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 :

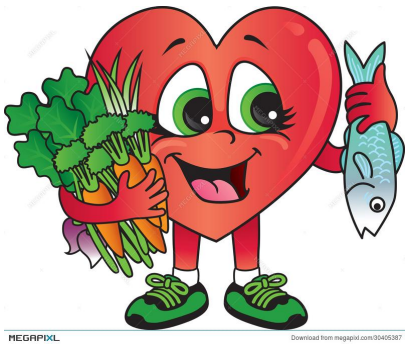
- Unhealthy diet,
- Physical inactivity
- Tobacco use
- Harmful use of alcohol

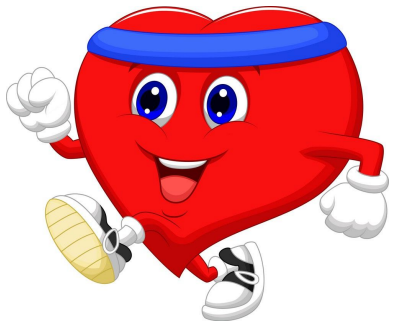
Effects of behavioral risk factors :

- 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 : 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 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()
```

```
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
|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?

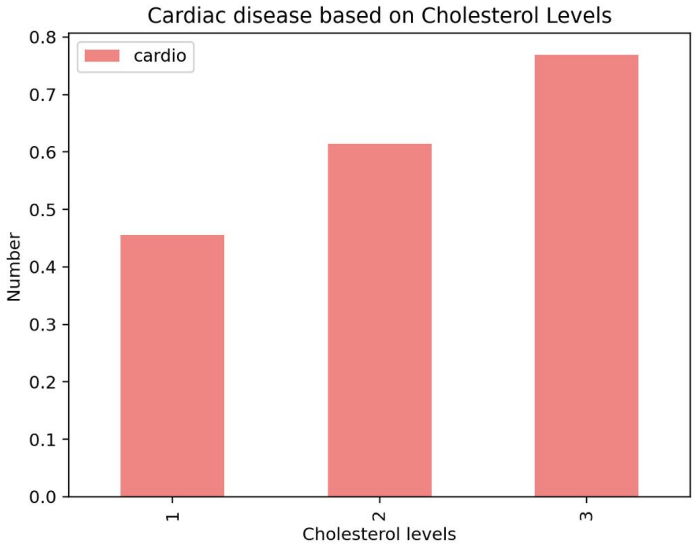
<u>Classification model to predict risk (Yes/No) of heart disease based on different factors</u>

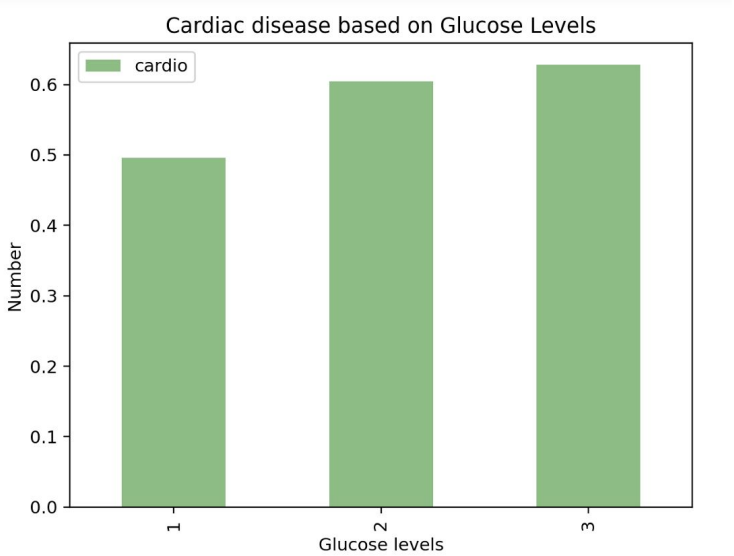
- ❖ Supervised Machine Learning
  - ➢ Logistic Regression
  - ➢ Random Forest
- ❖ Basic Neural Network
- ❖ Deep Neural Network

