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 :

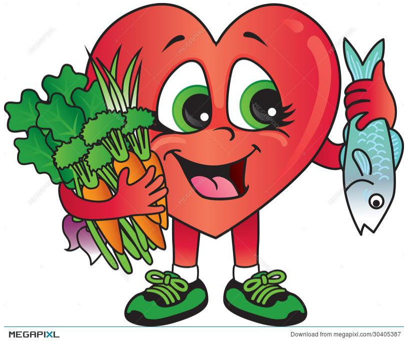
- Unhealthy diet,
- Physical inactivity
- Tobacco use
- Harmful use of alcohol

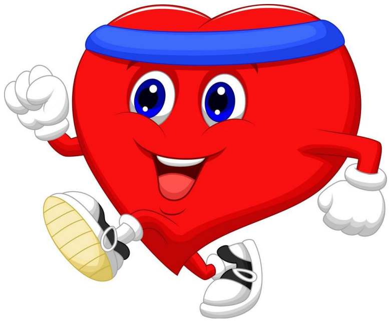
Effects of behavioral risk factors :

- 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

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?

Data repository

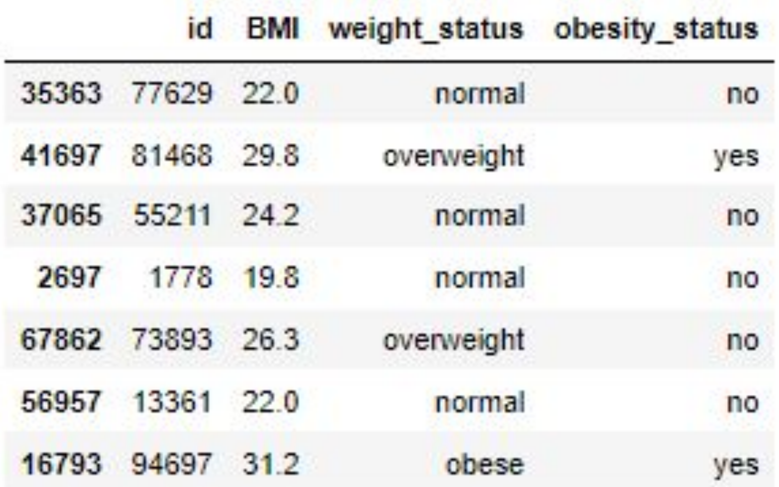
- ➔ Database : PostgreSQL

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

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

Cleaning, processing, feature engineering

- 70,000 observations. Cleaned in PySpark
  - Numbers not observed in adult human population removed (systolic BP 16,020)



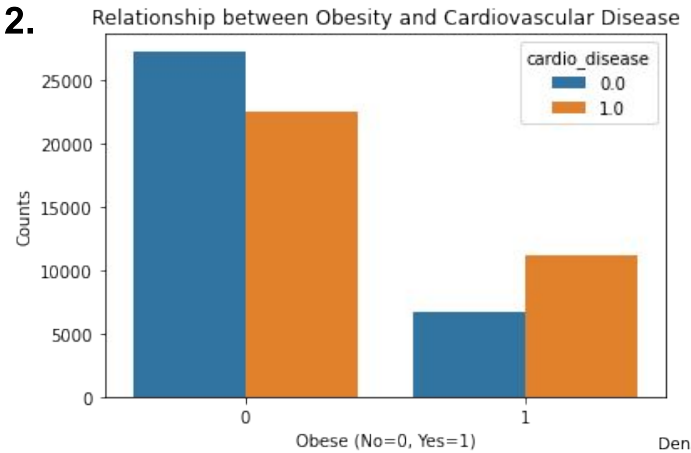
	id	BMI	weight_status	obesity_status
35363	77629	22.0	normal	no
41697	81468	29.8	overweight	yes
37065	55211	24.2	normal	no
2697	1778	19.8	normal	no
67862	73893	26.3	overweight	no
56957	13361	22.0	normal	no
16793	94697	31.2	obese	yes

- ❖ BMI (using information from CDC.gov)
  - BMI between 15-60
- ❖ Pulse Pressure (Difference between systolic and diastolic blood pressure numbers)
  - Positive and greater than 20

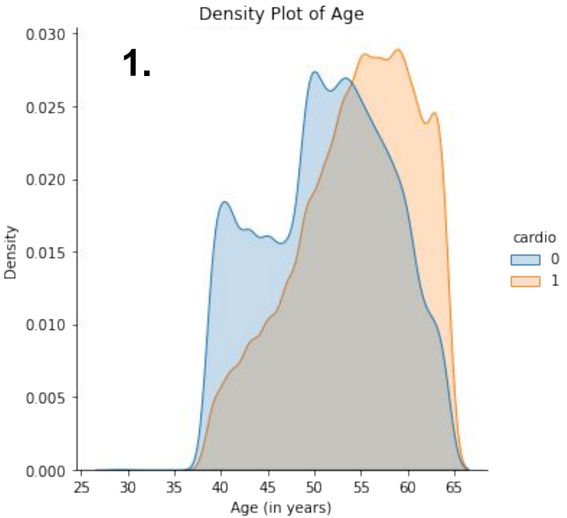
- ★ New features created - BMI, weight status and obesity, pulse-pressure
- ★ Total number of observation : 67466

