

# ***Be Heart Smart***

## **The Healthy Healthcare Enthusiasts (Collaborators):** (Final-Project Group 7)

- ❖ Ayse Ozgun
- ❖ Pam Noble
- ❖ Subhangi Ghosh
- ❖ Krishnakali Sarkar

# Cardiovascular Disease (CVDs)

Disorders of the heart and blood vessels including coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions.

Leading cause of death globally ~ 40% deaths in the US.

## Leading Behavioral Risk Factors :

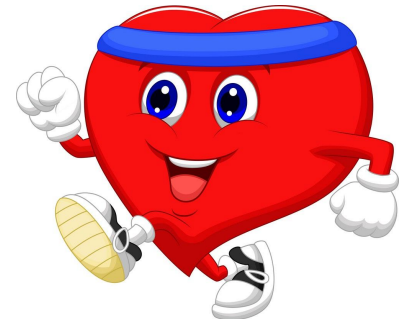
- Unhealthy diet,
- Physical inactivity
- Tobacco use
- Harmful use of alcohol

## A healthy heart is a happy heart

The purpose of this project is to spread awareness.  
Embracing a healthy lifestyle at any age can help prevent heart disease, and lower the risks for heart attack or stroke.

## Effects of behavioral risk factors :

- Raised blood pressure,
- Raised blood glucose,
- Raised blood lipids,
- Overweight and
- Obesity.



# About the data

Website : [Cardiovascular Disease dataset](#) (Kaggle)

## Description :

Three types of input features

- Objective
- Examination
- Subjective

Objective	Examination	Subjective
Age (days)	Systolic Blood Pressure	Smoking
Height (cm)	Diastolic Blood Pressure	Alcohol Intake
Weight (kg)	Cholesterol	Physical Activity
Gender	Glucose	

Target Variable : Presence or Absence of Cardiovascular Disease

# Database

- Amazon web service was employed to create a PostgreSQL server
- The server, PostgreRDS, hosts the Be-Heart-Smart database
- All team members can connect using PGAdmin
- SQL scripting was written to create tables to hold the Be-Heart-Smart project data
- BMI table and Cardio\_cleaned tables were joined on id to produce the Cardio\_cleaned\_with bmi table which contains data required to perform the analysis



The image shows a screenshot of a database management tool interface. At the top, there is a search bar with an eye icon. Below it, the 'public' schema is selected, indicated by a red diamond icon. The table 'cardio\_cleaned\_with\_bmi' is highlighted with a blue table icon. The table's structure is listed below, showing columns and their data types:

Column Name	Data Type
id	numeric
age	numeric
gender	numeric
height	numeric
weight	numeric
systolic_bp	numeric
diastolic_bp	numeric
cholesterol	numeric
glucose	numeric
smoker	numeric
alcohol_intake	numeric
active	numeric
cardio_disease	numeric
bmi	numeric
weight_status	character varying
obesity_status	character varying

# Initial Assessment of Data

- Downloaded data has values separated by semicolon. Converted to csv file using Microsoft Excel.
- 70000 observations
- 11 features

Descriptive stats on the continuous variables

```
# Summary statistics of the continuous variables
cardio_df.select("id", "age", "height", "weight", "systolic_bp", "diastolic_bp").describe().show()
```

summary	id	(in days) age	(in cm) height	(in kg) weight	systolic_bp	diastolic_bp
count	70000	70000	70000	70000	70000	70000
mean	49972.4199	19468.865814285713	164.35922857142856	74.20569000015259	128.8172857142857	96.63041428571428
stddev	28851.302323172928	2467.2516672413917	8.210126364538551	14.395756678414427	154.01141945609032	188.47253029639106
min	0	10798	55.0	10.0	-150.0	-70.0
max	99999	23713	250.0	200.0	16020.0	11000.0

# Data Pre-processing, Exploratory Data Analysis and Data Processing

## Data Pre-processing:

- 70,000 observations
  - ◆ Few observations have values not observed in human adults (eg. diastolic bp: 11000)
  - ◆ Negative values (eg. systolic bp: -150)
  - ◆ Categorical variables given values (eg. Glucose: 1-normal, 2-above normal, 3-well above normal)
- Various reasons for above numbers
- Observations with probable values for human adults will be retained
  - ◆ Height: 135 - 215 cm
  - ◆ Weight: 25 - 200 kg
  - ◆ Systolic bp: 80 - 180
  - ◆ Diastolic bp: 40 - 120
- Decision will taken with respect to negative numbers during Data Processing. May keep the absolute value but change sign, or may remove the datapoint entirely

Initial trial of data pre-processing in Excel had brought down the total number of observations to 60,510.

# Questions we hope to answer with the data:

- ★ Is a person at risk of heart disease?
- ★ What are the potential risk factors for heart disease--smoking, alcohol consumption, obesity, etc?
- ★ Which factors are the best predictors of heart disease?

