Predicting factors affecting presence or absence of heart disease and determining the gender at risk.

Sai Sravanthi Kilari

ABSTRACT

Objective:

The objectives of this project are to identify the factors which have maximum influence in the occurrence of the heart disease in the Cleveland population. And determining which gender is at high risk of developing heart disease and by what percentage.

Materials and Methods:

We have use Cleveland data set which contains 303 patients' information and included 14 attributes. We performed statistical (Chi square and Kruskal wallis test), regression (Logistic regression) and predictive analysis (Knn model and Random Forest mode) for all the attributes to predict the occurrence of heart diseases.

Results: Null hypothesis was rejected indicating that there is association between the factors and occurrence of the heart disease. According to Chi square test we can interpret that males have more chances of occurrence of heart disease compared to females.

Discussion and Conclusion: Sex as an association with the presence or absence of heart disease. Chi square indicates that males are at higher risk of getting heart disease than females. Among all the model's logistic regression has the highest accuracy and knn gave the least accuracy recorded. Chest pain as the highest association according to the Random Forest model

Key words:

Knn, AIC, AUC, 10-Fold cross validation, kappa.

INTRODUCTION

Cardiovascular diseases are one of the leading causes of death in men and women globally. About one third of death occur due to these diseases worldwide. Heart diseases refers to any condition that affects structure and function of heart. In most of heart diseases, a plaque or fatty deposits builds inside the arteries thereby narrowing the blood vessel. This leads to blockage of blood flow. Cardiovascular diseases are a series of diseases which include myocardial infraction (insufficient supply of oxygen to the tissues), arrhythmias (irregular heart diseases), coronary heart disease (damage to blood vessels to the heart), stroke (blockage of blood vessels to brain). Risk factors include high blood pressure, diabetes, obesity, cigarette smoking, lack of physical exercise, stress genetic predisposition. According to WHO, it is estimated that 17.9 million people died because of cardiovascular diseases in 2019, which is 32% of all global deaths. One in three women die from cardiovascular diseases. Early diagnosis and detection of heart diseases can help physicians to provide better treatment at the point of care.

OBJECTIVE AND HYPOTHESIS

The objectives of this project are to identify the factors which have maximum influence in the occurrence of the heart disease in the Cleveland population. And determining which gender is at high risk of developing heart disease and by what percentage. Hypothesis are as follows.

Null Hypothesis (H0): There is no association between the factors and occurrence of the heart disease.

Alternative Hypothesis (H1): There is association between the factors and occurrence of the heart disease.

2 .MATERIALS AND METHODS

2.1 Dataset

We have used the Cleveland dataset for the project. This dataset was downloaded from the University of California (UCI, Irvine) machine learning repository. The dataset was collected by Dr. Robert Detrano (Detrano et al., 1984) from the Cleveland Clinic Foundation in 1984 [1]. The original dataset was comprised of a total of 303 samples and 76 features, but nowadays, most of the research done on this dataset has only used 14 features and one target variable. Therefore, we have decided to focus on only those 14 features, these features include the results of non-invasive diagnostic tests along with the attributes such as Age, Sex,

Chest pain type, resting blood pressure (in mm Hg on admission to the hospital), Fasting blood sugar, etc. The target variable shows the result of the invasive coronary angiogram, where 0 denotes the absence of coronary heart disease and 1 denotes the presence of coronary heart disease. In the dataset, 138 samples have a target variable of 0, which means they don't have coronary heart disease, whereas 165 samples have a target variable of 0, which means they had coronary heart disease; there over 50% of people had CHD. Data showed that the number of males with a heart condition was greater than females. We also observed that more people are prone to heart disease after the age of 40.