Interesting insights

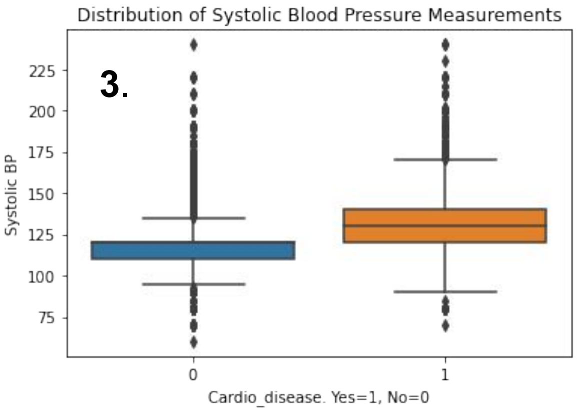
Orange is positive for CVD
Blue is negative for CVD



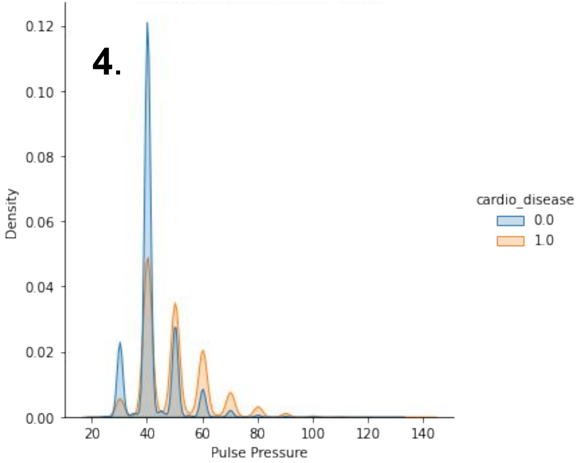
2. Obesity increases incidence of CVD



1 .Shift towards positive for CVD with increasing age



3. Patients with CVDs have higher blood pressure numbers

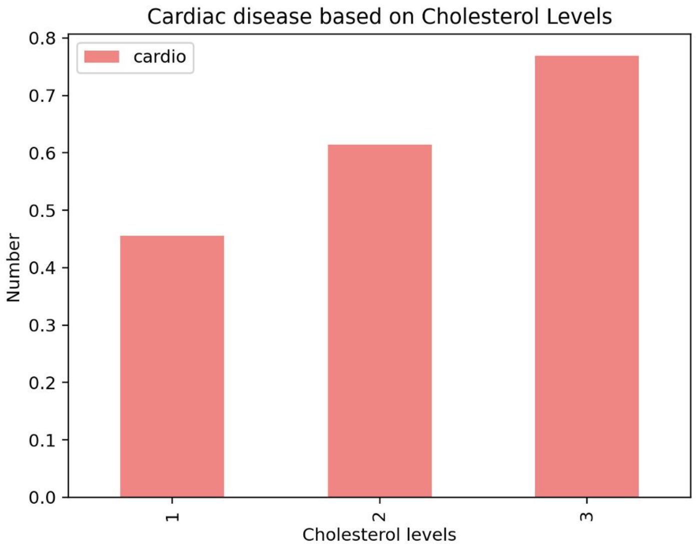


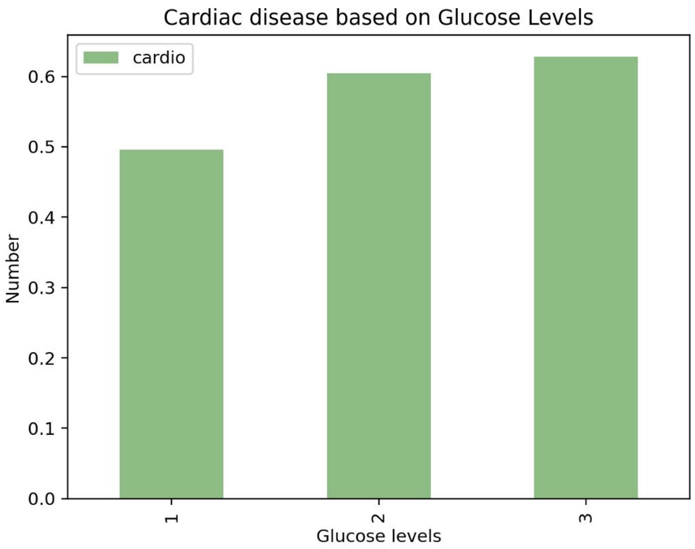
4. Higher pulse pressure shows increasing density for CVD