## Classification model to predict risk (Yes/No) of heart disease based on different factors

### ❖ Supervised Machine Learning

- Logistic Regression
- Random Forest

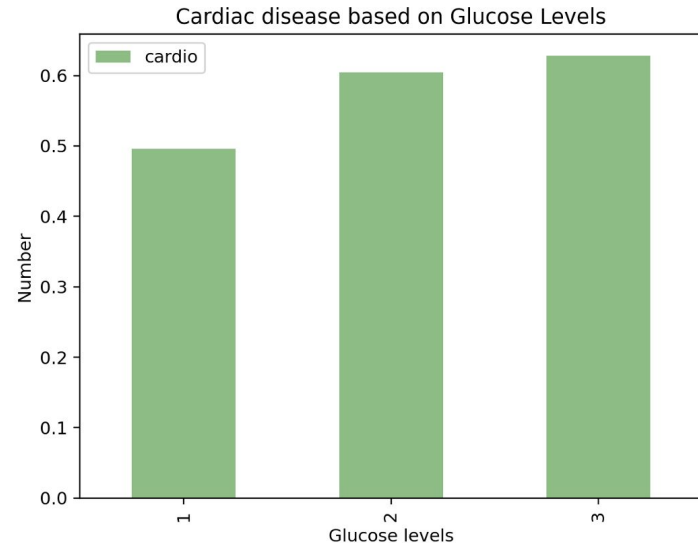
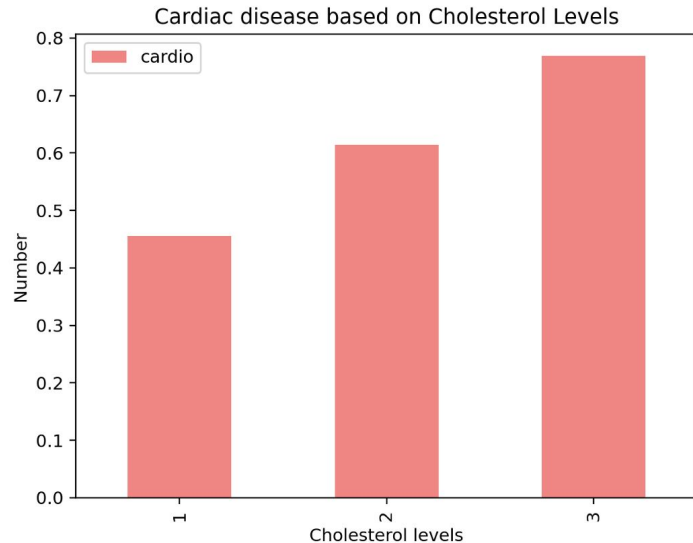
### ❖ Basic Neural Network

### ❖ Deep Neural Network

# Data Processing, and Exploratory Data Analysis

## Exploratory Data-Analysis:

Performed on the initial trial pre-processed data on Excel



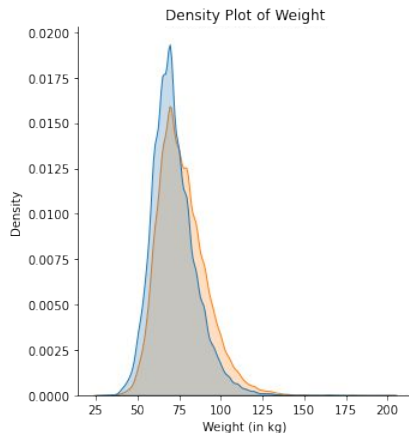


# Data Processing, and Exploratory Data Analysis

## Exploratory Data-Analysis:

Created a new table of BMI  
(using information from CDC.gov)

Performed on the cleaned up data

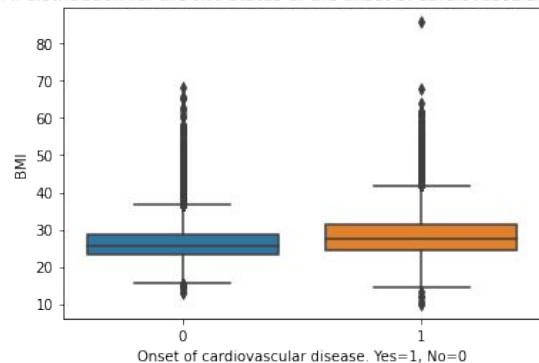


	id	BMI	weight_status	obesity_status	
	35363	77629	22.0	normal	no
	41697	81468	29.8	overweight	yes
	37065	55211	24.2	normal	no
	2697	1778	19.8	normal	no
	67862	73893	26.3	overweight	no
	56957	13361	22.0	normal	no
	16793	94697	31.2	obese	yes

```
BMI_df["BMI"].describe()

count    68297.000000
mean      27.420065
std        5.184147
min         9.900000
25%       23.900000
50%       26.300000
75%       30.100000
max       85.800000
Name: BMI, dtype: float64
```

BMI distribution for the two states of the onset of cardiovascular disease



# Supervised Machine Learning

## Logistic Regression

Comparison of preliminary logistic regression on raw and cleaned data

	Raw Data	Cleaned Data	Merged Data
Accuracy	69.1%	71.7%	72.6%

# Dashboard

Analysis of be-heart-smart data to predict the presence or absence of cardiovascular disease based on:

- Three types of input features and eleven data elements
  - Objective - Age (days), Height (cm), Weight (kg), Gender
  - Examination- Systolic Blood Pressure, Diastolic Blood Pressure, Cholesterol, Glucose
  - Subjective - Smoking, Alcohol Intake, Physical Activity
- Supervised Machine Learning to analyze different input features and data elements to predict presence or absence of cardiovascular disease will provide the data and graphics to highlight the outcomes from performing
  - Logistic Regression
  - Random Forest
  - Basic Neural Network
  - Deep Neural Network
- The dashboard will be created using Tableau