Heart Disease Prediction

Kavleen Kaur

A B S T R A C T

The diagnosis of heart disease in most cases depends on a complex combination of clinical and pathological data. Because of this complexity, there exists a significant amount of interest among clinical professionals and researchers regarding the efficient and accurate prediction of heart disease. In this paper, we have develop a heart disease prediction algorithm that can assist medical professionals in predicting heart disease status based on the clinical data of patients. Our approaches include two steps. Firstly, we select important factors from these 13 clinical factors., i.e., age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ecg, maximum heart rate, exercise induced angina, depression, slope, number of vessels colored, and form of thalassemia. Secondly, we have developed a model using binary logistic algorithm for the classifying of heart disease based on the selected important clinical factors. The results verify that the binary logistic regression algorithm has achieved a model which is 85.56% efficient.

1. INTRODUCTION

Human heart is the principal part of the human body, it regulates blood flow throughout our body. Any sort of disturbance to normal functioning of the heart can be classified as a heart disease. It is a tragic event which results the blocking of blood flow. People at risk of heart disease may show elevated blood pressure, glucose, and lipid levels as well as stress. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Looking at the general statistic data regarding heart disease, we can state that a person dies from heart disease every 33 seconds, every 20 seconds, one person has a heart attack and everyday millions of people die of heart disease. According to the World Health Organization more than 10 million die due to heart diseases every single year around the world. A healthy lifestyle and early detection are only ways to prevent the heart related diseases.

Prediction of heart disease is regarded as one of the most important subjects in the section of clinical data analysis as predicting heart disease is one of the most complicated tasks in medical field. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Records of large set of medical data created by medical experts are available for analysing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available that contribute to successful decision making. This plays a key role for healthcare professionals in making accurate decisions and providing quality services to the public. One of the main limitations of existing methods is the ability to draw accurate conclusions as needed. In our approach, we are using a data mining technique, binary logistic regression to predict the heart disease based on some health parameters.

1.1 HEALTH PARAMETERS CONTRIBUTING HEART DISEASE:

Genetics can play an role in cardiovascular health, but so can lifestyle changes. Some health parameters related to heart disease are chest pain (location and type), cholesterol levels, blood pressure (resting and while exercising), heart rate, flow of blood through the coronary

arteries. These are some major health parameters that are responsible for the presence of a heart disease in a patient. Apart from these, other factors can be the age or sex of the person. Heart diseases related to these health parameters are blood vessel disease, such as coronary artery disease, heart rhythm problems (arrhythmias), heart defects you're born with (congenital heart defects), heart valve disease, disease of the heart muscle, heart infection. Many of these forms of heart disease can be prevented or treated with healthy lifestyle choices.

1.2 OBJECTIVE:

In this paper, we develop a heart disease prediction model that can assist medical professionals in predicting heart disease status based on the clinical data of patients. This model generates prediction results using binary logistic regression technique. The aim of this is to determine the crucial factors for the presence of heart disease in a patient and build a model that can help us predict heart disease status in a patient by analysing a clinical dataset.

2. MATERIALS AND METHODS

2.1 DATA COLLECTION:

The dataset used was the Heart disease Dataset, which is a combination of 14 different database, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes, but all published experiments refer to using a subset of only 14 features. Therefore, we have used the UCI Cleveland dataset available in the Kaggle website for our analysis. The complete description of the 14 attributes used in the proposed work is mentioned in Table 1 shown below. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and pre-processed before it is submitted to the proposed algorithm.

The UCI Machine Learning Repository is freely available, is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

2.2 DATASET DESCRIPTION:

Given the discussion in the section 1.1, we focus on following health factors age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ecg, maximum heart rate, exercise induced angina, depression, slope, number of vessels colored, and form of thalassemia. We use the UCI Cleveland dataset available in the Kaggle website for our analysis. The dataset consist of 14 attributes in total and their description is mentioned in **Table 1**. We did perform descriptive analysis to explore our dataset. So, our dataset contains some categorical and numerical data. For categorical variable like chest pain type, fasting blood sugar, resting ECG, form of thalassemia, sex, and depression we can see the descriptive analysis in **figure1.** The figure 1 shows the effect of each type of these factors on the status of heart disease in the patient. It clearly illustrates the conditions that might be prevalent for a person with heart disease. Factors such as non-angina chest pain, the form of

thalassemia having fixed defects, Resting ECG showcasing a ST-T wave abnormality are some indicators for the presence of heart disease. For other numerical factors, we can see boxplots in **figure2** demonstrating the calculated medians, first and third quartiles.

Table 1.

