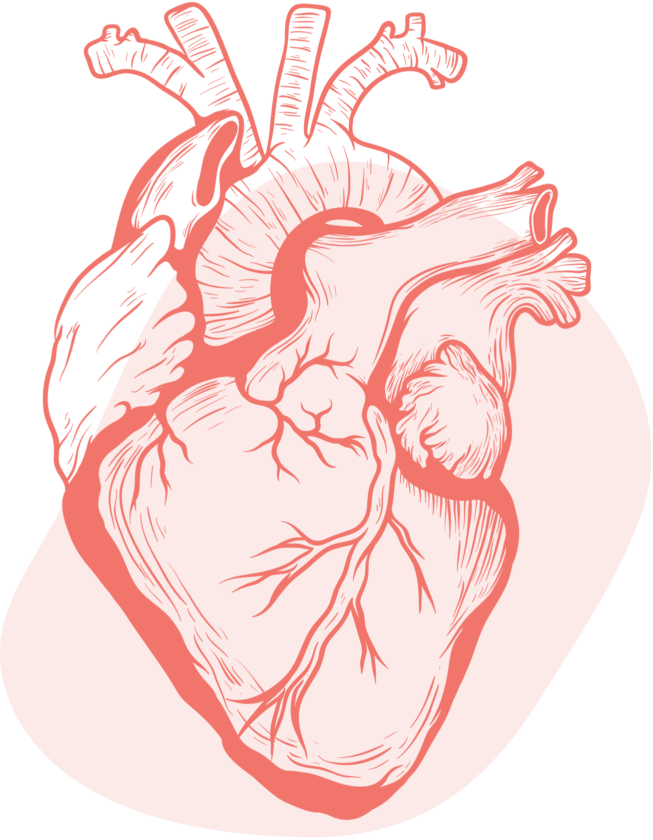


Cardiovascular Disease

STAT 660 ❤️

Esther & Irene

01

Introduction

- Topic
- Data Description

02

EDA

- Explore Variables
- Explore Features

03

Statistical Tests & Modeling

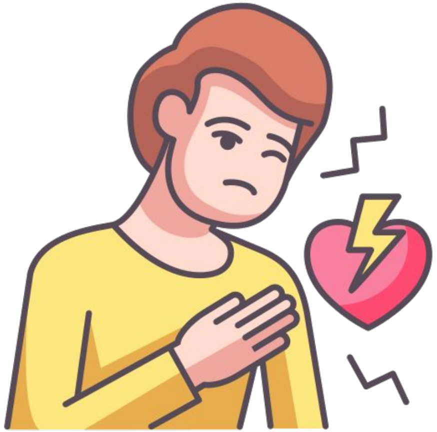
- T-tests
- Logistic Regression

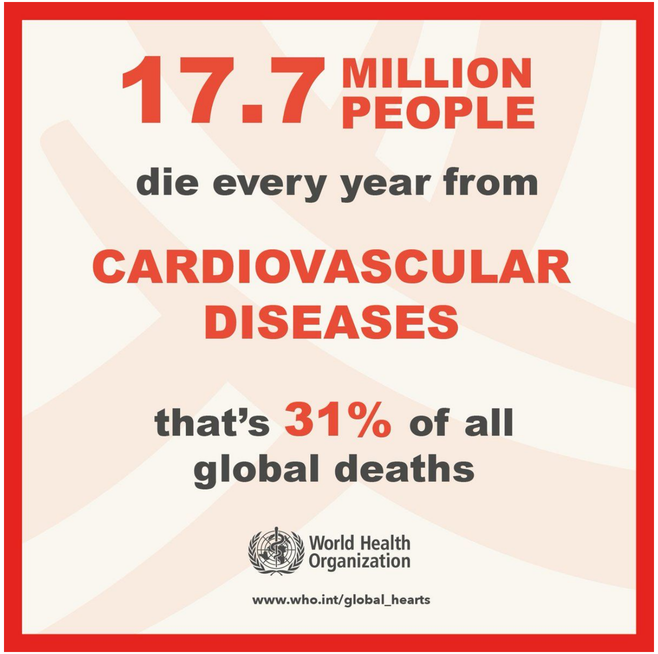
04

Conclusion

- Limitations
- Future Scope

Cardiovascular Disease (CVD)





17.7 MILLION PEOPLE

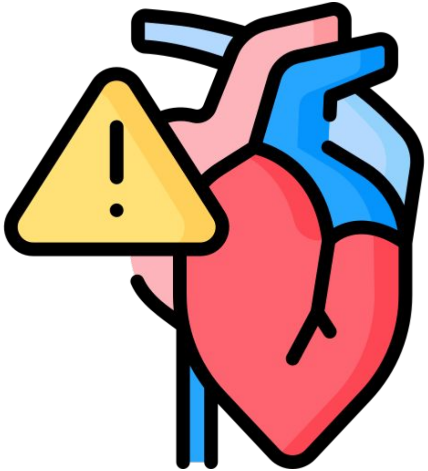
die every year from

CARDIOVASCULAR DISEASES

that's 31% of all global deaths

World Health Organization

www.who.int/global_hearts



Coronary Artery Disease, Heart Attack, Stroke, Heart Failure …

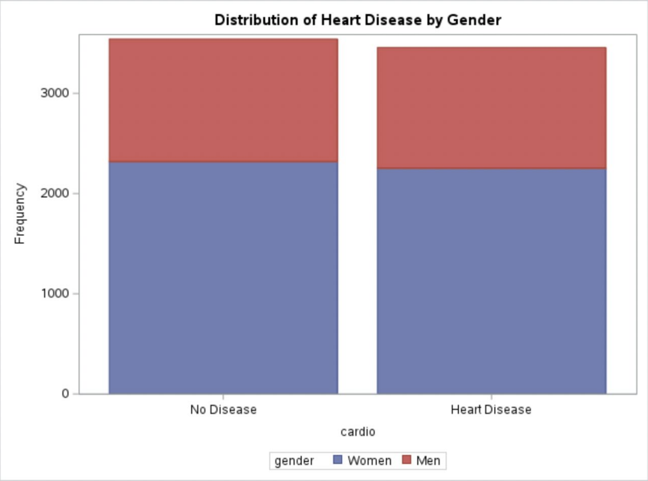
Building a model to predict cardiovascular disease is essential!

