**A**

*Research Paper*

*On*

# Statistical Analysis Of Heart Disease

**Submitted by**

*Gote Anuja Kedarnath*

*Pishte Apurva Shivaji*

**Roll No.: 23221,23248**

**MCA–I**

**SEM–II**

**Under the guidance of**

*Ankush Kudale*

**For the Academic Year 2023-24**


***Sinhgad Technical Education Society's***

**Sinhgad Institute of Management**

**Vadgaon Bk Pune 411041**

**(Affiliated to SPPU Pune & Approved by AICTE New Delhi)**

# Statistical Analysis of Heart Disease

# INDEX

# ABSTRACT

We are doing project on Heart Disease. We are collecting data from WWW.kaggle.com for this analysis ,we took a sample of 1000.There are 16 variable. Now a day's many peoples are suffer in heart disease problem. Therefor the main objective of this project is to find the factors which are affecting on the heart disease. We did our test & analysis on the basis of some measurements of patients like BMI, Gender, Age Category, Diabetes, General Health , Smoking , Alcohol Drinking, Stroke , Physical Health, Mental Health Diff Walking, Physical Activity, Sleep Time, Asthma, Kidney Disease ,Skin Cancer .

The main purpose is the study the factor which are affecting on the heart disease and reasons of heart failure. In that we are also test the :

1] Heart disease is may not be independent on gender.

2] Heart disease is may not be independent on age group.

3] Heart disease is may not be independent on diabetes.

4] Proportion of male and female is same for presence of heart diseases.

Statistical analysis has identified a number of risk factors for heart diseases including age, blood pressure, smoking, diabetes, asthma and mental health, heart disease background in family, obesity and lack of physical activity. The awareness of heart disease risk factors assists treatment and healthcare specialists to identify patients who are subject to high risk factors.

❖**Software used**

1] Ms –Excel

2] R- software

3] R- Studio

# INTRODUCTION

In this research project, the overall research and the project implementation is focused on predict a heart disease. The heart is an essential part of the human body. It pumps blood to every single part of the human body. The goal of the study is to develop a model that can help clinicians make better medical judgments using computerized and automated patient information. It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Any disorder that can lead to disturbing the functionality of the heart is called heart disease.

Heart disease in pregnancy is very rare but potential serious and complicate apex 1% of all pregnancies. The causes of heart disease depend on the type of disease. Some possible causes include lifestyle , genetics, infections, medicines, and other diseases. Diabetes (depending on gender) are two to four times more likely to die of coronary heart disease, and twice as likely to die of stroke, as person without diabetes. The prevalence of all these risk factors have been found to be higher in West Virginia than in the nation as a whole.

The variables being used in this project are BMI, Smoking, alcohol drinking, stroke, physical health, mental health, diffwalking, gender, age category, Diabetes, Physical activity, General health, Sleep time, Asthma, Kidney disease. Heart disease is the nation's No. 1 cause of death, killing about 650,000 people every year. Life expectancy is cut short by the disease and the health problems that stem from it.

Hope our project will help to understand the major causes of heart disease in our daily basis though the variables we have chosen.


# OBJECTIVE

The present work is intended to meet the following objectives:

- To determine the factor which are affect on the heart disease.
- To test heart disease is independent on age group.
- To test heart disease is independent on gender.
- To test heart disease is independent on diabetes.
- To test proportion of male and female is same for presence of heart disease.
- To find a high performance predictive model that classifies the heart disease.

❖ **Statistical Methods:**

We are going to use following Methods

- Chi- square test
- Regression analysis: Binary logistics regression
- Decision Tree

# PROBLEM STATEMENT

Chest pain (angina): Angina pectoris is a common symptom of heart disease characterized by discomfort, pressure, or pressure in the chest. It occurs when the heart muscle does not receive enough oxygen-rich blood.

Dyspnea (shortness of breath): Heart disease can lead to a build-up of fluid in the lungs, causing difficulty breathing, especially when exercising or lying down.

Fatigue: Heart disease can reduce the heart's ability to pump effectively, leading to reduced oxygen delivery to the tissues and leading to fatigue and weakness.

Swelling (edema): Fluid retention due to heart failure can lead to swelling of the legs, ankles, feet, or abdomen.

Irregular heartbeat (arrhythmia): Heart disease can disrupt the heart's normal electrical impulses, leading to an irregular heartbeat, palpitations, or skipped beats.

Dizziness or fainting (syncope): Decreased blood flow to the brain due to heart disease can cause dizziness, lightheadedness, or fainting.

Heart attack (myocardial infarction): A heart attack occurs when blood flow to part of the heart is blocked, leading to damage or death of heart muscle tissue. Symptoms include chest pain, shortness of breath, nausea, sweating, and discomfort in the arms, back, neck, or jaw.

Stroke: Heart disease increases the risk of blood clots forming in the heart and traveling to the brain, leading to stroke. Symptoms include sudden weakness or numbness in the face, arm, or leg, confusion, difficulty speaking or understanding speech, and difficulty walking.

Heart failure: Heart failure occurs when the heart's pumping ability is impaired, leading to a buildup of fluid in the lungs and other tissues. Symptoms include shortness of breath, fatigue, swelling and difficulty exercising.

Sudden cardiac arrest: In some cases, heart disease can lead to sudden cardiac arrest, when the heart suddenly stops beating. This can result in loss of consciousness, no pulse, and immediate collapse.

# LITERATURE REVIEW

Introduction

Define heart disease and its importance as a global health problem.

Provide a brief overview of heart disease prevalence and mortality worldwide.

Emphasize the importance of understanding the various factors contributing to heart disease for effective prevention and treatment.

Risk factors

Discuss the primary risk factors associated with heart disease such as hypertension, high cholesterol, diabetes, obesity, smoking, sedentary lifestyle, and family history.

Review epidemiological studies and meta-analyses that identify the strength of the association between these risk factors and the development of heart disease.