Data Processing and Exploratory Data Analysis

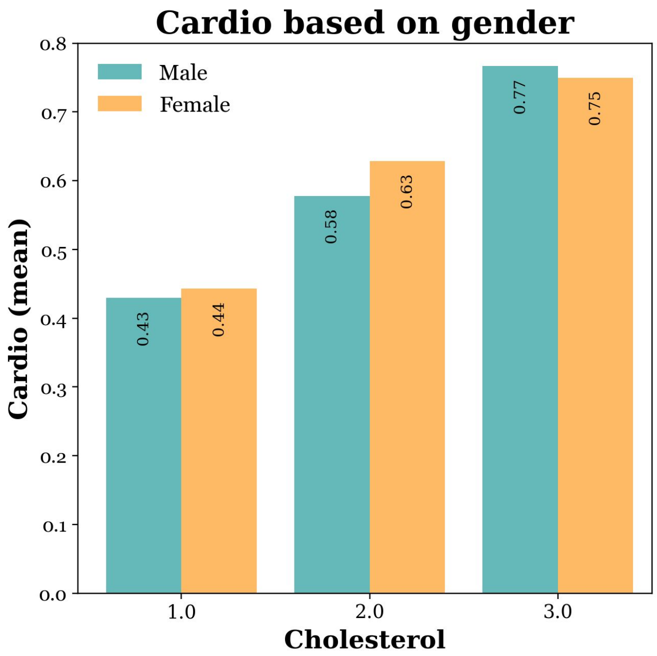
Exploratory Data-Analysis:

Performed on the initial trial pre-processed data on Excel

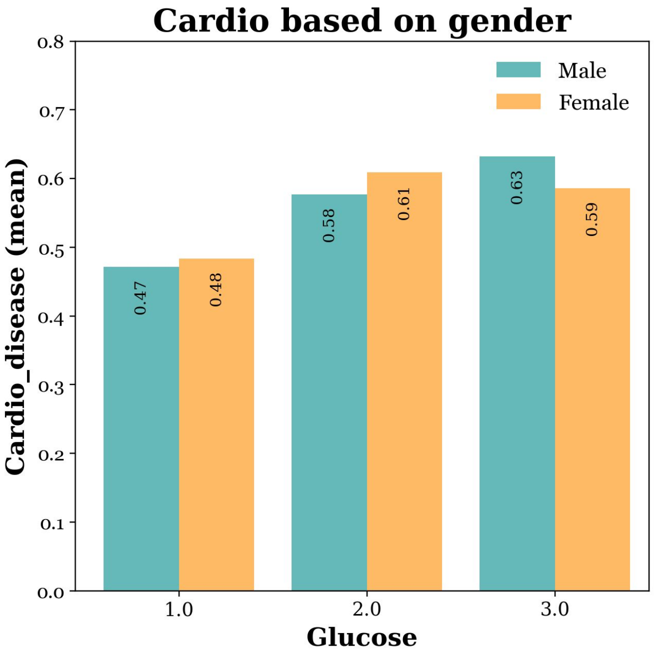




The effect of Cholesterol and Glucose on cardiac disease based on Gender



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

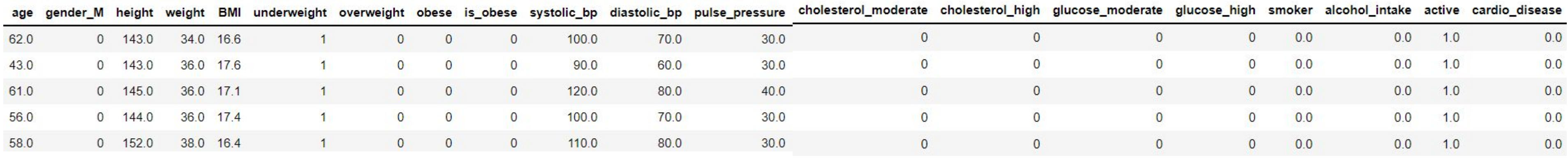


- Men with high Glucose levels have a higher chance of developing cardiac diseases.