Data Processing and Exploratory Data Analysis

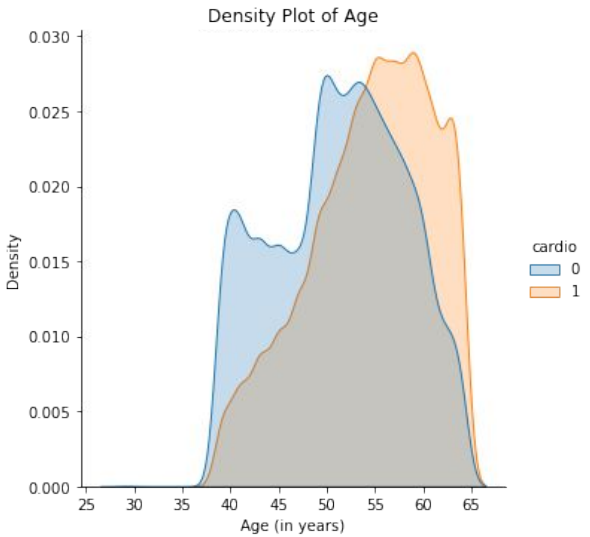
Exploratory Data-Analysis:

Performed on the initial trial pre-processed data on Excel

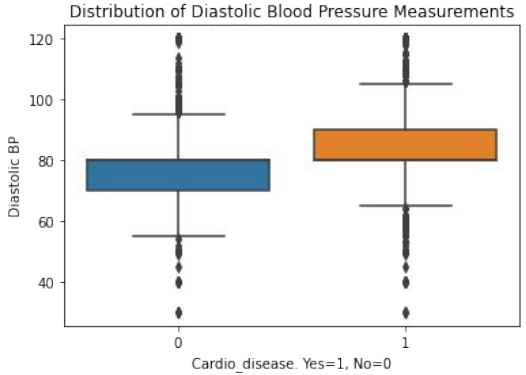


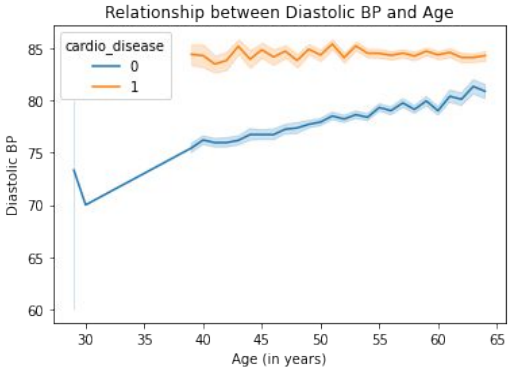


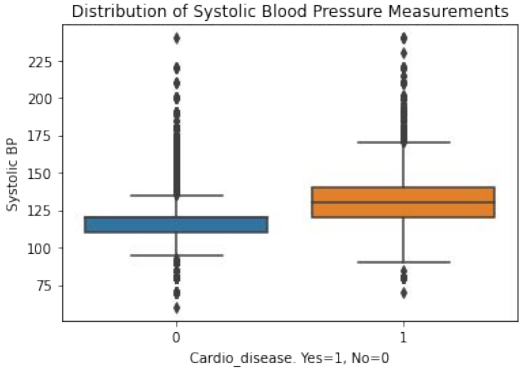
Data Processing and Exploratory Data Analysis (on cleaned up data)