Pathophysiology

Explore the underlying mechanisms and pathways involved in the development of various types of heart disease, including coronary artery disease, myocardial infarction, heart failure, arrhythmias, and valvular heart disease.

Highlight the key molecular, cellular and physiological processes involved in the pathogenesis of these conditions.

Diagnostic methods

Summarize conventional and advanced diagnostic techniques used in the evaluation of heart disease, such as electrocardiography (ECG), echocardiography, cardiac MRI, coronary angiography, stress testing, and biomarker tests.

Evaluate the accuracy, reliability, and limitations of these diagnostic modalities in various clinical scenarios.

Treatment strategies

Review pharmacologic and nonpharmacologic interventions used to treat heart disease, including lifestyle modifications, drug therapy (eg, antiplatelet agents, statins, beta-blockers, ACE inhibitors), percutaneous coronary intervention (PCI), coronary artery bypass grafting (CABG), cardiac resynchronization therapy (CRT), implantable cardioverter-defibrillators (ICD) and heart transplantation.

Discuss evidence from clinical trials and meta-analyses supporting the efficacy and safety of these interventions.

Prevention and public health initiatives

Explore population-based strategies to reduce the burden of heart disease, such as health education campaigns, promotion of healthy lifestyles, community-based interventions, and policy measures targeting tobacco control, dietary habits, and physical activity.

To assess the effectiveness of primary prevention measures in reducing the incidence of heart disease and related mortality.

Recent advances and future directions

Highlight emerging trends and innovations in the diagnosis, treatment, and prevention of heart disease, including new therapeutic targets, biomarkers, medical devices, and digital health technologies.

Discuss ongoing research efforts and clinical trials investigating promising strategies for

improving outcomes in patients with heart disease.

Conclusion

Summarize key findings and insights from the literature review.

Emphasize the importance of a multidisciplinary approach involving primary prevention, early detection, personalized treatment, and patient-centered care in addressing the challenges that heart disease presents.

# METHODOLOGY

We collect the data from the website www.kaggle.com .we took sample of 1000 patients. We did our test & analysis on the basis of some measurements ofpatients like that BMI, Gender, Age category, Diabetes, General health , Smoking , Alcohol drinking, Stroke , Physical health, Mental Health , Diff walking, Physical activity, Sleep time, Asthma, Kidney disease ,Skin cancer

## ❖ DESCRIPTION OF EACH TERM IN DATA

1) **Gender** : Female : 0,
Male : 1

2) Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs =100 cigarettes]

**Smoking** : No : 0,
Yes : 1

3) Heavy drinkers (adult men having more than 14 drinks per week and adultwomen having more than 7 drinks per week)

**Alcohol drinking** : No : 0,
Yes : 1

4) (Ever told) (you had) a stroke?

**Stroke** : No : 0,
Yes : 1

5) Do you have serious difficulty walking or climbing stairs?

**Diff walking**: No : 0,
Yes : 1

6) Now thinking about your **physical health**, which includes physical illnessand injury, for how many days during the past 30

7) Thinking about your **mental health**, for how many days during the past 30days was your mental health not good?

( coding for 6 and 7)

0 <= 10 : 0

10 to <= 20 : 1

20 to <= 25 : 2

25 to <= 30 : 3

**8) Age Category**

Fourteen-level age category

- young(30<=Age Category <35) : 0
- mature(35<=Age Category <50): 1
- senior(50<=Age Category <65):2
- old(65<=Age Category <80):3
- very old(80<=Age Category ):4

9) **Race** : Imputed race/ethnicity value

10) (Ever told) (you had) **diabetes**?

No:0
No, borderline diabetes : 1
Yes(during Pregnancy) : 2
Yes: 3

11) **General Health** : Would you say that in general your health is...

Poor : 0
Fair : 1
Good : 2
Very Good : 3
Excellent : 4

12) **Asthma:** (Ever told) (you had) asthma?

13) **Kidney Disease** : Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?

14) **Skin Cancer:** (Ever told) (you had) skin cancer?

NO: 0,

Yes: 1

**Chart for data presentation:**

The chart used for visual data presentation are given below.

- Bar Charts
- Pie Chart
- Multiple bar charts

## Test used:

### ❖ Chi-Square test for attributes :

### Test for independency of two attributes:

$H_O$: Two attributes A and B are independent

Against

$H_1$: Two attributes A and B are dependent.

Under $H_0$ test statistic is,

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(Oij - Eij)^2}{Eij} \sim \chi^2_{(m-1)(n-1)d.f.}$$

Where,

Oij= Observed frequency.

Eij= Expected frequency.

$N = \sum_{i=1}^{m} \sum_{j=1}^{n} Oij$

$A_i = \sum_{j=1}^{n} Oij$, $B_j = \sum_{i=1}^{m} Oij$

$$E_{ij} = \frac{Ai * Bj}{N} \quad ; \quad i=1,2,\ldots m,\ j=1,2\ldots n$$

## Decision rule:-

We reject $H_0$ at α% level of significance if, $\chi^2_{cal} \geq \chi^2_{(m-1)(n-1),\alpha}$ and accept

$H_0$ otherwise.

**Decision on p-value:** We reject $H_0$ at α% level of significance if, p-value < α , otherwise accept $H_0$.

## ❖ Binary Logistic Regresssion :

Binary logistic regression is a form of regression which is used when the dependent variable is binary and the independent variables are of any type. The goal of an analysis using logistic regression method is to find the best fitting and most reasonable model to describe the relationship between the outcome (dependent or response variable) and a set of independent (predictor or understand the impact of explanatory) variables and to determine the percent of variation in the dependent variable explained by the independent variables.

### Assumptions:

1. The dependent variable in binary logistic regression must be binary.
2. Independent variables in binary logistic regression can be nominal ,ordinal, radio and interval scale**.**
3. Logistic regression does not require a linear relationship between the dependent and independent variables.
4. The error terms (residuals) do not need to be normally distributed.
5. Homoscedasticity is not required.
6. Logistic regression requires there to be little or no multicolinearity among the independent variables.