Table 1: Data description

Attributes	**Description**	**Type**	**Value**
age	Age	Int	29 to 77
Sex	Sex	Int	Male: 1 Female: 0
cp	Chest pain type	Int	Typical angina: 0, Atypical angina: 1, non-anginal pain: 2, Asymptomatic: 3
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Int	94 to 200
chol	Serum cholesterol in mg/dl	Int	126 to 564
Fbs	Fasting blood sugar > 120 mg/dl	Int	True: 1, False: 0
restecg	Resting electrocardiographic results	Int	0: Nothing to note, 1: ST-T wave abnormality 2: Possible ventricular hypertrophy
thalach	Maximum heart rate achieved	Int	71 to 202

exang	Exercise induced angina	Int	Yes: 1, No:0
oldpeak	ST depression induced by exercise relative to rest	Real	0 to 6.2
slope	The slope of the peak exercise ST segment	Int	0,1,2
Ca	number of major vessels colored by flourosopy	Int	0,1,2,3,4
thal	thalium stress result	Int	Normal: 3, Fixed defect: 6, Reversable defect: 7
target	Presence or absence of disease	Int	1: Yes 0: No

2.2. Data Cleaning

Data cleaning is one of the most critical steps taken before starting any data analysis. It helps identify and remove nulls values and duplicate data to create an error-free dataset. Which further helps us in making accurate predictions with our data analysis models. For this dataset, first we looked for nulls values. We found that there are no null values in the dataset. Then we decided to assign each categorical variables to labels, for example, for sex column we assigned the 0 to female and 1 to male, fbs column we assigned 0 to false and 1 to true, more labels are explained in the figure 3. We also removed outliers in the data, it is shown in figure 4.

```
#sex column
cleve_data$sex[cleve_data$sex == 0] = "female"
cleve_data$sex[cleve_data$sex == 1] = "male"

#cp
cleve_data$cp[cleve_data$cp==0] = "typical angina"
cleve_data$cp[cleve_data$cp==1] = "atypical angina"
cleve_data$cp[cleve_data$cp==2] = "non-anginal pain"
cleve_data$cp[cleve_data$cp==3] = "asymptotic"

#fbs
cleve_data$fbs[cleve_data$fbs==0] = "false"
cleve_data$fbs[cleve_data$fbs==1] = "true"

#exang
cleve_data$exang[cleve_data$exang == 1] = "yes"
cleve_data$exang[cleve_data$exang == 0] = "no"

#restecg
cleve_data$restecg[cleve_data$restecg==0] = "Nothing to note"
cleve_data$restecg[cleve_data$restecg==1] = "ST-T Wave abnormality"
cleve_data$restecg[cleve_data$restecg==2] = " Definite left ventricular hypertrophy"

#slope
cleve_data$slope[cleve_data$slope == 0] = "upsloping"
cleve_data$slope[cleve_data$slope == 1] = "flat"
cleve_data$slope[cleve_data$slope == 2] = "downsloping"

#thal
cleve_data$thal[cleve_data$thal == 1] = "normal"
cleve_data$thal[cleve_data$thal == 2] = "fixed defect"
cleve_data$thal[cleve_data$thal == 3] = "reversible defect"
```