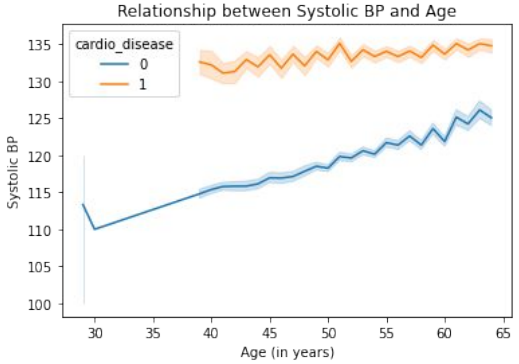


Age, and Blood Pressure appears to affect onset of cardiovascular diseases









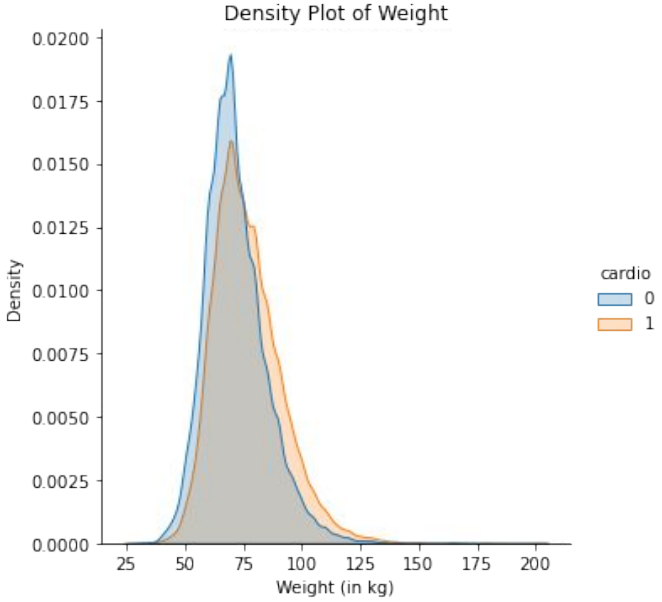
Creating the BMI table

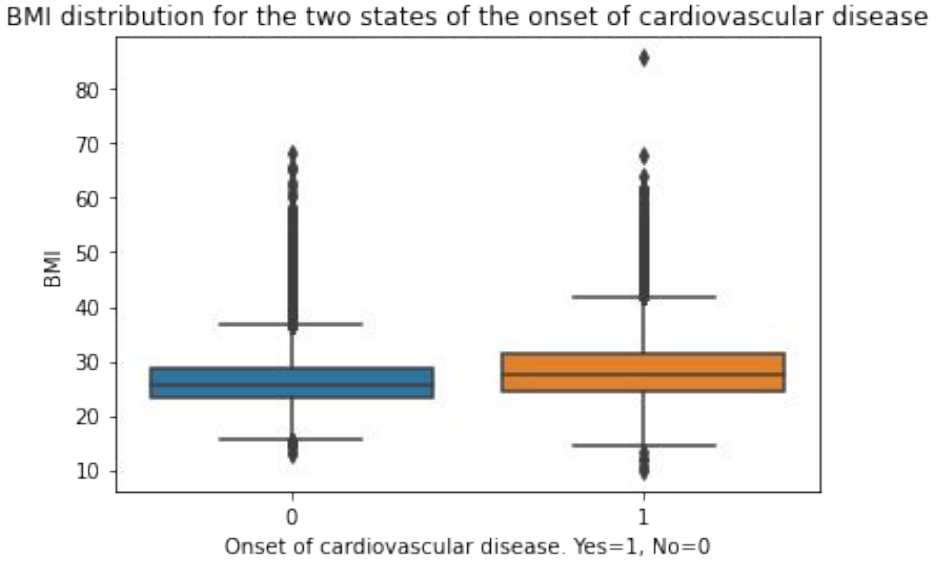
(using information from CDC.gov)