The Binary logistic regression model is,

$$P(Y)=\frac{e^{b0+b1x1+b2x2+\cdots+bnxn}}{1+e^{bo+b1x1+b2x2+\cdots+bnxn}}$$

Where,
P: probability of Y occuring
e: natural logarithm base
$b_0$: interception at y-axis
$b_1$: line gradient
$b_n$: regression coefficient $x_n$
$x_i$: predictor variable

We use the transformation called logit, which forces the prediction equation to predict the value between zero and one.

## ❖ Stepwise regression:

Stepwise regression includes regression models in which the choice of predictive variables is carried out by an automatic procedure.
Forward Selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most and repeating this process until none improves the model.

Backward elimination, which involves starting with all candidate variables, testing the deletion each variable using a chosen model comparison criterion ,deleting the variable(if any) that improves the model the most by being deleted and repeating this process until no further improvement is possible.

## ❖ AIC-Akien's Information Criteria:

The general for calculating AIC,
AIC=-2*ln (likelihood) +2*k
Here, ln is natural logarithm is the value of the likelihood k is the no. of parameter in the model.
AIC can also be calculated using residual sum of squares from regression,
AIC=n*ln (RSS /n) +2*k
Where, n is the no. of data points. RSS is the residual sum of squares. AIC requires a bias-adjustment small sample size.
Form the stepwise regression we get the regression models for our data. The model which has smallest value of AIC as compare to the other models gives the best regression model for our data.

## ❖ Decision Tree:-

Decision tree are powerful classification algorithm that are becoming increasing more popular with the growth of data mining in the field of information systems. A decision tree is a decision support tool that uses a tree-like graph or model of decision and their possible consequences include chances event outcome, resources cost and utility.

In the decision tree , there are two nodes while are the decision nodes

and leaf nodes. Decision nodes are used to make any decision & have multiple branches whereas, leaf node are the output of these decision and do not contain any further branches. It start with root node where expands on further branches and constant tree like structure. In order to built a tree, we use CART algorithm. A decision tree can contain categorical data i.e. binary representation data as well as numeric data.

**Terminologies:**

1. Root node
2. Leaf node
3. Splitting
4. Branch

## ❖ Confusion Matrix:-

A confusion matrix is a table that is often used to describe the performance of classification model on a set of data for which true values are known. It is a special table layout that allows visualization of performance an algorithm typically for supervised learning.

- **Sensitivity:-**

Sensitivity measure the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).In medical diagnosis, test sensitivity is the ability of a test to correctly identify those with the disease (true positive rate).

- **Specificity:-**

Specificity measure the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having condition).In medical diagnosis, test specificity is the ability of the test to correctly identify those without the disease (true negative rate)

| Predicted value | Actual value | |
|---|---|---|
| | True Positive | False Negative |
| | False Positive | True Negative |

$$\text{Sensitivity} = \frac{number\ of\ true\ positives}{number\ of\ true\ positive + number\ of\ false\ negatives} * 100$$

$$\text{Specificity} = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives} * 100$$

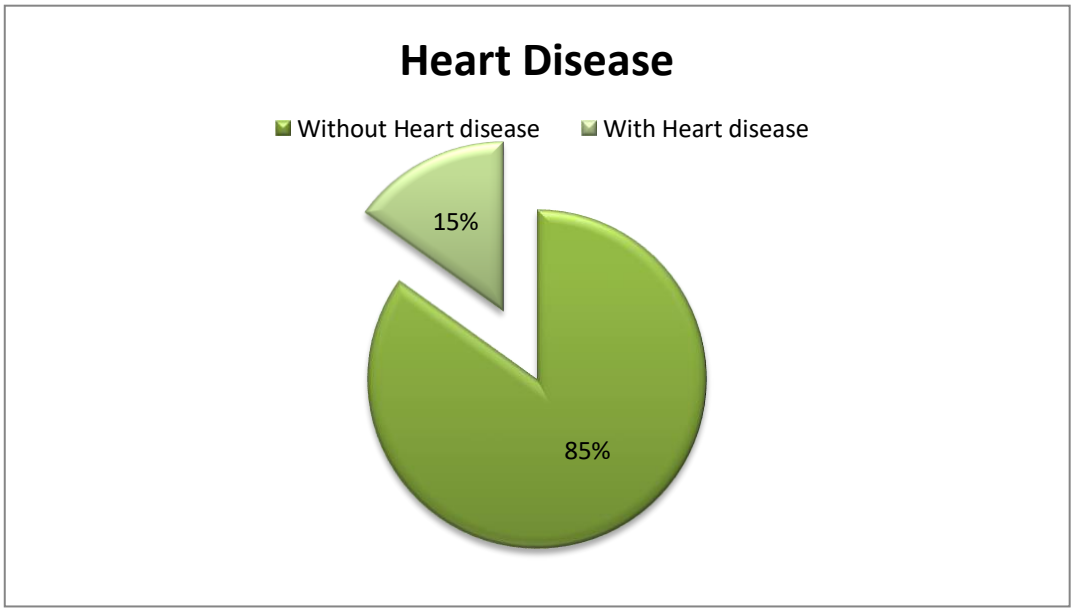
$$\text{Accuracy} = \frac{True\ positive + True\ negative}{True\ Positive + False\ Negative + True\ Negative + False\ Positive} * 100$$