Supervised Machine Learning : Logistic Regression

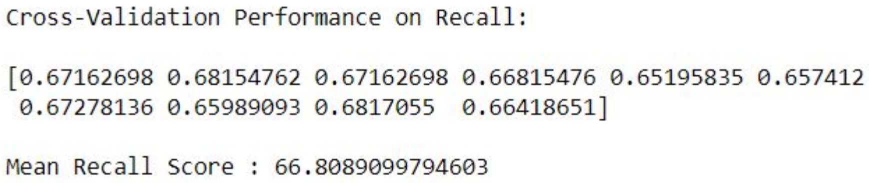
Purpose : Given a set of health and lifestyle conditions, the algorithm will be able to predict if the user has or does not have cardiovascular disease (CVD) -- a binary classification

Final_Table



age	gender_M	height	weight	BMI	underweight	overweight	obese	is_obese	systolic_bp	diastolic_bp	pulse_pressure	cholesterol_moderate	cholesterol_high	glucose_moderate	glucose_high	smoker	alcohol_intake	active	cardio_disease
62.0	0	143.0	34.0	16.6	1	0	0	0	100.0	70.0	30.0	0	0	0	0	0.0	0.0	1.0	0.0
43.0	0	143.0	36.0	17.6	1	0	0	0	90.0	60.0	30.0	0	0	0	0	0.0	0.0	1.0	0.0
61.0	0	145.0	36.0	17.1	1	0	0	0	120.0	80.0	40.0	0	0	0	0	0.0	0.0	1.0	0.0
56.0	0	144.0	36.0	17.4	1	0	0	0	100.0	70.0	30.0	0	0	0	0	0.0	0.0	1.0	0.0
58.0	0	152.0	38.0	16.4	1	0	0	0	110.0	80.0	30.0	0	0	0	0	0.0	0.0	1.0	0.0

- ★ Data divided into Train, Validation, and Test sets (60 : 20 : 20 %). Test set is unseen
- ★ Scaled using standard scaler. Fit on the Train set, and transformed Train, Validation, and Test
- ★ K Fold Cross Validation with k=10 (scoring on recall) :



```
Cross-Validation Performance on Recall:

[0.67162698 0.68154762 0.67162698 0.66815476 0.65195835 0.657412
 0.67278136 0.65989093 0.6817055  0.66418651]

Mean Recall Score : 66.8089099794603
```

Logistic Regression :

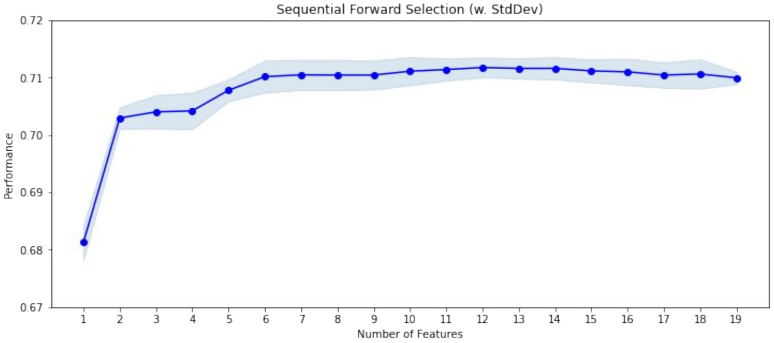
(On the Validation Set)

Accuracy : 72.7 %, Recall : 67 %

Precision : 75 %, F1-score : 71 %

Feature selection :

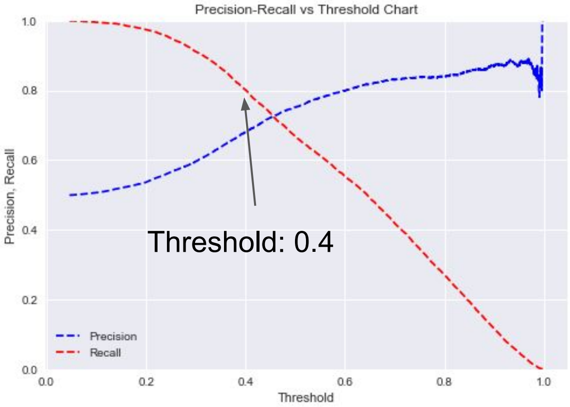
(Sequential Feature Selector was used)



Features:

1. age
2. underweight
3. is_obese
4. systolic_bp
5. pulse_pressure
6. cholesterol_high
7. active

It is important to not missing patients with cardiovascular disease, therefore recall is maximized



Threshold : 0.4 (on Test Set)

Accuracy	70.9 %
Recall	**80 %**
Precision	67.2 %
F1-score	73 %

Key Takeaways

➔ LR model was hypertuned to detect cardiovascular risk, even at the expense of including some false positives