	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

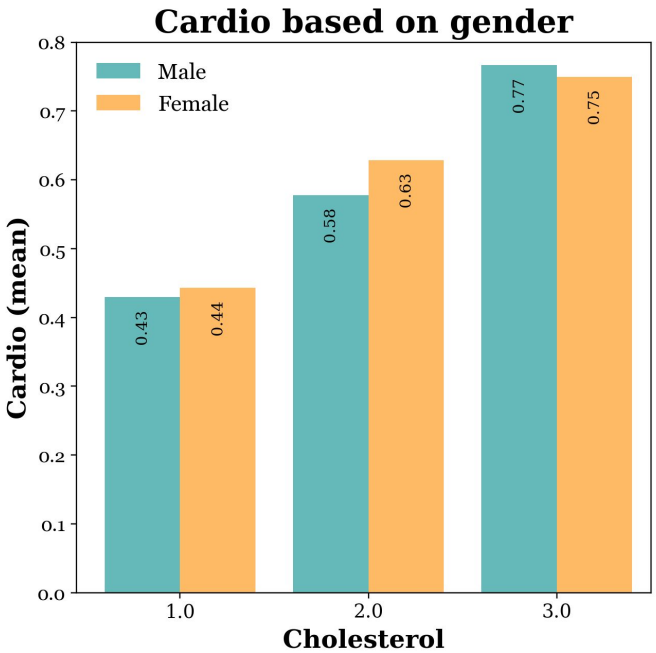
```
BMI_df["BMI"].describe()
```

```
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
```

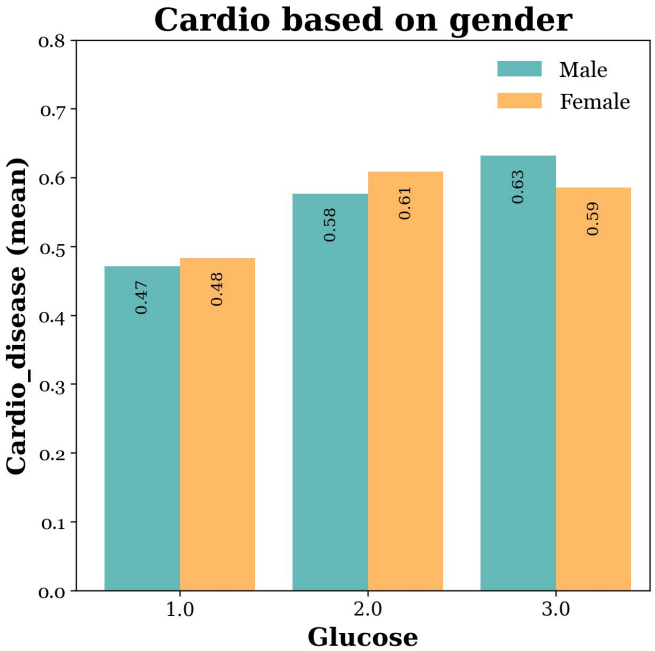




The effect of Cholesterol and Glucose on cardiac disease based on Gender



- 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.

Supervised Machine Learning

Logistic Regression

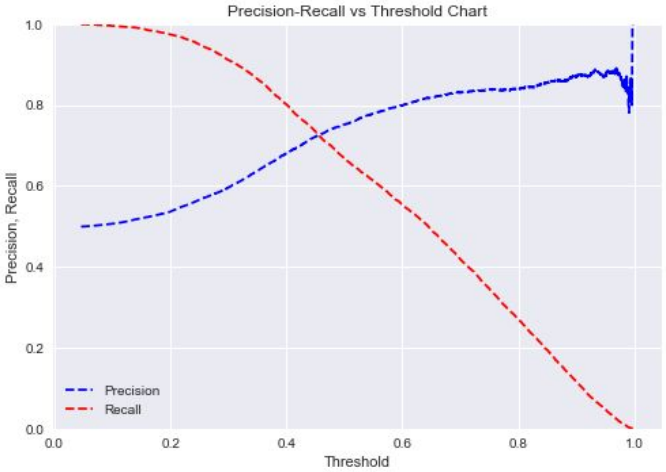
Comparison of logistic regression on raw and cleaned data

