Luvleen Rehal
1005770131

Introduction

Heart disease is a variety of conditions that affect the heart and blood vessels. The most common type is coronary artery disease which affects the blood flow to the heart, but other diseases include arrhythmia, heart failure and valve problems. Understanding the impact of heart disease and its influencing factors is crucial for individuals who have witnessed its effects. Several factors that increase heart disease risk include blood sugar, age, blood pressure, st segment, cholesterol, max heart rate, sex, major vessels, old peak. Predicting the relationship between these factors can help individuals detect the disease earlier. A healthy lifestyle is crucial in maintaining a healthy heart and recovering from heart disease. Regular medical checkups and awareness of known risk factors also play a significant role in preventing heart disease. This topic is essential for everyone to grasp as it affects a large population, and raising awareness can prevent this from increasing.

The association between chest pain and heart disease is essential to understanding the relationship because it can lead to more significant risks. Individuals who experience chest pain are likely to get a heart attack. If it is noticed earlier on, timely medical attention can help. According to the "British Journal of General Practice," chest pain is the most crucial symptom of ischaemic heart disease (Nilsson et al., 2003). They completed a study which resulted in 8 percent of patients being diagnosed with ischaemic heart disease that experience some heart pain. This number may not seem significant, and most patients were not diagnosed, but it can still help in interpreting the chest pain level with a disease as it does occur.

The relation between blood pressure and the risk of heart disease is analyzed with research in which 2000 coronary deaths occurred in 6 years after these men were researched who had no prior history of heart attacks (Stamler et al., 1989). The risk increased among hypertensive men with high cholesterol levels (Stamler et al., 1989). This number could have been reduced if the correct treatment had been given and the prediction of heart diseases had been given.

These two academic papers provide a deep insight into the impact of heart disease on individuals. They share the importance of chest pain which most people may ignore at first but can cause severe damage and influence high blood pressure and cholesterol. This paper will answer the prediction of heart disease with chest pain.

Methods

Dataset

Addressing multicollinearity and splitting the data into training and test sets are essential steps in ensuring the validity and reliability of the predictive model. Variables that are highly correlated or uncorrelated with one another that are not analyzed in the prediction are removed. The data is cleaned, and variables that do not have an impact on the data or assessed in this study are removed. The prediction of heart disease is made with 270 samples with 10 independent variables that age, gender, chest pain, resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate, old peak, ST segment, thal, and major vessels as seen in Table 1.

Also, the data is split into two sets which are the training and test set. The training set trains the predictive model, while the test set evaluates the model's performance. This will allow us to assess the model’s performance and is useful in determining a prediction. It is crucial to address multicollinearity to achieve more precise results because it is a factor in inflating the variance and errors of data. This can influence variables which are important in the model, and which are not. Multicollinearity can also lead to a misleading interpretation of the relationship between predictors, and the outcome can be overfitted, which can change the whole result of the correlation between heart disease and its predictors.

Table 1: Summary Statistics of Variables in the “Heart Disease Prediction Dataset” (Singh et al., 2023)

Variable Name	**Code**	**Description**
Age	age	Age of the person
Sex	sex	Sex of the person
Chest pain type	chest.pain.type	Chest pain type (4 values)
Serum cholesterol	serum.cholestrol	Serum cholesterol in mg/dl
Fasting blood sugar	resting.blood.sugar	If its bigger than 120 mg/dl
Maximum heart rate	max.heart.rate	Maximum heart rate achieved
oldpeak	oldpeak	ST depression caused by activity in comparison to rest.
ST segment	ST.segement	Slope of the peak exercise ST segment. Part of the ECG that is used to find electrical activity of the heart.
Number of major vessels	major.vessels	The number of vessels colored by fluoroscopy, a medical device. Used as an indicator for coronary heart disease.
Thal	thal	A cardiac imaging test used to measure the blood flow to the heart.
Heart Disease	heart.disease	1. Absense 2. Presence

Linear regression models

Linear regression models are used to arrive at the prediction. In this case, heart disease prediction is based on factors such as age, gender, chest pain, resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate, old peak, ST segment, and major vessels. Eliminating a predictor in each model with the highest p-value allows for a model that best fits the data. The final model has the most crucial predictor and is an important method to validate the prediction of heart disease. Comparing all the linear models in one graph helps easily see which predictors have the most significant impact.

AIC and BIC

Analyzing the AIC and BIC is also important to compare different models based on their goodness of fit. A lower AIC and BIC usually indicates that the model has avoided overfitting and has captured the correct relationship between predictors. In the report, the AIC and BIC are used to determine which model best fits the prediction of heart disease and which predictors have the most impact on the data.

P - Value

The p-value helps to measure the significance of the relationship or prediction, in this case is the prediction of heart disease with chest pain. To determine the relationship, the p-value allows to either accept or reject the null hypothesis.