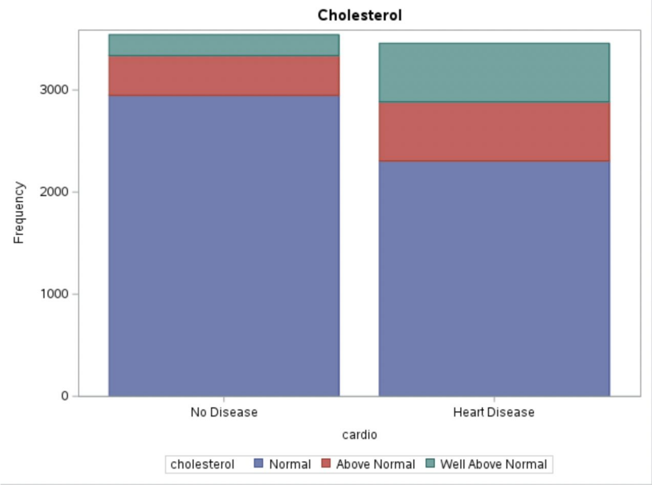
Our Data

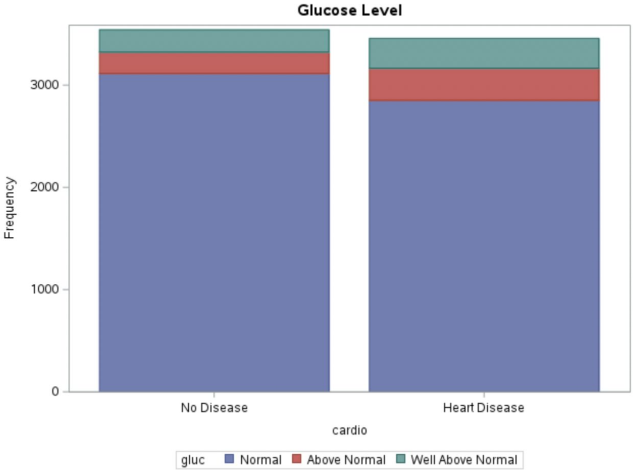
- From Kaggle
- Consists of 70,000 records of patients data
- 12 variables (5 numerical, 7 categorical)
- SRS - 7,000 observations

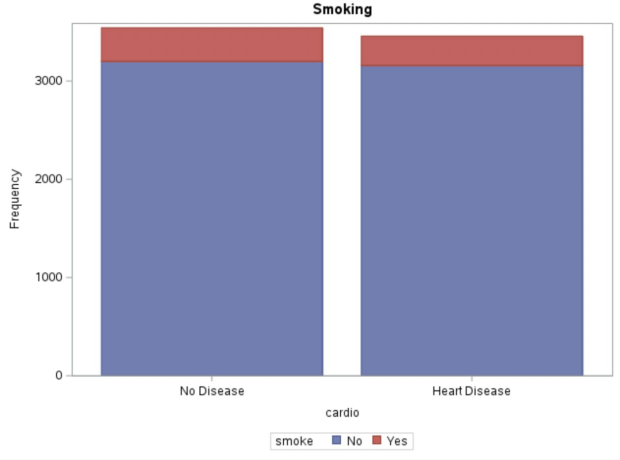
1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

