

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

Effects of behavioral risk factors:

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

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.





About the data

Website : <u>Cardiovascular Disease dataset</u> (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 PosgreSQL 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



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()
                                                                (in kg) weight| systolic bp|
                       id
                                  (in days) age
                                                 (in cm) height
                                                                                                         diastolic bp
 summary
                     70000
                                       70000
                                                         70000
                                                                           70000
                                                                                              70000
                                                                                                                70000
  count
                49972.4199 19468.865814285713 164.35922857142856
                                                               74.20569000015259
 stddev
        28851.302323172928 2467.2516672413917
                                       10798
                                                          55.0
                                                                            10.0
    min
                                                                                             -150.0
                                                                                                                -70.0
                                       23713
                                                         250.0
                                                                           200.0
                                                                                            16020.0
```

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?

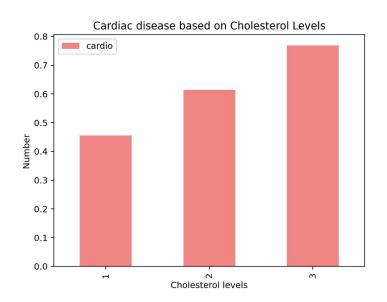
Classification model to predict risk (Yes/No) of heart disease based on different factors

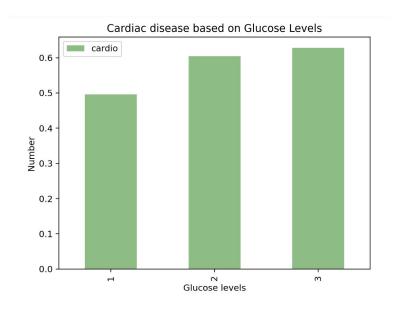
- Supervised Machine Learning
 - Logistic Regression
 - > Random Forest
- Basic Neural Network
- Deep Neural Network

Data Processing and Exploratory Data Analysis

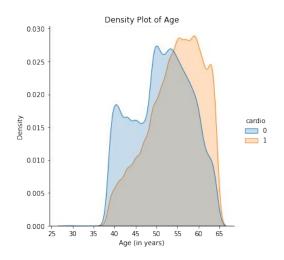