# OBSERVATION AND FINDING

## Graphical Representation

### 1. Heart Disease

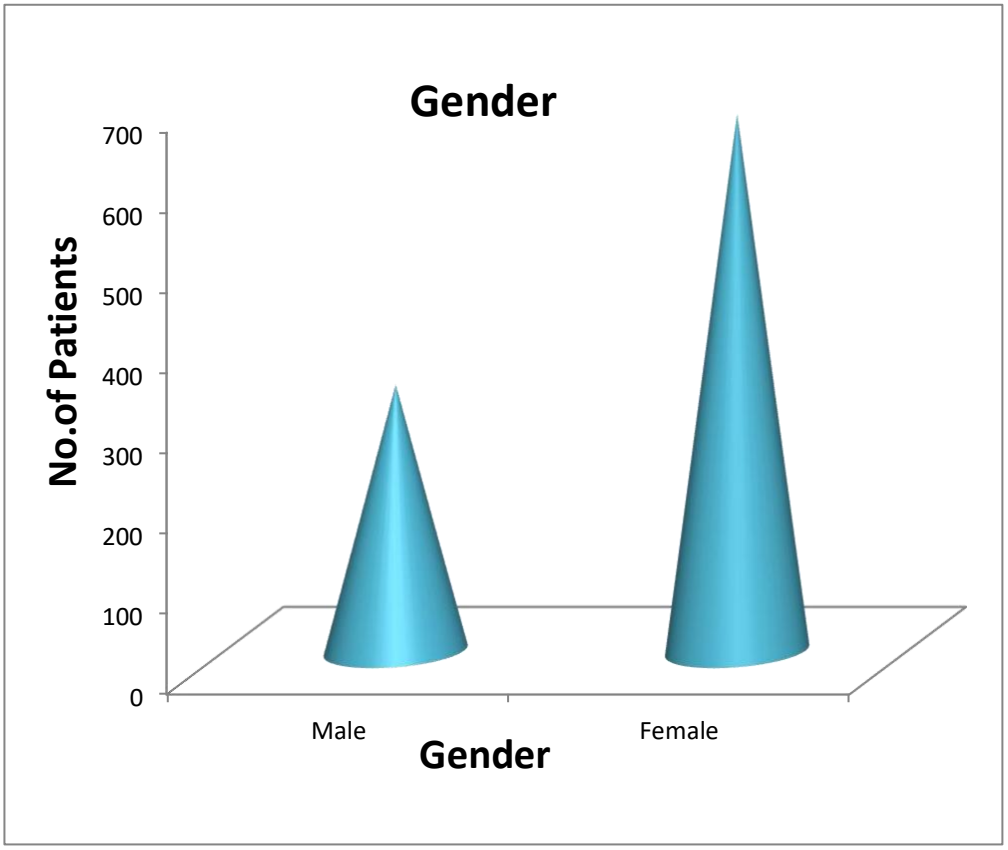
| Heart disease | No. of Patients |
|---|---|
| Without Heart disease | 849 |
| With Heart disease | 151 |
| Total | 1000 |




**Conclusion** : From the above pie chart, we can say that there are more patients without heart disease.

## 2. Gender:

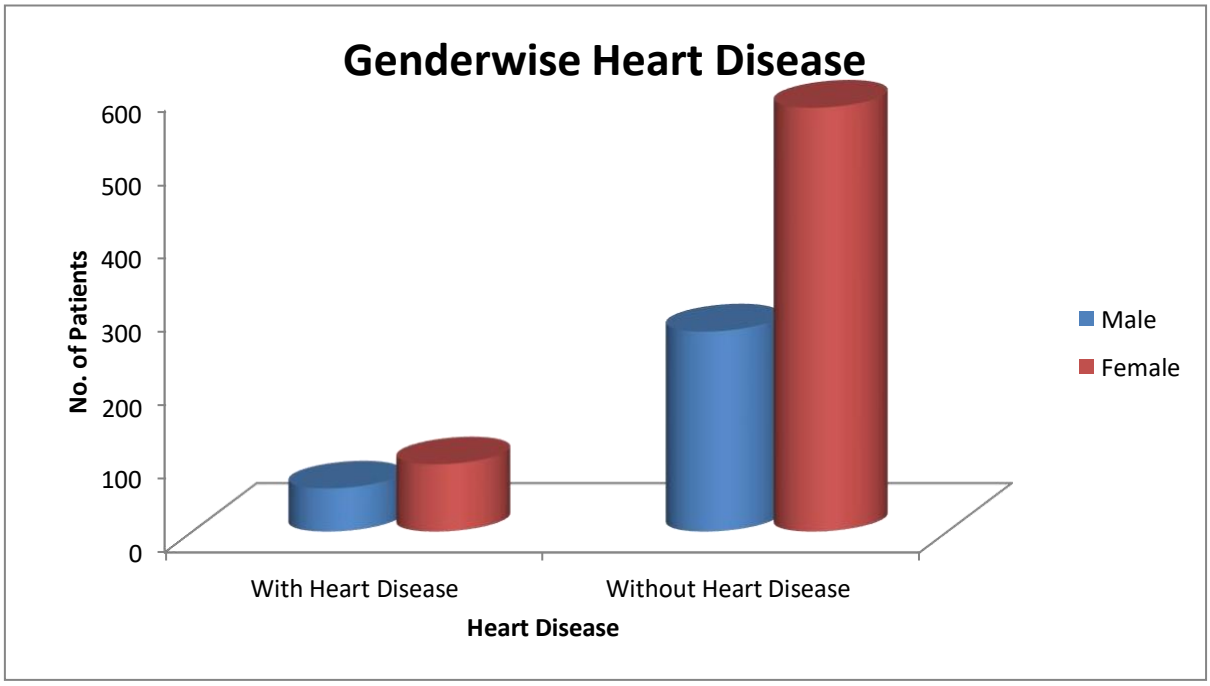
| Gender | No. of Patients |
|---|---|
| Male | 332 |
| Female | 668 |
| Total | 1000 |