Figure 3: Assigning labels to categorical variables

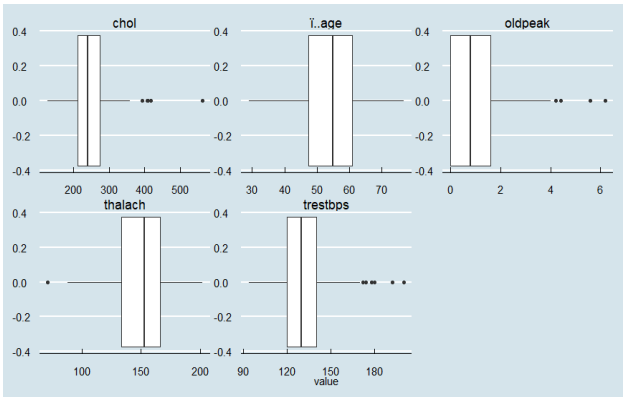

Figure 4: Box plots showing outliers

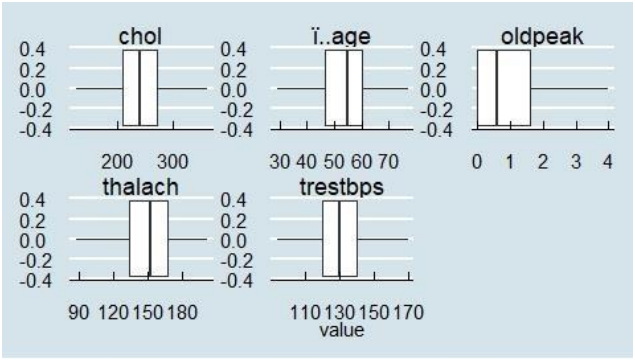

Figure 5: Box plots with outliers removed

2.3. Exploratory data analysis

Exploratory data analysis (EDA) is mainly used to understand the patterns of within the data, and to detect outliers from the data. To better understand the distribution of numerical data in our dataset, we decided to employ bar plots for categorical variables and box plots for numerical variables.