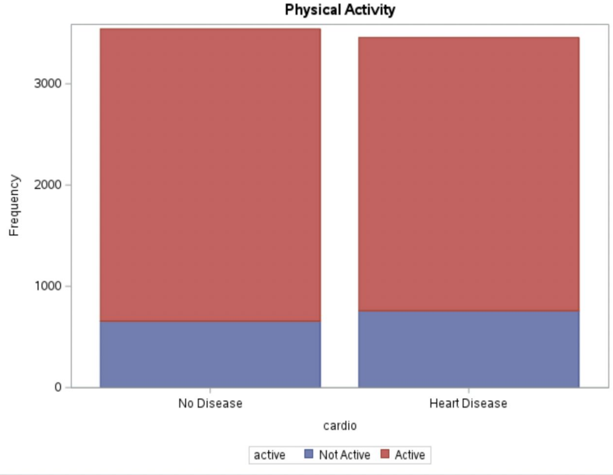
EDA

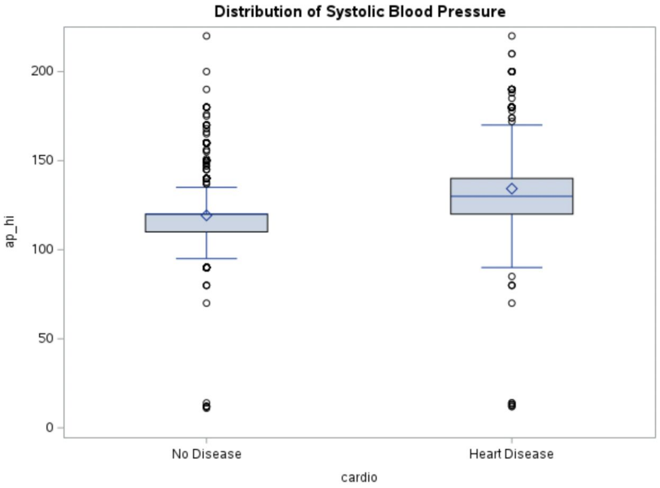

Distribution of Heart Disease by Gender

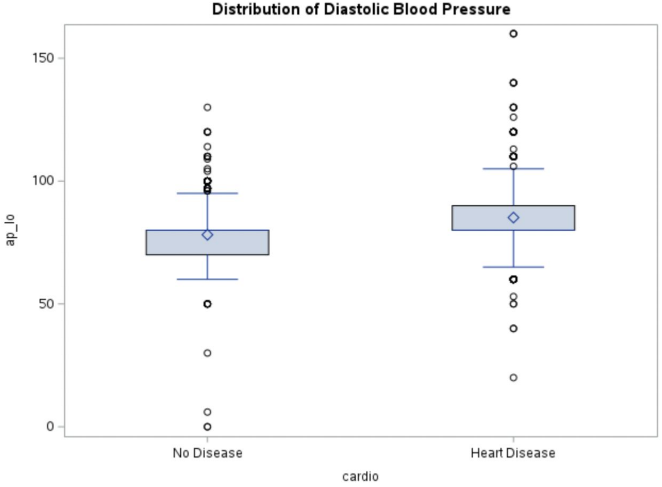

Cholesterol

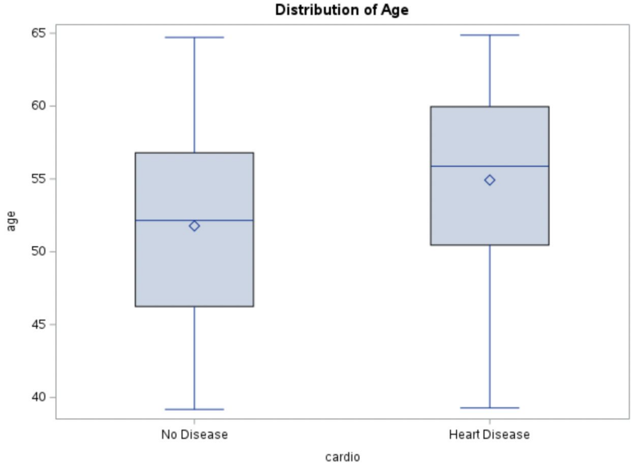

Glucose Level


Smoking

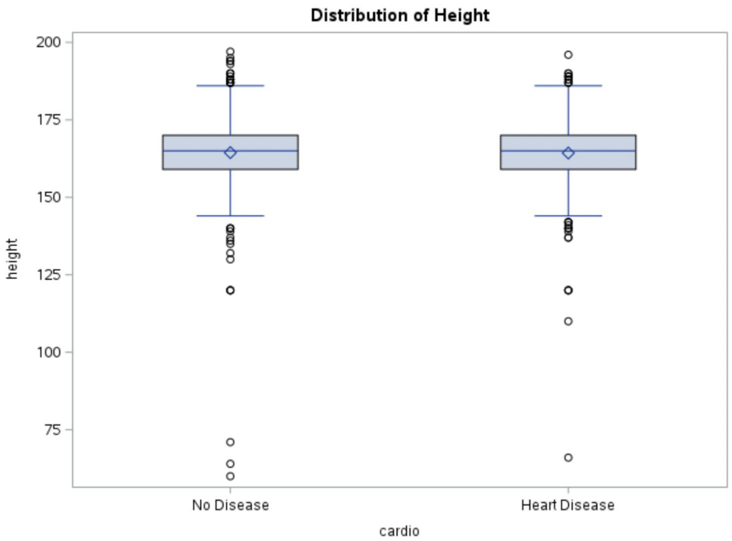

Physical Activity

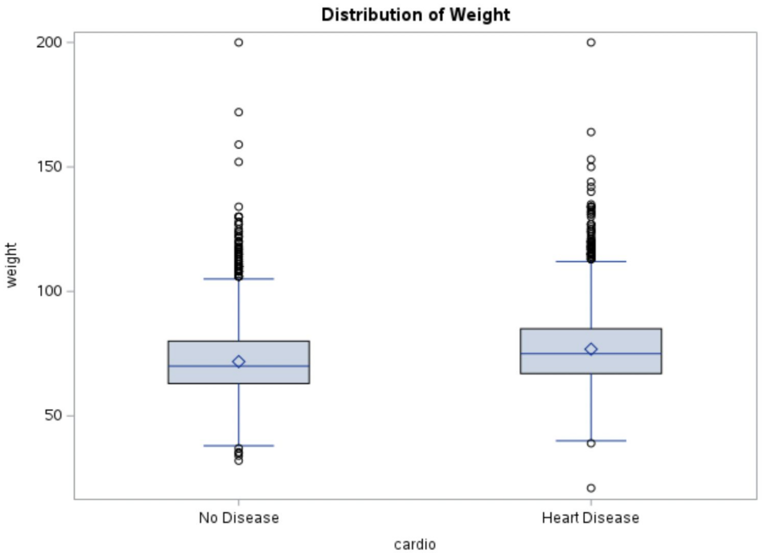




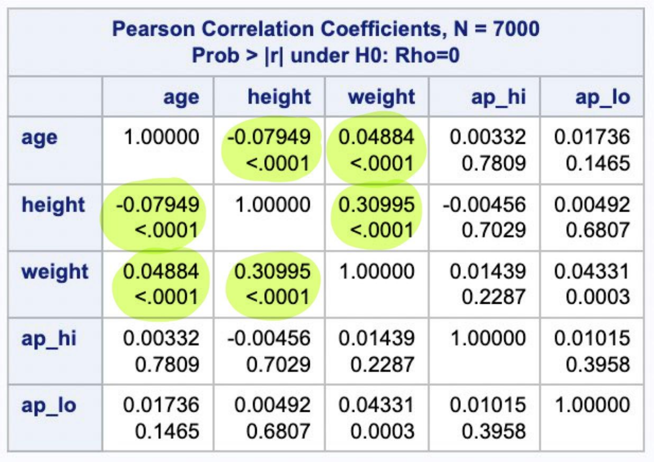


	No Disease	**Heart Disease**
Mean Age	51.7	54.9
Mean Systolic BP	119.3	134.3
Mean Diastolic BP	78.1	85.1



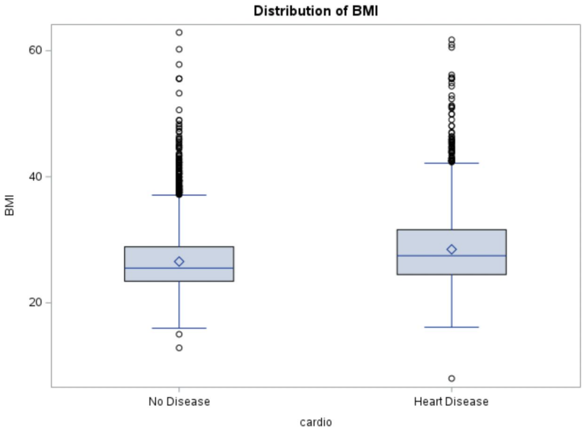


	No Disease	Heart Disease
Mean Height	164.4	164.2
Mean Weight	71.8	76.8



Pearson Correlation Coefficients, N = 7000
Prob > |r| under H0: Rho=0

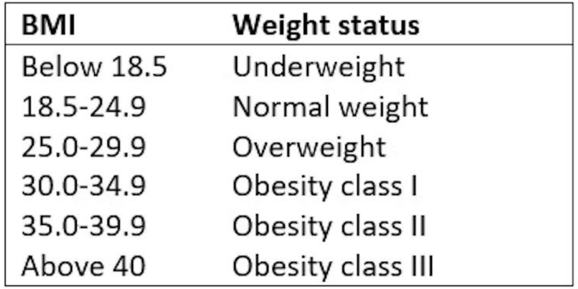
	age	height	weight	ap_hi	ap_lo
age	1.00000	-0.07949 <.0001	0.04884 <.0001	0.00332 0.7809	0.01736 0.1465
height	-0.07949 <.0001	1.00000	0.30995 <.0001	-0.00456 0.7029	0.00492 0.6807
weight	0.04884 <.0001	0.30995 <.0001	1.00000	0.01439 0.2287	0.04331 0.0003
ap_hi	0.00332 0.7809	-0.00456 0.7029	0.01439 0.2287	1.00000	0.01015 0.3958