➔ Out of the 19 original features, the model returned 7 as key determinants

Supervised Machine Learning

Random Forest Classifier

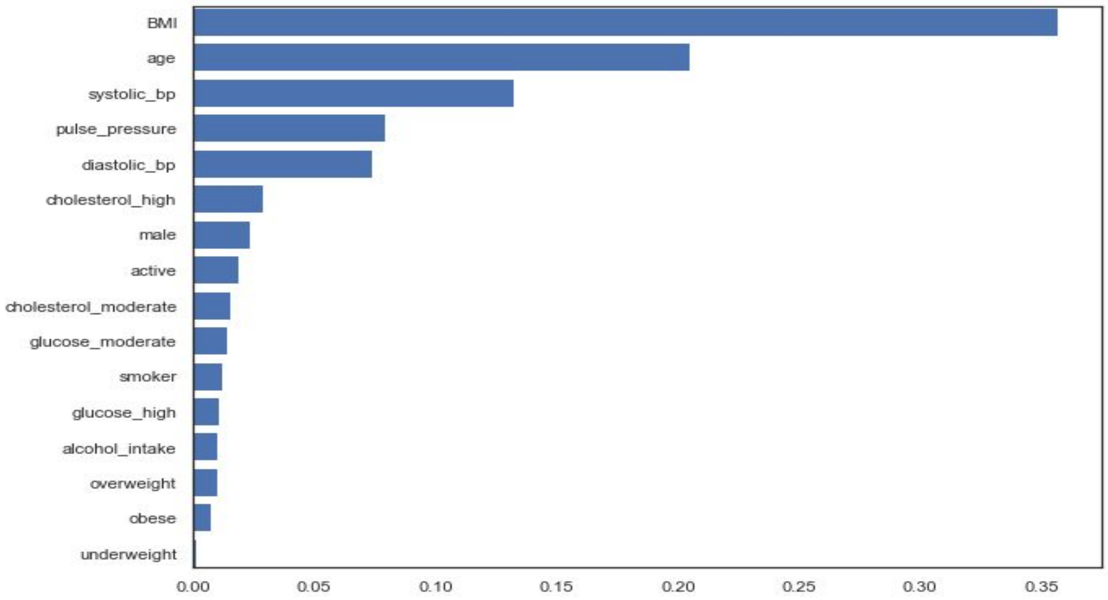
- Split the data into ‘Train: Validation: Test’ sets → 70: 20: 10
- Perform 10-fold Cross-validation
- Predict Random Forest classifier model--with default hyperparameters

	Initial model
Accuracy	69.1%
Recall	69.3%
F1	69.2%

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



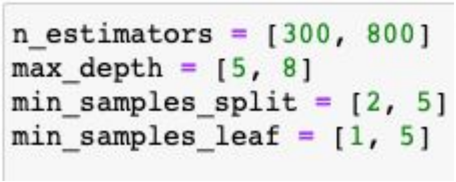
	Model with feature selection
Accuracy	66.19%
Recall	64.0%
F1	66.0%

Supervised Machine Learning

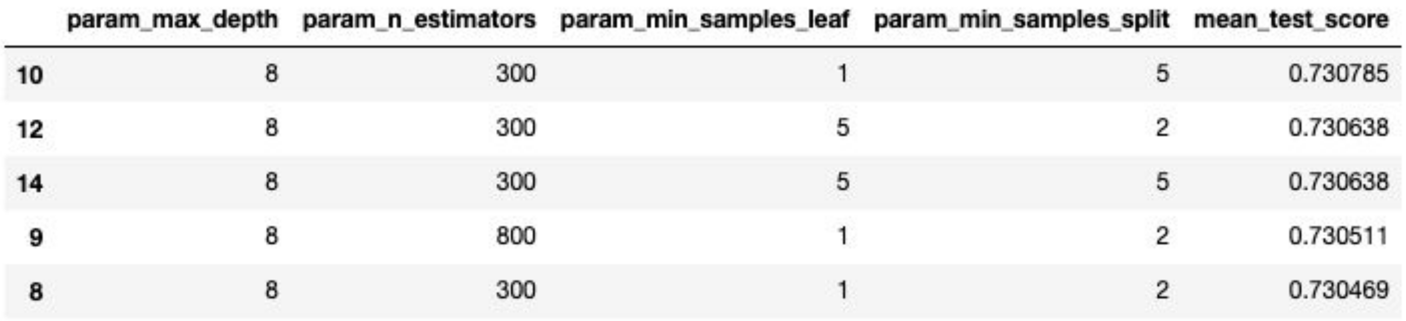
Random Forest Classifier

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

--Top 5 Hyperparameters--

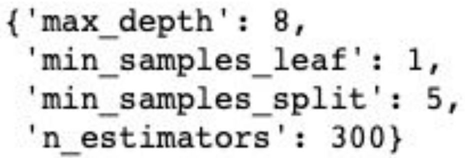


```
n_estimators = [300, 800]
max_depth = [5, 8]
min_samples_split = [2, 5]
min_samples_leaf = [1, 5]
```



	param_max_depth	param_n_estimators	param_min_samples_leaf	param_min_samples_split	mean_test_score
10	8	300	1	5	0.730785
12	8	300	5	2	0.730638
14	8	300	5	5	0.730638
9	8	800	1	2	0.730511
8	8	300	1	2	0.730469