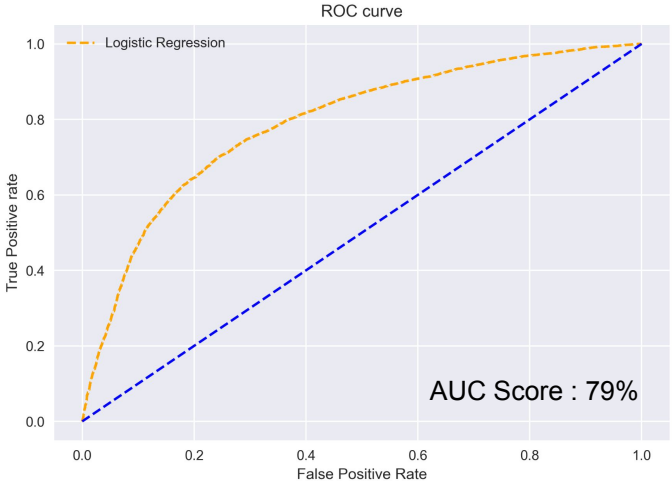
	Raw Data	Cleaned Data	Merged Data
Accuracy	69.1%	71.7%	72.5%
Recall	66 %	66 %	67 %

Feature selection : age, underweight, is_obese, systolic_bp, pulse_pressure, cholesterol_high, active

Supervised Machine Learning

Logistic Regression





Threshold : 0.4

accuracy	0.708759
recall	0.810176
precision	0.672346
roc_auc_score	0.709134

Supervised Machine Learning

Random Forest Classifier

- Split the data into ‘Train: Validation: Test’ sets → 70: 20: 10
- Scale with StandardScaler ()
- Create a random forest classifier instance--with default hyperparameters
- Perform Kfold Cross-validation, k=10 on the scaled training set→ no overfitting

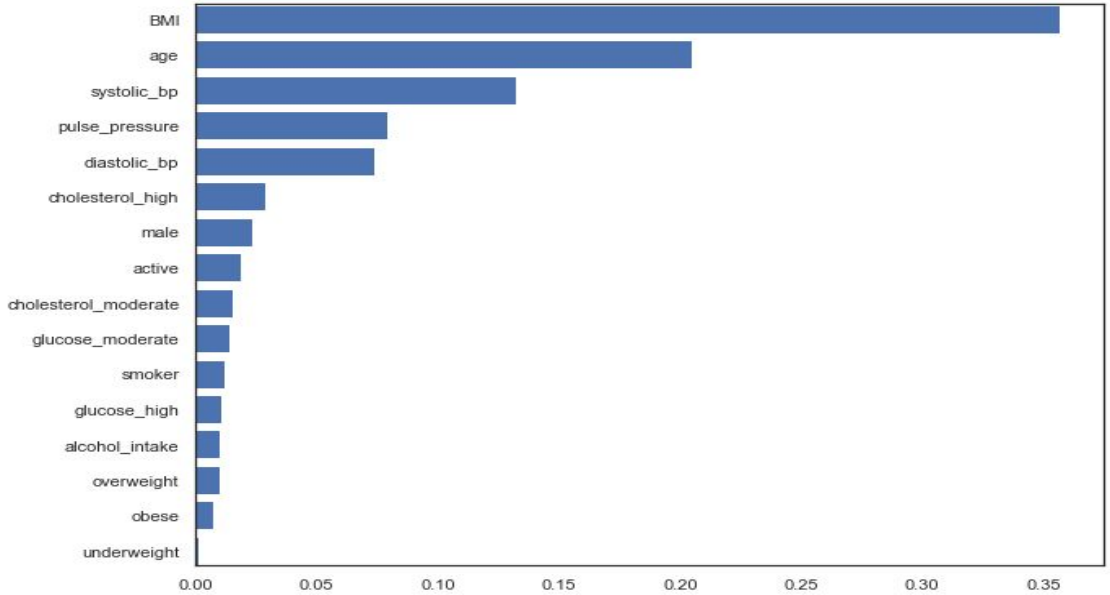
	Initial model
Accuracy	68.9%
Mean roc_auc	74.5%
Recall	68.6%
F1	68.7%

```
[0.73708368 0.73620492 0.75612075 0.74501485 0.74357571 0.7518145
 0.74286975 0.75980342 0.73852016 0.74019497]
```