Attributes	Variable name (in dataset)	Description	Values
Age	age	Patient's age in years	Continuous Values
Sex	sex	Sex of the Patient	1- Male 2-Female
Chest pain type	cp	Chest pain type and location	1-Typical angina 2-Atypical angina 3-non-angina pain 4-Asymptomatic
Rest Blood Pressure	trestbps	Resting blood pressure	Continuous Values in mm/Hg
Serum Cholesterol	chol	Serum cholesterol	Continuous Values in mg/dl
Resting ECG	restcg	Resting electrocardiographic results	0-Normal 1-Having ST-T wave abnormality 2-Left ventricular hypertrophy
Fast Blood Pressure	fbs	Fasting blood sugar	1. >= 120 mg/dl 2. <= 120 mg/dl
Max Heart Rate	thalach	Max heartbeat of patient	Continuous Values
Exercise Induced Pain	exang	Exercise induced angina	1- Yes 2- 2-No
Depression	oldpeak	Patients' depression level.	Continuous Values
Slope	slope	Patient condition during peak exercise	1- Upsloping 2-Flat 3-Down sloping
Vessels Under Fluro	ca	Number of vessels under Fluro	Continuous Values
Form of thalassemia	thal	Defect type	3- Normal 6- Fixed defects 7 Reversible defects
Heart disease	target	Heart disease status	1- Absence of heart disease 2- Presence of heart disease

Figure1

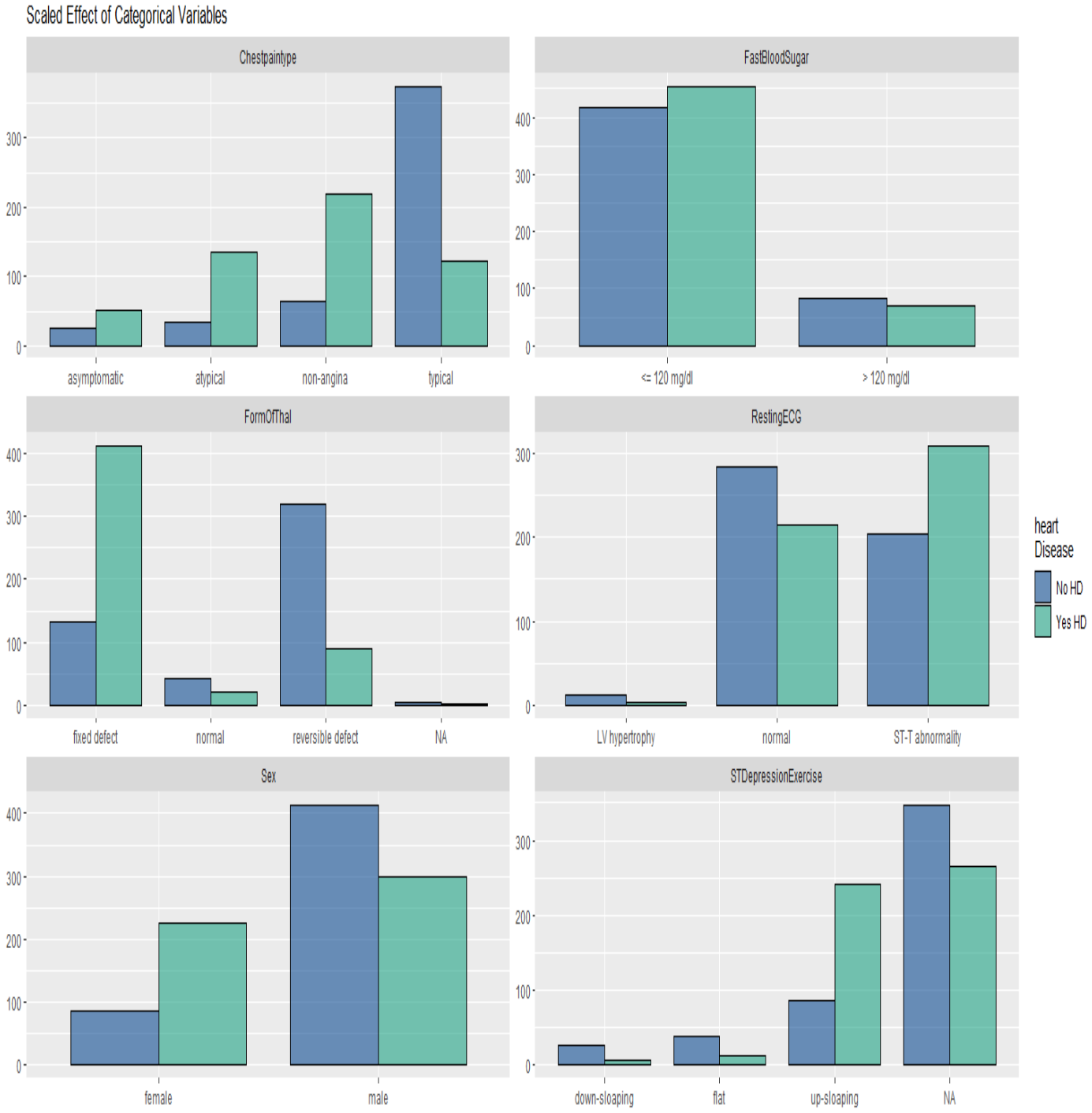


Figure1: shows the effect of each type of these factors on the status of heart disease in the patient. It clearly illustrates the conditions that might be prevalent for a person with heart disease. Factors such as non-angina chest pain, the form of thalassemia having fixed defects, Resting ECG showcasing a ST-T wave abnormality are some indicators for the presence of heart disease