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

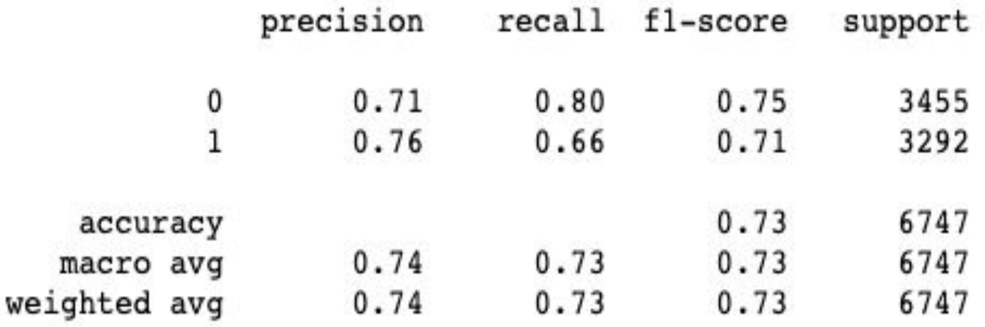


```
{'max_depth': 8,
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'n_estimators': 300}
```

Supervised Machine Learning

Random Forest Classifier

Classification Report

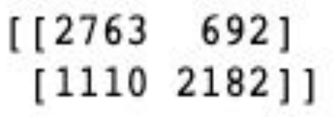


```
              precision    recall  f1-score   support

           0       0.71      0.80      0.75      3455
           1       0.76      0.66      0.71      3292

    accuracy                           0.73      6747
   macro avg       0.74      0.73      0.73      6747
weighted avg       0.74      0.73      0.73      6747
```

Confusion Matrix

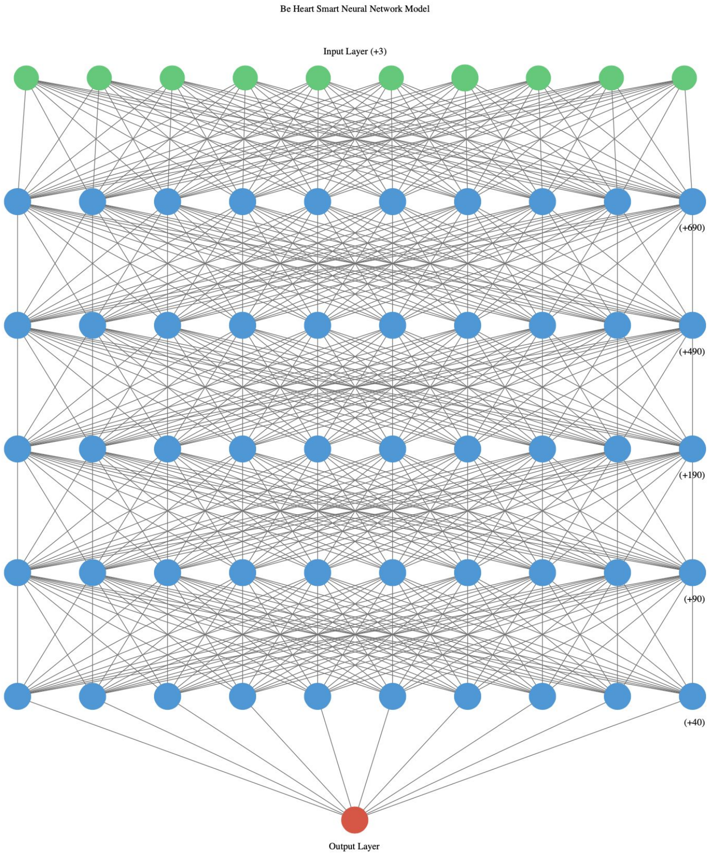


```
[[2763  692]
 [1110 2182]]
```

	Final model
Accuracy	73.3%
Recall	66.0%
Precision	76.0%
F1	71.0%

Neural network Model

- The deep neural network model was 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%



Comparison of Machine Learning

	Logistic Regression	Random Forest	Deep Neural network
Accuracy	70%	73.3%	73.0%
Precision	67%	76.0%	-
Recall	80%	66.0%	-
F1	73%	71.0%	-