ap_lo	0.01736 0.1465	0.00492 0.6807	0.04331 0.0003	0.01015 0.3958	1.00000



- Correlation between height, weight, or age variables are significant.

- Created a new variable, BMI (Body Mass Index)

$$\text{BMI} = \frac{\text{Weight (in kilograms)}}{\text{Height}^{2}\text{ (in meters)}}$$



BMI	Weight status
Below 18.5	Underweight
18.5-24.9	Normal weight
25.0-29.9	Overweight
30.0-34.9	Obesity class I
35.0-39.9	Obesity class II
Above 40	Obesity class III

Research Questions

1. Is there a significant difference in the mean BMI between people with cardiovascular disease and those without disease?
    a. Compare the results with the height and weight variables in the dataset

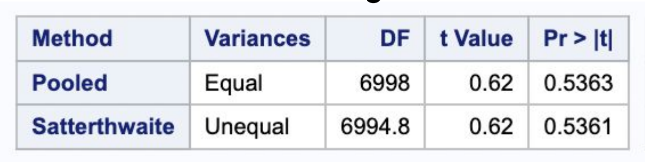
2. What factors contribute to the presence of cardiovascular disease?
    a. Build a model to assess individual's risk of cardiovascular disease

Statistical Tests

H0: There is no significant difference in the mean
Height
Weight
BMI
between the people with heart disease and those with no disease.

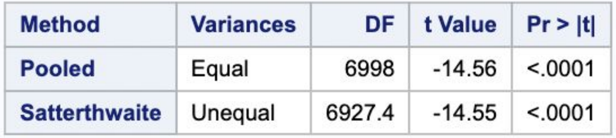
H1: There is a significant difference.