**Conclusion**: Form the above chart, we can say that there are more females patients than male patients.

## 3. Gender wise Heart Disease:

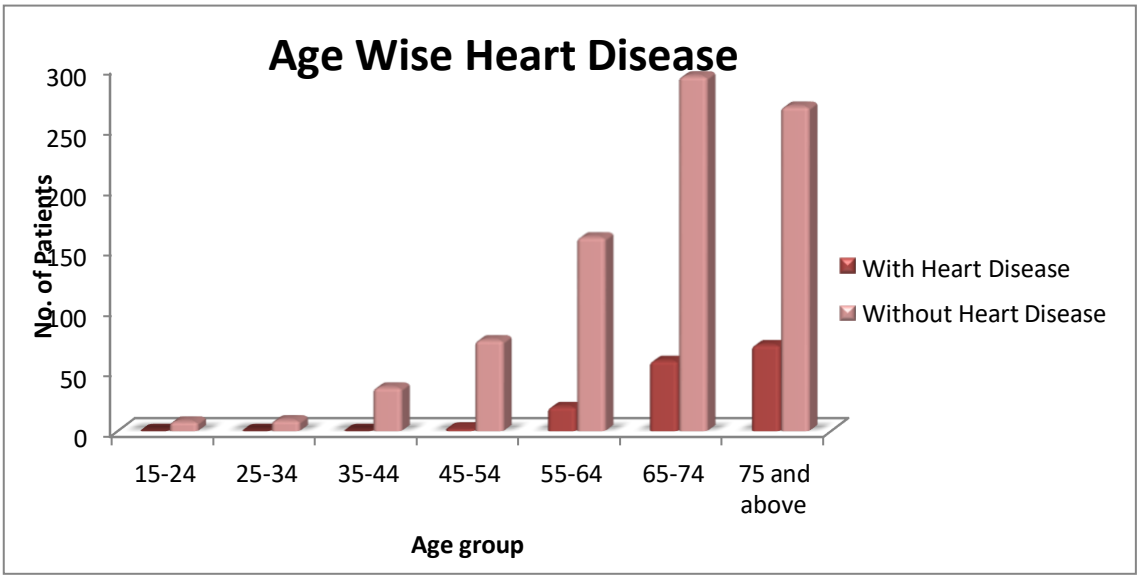
| Gender / Heart Disease | Male | Female | Total |
|---|---|---|---|
| **With Heart Disease** | 59 | 92 | 151 |
| **Without Heart Disease** | 273 | 576 | 849 |
| **Total** | 332 | 668 | 1000 |




**Conclusion** : From the above chart,we can say that there are more females without heart disease as compared to males.

## 4. Age wise Heart Disease:

| Heart Disease / Age Group | With Heart Disease | Without Heart Disease |
|---|---|---|
| 15-24 | 0 | 8 |
| 25-34 | 0 | 9 |
| 35-44 | 0 | 36 |
| 45-54 | 2 | 75 |
| 55-64 | 20 | 160 |
| 65-74 | 58 | 293 |
| 75 and above | 71 | 268 |
| Total | 151 | 849 |




**Conclusion** : From the above chart, we can say that the age range 65-74 has more patients without heart disease, while 75 and older has more heart disease patients.

# Descriptive Statistics

| Statistics | Skin Cancer | BMI | Smoking | Alcohol Drinking | Stroke | Physical Health |
|---|---|---|---|---|---|---|
| Mean | 0.19 | 28.6332 | 0.421 | 0.031 | 0.088 | 4.624 |
| Standard Error | 0.012412 | 0.23557 | 0.01562 | 0.005484 | 0.00896 | 0.287192 |
| Median | 0 | 28.02 | 0 | 0 | 0 | 0 |
| Mode | 0 | 25.06 | 0 | 0 | 0 | 0 |
| Standard Deviation | 0.392497 | 7.4495 | 0.493967 | 0.173404 | 0.283437 | 9.081801 |
| Sample Variance | 0.154054 | 55.49539 | 0.244003 | 0.030069 | 0.080336 | 82.4791 |
| Kurtosis | 0.506247 | 3.33136 | -1.90109 | 27.43302 | 6.498569 | 2.668007 |
| Skewness | 1.582795 | 0.303628 | 0.320501 | 5.420162 | 2.913001 | 2.020777 |
| Range | 1 | 71.13 | 1 | 1 | 1 | 30 |
| Minimum | 0 | 4.69 | 0 | 0 | 0 | 0 |
| Maximum | 1 | 75.82 | 1 | 1 | 1 | 30 |
| Sum | 190 | 28633.25 | 421 | 31 | 88 | 4624 |
| Count | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |

| Statistics | Mental Health | DiffWalking | Gender | Age Category | Diabetes | Physical Activity |
|---|---|---|---|---|---|---|
| Mean | 3.316 | 0.275 | 0.332 | 2.783 | 0.763 | 0.673 |
| Standard Error | 0.242778 | 0.014127 | 0.0149 | 0.028191 | 0.0407 | 0.014842 |
| Median | 0 | 0 | 0 | 3 | 0 | 1 |
| Mode | 0 | 0 | 0 | 3 | 0 | 1 |
| Standard Deviation | 7.67731 | 0.446738 | 0.471167 | 0.891463 | 1.287046 | 0.469352 |
| Sample Variance | 58.94109 | 0.199575 | 0.221998 | 0.794706 | 1.656487 | 0.220291 |
| Kurtosis | 5.788207 | -0.98324 | -1.4924 | 0.512506 | -0.66498 | -1.45729 |
| Skewness | 2.616394 | 1.009321 | 0.714553 | -0.70781 | 1.139874 | -0.73866 |
| Range | 30 | 1 | 1 | 4 | 3 | 1 |
| Minimum | 0 | 0 | 0 | 0 | 0 | 0 |
| Maximum | 30 | 1 | 1 | 4 | 3 | 1 |
| Sum | 3316 | 275 | 332 | 2783 | 763 | 673 |
| Count | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |

| Statistics | GenHealth | SleepTime | Asthma | KidneyDisease |
|---|---|---|---|---|
| Mean | 2.16 | 7.242 | 0.144 | 0.069 |
| Standard Error | 0.03452 | 0.05204 | 0.011108 | 0.00802 |
| Median | 2 | 7 | 0 | 0 |
| Mode | 2 | 8 | 0 | 0 |
| Standard Deviation | 1.0916 | 1.64564 | 0.351265 | 0.25358 |
| Sample Variance | 1.19159 | 2.70814 | 0.123387 | 0.0643 |
| Kurtosis | -0.5753 | 4.73533 | 2.129294 | 9.6209 |
| Skewness | -0.1723 | 0.68006 | 2.031019 | 3.40612 |
| Range | 4 | 17 | 1 | 1 |
| Minimum | 0 | 1 | 0 | 0 |
| Maximum | 4 | 18 | 1 | 1 |
| Sum | 2160 | 7242 | 144 | 69 |
| Count | 1000 | 1000 | 1000 | 1000 |