Model Evaluation

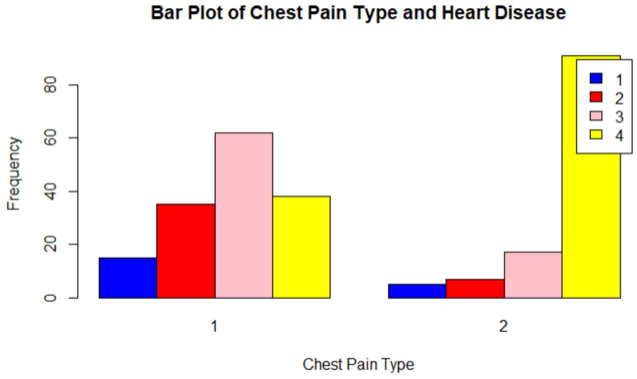
To validate the model, several techniques can be used. This includes splitting the data into a training set and a testing set. Then, fitting the model allows us to achieve a good model that answers the question and has a relation—making predictions on the testing data. Also, evaluating the model performance in which the predicted values are compared with the actual values from the testing set can be done using mean square error and R^2.

Results

Data Preprocessing and Multicollinearity

In the prediction of heart disease, certain factors such as electrocardiographic results and exercise-induced angina are taken into account in but will not be specifically addressed in our study. The model mainly considers medical conditions that lead to heart disease. Electrocardiographic results include heart rate, another predictor in the data, so there is overlapping with this. As for exercise-induced angina, it does not accurately represent a prediction for all individuals concerning their age because some people may not have regular physical exercise. The prediction of heart disease can still be concluded without these two variables. Chest Pain Type is associated with heart disease by Figure 1. It highlights the frequency of different types of chest pain in people with no heart disease (1) and with heart disease (2). Individuals that have heart disease have a higher chest pain of type 4 which proves the research question. The higher the chest pain, the more likely the individual is to have heart disease.

Figure 1: Bar Plot of Chest Pain Type and Heart Disease

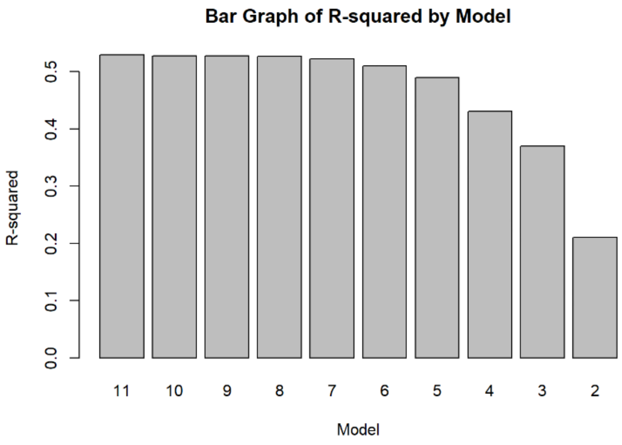


Linear Regression model

After validating and removing predictors from each model, it concluded that chest pain is the predictor with the most significant impact on heart disease. Through a stepwise approach, one predictor with the highest p-value was eliminated at each stage, ultimately leading to the identification of the most significant model. This answers the question of the relationship between chest pain and heart disease because it is an essential predictor in achieving the correct results.

Looking at figure 2, it is visible that the data had an impact three times, with one that was very significant. Models eight and nine were impacted, but the most significant is model 10, which is the linear model concerning chest pain and heart disease. Model eight is lm(heart.disease~chest.pain.type+oldpeak+major.vessels, adisease) and model nine is lm(heart.disease~chest.pain.type+oldpeak, adisease). Major vessels contribute to the data because removing them causes the model to drop. However, model 10 removes the old peak, another contributor to the relationship, but removing it gives the chest pain predictor high significance. These are the most important factors that can help predict heart disease in an individual.

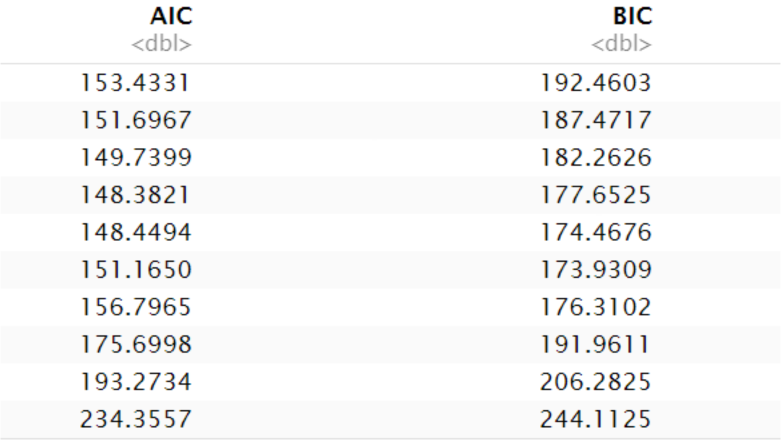
Figure 2: Bar Graph of All Models with R-Squared Values