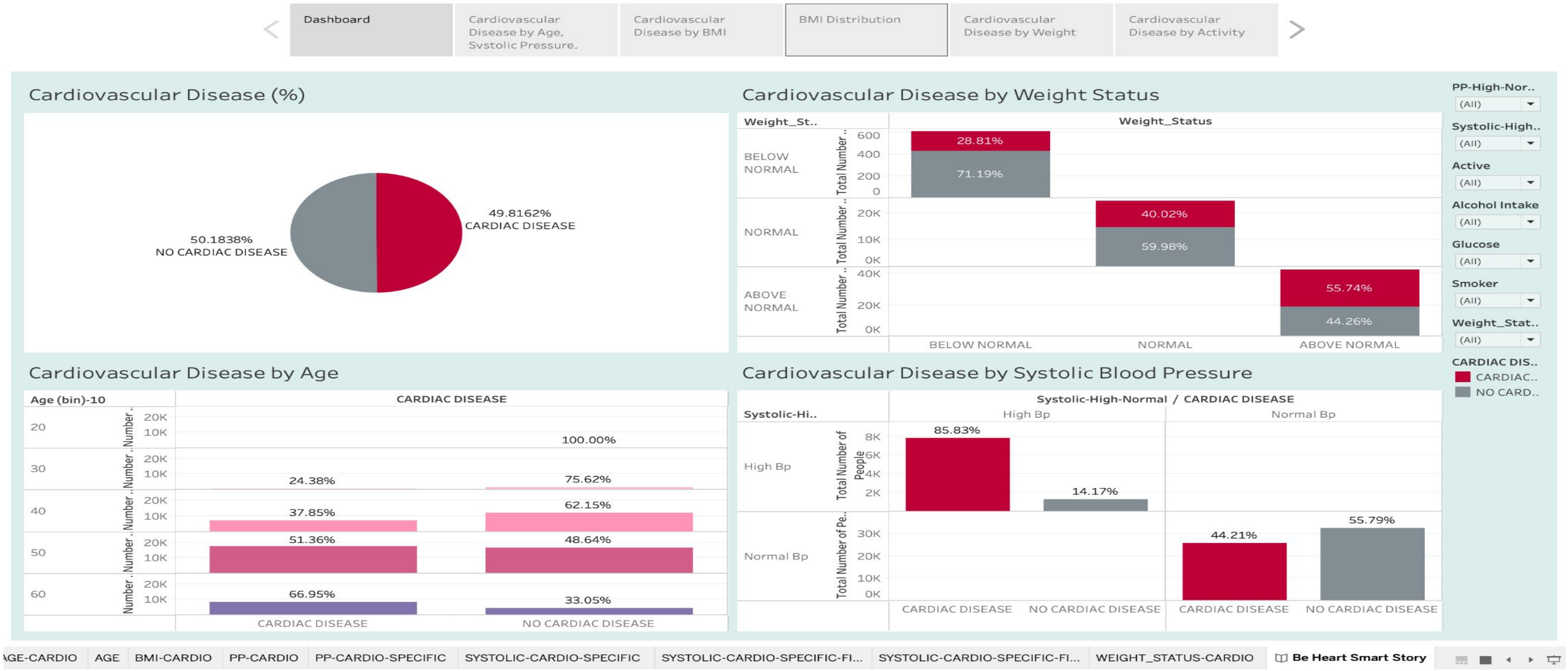
We want to maximize recall without compromising accuracy