Exploratory Data-Analysis:

Performed on the initial trial pre-processed data on Excel

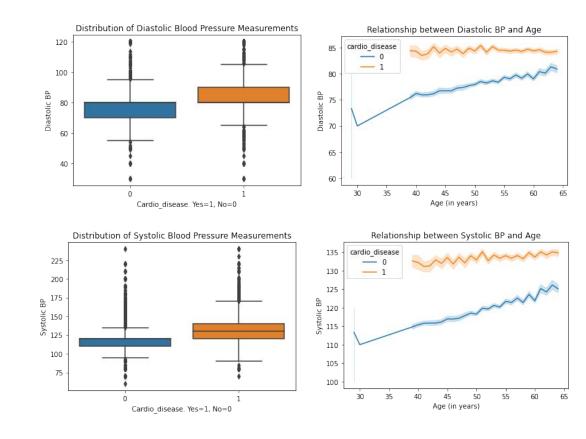




Data Processing and Exploratory Data Analysis (on cleaned up data)

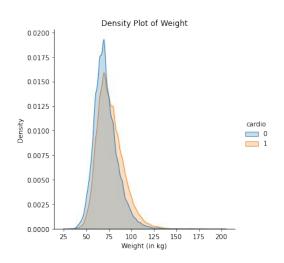


Age, and Blood Pressure appears to affect onset of cardiovascular diseases



Creating the BMI table

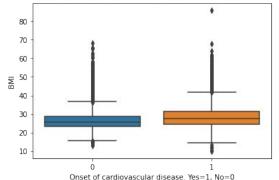
(using information from CDC.gov)



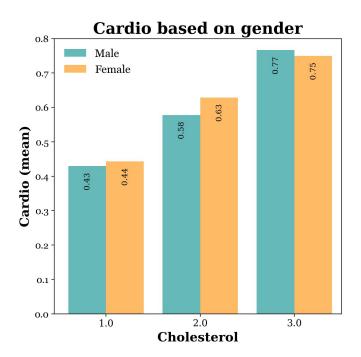
	id	BMI	weight_status	obesity_status
35363	77629	22.0	normal	no
41697	81468	29.8	overweight	yes
37065	55211	24.2	normal	no
2697	1778	19.8	normal	no
67862	73893	26.3	overweight	no
56957	13361	22.0	normal	no
16793	94697	31.2	obese	yes

BMI_df	f[" <mark>BMI</mark> "].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

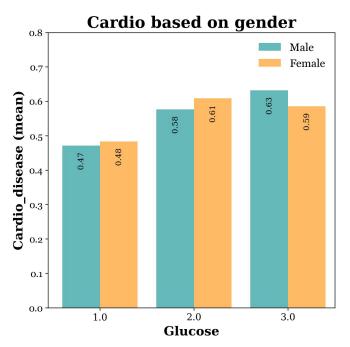
BMI distribution for the two states of the onset of cardiovascular disease



The effect of cholesterol and glucose on cardiac disease Based on gender



 Men with high Cholesterol have a higher chance of developing cardiac disease.



 Men with high Glucose levels have a higher chance of developing cardiac diseases.

Logistic Regression

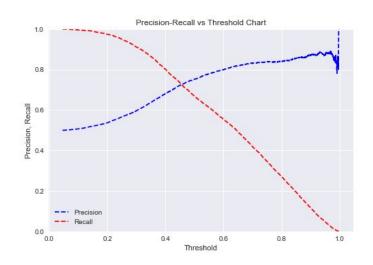
Comparison of logistic regression on raw and cleaned data

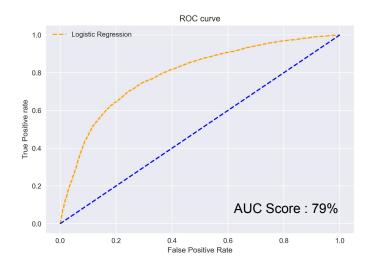
age -	1	-0.025	0.1	-0.029	0.017	0.21	0.022	0.15	0.025	0.094	-0.049	-0.029	-0.011	0.24
gender_M -	-0.025	1	-0.11	-0.0013	0.058	0.06	-0.017	-0.031	-0.0055	-0.021	0.34	0.17	0.0058	0.0058
ВМІ -	0.1	-0.11	1	-0.18	-0.043	0.27	0.083	0.14	0.11	0.078	-0.034	0.017	-0.014	0.19
underweight -	-0.029	-0.0013	-0.18	1	-0.072	-0.061	-0.0082	-0.018	-0.0051	-0.017	0.012	0.00055	0.0078	-0.04
overweight -	0.017	0.058	-0.043	-0.072	1	0.0083	0.006	-0.017	-0.0093	-0.014	0.024	0.014	0.0069	0.014
systolic_bp -	0.21	0.06	0.27	-0.061	0.0083	1	0.1	0.15	0.078	0.063	0.029	0.036	-0.0031	0.43
cholesterol_moderate -	0.022	-0.017	0.083	-0.0082	0.006	0.1	1	-0.14	0.27	-0.053	0.025	0.038	-0.0065	0.081
cholesterol_high -	0.15	-0.031	0.14	-0.018	-0.017	0.15	-0.14	1	-0.014	0.48	-0.0034	0.016	0.012	0.19
glucose_moderate -	0.025	-0.0055	0.11	-0.0051	-0.0093	0.078	0.27	-0.014	1	-0.081	0.016	0.03	-0.012	0.053
glucose_high -	0.094	-0.021	0.078	-0.017	-0.014	0.063	-0.053	0.48	-0.081	1	-0.015	-0.0044	-0.003	0.069
smoker -	-0.049	0.34	-0.034	0.012	0.024	0.029	0.025	-0.0034	0.016	-0.015	1	0.34	0.024	-0.016
alcohol_intake 1	-0.029	0.17	0.017	0.00055	0.014	0.036	0.038	0.016	0.03	-0.0044	0.34	1	0.024	-0.0086
active =	-0.011	0.0058	-0.014	0.0078	0.0069	-0.0031	-0.0065	0.012	-0.012	-0.003	0.024	0.024	1	-0.038
