Heart Disease
Diagnosis & Prediction

Group 6

I035 Tanish Patwari

I033 Sahil Pariani

I027 Saket Lakhotia

I024 Uzair Khan

I012 Sanskruti Dani

I010 Aman Chowhan

Table of Contents

1. Problem Statement

To predict based on the given attributes of a patient that whether that particular person has a heart disease or not. Another goal is to perform the experimental task that will help diagnose and find out various insights from this dataset which could help in understanding the problem more.

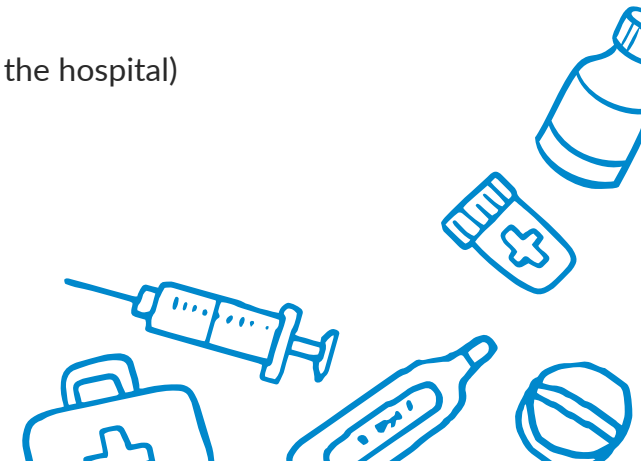
2. Analysis Steps

2.1 About the Dataset

The selected dataset is of multivariate type which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. The dependent variable “Presence of Heart disease” is coded as (1 = no, 0= yes).

2.2 Definition and Coding of Variables

- Age: The person's age in years
- Sex: The person's sex (1 = male, 0 = female)
- cp: chest pain type
  — Value 0: asymptomatic
  — Value 1: atypical angina
  — Value 2: non-anginal pain
  — Value 3: typical angina
- trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
- chol: The person's cholesterol measurement in mg/dl
- fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- thalach: The person's maximum heart rate achieved
- ca: The number of major vessels (0–3)



restecg: resting electrocardiographic results
— Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria
— Value 1: normal
— Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

exang: Exercise induced angina (1 = yes; 0 = no)

oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)

slope: the slope of the peak exercise ST segment — 0: downsloping; 1: flat; 2: upsloping
0: downsloping; 1: flat; 2: upsloping

thal: A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously
Value 1: fixed defect (no blood flow in some part of the heart)
Value 2: normal blood flow
Value 3: reversible defect (a blood flow is observed but it is not normal)

Dependent Variable

target: Heart disease (1 = no, 0= yes)

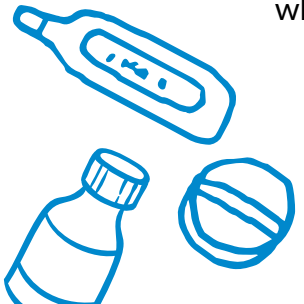
2.3 Method And Significance Test

2.3.1 Forward Logistic Regression

- In forward method, logistic regression starts with single variable and adds one variable at a time and tests its significance and removes the insignificant variables from the model.

2.3.2 Significance Test

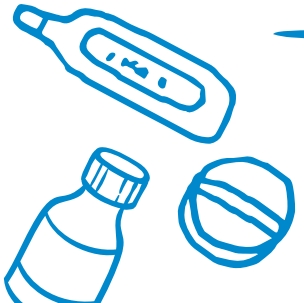
- Hosmer and Lemeshow chi square test used to test the overall model of goodness-of fittest (It indicates a poor fit if the significance value is less than 0.05) It is the modified chi-square test , which is better than the traditional chi-square test.



2.4 Hypothesis

Null Hypothesis
All the coefficients in the regression equation take the value zero.

Alternate Hypothesis
The model with predictors currently under consideration is accurate and differs significantly from the null or zero.