Figure2:

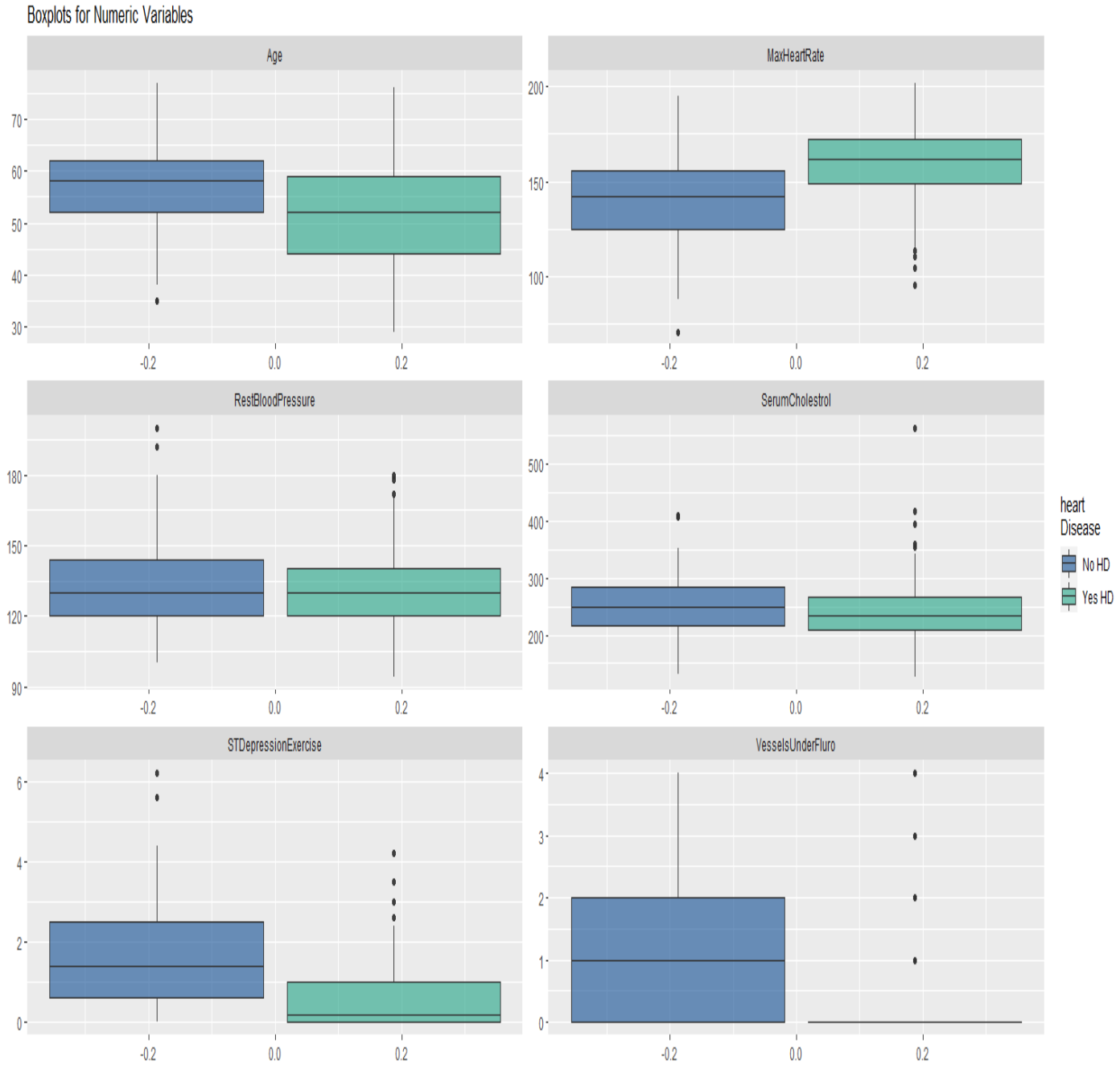
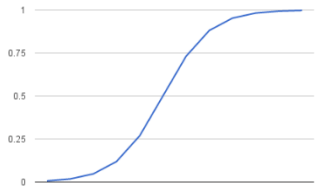


Figure2: shows boxplots for numerical variables demonstrating the calculated medians, first and third quartiles.