**Conclusion:** From the above table we can say that average BMI is 28.6332. Average age of patients is 50-65 years.

# Statistical Analysis

## Test for independence of attribute:

**1.** $H_0$: Heart disease are independent of gender.

v/s

$H_1$: Heart disease are dependent of gender.

| Gender / Heart Disease | Male | Female | Total |
|---|---|---|---|
| **With Heart Disease** | 59 | 92 | 151 |
| **Without Heart Disease** | 273 | 576 | 849 |
| **Total** | 332 | 668 | 1000 |

## Test statistic under $H_o$:

$$\chi 2 = \frac{N(ad-bc)\text{^}2}{(a+b(a+c)(b+d)(c+d)}$$

## Decision:

Pearson's Chi-squared test
X-squared = 89.142
df = 19
level of significance (α)=0.05
P-value = 0.0000
Here, P-value < level of significance(α)
Hence, we reject $H_0$ at 5 % level of significance.

**Conclusion:** Heart disease are may not be independent of gender.

**2.** $H_0$: Heart disease are independent on age group.

v/s

$H_1$: Heart disease are dependent on age group.

| Heart Disease / Age Group | With Heart Disease | Without Heart Disease |
|---|---|---|
| 15-24 | 0 | 8 |
| 25-34 | 0 | 9 |
| 35-44 | 0 | 36 |
| 45-54 | 2 | 75 |
| 55-64 | 20 | 160 |
| 65-74 | 58 | 293 |
| 75 and above | 71 | 268 |
| Total | 151 | 849 |

## Test statistics under $H_0$:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(Oij - Eij)\text{^}2}{Eij} \sim \chi^2_{(m-1)(n-1),d.f.}$$

## Decision:

Pearson's Chi-squared test
X-squared = 30.635
df = 7
level of significance(α)=0.05
P-value = 0.00007
Here, P- value < level of significance (α)
Hence, we reject $H_0$ at 5% level of significance.

**Conclusion:** Heart disease may not be independent of age group.

**3.** $H_o$: Heart disease are independent of diabetes.

v/s

$H_1$: Heart disease are dependent of diabetes.

| Diabetes \ Heart Disease | Yes | No |
|---|---|---|
| **No** | 85 | 644 |
| **No, borderline diabetes** | 1 | 23 |
| **Yes(during pregnancy)** | 0 | 2 |
| **Yes** | 65 | 180 |

**Test statistics under $H_0$:**

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(Oij - Eij)\text{^}2}{Eij} \sim \chi^2_{(m-1)(n-1)d.f.}$$

**Decision:**

Pearson's Chi-squared test
X-squared = 34.294
df = 3
level of significance(α)=0.05
P-value = 0.0000
Here, P- value < level of significance (α)
Hence, we reject $H_0$ at 5% level of significance.

**Conclusion:** Heart disease may not be independent of diabetes.

## ➢ Proportion Test

**4.** $H_0$: Proportion of male and female is same for presence of heart disease.

v/s

$H_1$: Proportion of male and female is not same for presence of heart disease.

| Gender / Heart Disease | Male | Female | Total |
|---|---|---|---|
| **With Heart Disease** | 59 | 92 | 151 |
| **Without Heart Disease** | 273 | 576 | 849 |
| **Total** | 332 | 668 | 1000 |

**statistic under $H_0$:**

$$Z=\frac{(p1-p2)}{\sqrt{PQ(\frac{1}{n1}+\frac{1}{n2})}} \sim N(0,1)$$

X-squared=2.4629

df =1

P-value =0.1166

Level of significance (α) =0.05

Here P-value > level of significance (α)

So, we accept Ho at 5% level of significance.

**Conclusion :** Proportion of male and female may be same for presence of heart disease.

# Logistic Regression

We test,
$H_0$: Regression coefficient is not significant.
v/s
$H_1$: Regression coefficient is significant.

glm (Heart Disease ~ BMI + Smoking + Alcohol Drinking + Stroke + Physical Health + Mental Health + Diff Walking + Gender + Age Category + Diabetes + Physical Activity + General Health + Sleep Time + Asthma + Kidney Disease + Skin Cancer , family=binomial, data=test data)

**Deviance Residuals:**

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -2.5238 | -0.5113 | -0.2574 | -0.1224 | 2.8944 |

**Coefficients:**

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.9599 | 2.5035 | -2.78 | 0.0054 ** |
| BMI | 0.0490 | 0.0349 | 1.403 | 0.1606 |
| Smoking | -0.005121 | 0.528712 | -0.010 | 0.99227 |
| Alcohol Drinking | 0.4204 | 1.2200 | 0.345 | 0.7303 |
| Stroke | 0.9905 | 0.6815 | 1.453 | 0.1461 |
| Physical Health | 0.0091 | 0.0318 | 0.287 | 0.7737 |
| Mental Health | -0.0374 | 0.0485 | -0.771 | 0.4406 |
| Diff Walking | 0.4158 | 0.624 | 0.666 | 0.5051 |
| Gender | 1.5314 | 0.5621 | 2.724 | 0.0064** |
| Age Category | 1.0468 | 0.4292 | 2.439 | 0.0147* |
| Diabetes | 0.3616 | 0.1896 | 1.907 | 0.0165. |
| Physical Activity | 0.7530 | 0.5745 | 1.311 | 0.1899 |
| General Health | -0.7311 | 0.3521 | -2.076 | 0.0378* |
| Sleep Time | -0.0315 | 0.169 | -0.186 | 0.8521 |
| Asthma | 0.6494 | 0.5807 | 1.119 | 0.2633 |
| Kidney Disease | -0.3029 | 0.8647 | -0.35 | 0.726 |
| Skin Cancer | 0.6109 | 0.569114 | 1.074 | 0.283 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter of binomial family taken to be 1 )