Height



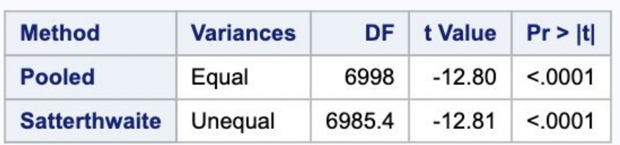
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6998	0.62	0.5363
Satterthwaite	Unequal	6994.8	0.62	0.5361

Weight



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6998	-14.56	<.0001
Satterthwaite	Unequal	6927.4	-14.55	<.0001

BMI



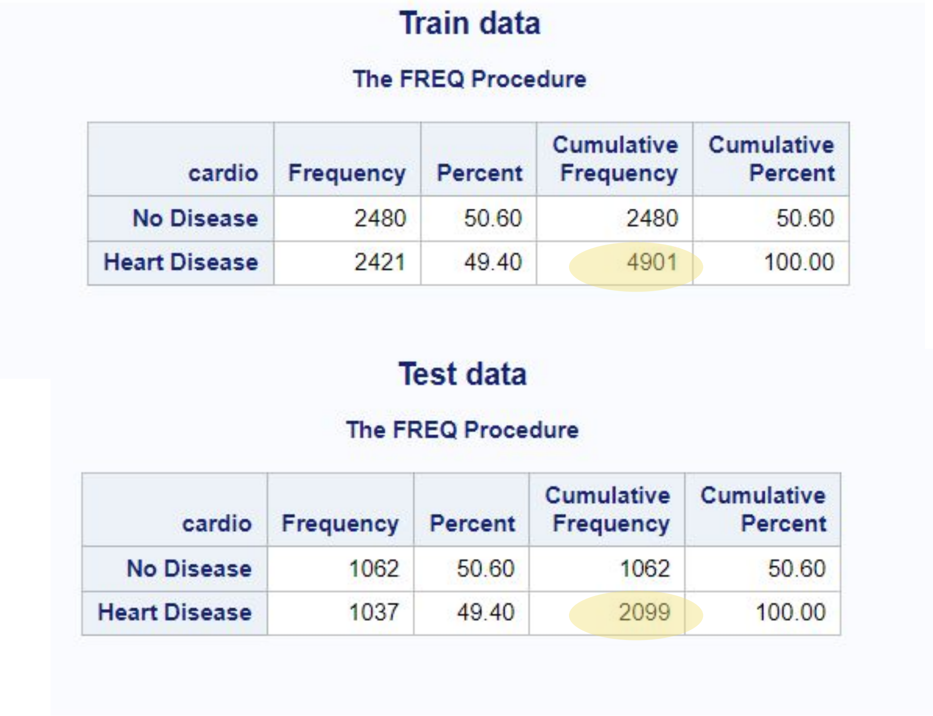
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6998	-12.80	<.0001
Satterthwaite	Unequal	6985.4	-12.81	<.0001

There is no significant difference in the mean height, but there is a significant difference in the mean weight and mean BMI.

Decision:

Replace the height and weight variables with BMI in order to reduce redundancy while we keep all the information from the original dataset.

Training and test Datasets



Train data

The FREQ Procedure

cardio	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No Disease	2480	50.60	2480	50.60
Heart Disease	2421	49.40	4901	100.00

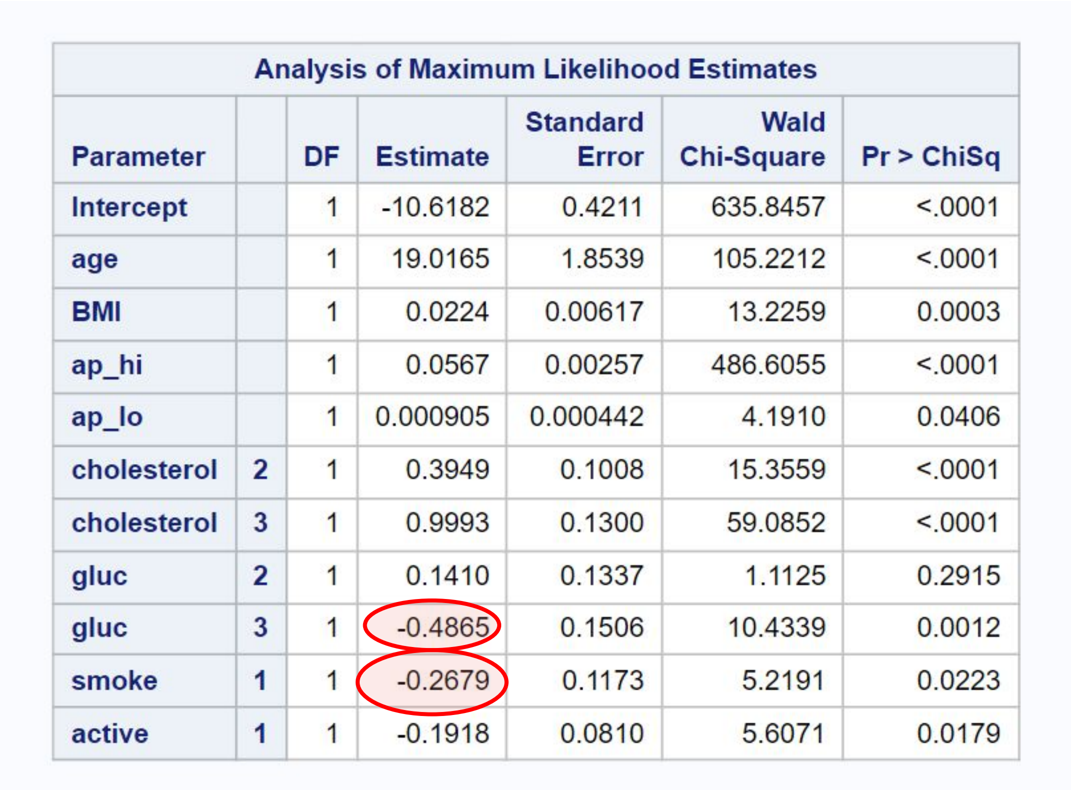
Test data

The FREQ Procedure

cardio	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No Disease	1062	50.60	1062	50.60
Heart Disease	1037	49.40	2099	100.00