3. SPSS Output & 4. Interpretation

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Block 0: Beginning Block

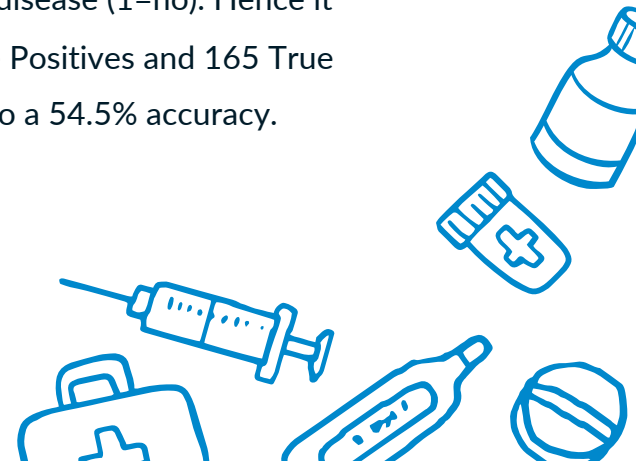
Classification Table[a,b]

			Predicted		
			target		
	Observed		0	1	Percentage Correct
Step 0	target	0	0	138	.0
		1	0	165	100.0
	Overall Percentage				54.5

a. Constant is included in the model.

b. The cut value is .500

Initially Block 0 is created with no independent variables and only the constant term in the equation. It predicts all patients to not have heart disease (1=no). Hence it gives us 138 False Positives and 165 True Positives leading to a 54.5% accuracy.



Variables added at each step can be seen on the right.

Only 8 out of the initial 14 predictors have been selected and added to the model as the rest have been deemed statistically insignificant using logistic regression.

a. Variable(s) entered on step 1: exang.

b. Variable(s) entered on step 2: ca.

c. Variable(s) entered on step 3: oldpeak.

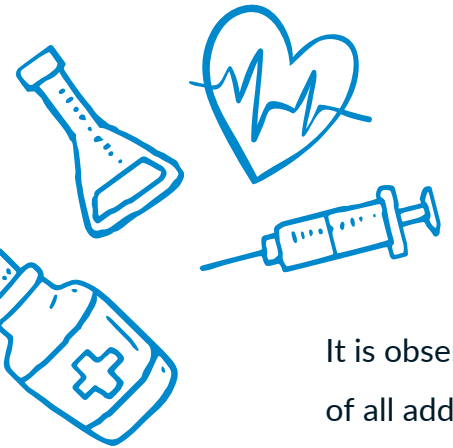
d. Variable(s) entered on step 4: cp.

e. Variable(s) entered on step 5: thal.

f. Variable(s) entered on step 6: sex.

g. Variable(s) entered on step 7: thalach.

h. Variable(s) entered on step 8: trestbps.



It is observed that significance value of all added variables at each step is below 0.05 and is hence significant. B Value in the table is the coefficient of each independent variable and can be considered as a measure of impact each has on the probability of classifying any data point to the target group.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	exang	-2.024	.283	51.327	1	<.001	.132
	Constant	.829	.152	29.637	1	<.001	2.290
Step 2[b]	exang	-2.084	.303	47.320	1	<.001	.124
	ca	-.910	.152	35.788	1	<.001	.403
	Constant	1.505	.204	54.554	1	<.001	4.504
Step 3[c]	exang	-1.807	.319	32.015	1	<.001	.164
	oldpeak	-.760	.155	24.068	1	<.001	.468
	ca	-.822	.156	27.841	1	<.001	.439
	Constant	2.111	.256	67.745	1	<.001	8.253
Step 4[d]	cp	.824	.165	24.952	1	<.001	2.279
	exang	-1.313	.349	14.166	1	<.001	.269
	oldpeak	-.916	.173	27.879	1	<.001	.400
	ca	-.820	.163	25.409	1	<.001	.440
	Constant	1.279	.294	18.854	1	<.001	3.592
Step 5[e]	cp	.819	.170	23.314	1	<.001	2.268
	exang	-1.268	.362	12.261	1	<.001	.282
	oldpeak	-.897	.181	24.407	1	<.001	.408
	ca	-.822	.171	23.262	1	<.001	.439
	thal	-.994	.267	13.808	1	<.001	.370
	Constant	3.514	.685	26.275	1	<.001	33.574