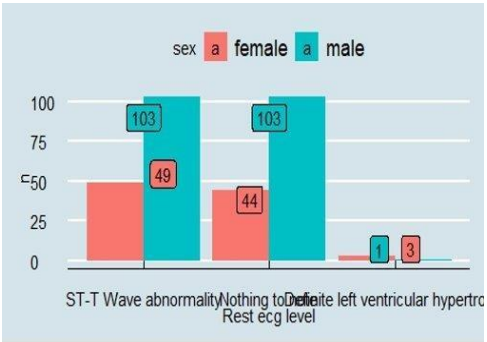

Figure 6.1 Distribution of Females and males

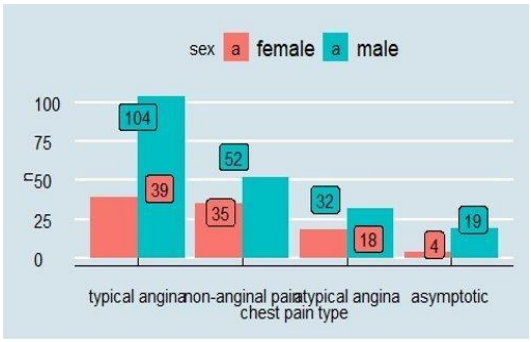

Figure 6.2 Chest pain among males and females

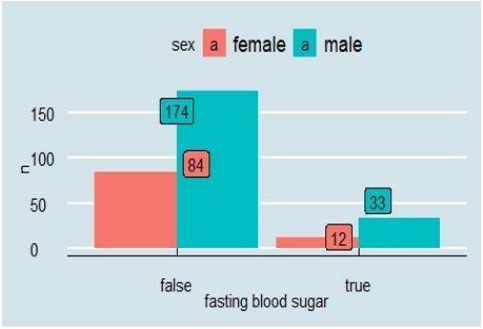

Figure 6.3 Distribution of Fbs among males and females

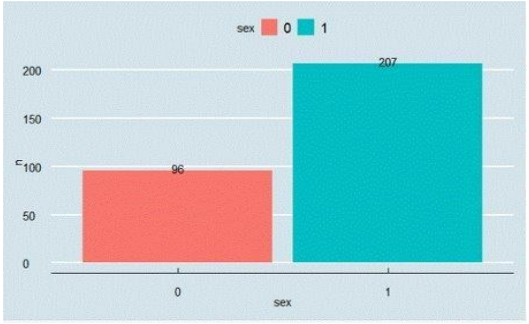

Figure 6.4 ECG slope among males and females

Figure 6: Data distribution of categorical variables

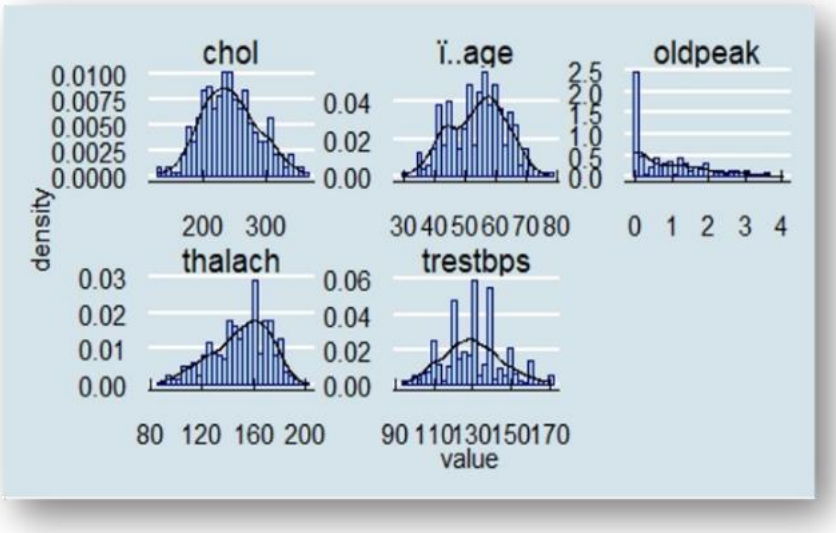

Figure 7: Data distribution of numerical variables

Table: 2 Data distribution each variable among male and female

Characteristic	Female, N= 85	Male, N = 199
age	54.94 (9.64)	53.59(8.94)
cp		
asymptomatic	4/85 (4.7%)	18/199 (9.0%)
atypical angina	18/85 (21%)	31/199 (16%)
non anginal pain	33/85 (35%)	50/199(25%)

typical angina	30/85 (39%)	100/199 (50%)
trestbps	129.71 (15.60)	130.00 (15.31)
chol	250.74 (48.39)	238.38 (42.70)
fbs		
false	76/85 (89%)	168/199 (84%)
true	9/85 (111%)	31/199 (16%)
restecg		
definite left ventricular hypertrypohy	2/85 (2.4%)	0/199(0%)
nothing to note	37/85 (44%)	100/199 (50%)
st-t wave abnormality	46/85 (54%)	99/199 (50%)
thalach	151.45 (19%)	149.44 (23.49)
exang	16/85 (52%)	74/199 (37%)
oldpeak	0.71 (0.84)	1.05(1.08)
slope		
downsloping	44/85 (69%)	94/19 (47%)
flat	38/85 (45%)	92/199 (46%)
upsloping	3/85 (3.5%)	13/199 (7.0%)
ca		
0	59/85 (69%)	106/199 (53%)
1	14/85 (16%)	49/199(25%)
2	10/85 (12%)	25/199(13%)
3	2/85 (2.4%)	14/199(7.0%)
thal		
0	1/85 (1.2%)	1/199(0.5%)
fixed defect	74/85 (87%)	86/199 (43%)
normal	1/85 (1.2%)	16/199(8%)

reversible defect	9/85 (11%)	96/199(48%)
heart disease		
0	17/85 (20%)	108/199(545)
1	68/85 (80%)	91/199(46%)

2.4. Correlation Analysis

Correlation analyses are performed to measure the relationship between two continuous variables, for example, relationship between independent and dependent variables or two independent variables. In correlation analysis we look for the value of correlation coefficient which is also known as Pearson Product Moment correlation coefficient. For example, If the value of correlation coefficient r is 0.8 that means there is strong relation between the target variable and that attribute. For our correlation analysis, found that there is highly positive correlation between cp (0.77), thalach (0.8), and slope (0.71) and our target variable (which is the dependent variable), whereas there is highly negative correlation between exang (-0.81), oldpeak (-0.8) and thal (-0.69) and the target variable and the results were plotted using heat map.