2.2 METHOD DESCRIPTION:

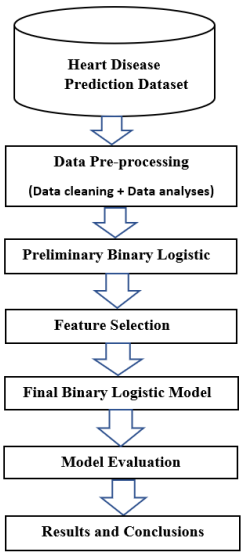
To elucidate the role of the independent variables on the response variable (presence of heart disease), we will deploy a binary logistic model. Binary logistic regression is a classification algorithm that measures the relationship between the categorical target variable and one or more independent variables. It is useful for situations in which the outcome for a target variable can have only two possible types. Binary logistic regression classification makes use of one or more predictor variables that may be either be continuous or categorical to predict the target variable classes. This technique helps to identify important factors (Xi) impacting the target variable (Y) and the nature of the relationship between each of these factors and the dependent variable. In logistic regression, instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1(**Fig 2.2.1**). Binary logistic regression good for this classification as the dependent variable here is dichotomous or binary in nature (presence or absence of heart disease in a patient) also this model helps us assess how well our set of variables predicts our categorical dependent variable and determine the "goodness-of-fit" of our model, also it provides a summary of the accuracy of classification of cases, which helps us determine the percent of predictions made from this model that will be correct.

Fig 2.2.1



Considering the assumptions for a binary logistic model, the dataset is feasibly small and there is no or less multicollinearity between the variable, the method adopted for this project is to perform explanatory analysis on a clean dataset followed by applying preliminary modelling to determine key factors effecting the status of the heart disease in a patient using the backward model selection technique, eliminating the non-important factor to develop a final model. At the end, the project aims to determine the accuracy and efficiency of the model.

Fig 2.2.2



2.3 ANALYTICAL APPROACH:

2.3.1 DATA ANALYSIS:

After cleaning the dataset by checking the missing values and transforming the dataset as required, we performed the Shapiro-wilk normality test to test whether the given variables were normally distributed or not. Once the test yielded statistically significant, we could infer that almost all variables showed skewness in their distribution i.e., were not normally distributed. After normally distributing the variable, we ran the binary logistic model which resulted with higher AIC value (Akaike Information Criterion). Hence, ran the model without normally distributing the variable which resulted in developing a much better model with very low AIC value compared to the model with normally distributed variables.

2.3.2 CORRELATION ANALYSIS:

To quantify the association between the dependent and the independent variables, we have computed the pairwise correlation coefficients for all the variables. The coefficient matrix **Table 2.,** generated is a summarized comparison between the variables which demonstrates not only the correlation of dependent variables with the independent variables but also illustrates that there is no multicollinearity between the variables. As discussed above, low correlation (absence of multicollinearity) among the variables is one of the major assumptions for developing a binary logistic model.

2.3.2 CONSTRUCTING BINARY LOGISTIC MODEL WITH HEART DISEASE AS A RESPONSE VARIABLE:

As discussed in section 2.2, the reason of choosing and the basic methodology of this model, here we will discuss in detail the steps to develop this model. The correlation analysis though gives us a basic understanding of the relationship of the dependent and the independent variables, here we will determine a preliminary model to identify key important factors affecting the status of heart disease. Before proceeding, let us know the dependent and the independent variables. Here, the model initially will take dependent variable as heart disease, as we will be studying the effects on other variables-the independent variables on the status of heart disease stored in the variable “heart disease” (0- absence of heart disease 1-presence of heart disease). The independent variables are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ecg, maximum heart rate, exercise induced angina, depression, slope, number of vessels colored, and form of thalassemia.

Starting from the preliminary model, we propose procedure outlined below for our final model:

1) The preliminary model developed is not the best model given the set of independent variables.
2) Thus, for model selection as there are multiple methodologies for variable selection, we will use the backward model selection (step AIC function in R: the ‘stepAIC’ function in R performs a stepwise model selection with an objective to minimize the AIC value.). model selection based on Akaike's Information Criterion, which is calculated as AIC = 2k - 2lnL, where k is the number of parameters estimated by the model and lnL is the log likelihood of the data given the model. Basically, model

selection with AIC attempts to select the model that best explains the data (highest likelihood), while still not fitting too many parameters. This basic approach is agnostic to Cross Validation.

3) The process is repeated until, we can no longer achieve a model with lower AIC value, thus obtaining the final model.
4) The goodness of the model is then compared by the AIC value of the models.
5) The final model is then analysed by calculating its efficiency and the model's predictive power.

2.3.3 CODING LANGUAGE AND LIBRARIES USED:

For our coding, we used R language along with the following libraries in our coding: dplyr, tidyr, ggplot2, corrplot, rcompanion, car, mlbench, MASS, pROC, car.

