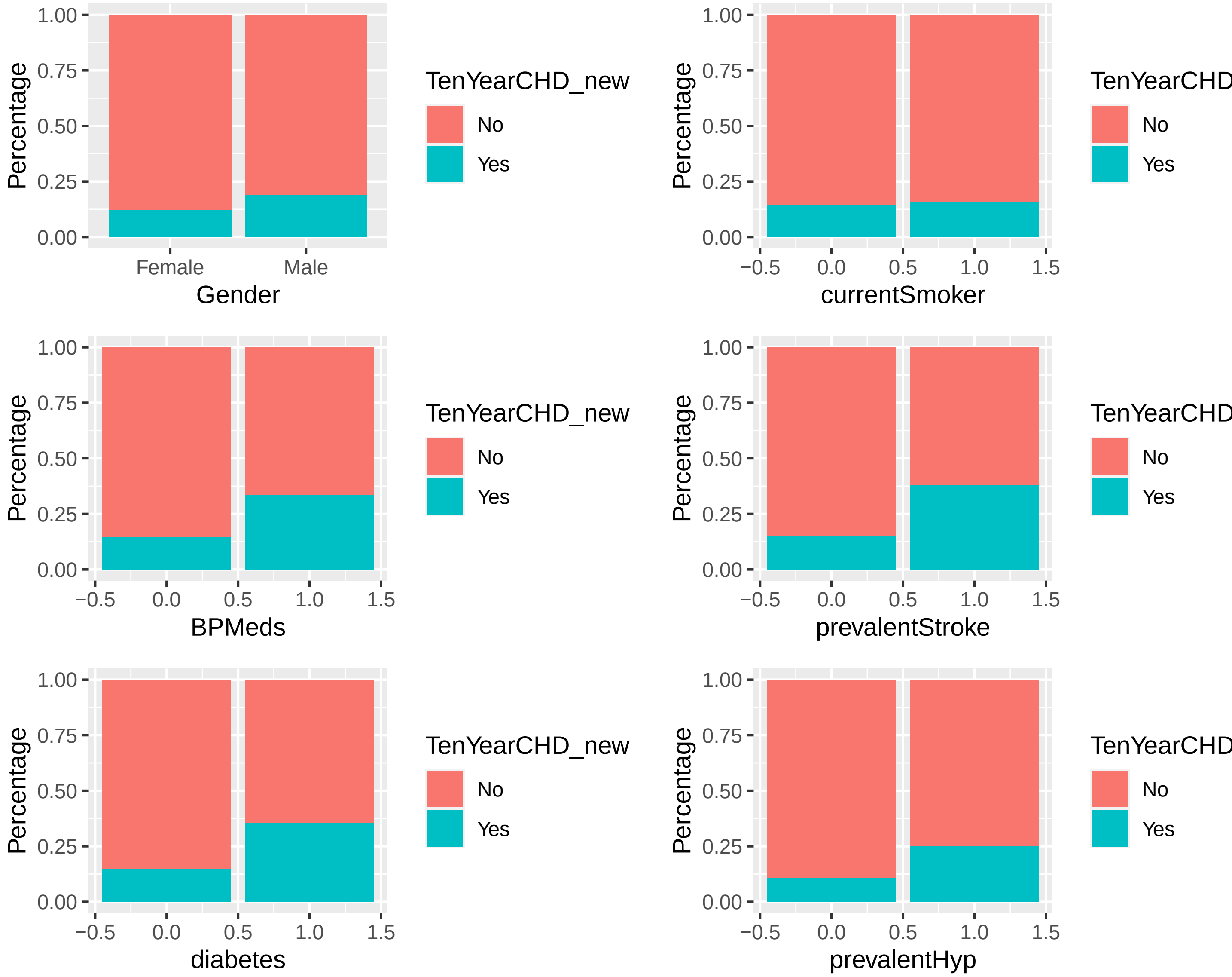


Table 3: LASSO vs Stewise Variable Selection Results

Variable Name	LASSO Coefficient	Stepwise Coefficient
GenderMale	0.181	0.497
age	0.042	0.071
cigsPerDay	0.002	0.019
glucose	0.002	0.007
sysBP	0.0114	.
prevalentHyp	0.0173	0.667
totChol	.	0.003
prevalentStroke	.	0.782