Step 6[f]	sex	-1.273	.398	10.240	1	.001	.280
	cp	.865	.173	24.874	1	<.001	2.376
	exang	-1.279	.372	11.796	1	<.001	.278
	oldpeak	-.861	.181	22.636	1	<.001	.423
	ca	-.772	.171	20.331	1	<.001	.462
	thal	-.858	.268	10.242	1	.001	.424
	Constant	4.026	.727	30.649	1	<.001	56.036
Step 7[g]	sex	-1.390	.406	11.727	1	<.001	.249
	cp	.787	.175	20.299	1	<.001	2.197
	thalach	.024	.009	7.210	1	.007	1.024
	exang	-1.045	.389	7.212	1	.007	.352
	oldpeak	-.741	.182	16.491	1	<.001	.477
	ca	-.713	.174	16.731	1	<.001	.490
	thal	-.896	.275	10.658	1	.001	.408
	Constant	.464	1.482	.098	1	.754	1.590
Step 8[h]	sex	-1.505	.420	12.824	1	<.001	.222
	cp	.829	.177	21.828	1	<.001	2.291
	trestbps	-.020	.010	4.369	1	.037	.980
	thalach	.026	.009	8.175	1	.004	1.026
	exang	-.989	.397	6.195	1	.013	.372
	oldpeak	-.703	.186	14.291	1	<.001	.495
	ca	-.703	.176	15.904	1	<.001	.495
	thal	-.905	.280	10.426	1	.001	.405
	Constant	2.794	1.884	2.199	1	.138	16.338

So, in the table on step 8 it is observed that thalach which is basically maximum heart rate achieved has a positive b value of 0.026 which signifies that higher values of thalach will mean higher chance of falling in the target group of “Heart Disease Absent”. Sex has negative B value of -1.505 which means that males (Coded as 1) have a lesser probability of belonging to the target group which is “Heart Disease Absent” (Coded as 1).

We can also look at Exp(B) which are the odds ratios for the predictors. They are the exponentiation of the coefficients and makes our interpretation easier. Exp (B) value of 1 means no impact , values between 0 and 1 signify negative impact and values greater than 1 signify positive relation between the independent and dependent variables. Trestbps (Resting Blood Pressure) has Exp (B) value of 0.980 which is slightly less than 1 and tells us that there is a slight decrease in probability of having no heart disease when there is increase in resting blood pressure levels.

Nagelkerke R Squared is an adjusted version of Cox and Snell R Squared. The range of values for Nagelkerke fall between 0 and 1. It measures the proportion of the total variation of the dependent variable can be explained by independent variables in the current model. Here it is observed that R^2 Value keeps on increasing ranging from) 0.239 in step 1 to 0.643 in Step 8 which shows that at each step the model can explain the variance in the dependent variable better than the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	357.903[a]	.179	.239
2	313.112[a]	.292	.390
3	284.590[b]	.355	.475
4	256.661[b]	.412	.551
5	242.224[c]	.439	.588
6	231.037[c]	.460	.615
7	223.312[c]	.473	.633
8	218.836[c]	.481	.643

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

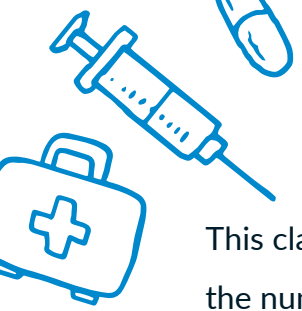
b. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

c. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	0	.
2	9.620	5	.087
3	9.223	8	.324
4	13.337	8	.101
5	8.529	8	.384
6	8.789	8	.360
7	10.166	8	.254
8	7.060	8	.530

- The Hosmer-Lemeshow statistic indicates a poor fit if the significance value is less than 0.05.
- The significance value fluctuates from step 1 to 8 but it's observed that step 8 has the highest significance value at 0.530 which tells us that it gives us the best fit compared to all other models.