3. RESULTS:

3.1 RESULT OF CORRELATION ANALYSIS:

In this section, we provide the results of correlation analysis between the variables. In **Table 2,** we see the Pearson correlation between the variables. From the table, we can see that most of the variables are negatively correlated. To be precise, for variable heart disease, we can say that it is strongly correlated with chest pain, exercise induced pain, slope, and the form of thalassemia. Thus, according to the correlation analysis, we can infer that the status of heart disease in a patient is dependent on sex of the patient, chest pain type, exercise induced angina, peak during exercise (slope) and the form of thalassemia. To understand the relationship and perform an in-depth analysis, we will run a binary logistic regression model technique to accurately determine the relationships between the dependent and the independent variables. The correlation matrix also shows that there is no multicollinearity between the variables. As discussed above in section 2.3, low correlation (absence of multicollinearity) among the variables is one of the major assumptions for developing a binary logistic model.

3.2 RESULTS OF THE PRELIMNARY BINARY LOGISTIC MODEL WITH HEART DISEASE AS A RESPONSE VARIABLE.

Table 3., illustrates the results of the preliminary binary logistic regression. In this model, we can see that for every continuous variable, the interpretation would be, for every one-unit increase, the log odds of having positive status of heart disease versus the negative status. would increase or decrease by the coefficient values given. If the resultant p value is less than 0.05, then we can say that those results are statistically significant at alpha=0.05. From this initial model we can see that the factors affecting heart disease are sex, chest pain type, rest blood pressure, serum cholesterol, maximum heart rate, exercise induced angina, depression, slope, number of vessels colored, and form of thalassemia. The resulting AIC (Akaike's Information Criterion) is 746.9.

The preliminary model developed is not the best model given the set of independent variables Thus, for model selection as there are multiple methodologies for variable selection, we will use backward model selection. Basically, this model selection with AIC, attempts to select the model that best explains the data (highest likelihood), while still not fitting too many parameters. The approach is agnostic to Cross-Validation. The process is repeated until, we

can no longer achieve a model with lower AIC value by reducing the least significant variables one after another, thus obtaining the final model.

3.3 RESULTS OF THE FINAL BINARY LOGISTIC MODEL WITH HEART DISEASE AS A RESPONSE VARIABLE.

Table 4 illustrates the results of the final binary logistic model which resulted in decreased AIC value than the preliminary model. The interpretation of the final model is that the **key *important factors determining the status of heart disease are sex, chest pain type, resting blood pressure, serum cholesterol, maximum heart rate, exercise induced angina, depression, slope, number of vessels colored, and form of thalassemia*.** From the numerical data in the result table, we can interpret that for categorical variables like sex, we can interpret results as sex being with base category male, versus the category female, changes the log odds of heart disease being positive (versus being negative) by 3.181. For numeric variable like resting blood pressure, we can say that for every one unit increase in the resting blood pressure, the log odds of being positive status of heart disease decreases by 0.019. For maximum heart rate, we can say that for every one unit increase in the maximum heart rate of the patient, the log odds of being positive status of heart disease increases by 0.025.

As the resulting p-values for these key important factors is less than 0.05, we can say that the results are statistically significant at alpha 0.05. The resulting AIC value is 743.5 which is lower than 746.9 (AIC value for preliminary binary logistic model). The AIC value stands for Akaike Information Criteria. It is analogous to adjusted R^2 and is the measure of fit which penalizes model for the number of independent variables hence models with lower AIC values are preferred.

In further section 3.5, we will compare the models and the efficiency and the accuracy of the model by using confusion matrix and calculating the area under the ROC curve which will help us determine the predictivity of the model. Also, will also look at the distribution of the fitted values by plotting a histogram of the fitted values of this model.

Table 2.

	Age	Sex	Chest pain	blood pressure	Serum Cholestrol	Resting ECG	Fast Blood Sugar	Max Heart Rate	Exercise Induced Pain	Depression	Slope	Vessels under Fluro	Thal
Age	1												
Sex	0.01	1											
Chest pain type	0.00	0.04	1										
Blood Pressure	0.04	0.01	0.04	1									
Serum Cholesterol	-0.02	0.00	-0.01	-0.02	1								
Resting ECG	-0.03	-0.06	0.04	0.03	-0.04	1							
Fast Blood Pressure	0.02	0.03	0.08	-0.05	0.02	-0.10	1						
Maximum Heart Rate	0.01	-0.02	-0.01	-0.05	-0.01	-0.02	0.06	1					
Exercise Induced Pain	-0.02	0.14	-0.40	-0.03	-0.02	-0.07	0.05	0.00	1				
Depression	0.02	0.02	0.01	0.04	-0.02	-0.04	-0.02	-0.03	-0.01	1			
Slope	-0.03	-0.03	0.13	0.04	0.02	0.09	-0.06	0.05	-0.27	0.05	1		
Vessels under Fluro	-0.04	0.01	-0.01	-0.02	0.00	-0.04	-0.03	0.06	-0.01	0.02	-0.02	1	
Form of Thal	-0.09	0.02	-0.16	-0.05	0.00	-0.02	-0.04	-0.02	0.20	-0.03	-0.09	0.02	1
Heart disease	-0.01	-0.28	0.43	0.06	0.00	0.13	-0.04	0.01	-0.44	0.05	0.35	-0.05	-0.34

Table 3.