Null deviance: 156.74 on 207 degrees of freedom
Residual deviance: 118.05 on 191 degrees of freedom

**AIC: 152.05**

Number of Fisher Scoring iterations: 6

Here n=208, p=16,

Level of significance ($\alpha$) = 0.05

Model Deviance = 118.05

$\chi^2_{\alpha,n-p}$ = 132.2582

Decision rule: Model deviance < $\chi^2_{\alpha,n-p}$

So, we may accept the $H_0$ 5% level of significance.

**Conclusion:** From the above table we can conclude that regressor Gender, Age Category, Diabetes, General Health are significant. While regressor BMI, Smoking, Alcohol Drinking, Stroke, Physical Health, Mental Health, Diff Walking, Physical Activity, Sleep Time, Asthma, Kidney Disease, Skin Cancer are not significant.

# Stepwise Regression

## Model : 1

Start: **AIC=152.05**

Heart Disease ~ BMI + Smoking + Alcohol Drinking + Stroke + Physical Healt h + Mental Health + DiffWalking + Gender + Age Category + Diabetes + Physical Activity + General Health + Sleep Time + Asthma + Kidney Disease + Skin Cancer

| Regressors | df | Deviance | AIC |
|---|---|---|---|
| Smoking | 1 | 118.05 | 150.05 |
| Sleep Time | 1 | 118.09 | 150.09 |
| Physical Health | 1 | 118.14 | 150.14 |
| Alcohol Drinking | 1 | 118.17 | 150.16 |
| Kidney Disease | 1 | 118.18 | 150.18 |
| DiffWalking | 1 | 118.49 | 150.49 |
| Mental Health | 1 | 118.75 | 150.75 |
| Skin Cancer | 1 | 119.18 | 151.18 |
| Asthma | 1 | 119.26 | 151.26 |
| Physical Activity | 1 | 119.88 | 151.88 |
| BMI | 1 | 119.90 | 151.90 |
| <none> | | 118.05 | 152.05 |
| Stroke | 1 | 120.08 | 152.08 |
| Diabetes | 1 | 121.68 | 153.68 |
| General Health | 1 | 122.82 | 154.82 |
| Age Category | 1 | 125.20 | 157.20 |
| Gender | 1 | 125.91 | 157.91 |

Stepwise regression was followed by a model.

**Model : 2**

Step: **AIC=136.6**
Heart Disease ~ Stroke + Diabetes + Gen Health+ Gender + Age Category

| | Df | Deviance | AIC |
|---|---|---|---|
| <none> | | 124.60 | 136.60 |
| Stroke | 1 | 127.12 | 137.12 |
| Diabetes | 1 | 129.72 | 139.72 |
| General Health | 1 | 131.88 | 141.88 |
| Gender | 1 | 132.83 | 142.83 |
| Age Category | 1 | 133.02 | 143.02 |

**Conclusion**: By comparing the AIC values in the above models the value of the AIC is reduced when the variables are eliminated from the model. We conclude that regressor Stroke, Diabetes, General Health, gender, Age category are most significant.

**Confusion matrix:**

| Actual value / Predicted Value | With Heart disease | Without Heart disease |
|---|---|---|
| With Heart disease | 180 | 2 |
| Without Heart disease | 23 | 3 |

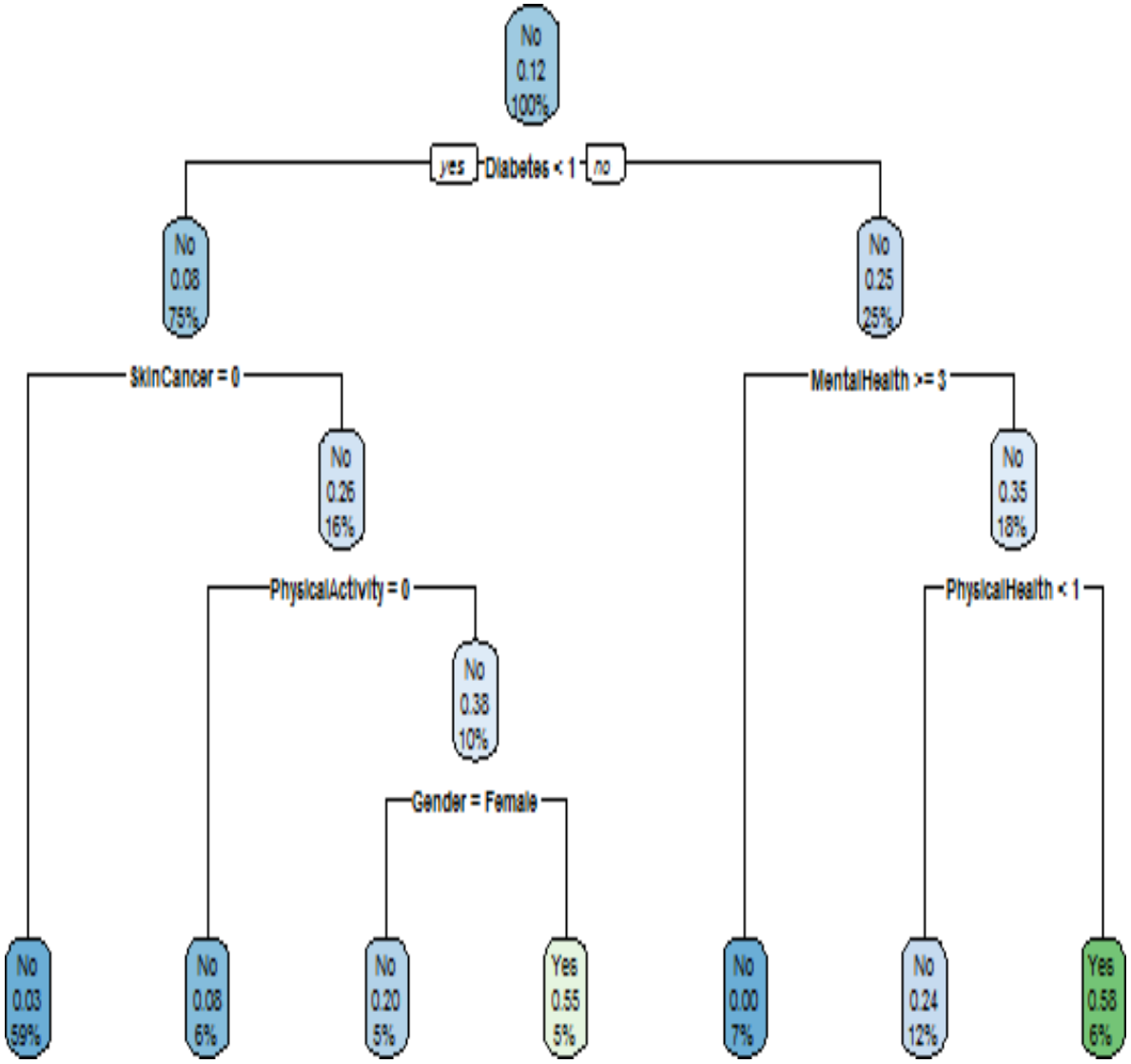
Accuracy : 87.98%
Sensitivity : 98.90%
Specificity : 11.53%
**Conclusion:** From the above table, we can conclude that accuracy of the model is 87.98%.