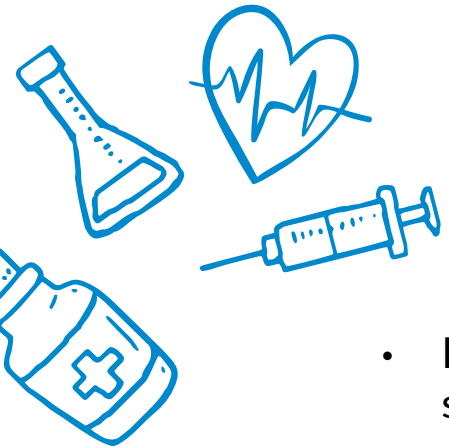


This classification table shows us the changes observed in the number of correct and incorrect classifications after each iteration. In step 1 when we have only 1 predictor (Exang) the number of True Positives and True Negatives are 76 and 142 respectively leading to an overall accuracy of 86.1%. It is also observed that the overall accuracy of the models keep on increasing till step 7 as we keep adding predictors. At step 8 there is a slight drop in overall accuracy in the model from 84.5% to 83.8%. This is because the added predictor trestbps is statistically significant (p value(0.037) is also slightly on the higher side) but has a very low coefficient value of -0.020.

Classification Table[a]

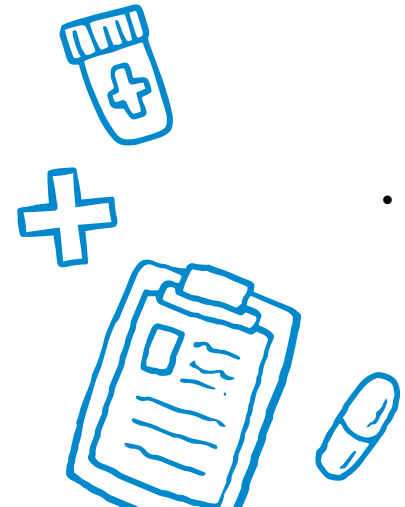
			Predicted		
			target		
	Observed		0	1	Percentage Correct
Step 1	target	0	76	62	55.1
		1	23	142	86.1
	Overall Percentage				71.9
Step 2	target	0	100	38	72.5
		1	37	128	77.6
	Overall Percentage				75.2
Step 3	target	0	98	40	71.0
		1	28	137	83.0
	Overall Percentage				77.6
Step 4	target	0	104	34	75.4
		1	26	139	84.2
	Overall Percentage				80.2
Step 5	target	0	104	34	75.4
		1	15	150	90.9
	Overall Percentage				83.8
Step 6	target	0	106	32	76.8
		1	20	145	87.9
	Overall Percentage				82.8
Step 7	target	0	106	32	76.8
		1	15	150	90.9
	Overall Percentage				84.5
Step 8	target	0	107	31	77.5
		1	18	147	89.1
	Overall Percentage				83.8

a. The cut value is .500



5. Conclusion

- It is observed that 8 out of initially selected 14 predictors are statistically significant.
- It is also observed that all statistically significant 8 variables have nonzero coefficients and hence positively/negatively impact the probability of “absence of heart disease” in each patient.
- Hence, Null hypothesis is rejected.



Thank You

I035 Tanish Patwari
I033 Sahil Pariani
I027 Saket Lakhotia
I024 Uzair Khan
I012 Sanskruti Dani
I010 Aman Chowhan