	Estimated Std.	Error	Z Value	P value	
intercept	3.690	1.401	2.633	< 0.05	***
Age	-0.008	0.013	-0.650	0.516	
Sex	-1.847	0.257	-7.020	< 0.05	***
Chest pain type	0.847	0.100	8.516	< 0.05	***
Resting Blood Pressure	-0.018	0.006	-3.245	< 0.05	***
Serum Cholesterol	-0.006	0.002	-2.757	< 0.05	***
Resting ECG	-0.102	0.285	-0.355	0.723	
Fasting Blood Sugar	0.413	0.189	2.187	0.029	
Maximum Heart Rate	0.024	0.006	4.158	< 0.05	***
Exercise Induced Pain	-0.991	0.224	-4.418	< 0.05	***
Depression	-0.571	0.116	-4.920	< 0.05	***
Slope	0.534	0.189	2.831	< 0.05	***
Vessels under Fluro	-0.754	0.103	-7.321	< 0.05	***
Form of Thal	0.877	0.156	-5.693	< 0.05	***

Significance at p < 0.05 "***"

AIC value: 746.9

N=13

Table 4.

	Estimated Std.	Error	Z value	P value	
intercept	3.181	1.159	2.746	< 0.05	***
Sex	-1.833	0.254	-7.216	< 0.05	***
Chest pain type	0.846	0.098	8.592	< 0.05	***
Resting Blood Pressure	-0.019	0.005	-3.553	< 0.05	***
Serum Cholesterol	-0.006	0.002	-2.928	< 0.05	***
Resting ECG	0.428	0.188	2.277	0.023	
Maximum Heart Rate	0.025	0.005	4.808	< 0.05	***
Exercise Induced Pain	-0.990	0.223	-4.441	< 0.05	***
Depression	-0.564	0.115	-4.894	< 0.05	***
Slope	0.541	0.188	2.885	< 0.05	***
Vessels under Fluro	-0.766	0.102	-7.545	< 0.05	***
Form of Thal	-0.877	0.152	-5.750	< 0.05	***

Significance at p < 0.05 "***"

AIC value: 743.5

N=13

Table 5.

	PRELIMNARY MODEL				FINAL MODEL			
	Estimated Std.	**P value**		**VIF**	**Estimated Std.**	**P value**		**VIF**
intercept	3.690	< 0.05	***		3.181	0.006		
Age	-0.008	0.516		1.399				
Sex	-1.847	< 0.05	***	1.365	-1.833	< 0.05	***	1.341
Chest pain type	0.847	< 0.05	***	1.287	0.846	< 0.05	***	1.237
Resting Blood Pressure	-0.018	< 0.05	***	1.137	-0.019	< 0.05	***	1.066
Serum Cholesterol	-0.006	< 0.05	***	1.278	-0.006	< 0.05	***	1.235
Resting ECG	-0.102	0.723		1.154	0.428	0.023		
Fasting Blood Sugar	0.413	0.029		1.092				
Maximum Heart Rate	0.024	< 0.05	***	1.421	0.025	< 0.05	***	1.191
Exercise Induced Pain	-0.991	< 0.05	***	1.143	-0.990	< 0.05	***	1.129
Depression	-0.571	< 0.05	***	1.408	-0.564	< 0.05	***	1.391
Slope	0.534	< 0.05	***	1.494	0.541	< 0.05	***	1.476
Vessels under Fluro	-0.754	< 0.05	***	1.120	-0.766	< 0.05	***	1.095
Form of Thal	-0.877	< 0.05	***	1.095	-0.877	< 0.05	***	1.044
	AIC: 746.9				**AIC: 743.5**			
	Significance at p < 0.05 "*"**				**Significance at p < 0.05 "***"**			
N=13								

3.4 COMPARING THE RESULTS OF THE PRELIMNARY AND THE FINAL BINARY LOGISTIC MODEL.

In **Table 5,** we can see that the final binary logistic model has the resulting AIC value of 743.5 which is lower than 746.9, the AIC value for preliminary binary logistic model. The AIC value stands for Akaike Information Criteria. It is analogous to adjusted R^2 and is the measure of fit which penalizes model for the number of independent variables hence models with lower AIC values are preferred.

Also, the **Table 5** illustrates the VIF (Variance Inflation Factors) values of all the variables individually. As VIF starts at 1 and has no upper limit, we can say that our both models have VIF values less than 5 which is considerably good. Also, there is no valid method of detecting collinearity in logistic regression. It is problematic figuring out how much collinearity is too much for logistic regression, but David Belslely did extensive work with condition indexes. He found that indexes over 30 with substantial variance accounted for in more than one variable was indicative of collinearity that would cause severe problems in the model. Hence, here we can see that *both models have low collinearity* but by comparing models as done in **Table 5**, we can see that *the multi-collinearity is even lower in the final binary logistic model*.