Supervised Machine Learning

Random Forest Classifier

- Perform feature selection using sklearn's feature_importances_
- Create a RF classifier with selected features only
- Predict the model



Accuracy	Model with feature selection
Mean roc_auc	66.4%
Recall	72.1%
F1	64.1%
Accuracy	65.5%

Supervised Machine Learning

Random Forest Classifier

- Optimize our model with hyperparameter tuning
- Search for best parameters using scikit_learn's GridSearchCV function

```
n_estimators = [300, 800]
max_depth = [5, 8]
min_samples_split = [2, 5]
min_samples_leaf = [1, 5]
```

--Top 5 Hyperparameters--

	param_max_depth	param_n_estimators	param_min_samples_leaf	param_min_samples_split	mean_test_score
12	8	300	5	2	0.732411
14	8	300	5	5	0.732411
9	8	800	1	2	0.732369
11	8	800	1	5	0.732285
8	8	300	1	2	0.732179

- Plug in the best hyperparameters into our Random Forest Classifier, retrain our model and predict it on the unseen test set.

```
{'max_depth': 8,
 'min_samples_leaf': 5,
 'min_samples_split': 2,
 'n_estimators': 300}
```

Supervised Machine Learning

Random Forest Classifier

Roc_auc scores for the

```
[0.81080998 0.8156825  0.7829488  0.82300446 0.77963021 0.79906238
 0.79707829 0.78037387 0.77874049 0.81119241]
```

	Initial model
Accuracy	73.4%
Mean roc_auc	79.78%
Recall	66.0%
F1	71.0%

Classification Report

	precision	recall	f1-score	support
0	0.71	0.80	0.75	3385
1	0.77	0.66	0.71	3362
accuracy			0.73	6747
macro avg	0.74	0.73	0.73	6747
weighted avg	0.74	0.73	0.73	6747