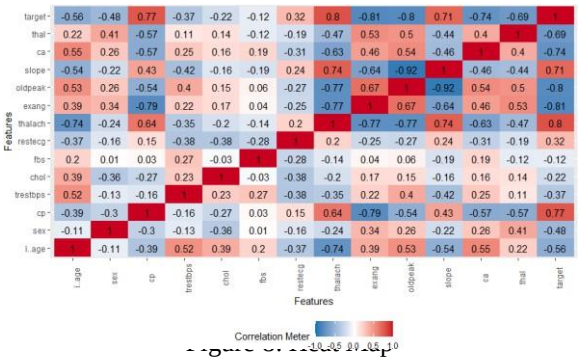

Figure 3. Heat Map

2.5. Normality assumption

For normality assumption, we decided to employ Shapiro-Wilk test. It is a test to check if the sample is normally distributed or not. The test calculates value of W if the value of W is small that means the sample is not normally distributed, and if it is normally distributed then we can reject the null hypothesis. For our dataset, we found out that, All the variables p values are less than the significance level (<0.05) which indicates that the data did not pass the normalcy assumptions. We can also say based on normality assumptions that Shapiro tests are not

statistically significant. Therefore, there is evidence for rejecting null hypothesis (i.e., data is normally distributed).

Table 3: Shapiro-Wilk Test results

Variables	**P values**
chol	5.365 e-19
age	0.05
ca	< 2.2 e-16
thal	< 2.2 e-16
sex	< 2.2 e-16
cp	< 2.2 e-16
thalach	< 3.76 e-17
oldpeak	<2.2 e-16
Slope	<2.2 e -16
exang	<2.2 e-16
trestbps	< 1.458e -06
restecg	< 2.2e-16
target	<2.2 e-16

2.6. Statistical analysis

Chi-square Test

For statistical analysis, we decided to employ Chi-squared test. When the null hypothesis is true, a chi-squared test is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution. We used Chi-square test to determine the association between the sex and target (dependent) variable. We considered null hypothesis as there is no

statistical association between the sex and target. Alternative hypothesis include that there is statistical association between the sex and target. The p value obtained is < 0.05 i.e., less than significance level. This indicates that there is a statistical association between the sex and target. Therefore, we have enough evidence to reject the null hypothesis. We have done odd's ratio and risk ratio to compare the occurrence of the outcome (target) with the variable of interest (sex). The results of odd's ratio obtained conclude that the odds of the male having heart disease is 73% more in men than female. And risk ratio concluded that males are at 55% risk of getting heart diseases when compared to the females.

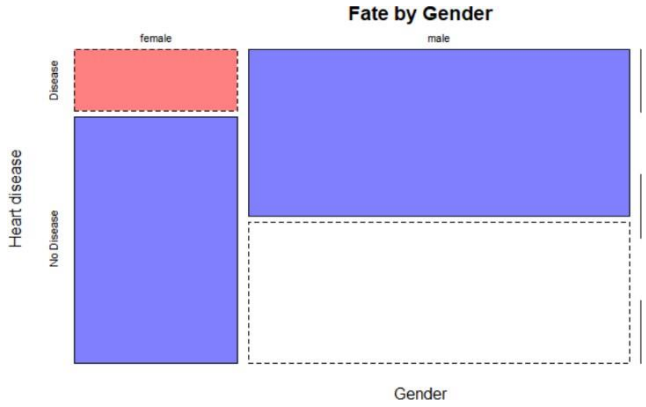

Figure 9: Percentage of heart disease in male and female

Kruskal Wallis Rank Sum Test

As our data did not pass the test of normalcy, we have done Kruskal Wallis test to test if there are differences between the groups. We considered null hypothesis that the group medians are same and alternative hypothesis as not all group medians are equal. The p value obtained is less than the significance level indicating that the group means are different. Therefore, there is enough evidence for rejecting the null hypothesis.