Dividing the data into ***70% training*** and ***30% testing*** datasets using ***Random Sampling***

Logistic Regression



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-10.6182	0.4211	635.8457	<.0001
age		1	19.0165	1.8539	105.2212	<.0001
BMI		1	0.0224	0.00617	13.2259	0.0003
ap_hi		1	0.0567	0.00257	486.6055	<.0001
ap_lo		1	0.000905	0.000442	4.1910	0.0406
cholesterol	2	1	0.3949	0.1008	15.3559	<.0001
cholesterol	3	1	0.9993	0.1300	59.0852	<.0001
gluc	2	1	0.1410	0.1337	1.1125	0.2915
gluc	3	1	-0.4865	0.1506	10.4339	0.0012
smoke	1	1	-0.2679	0.1173	5.2191	0.0223
active	1	1	-0.1918	0.0810	5.6071	0.0179

Training our model into training dataset, using *stepwise selection*

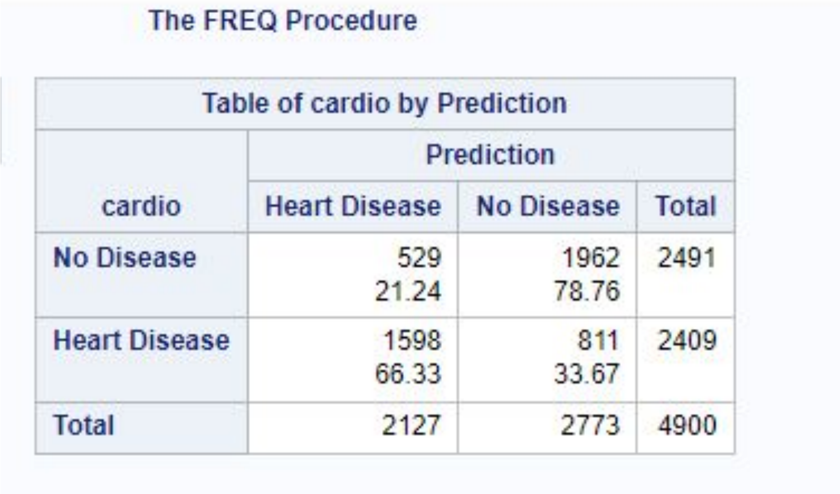
- Gender and Alcohol are not good predictors

- Glucose and Smoke have unexpected result !

"The more you smoke..
The higher your blood sugars are..
The chance of you getting
heart disease increase" - CDC

Logistic Regression

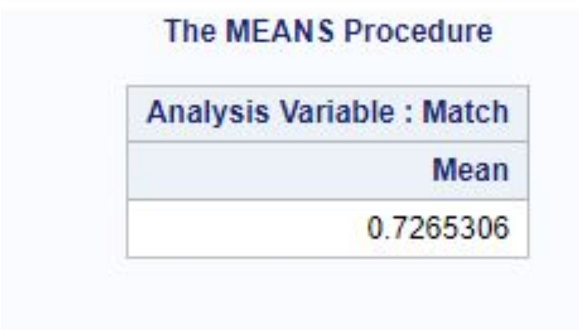
Confusion Matrix



The FREQ Procedure

Table of cardio by Prediction			
	Prediction		
cardio	Heart Disease	No Disease	Total
No Disease	529 21.24	1962 78.76	2491
Heart Disease	1598 66.33	811 33.67	2409
Total	2127	2773	4900

Accuracy



The MEANS Procedure

Analysis Variable : Match
Mean
0.7265306

Association between Categorical Variables

Chi-Square Test of Independence

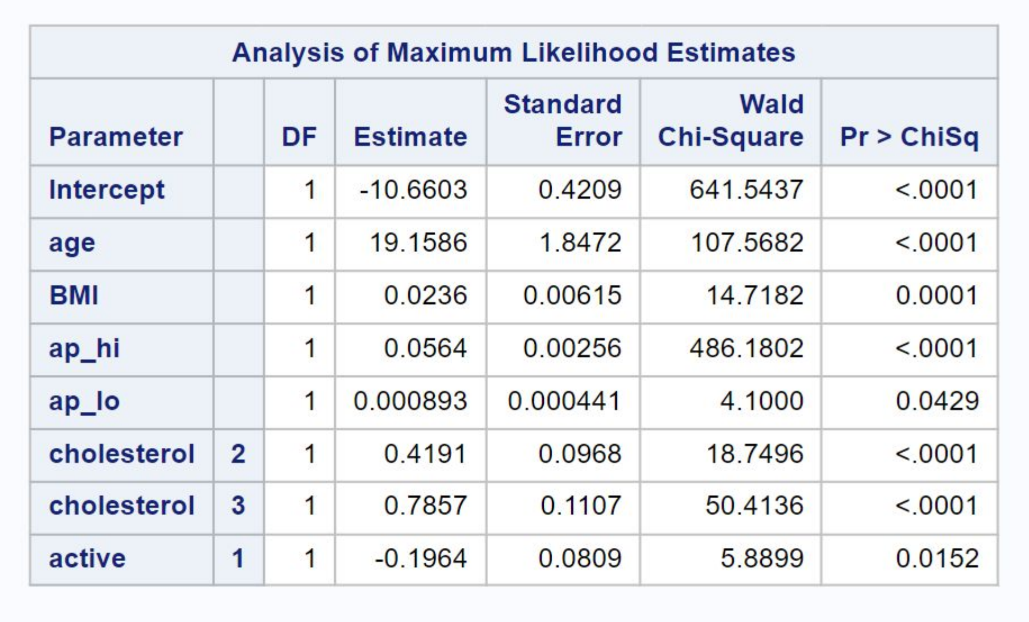
Variables	P-value
Cholesterol & Glucose	<.0001
Cholesterol & Smoke	0.0298
Smoke & Gender	<.0001

H0: Two variables are independent.
H1: Two variables are not independent .
With p-value < 0.05, these 3 pairs of variables are associated and are not independent of each other.