3.5 MODEL EVALUATION OF THE FINAL BINARY LOGISTIC MODEL WITH HEART DISEASE AS A RESPONSE VARIABLE.

To evaluate the final binary logistic model, we will need basic understanding of the parameters that will help us understand the efficiency and the accuracy of the model. For evaluating the model, we will be using confusion matrix and calculating the area under the ROC curve which will help us determine the predictivity power of the model. Also, will look at the distribution of the fitted values by plotting a histogram of the fitted values of this model. Before we start

explaining the results, we will first understand the terms listed above and according to the definitions, will evaluate our model alongside.

3.5.1 AIC VALUE:

The AIC value stands for Akaike Information Criteria. It is analogous to adjusted R^2 and is the measure of fit which penalizes model for the number of independent variables hence models with lower AIC values are preferred.

The resulting ***AIC value of the final logistic regression is 743.5*** (**Table 4**) which is lower than 746.9, the AIC value for preliminary binary logistic model (**Table 3**) illustrating that the measure of fit after penalizing the model for number of independent variables is 743.5 which is higher than the preliminary model.

3.5.2 CONFUSION MATRIX:

It is a tabular representation of Observed vs Predicted values. It helps to quantify the efficiency (or accuracy) of the model. The confusion matrix for our model and the accuracy of the model is give in the **Table 6.**

Table 6.

	Predicted Negative	Predicted Positive
Observed Negative	398	101
Observed Positive	47	479
Efficiency = sum of diagonals/sum of all values= (398+479)/1025 =85.56.		
Thus, from confusion matrix, the accuracy is 85.56%		

From the confusion matrix, we can see the calculation for the accuracy of the model. Thus, we can say that the ***accuracy of the final binary logistic model is 85.56%.***

3.5.3 ROC CURVE

ROC stands for Receiver Operating Characteristic. It explains the model's performance by evaluating Sensitivity vs Specificity. The area under the ROC Curve is an index of accuracy. Higher the area under the curve, better the prediction power of the model.

The **Figure 3** shows the ROC curve for the binary logistic model. The ***area under the curve is 92.46 illustrating that the predictivity power of the model is 92.46%.***

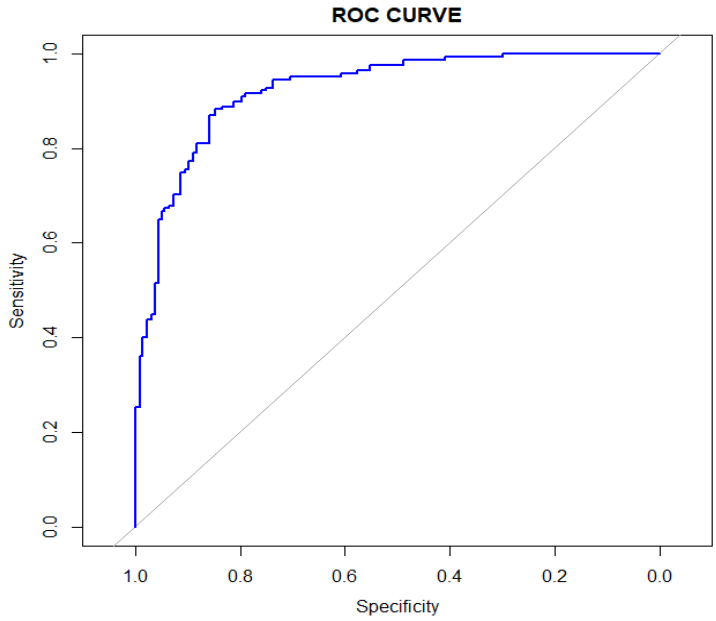
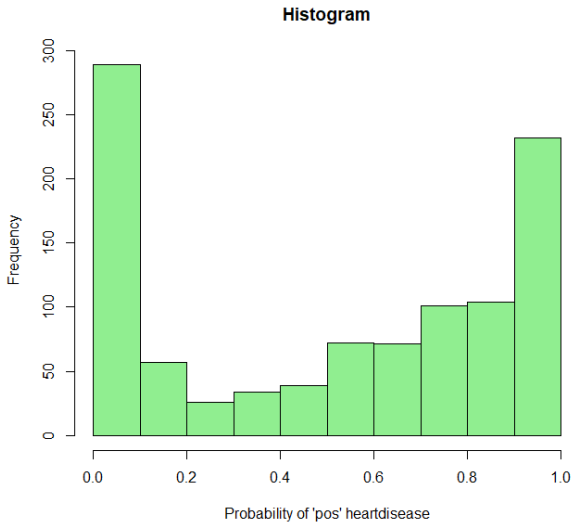