# Decision Tree

n= 208

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 208 26 0 (0.87500000 0.12500000)
2) Diabetes< 0.5 157 13 0 (0.91719745 0.08280255)
4) Skin Cancer< 0.5 123 4 0 (0.96747967 0.03252033) *
5) Skin Cancer>=0.5 34 9 0 (0.73529412 0.26470588)
10) Physical Activity< 0.5 13 1 0 (0.92307692 0.07692308) *
11) Physical Activity>=0.5 21 8 0 (0.61904762 0.38095238)
22) Gender=Female 10 2 0 (0.80000000 0.20000000) *
23) Gender=Male 11 5 1 (0.45454545 0.54545455) *
3) Diabetes>=0.5 51 13 0 (0.74509804 0.25490196)
6) Mental Health>=2.5 14 0 0 (1.00000000 0.00000000) *
7) Mental Health< 2.5 37 13 0 (0.64864865 0.35135135)
14) Physical Health< 1 25 6 0 (0.76000000 0.24000000) *
15) Physical Health>=1 12 5 1 (0.41666667 0.58333333) *




**Conclusion**: From above decision tree, we can conclude that at the top is the overall probability of presence of heart disease. It shows the proportion of patients with heart disease. 12% of patients have heart disease.

This root node ask whether the diabetes <1. If yes, then you go down to the decision node. With an 8% chance of having heart disease, 75% of people do not have diabetes.

In the second decision node, if no diabetes patients skin cancer value is 0.If, yes then chance of presence of heart disease is 3%.

## Confusion Matrix:

| Actual value / Predicted value | With Heart disease | Without Heart disease |
|---|---|---|
| With Heart disease | 172 | 13 |
| Without Heart disease | 10 | 13 |

Accuracy = 88.94%
Sensitivity = 92.97%
Specificity = 56.52%

**Conclusion**: From the above table we can say that accuracy of the decision tree is 88.94% .

The performance of above all classifiers is compared in following:

| **Performance measure / Classifiers** | **Accuracy(%)** | **Sensitivity(%)** | **Specificity(%)** |
|---|---|---|---|
| Logistic regression | 87.98 | 98.90 | 11.53 |
| Decision tree | 88.94 | 92.97 | 56.52 |

**Conclusion**: From the above table, we can conclude that the decision tree is a high performance predictive model that classifies heart diseases.

# Limitations

The study covers secondary data pertaining to the heart disease on male and female.

- We consider the age group 18-80 and above for the project.
- The number of females in the data are more as compare to male.
- We develop model only with available variables but if add other variables then we expect that our models gives better result.
- The study is subject to common limitation of sample survey.

# Scope

- This analysis is useful for other problems related with medical studies like predicting diabetes.
- By using this analysis we can predict the patients has the presence or absence of heart disease.
- Using statistical analysis, we can predict patients at risk of disease or health conditions.

# Conclusion

## ❖ GRAPHICAL REPRESENTATION

- There are more patients without heart disease.
- There are more female patients than male patients.
- There are more females without heart disease as compared to males.
- The age range 65-74 has more patients without heart disease, while 75 and older has more heart disease patients.

## ❖ TESTING

- Heart diseases are may not be independent on gender.
- Heart diseases are may not be independent on age group.
- Heart diseases are may not be independent on diabetes.
- Proportion of male and female may be same for presence of heart disease.
- From the logistic regression analysis we can conclude that regressor Gender, Age Category, Diabetes, General Health are significant. While regressor BMI, Smoking, Alcohol Drinking, Stroke, Physical Health, Mental Health, Diff Walking, Physical Activity, Sleep Time, Asthma, Kidney Disease, Skin Cancer are not significant.
- From the stepwise regression we can conclude that regressor Stroke, Diabetes, General Health, Gender, Age Category are most significant.
- From the decision tree we, can say that the variables of diabetes, physical activity, gender, mental health are more important and classifies heart disease.

# References

Referred Books:

1. Fundamental of statistics : Das Gupta & Goon Gupta
2. Statistical Method : Dr. Jayant Tatke
3. Fundamentals of Applied statistics: S.C.Gupta

Visited websites:

1. Google. (kaggle.com)

# Thank you