Decision:

Try dropping glucose and smoke from the model and see if the accuracy improves.

Logistic Regression



Analysis of Maximum Likelihood Estimates

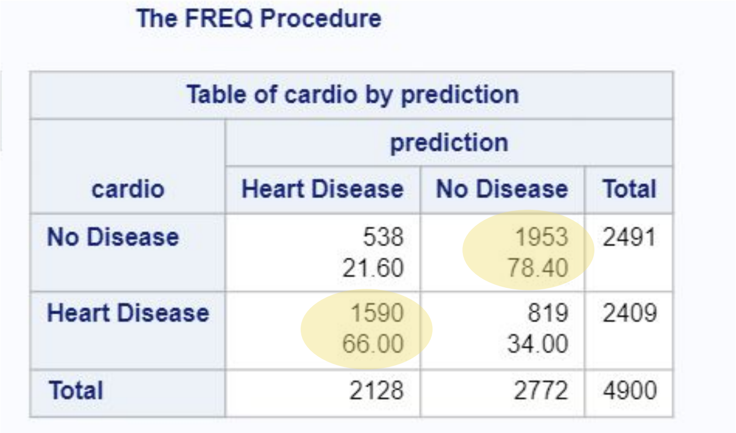
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-10.6603	0.4209	641.5437	<.0001
age		1	19.1586	1.8472	107.5682	<.0001
BMI		1	0.0236	0.00615	14.7182	0.0001
ap_hi		1	0.0564	0.00256	486.1802	<.0001
ap_lo		1	0.000893	0.000441	4.1000	0.0429
cholesterol	2	1	0.4191	0.0968	18.7496	<.0001
cholesterol	3	1	0.7857	0.1107	50.4136	<.0001
active	1	1	-0.1964	0.0809	5.8899	0.0152

Fitting logistic regression into training dataset without Glucose and Smoke

- Age, BMI, Ap_hi, Ap_lo, cholesterol, active are good predictors for heart disease

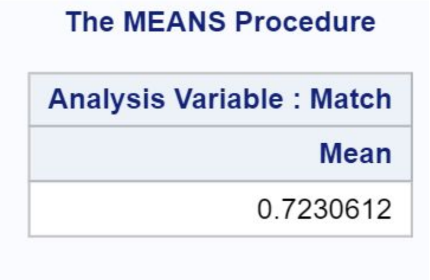
Logistic Regression

Training



The FREQ Procedure

Table of cardio by prediction			
	prediction		
cardio	Heart Disease	No Disease	Total
No Disease	538 21.60	1953 78.40	2491
Heart Disease	1590 66.00	819 34.00	2409
Total	2128	2772	4900

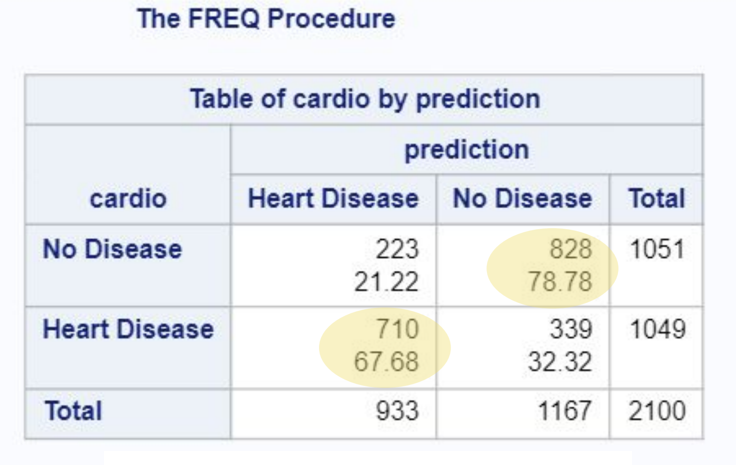


The MEANS Procedure

Analysis Variable : Match
Mean
0.7230612

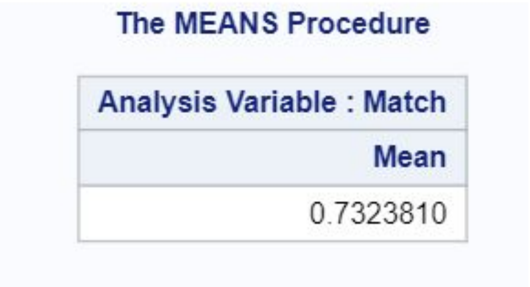
Vs 0.7265

Testing



The FREQ Procedure

Table of cardio by prediction			
	prediction		
cardio	Heart Disease	No Disease	Total
No Disease	223 21.22	828 78.78	1051
Heart Disease	710 67.68	339 32.32	1049
Total	933	1167	2100