Figure 3

Figure 3: The ***area under the curve is 92.46 illustrating that the predictivity power of the model is 92.46%.***

3.5.3 HISTOGRAM OF THE FITTED VALUE:

In **Figure 4, we will see the histogram of the fitted values of the final model.**



4. DISCUSSION:

Our analysis of the dataset revealed the important factors affecting the status of heart disease and these factors have helped us develop the model which is 85.6% accurate. In short, the procedure for developing this model includes ***pre-processing the dataset*** by *checking* ***for missing values (NA)*** which **are** ***none in the dataset*** and ***then transforming the dataset*** as required. As the dataset consist of both categorical variables and numerical variables, it is ***imperative to convert the categorical variables into factors.*** After pre-processing the data, we tried normally distributing the data ***which resulted in a model with AIC value of 926 hence proceeded without normally distributing the dataset.*** Build the initial model with all independent variable and then used backward model selection to minimize AIC value to build the final model. The evaluation of the final model can be read under section 3.5.

The dataset originally had 76 attributed, Various researchers have been done their studies on the dataset to reduce the dataset to this dataset available on Kaggle. This paper would be the first to explore in-dept analysis of this dataset through Binary Logistic Regression Model. Much research has been done using SVM, KNN, Decision tree, Naïve Bayes exploring the nature of this dataset. Studies have proven that Naive Bayes is the best classification model for this problem with 85% accuracy. Here, we can state the model developed here is 85.56% accurate. The research papers for this research have been listed in the REFERENCES.

4.1 LIMITATIONS:

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions but considering the fact stated above, the use of this model for making decision for such intricate task would not be suggested until the model is much better as the main limitations of this developed model is the ability to draw accurate conclusions as needed. Considering the model is 85.56% accurate would be good in stats but it is paramount to understand that using the model for predicting the status could result in millions of more death by not accurately predicting the heart disease even in some cases. Also, it is vital to note, as discussed earlier that, there are 76 attributes in total to properly determine the factors effecting the status of heart disease in a patient and here we are only considering 14 of them.

5. CONCLUSION

This study shows the effect on some key important health parameters like sex of the patient, chest pain type representing the location or the valve where the pain is occurring, the blood pressure of the patient when the patient is resting, maximum heart rate of the patient, cholesterol levels, pain during exercise, depression state of the patient, peak heart rate value during exercise(slope),number of coloured vessels and the form of thalassemia representing the condition of patient's heart. Based on these factors, the study also focuses on building a Binary Logistic Regression model that predicts the status of heart disease in patient based on

these key important factors resulting in a model accuracy of 85.56% and predictability of the model is 92. 46%.

DATA AVAILIBITY STATEMENT

Publicly available dataset was analysed in this study. This data can be found here: https://www.kaggle.com/datasets/yasserh/heart-disease-dataset.

REFERENCES

[1] Wu R, Peters W, Morgan MW. The next generation clinical decision support: linking evidence to best practice. J Healthc Inf Manag, 2002;16:50-5.

[2] Thuraisingham BM. A Primer for Understanding and applying data mining. IT Professional 2000;1:28-31.

[3] Rajkumar A, Reena GS. Diagnosis of heart disease using datamining algorithm. Global Journal of Computer Science and Technology 2010; 10:38-43.

Mackay, J., Mensah, G. 2004 "Atlas of Heart Disease and ISBN-10 9241562765.

[4] Robert Detrano 1989 "Cleveland Heart Disease Cleveland Clinic Foundation.

[5] Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong medical data to predicting outcome of Coronary Heart International Conference November 2007, pp 868-872.

[6] Jianxin Chen, Guangcheng Xi, Yanwei Xing, Jing Chen, Specifications: A Comparison of Five Data Mining Modelling and Simulation Lecture Notes in Computer Science, pp 129-135.

[7] Jyoti Soni, Ujma Ansari, Dipesh Sharma 2011 Journal of Computer Applications, doi 10.5120/223

[8] A systematic literature review by H. Benhar and J.L Fernandez-Aleman (Discussed ANN+PCA,SVM+PCA,ANN+CHI)

[9] Heart Disease Prediction System Using Data Mining Technique by KNN approach by V. Krishnaiah, G. Narsimha, N. Subhash

[10] Heart Disease Prediction System Using Data Mining Technique by Naive Bayes and Decision Tree by R. Fadnavis,K. Dhore, D. Gupta,

[11] An improved feature selection approach for chronic heart disease detection by S.J Sushma, Tsehay, Admassu Assegie, D.C Vinutha and S.Padmashri.

[12] Improving the Accuracy for Analyzing Heart Disease Prediction Based On Ensemble Model- used Feature Extraction methods like LDA and PCA

[13]Heart disease Prediction Using Exploratory Data Analysis by Indra Kumari Ranganathan and Soumya Jenna.