AIC and BIC

Table 2 is used to analyze the AIC of each model after removing a predictor, if the AIC is higher, it indicates that the predictor contributed to the model fit. The AIC of model six increases, proving that the predictor removed previously, the ST segment has a good relation with heart disease. All the other predictors after this have the same impact but model ten has the greatest increase in AIC with the removal of the predictor's old peak. Relating the concept of AIC with the prediction of heart disease, it is clear that blood sugar, age, blood pressure, st segment, and cholesterol are less crucial in predicting heart disease. However, max heart rate, sex, major vessels, and old peak are good indicators in the prediction based on the models. The BIC is also a factor in determining which is the most helpful predictor for heart disease. The BIC consistently decreases until model seven. The predictor that is removed at this step is max heart rate. Followed by this are sex, major vessels, and old peak, which have the greatest increase in BIC.

Table 2: AIC and BIC Values of all Models



AIC <dbl>	BIC <dbl>
153.4331	192.4603
151.6967	187.4717
149.7399	182.2626
148.3821	177.6525
148.4494	174.4676
151.1650	173.9309
156.7965	176.3102
175.6998	191.9611
193.2734	206.2825
234.3557	244.1125

P - Value

There is a strong association between chest pain and heart disease, as evidenced by the low p-value of 2.604e-11. The confidence interval can be calculated as 1-2.604e^-11 = 0.9999 = 99 percent confidence interval. The null hypothesis is no relationship between chest pain and heart disease. Given the confidence interval, this is rejected as there is strong evidence against it. Thus, the p–value indicates a significant relationship between chest pain and heart disease and can be used as an indicator for prediction.

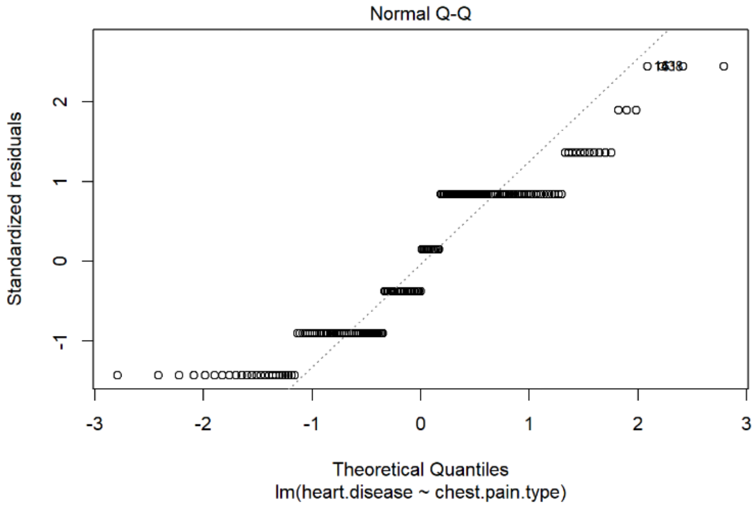
Model Evaluation

The data has been split into the training set which is adisease and the testing set which is adisease.testing. Fitting the model using the training data is though lm() for linear regressions and because the prediction is between chest pain and heart disease, it came to a conclusion that chest pain had the largest significance. Evaluating the model's performance, the mean square error (MSE) is calculated to be 0.2207785. A lower MSE indicates a better performance, suggesting that the model provides relatively accurate predictions. The R^2 is found to be 0.0802825 and indicates a good fit.

Normal QQ plot

The normal qq plot assesses the assumption of normality and compares the theoretical quantiles with the standardized residuals. The points follow a straight line which indicates that the data is close to a normal distribution of data and the points are near the line. There are a few outliers that should be accounted for.

Figure 3: The Normal QQ plot of Heart Disease and Chest Pain Type



Discussion

The data collected from this study suggests that chest pain has a strong relationship with heart disease. As the chest pain type increased so did the population of individuals that have heart disease. The other models all provide a deep correlation with heart disease as well such as max heart rate and sex which was determined by calculating the AIC and BIC. By eliminating each predictor in a model and arriving at the final model of chest pain, it was a clear indicator that this was the most significant model and analyzing this is my study can help others realize the important of chest pain.

Bibliography:

Nilsson, S., Scheike, M., Engblom, D., Karlsson, L. G., Mölstad, S., Åkerlind, I., Åkerlind, K., & Nylander, E. (2003, May). Chest pain and ischaemic heart disease in primary care. https://bjgp.org/content/bjgp/53/490/378.full.pdf

Singh, U. (2023, May 26). *Heart disease prediction dataset*. Kaggle. https://www.kaggle.com/datasets/utkarshx27/heart-disease-diagnosis-dataset?resource=download

Stamler, J., Neaton, J. D., & Wentworth, D. N. (1989). Blood pressure (systolic and Diastolic) and risk of fatal coronary heart disease. *Hypertension*, *13*(5_supplement). https://doi.org/10.1161/01.hyp.13.5_suppl.i2