Table 4: P values of Kruskal Wallis Test

Variables	**P values**
age	3.429 e-05
ca	1.83 e -15

cp	1.157 e-15
thal	2.407 e-12
oldpeak	2.395 e-13
trestbps	0.0346
chol	0.03566
sex	1.049 e-06
restecg	0.009806
slope	1.08 e -10
exang	3.198 e-14
thalach	9.748 e- 14

Post Hoc Test

Dunn- Kruskal Wallis pair comparison test is a non-parametric analog to multiple pairwise test following the rejection of the Kruskal-Wallis test. We have done this test determine if there are any differences with the group. P values are adjusted by Benjamini-Hocberg method. P values are found be less than the significance level.

3.1. Regression analysis

LOGISTIC REFRESSION

Regression analysis is a technique used to estimate the relationship between a target variable or dependent variable and one or more confounding variables. The target variable is also known as dependent variable. For our regression analysis, we have decided to employ Logistic regression, Logistic regression is one of the most common and used algorithm to do regression analysis. We have used logistic regression with step wise AIC backward regression. The lower the AIC better is the performance of the model. We got the highest AIC with the step wise

model which is 203 and the accuracy if the model was 0.94 indicating that the model was well fitted.

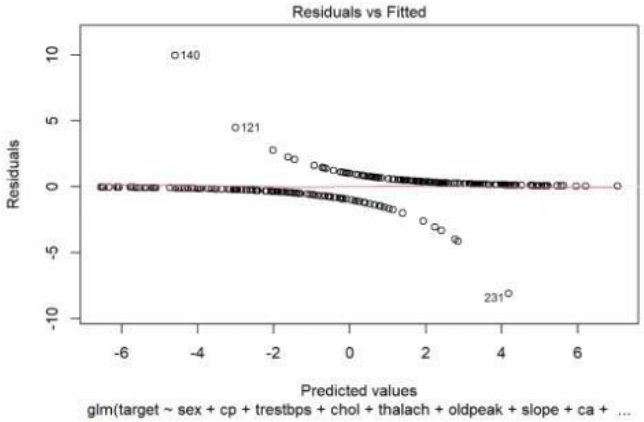

Fig 10: Residual vs fitted plot

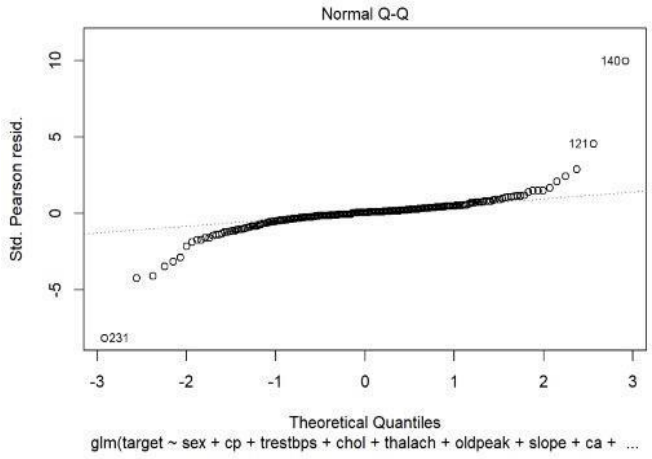

Figure 11: Normal QQ plot

Table 4: Results of logistic regression

Characteristic	N	log (0R)	95% Cl	p-value
age	303	0.03	-0.02, 0.08	0.3
sex	303			<0.001
0		-	-	

1		-1.9	-3.0, -0.78	
cp	303			<0.001
0		-	-	
1		0.86	-0.24, 2.0	
2		2.0	1.0, 3.1	
3		2.4	1.1, 3.9	
trestbps	303	-0.03	-0.05,0.00	0.025
chol	303	0.00	-0.01, 0.00	0.3
fbs	303			0.4
0		-	-	
1		0.45	-0.69, 1.6	
restecg	303			0.5
0		-	-	
1		0.46	-0.32,1.3	
2		-0.17	-5.5,3.4	
thalach	303	0.02	0.00,0.04	0.83
exang	303			0.085
0		-	-	
1		-0.78	-1.7,0.11	
oldpeak	303	-0.40	-0.89, 0.07	0.0094
slope	303			
0				
1		-0.78	-2.6, 0.92	
2		0.69	-1.2, 2.5	
ca	303			<0.001