cardio_disease -	0.24	0.0058	0.19	-0.04	0.014	0.43	0.081	0.19	0.053	0.069	-0.016	-0.0086	-0.038	1
Merged Data	- age	gender_M -	BMI -	underweight –	overweight –	systolic_bp -	esterol_moderate -	dolesterol_high -	lucose_moderate	glucose_high	smoker	alcohol_intake	active -	cardio_disease

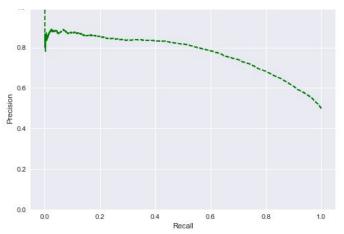
	Raw Data	Cleaned Data	Merged Data
Accuracy	69.1%	71.7%	72.5%
Recall	66 %	66 %	67 %

Logistic Regression

Threshold: 0.4, Recall: 80%, Accuracy: 71.4%





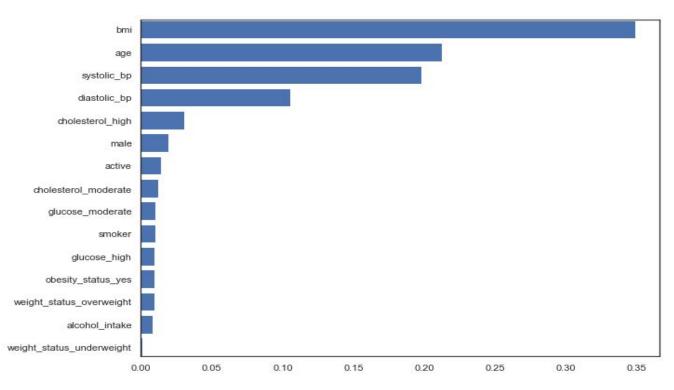


Random Forest Classifier

Comparison of preliminary random forest classifier model on raw, cleaned, merged and hypertuned data.

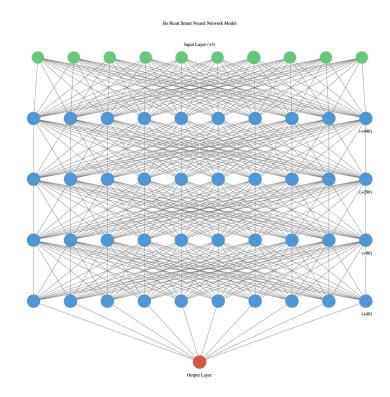
	Raw Data	Cleaned Data	Merged Data	Merged Data after feature engineering and hyperparameter tuning
Test Accuracy	70.2%	70.3%	69.3%	73.0%
Mean ROC-AUC	76.1%	66.4%	74.5%	79.8%

Random Forest Classifier: Feature importances



Neural network Model

- The deep neural network model is run on the final merged dataset.
- Activation function for input: Relu
- Output function: Sigmoid
- No of hidden layers: 5
- The loss function:binary_crossentrophy
- Optimizer: rmsprop
- The accuracy of this model is 73%



Deep Neural Network Model Visualization

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

Dashboard in progress



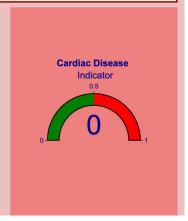


Patients ID No:

Demographic Info

ID: 0
AGE: 50
GENDER: 2
HEIGHT: 168.0
WEIGHT: 62.0
SYSTOLIC_BP: 110.0
DIASTOLIC_BP: 110.0
CHOLESTEROL: 1
GLUCOSE: 1
SMOKER: 0
ALCOHOL_INTAKE: 0
ACTIVE: 1
CARDIO_DISEASE: 0
BMI: 22.0
WEIGHT_STATUS: normal

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age. The most important behavioral risk factors of heart disease and stroke are: • unhealthy diet, • physical inactivity, • tobacco use and • harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as • raised blood pressure, • raised blood glucose, • raised blood lipids, • overweight and • obesity. A healthy heart is central to overall good health. The purpose of this project is to spread awareness among individuals that embracing a healthy lifestyle at any age can prevent heart disease and lower the risks for heart attack or stroke.



Dashboard Details

- The dashboard is interactive
 - Patient whose ID is selected, their demographic information is displayed.
 - (At this time, indicator changes to reflect the cardiovascular health of the Patient whose
 ID is selected.)
 - An input option will be created to accept health numbers from users.
 - The indicator will then change to reflect the cardiac health of the user.

 \triangleright