The MEANS Procedure

Analysis Variable : Match
Mean
0.7323810

Logistic Regression

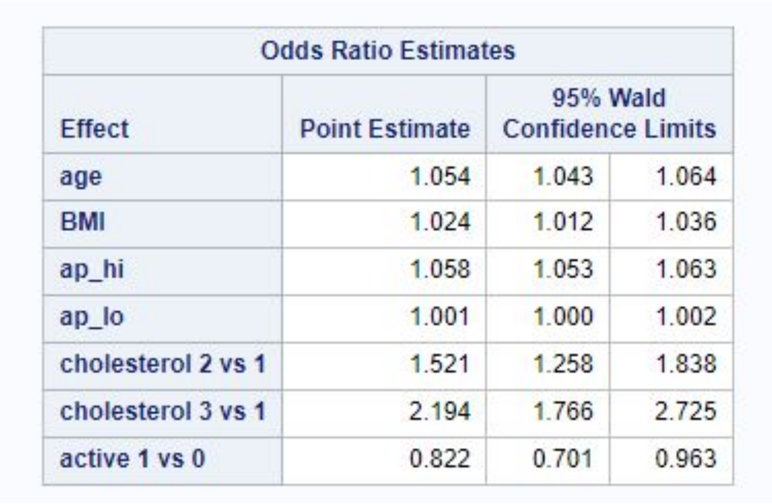
Model

$$\boldsymbol{Log}\left(\frac{\boldsymbol{p}}{\boldsymbol{1-p}}\right) = -10.6603 + 19.1586\boldsymbol{Age} + 0.0236\boldsymbol{BMI} + 0.0564\boldsymbol{Ap-hi} + 0.0008\boldsymbol{Ap-lo} + 0.4191\boldsymbol{Cholesterol2} + 0.7857\boldsymbol{Cholesterol3} - 0.1964\boldsymbol{Active}$$

P = “Heart disease”

Odds ratio

- Every unit increase in **BMI** associated with a **2.4 %** *(1.024 - 1)* increase in the odds of getting heart disease
- Being active will decrease the odds of getting heart disease by **17.8%** *(1 - 0.822)*



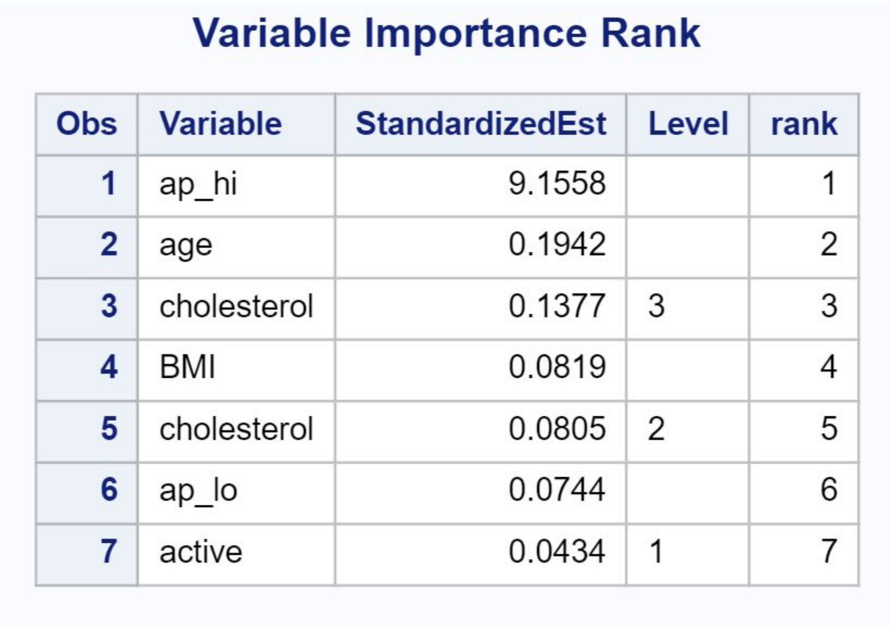
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.054	1.043	1.064
BMI	1.024	1.012	1.036
ap_hi	1.058	1.053	1.063
ap_lo	1.001	1.000	1.002
cholesterol 2 vs 1	1.521	1.258	1.838
cholesterol 3 vs 1	2.194	1.766	2.725
active 1 vs 0	0.822	0.701	0.963

Logistic Regression

Standardized coefficients are coefficients adjusted so that they may be interpreted as having the same, standardized scale and the ***magnitude of the coefficients can be directly compared (ranked)****.*

(Menard S. 2004)



Variable Importance Rank

Obs	Variable	StandardizedEst	Level	rank
1	ap_hi	9.1558		1
2	age	0.1942		2
3	cholesterol	0.1377	3	3
4	BMI	0.0819		4
5	cholesterol	0.0805	2	5
6	ap_lo	0.0744		6
7	active	0.0434	1	7

Conclusion

1. There is significant difference in the mean of BMI between people with and without heart disease
2. Logistic regression is a pretty good model in predicting heart disease. (73.23% accuracy)
3. Important Factors contributing to the likelihood of getting heart disease: Ap_hi, age , cholesterol, BMI, Ap_lo, active.

Implication

“ Be and active person, keep your blood pressure, cholesterol levels and BMI normal to decrease the likelihood of getting a heart disease ! “

Limitations

1. The proportion of people with and without heart disease in data is nearly equal which does not accurately reflect the reality
2. Limitation in using Accuracy as sole metric to evaluate model

Recommendation

1. Adding interaction term
2. Building other classifications models and compare models with other metrics as well

Thank you