0		-	-	
1		-2.3	-3.4, -1.3	
2		-3.5	-5.2, -2.0	
3		-2.2	-4.3, -0.55	
4		1.3	-2.0,5.0	
thal	303	1.3		0.003
0		-	-	
1		2.6	-2.0, 7.6	
2		2.4	-2.2.7.1	
3		0.92	-3.7,5.6	

OR = Odds Ratio, Cl= confidence interval

Figure 12: Confusion matrix of logistic regression

3.2 Predictive analysis

For predictive analysis, we decided to use two of the most used machine learning algorithms, namely k-nearest neighbours' algorithm and Random Forest.

3.2.1. K-nearest neighbours' algorithm

We have used K nearest neighbour machine learning algorithm for prediction of the presence or absence of the heart disease. It a supervised machine learning model and does not assume that the data is normally distributed. Since the data set had many categorical values and the data was not normally distributed, we have used this mode for prediction. 10-fold cross validation was done to identify the best K value to get for the model achieve the highest fit accuracy. The highest accuracy obtained was 0.7 with K=23. Ans the least accuracy was with K=7.

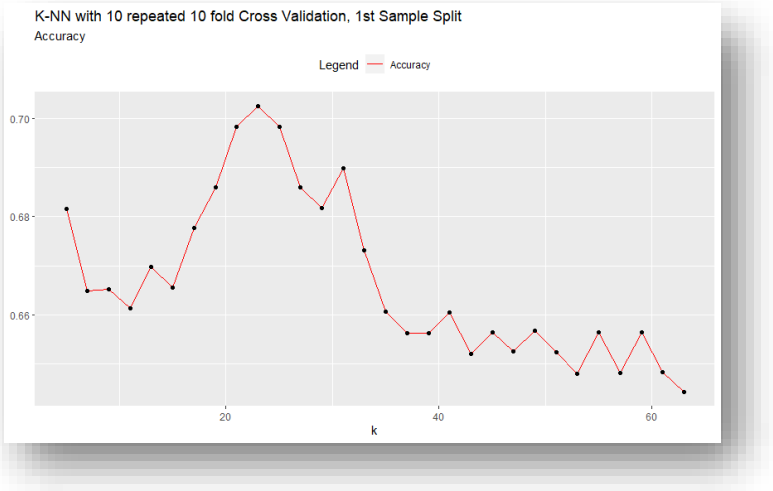

Figure 13: plot showing the accuracy levels at each k value

3.2.2. Random Forest

We have used Random Forest supervised machine learning algorithm for the prediction analysis. It consists of multiple individual decision trees that work together as an ensemble to solve both classification and regression problems. In the random forest, every tree makes a prediction about the class, and the class with the most votes make the prediction. Best thing about random forest is that it gives great results even when we are using its default parameters.

We have plotted the variable importance post random forest model and the variable that had highest influence on the presence of the heart disease was chest pain and fasting blood glucose had least importance.