Confusion Matrix

```
[[2716  669]
 [1127 2235]]
```

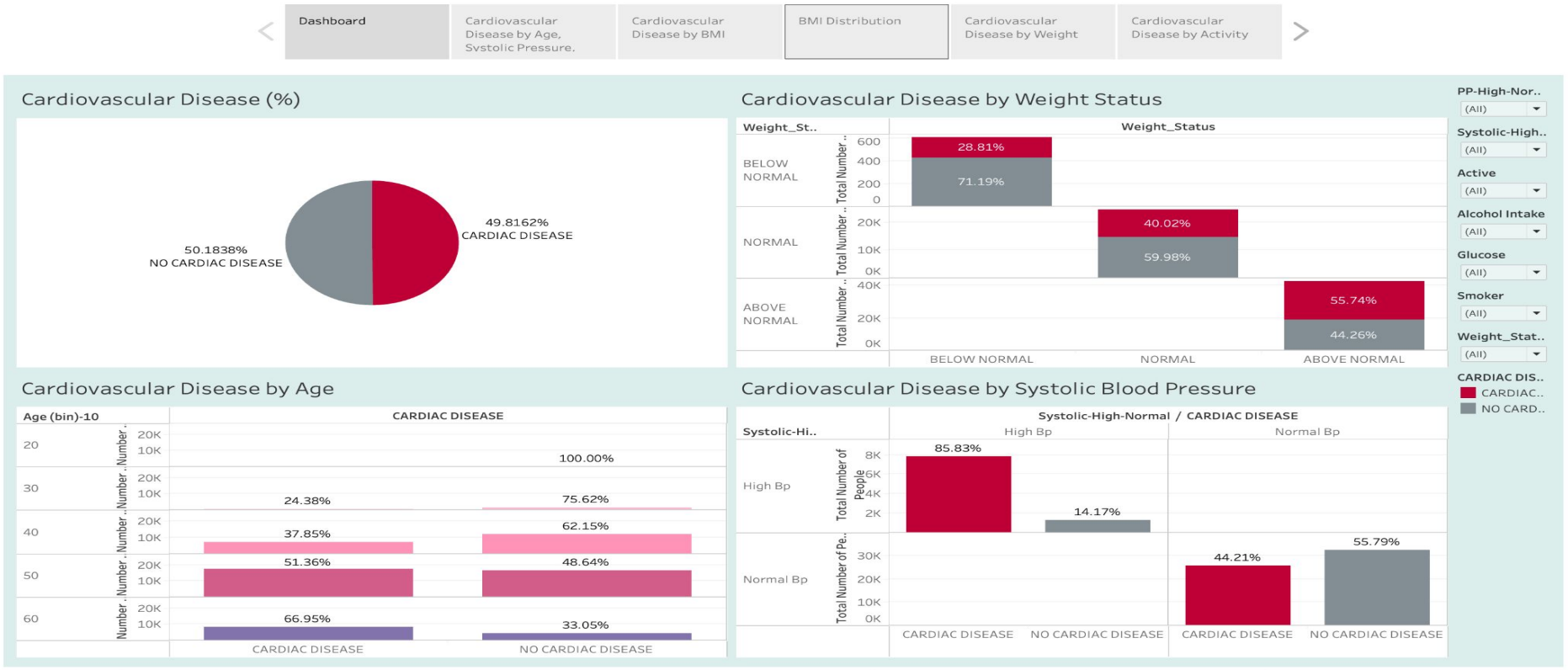
Supervised Machine Learning

Random Forest Classifier

- Comparison of the random forest classifier model at different stages of the analysis.

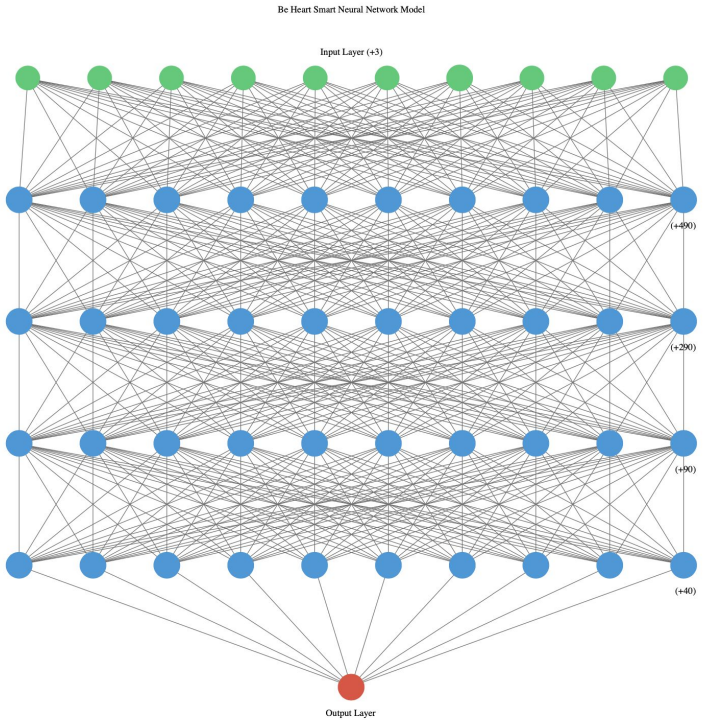
	Initial model	Model with feature selection	After hyperparameter tuning on the test set
Accuracy	68.9%	66.4%	73.4%
Mean roc_auc	74.5%	72.1%	79.8%
Recall	68.6%	64.1%	66.0%
F1	68.7%	65.5%	71.0%