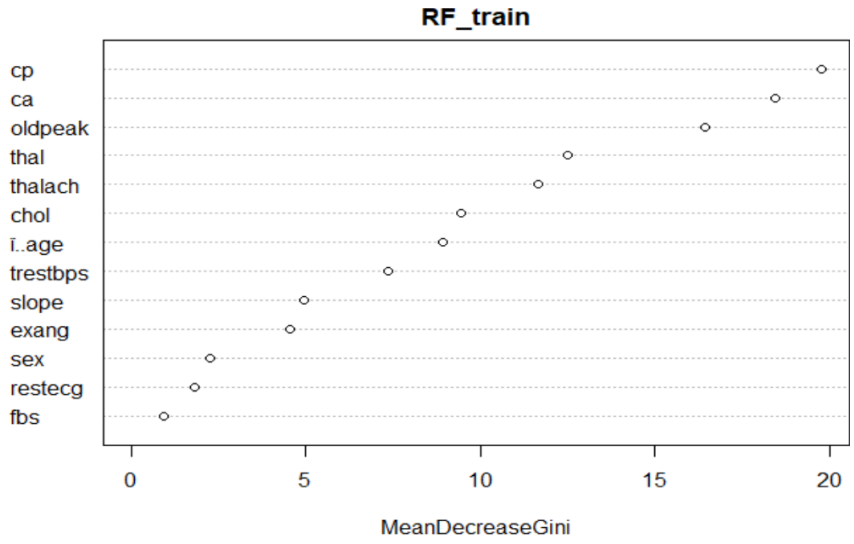

Figure 14: Variable importance plot using random forest model

RESULTS:

Table 5: Results of all the statistical tests

	Shapiro Wilk test	Kruskal Wallis Test	Chi Square	Logistic Regression
Significance value	0.05	0.05	0.05	0.05
P value	<0.05 for all the variables	<0.05 for all the variables	2.091e	<0.05

Table 6: Results of machine learning models

	Logistic regression	KNN	Random Forest
Accuracy	0.94	0.67(K=5) 0.70(K=23)	0.803
Kappa	-	0.33	0.6

The results from the table 5 show that the data was not normally distributed from the Shapiro wilk test, Kruskal Wallis test results indicate that there significant association between the variables selected and the presence of heart disease, results of the chi square test indicate that males and females are not equally affected by heart disease. Table 6 shows the accuracy and kappa values of the machine learning models where the highest accuracy is shown by logistic regression and the least by KNN.

DISCUSSION

Most of the variables of the data were categorical in nature. With correlation analysis, the highest positive correlation was found among thalach, but random forest variable importance has shown that chest pain had the highest influence in presence of heart disease. Chi-square results indicate that sex have an association with the presence or absence of heart disease. Among all the machine learning models the logistic regression has the highest accuracy of 0.94 and KNN gave the least (0.67 with K=5).

CONCLUSION

Significant association was found among the variables and the target (which is the presence or absence of heart disease. The variable with highest association is chest pain obtained from random forest. Chi square results indicate that males are at a higher risk than females. Among the machine learning models logistic regression gave the highest accuracy value of 0.94. Hence, we can conclude that there is an association between the factors and occurrence of heart disease. Therefore, we reject our null hypothesis i.e., there is no association between the factors and the occurrence of heart disease.

REFERENCES

Cardiovascular diseases (CVDs). (2021). Retrieved 9 December 2021, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Centers for Disease Control and Prevention. (2021, January 19). *Heart disease*. Centers for Disease Control and Prevention. Retrieved December 11, 2021, from https://www.cdc.gov/heartdisease/index.htm.

Comprehensive R Archive Network (CRAN). (2021, October 16). *Presentation-ready data summary and analytic result tables [R package gtsummary version 1.5.0]*. The Comprehensive R Archive Network. Retrieved December 11, 2021, from https://cran.r-project.org/web/packages/gtsummary/index.html.

Detrano, R., Yiannikas, J., Salcedo, E., Rincon, G., Go, R., Williams, G., & Leatherman, J. (1984). Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. *Circulation*, *69*(3), 541-547. doi: 10.1161/01.cir.69.3.541

Rawat, S. (2021, June 28). *Heart disease prediction*. Medium. Retrieved December 11, 2021, from https://towardsdatascience.com/heart-disease-prediction-73468d630cfc.

R]eliability. (2019, September 29). *Heart disease prediction from patient data in R: R-bloggers*. R. Retrieved December 11, 2021, from https://www.r-bloggers.com/2019/09/heart-disease-prediction-from-patient-data-in-r/.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*(18), 1837–1847. https://doi.org/10.1161/01.cir.97.18.1837