Tableau Dashboard



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%

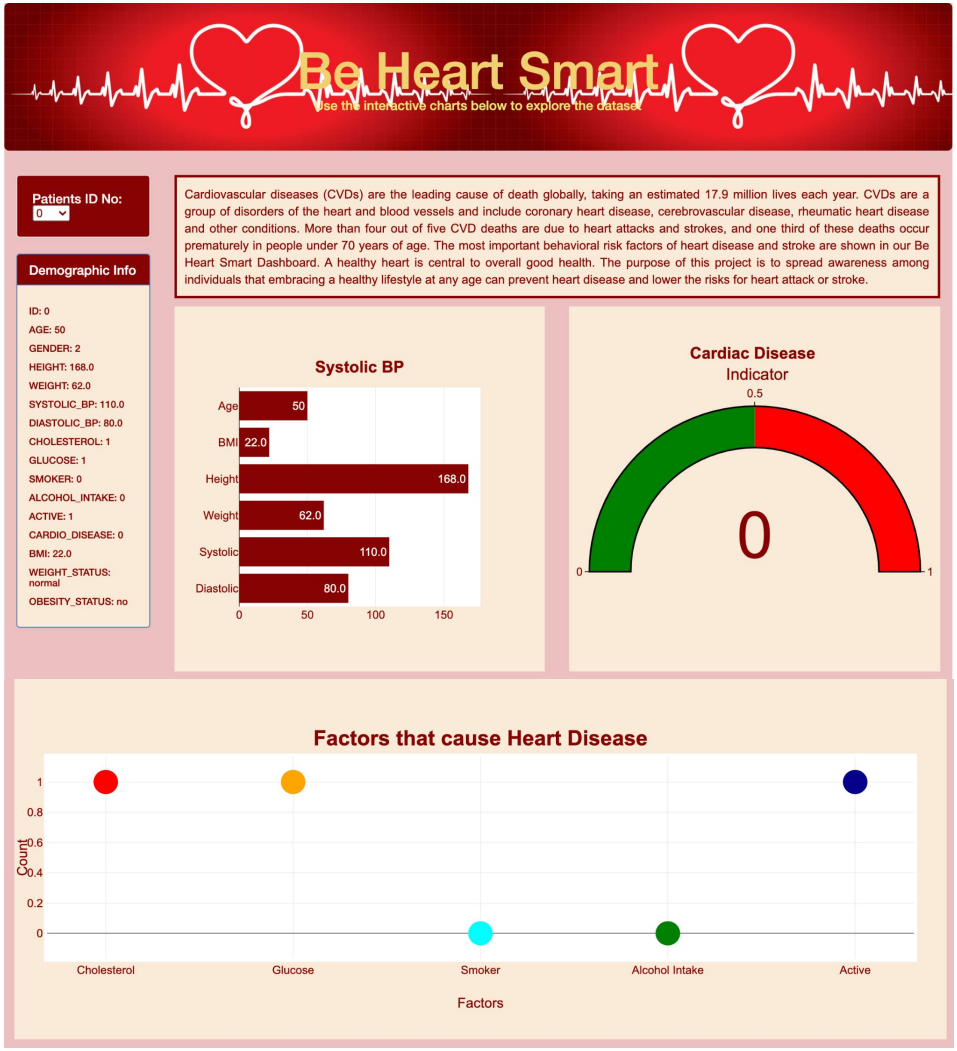


Deep Neural Network Model Visualization

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
  - Deep Neural Network
- An Interactive dashboard was developed and allows users to select a participants ID number and a gauge will show if the chosen participant is at risk of developing heart disease.
  - If the gauge reads “0” then the patient is not at risk of developing Heart disease
  - If the gauge reads “1” then there is risk that the participant may develop Heart disease based on analysis the data features collected
  - Bar graph of the Important features
  - Bubble graph of behavioral features
- Web scrap of latest news from the American Heart Association



Dashboard

❤️ Click Me