Thank you

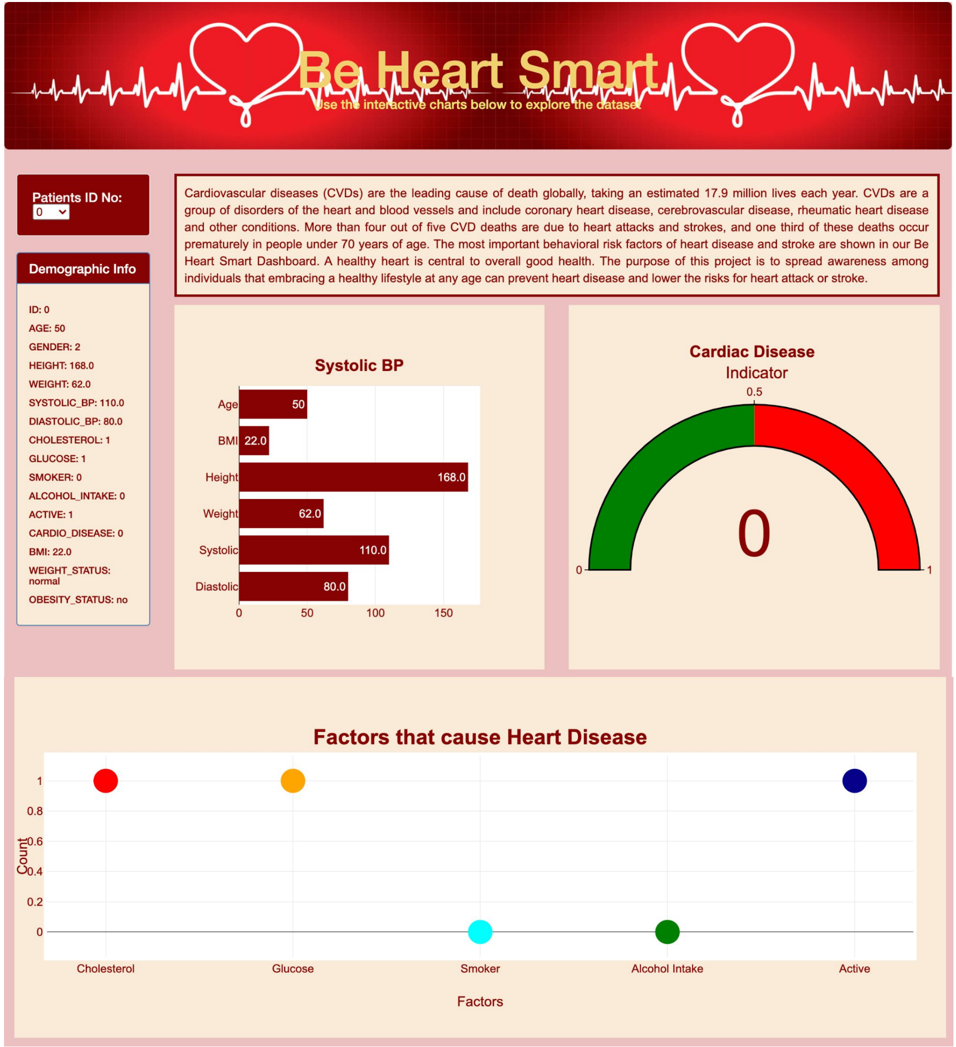
We would like to thank

- Klaus
- Artem
- Jacob
- Trent
- Jackson
- Geoff
- Gael
- Tutors
- All our classmates
- Last but not the least the Amazing Group 7 members .

Tableau Dashboard



Dashboard



Be Heart Smart

Use the interactive charts below to explore the dataset

Patients ID No: 0

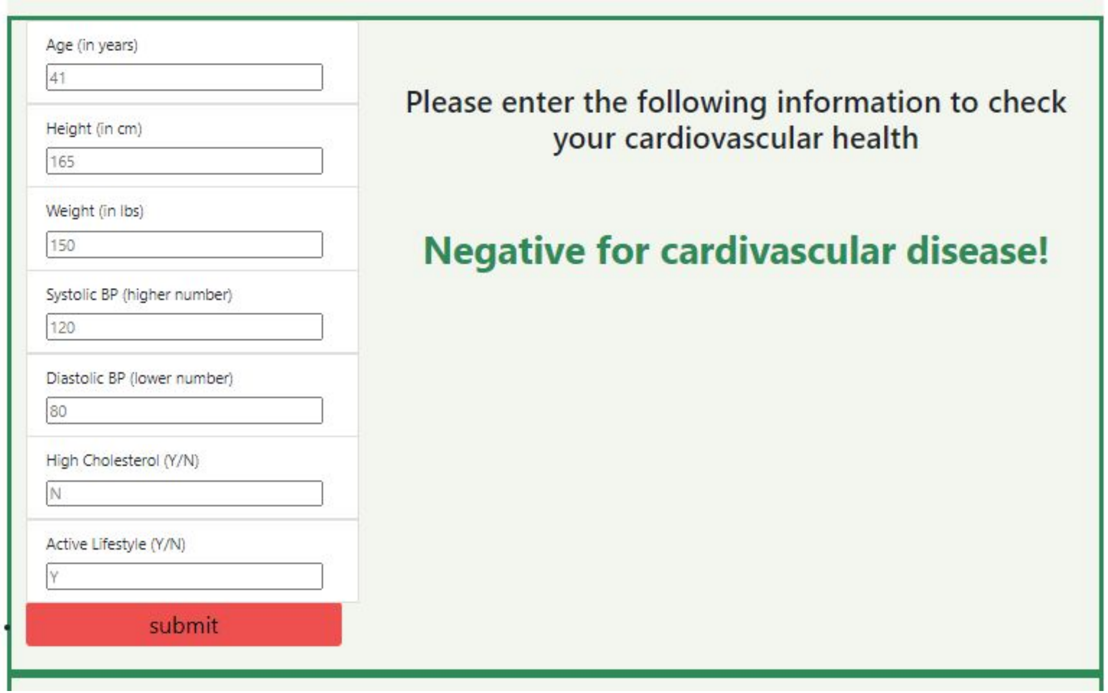
Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four 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 shown in our Be Heart Smart Dashboard. 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.

Demographic Info

ID: 0
AGE: 50
GENDER: 2
HEIGHT: 168.0
WEIGHT: 62.0
SYSTOLIC_BP: 110.0
DIASTOLIC_BP: 80.0
CHOLESTEROL: 1
GLUCOSE: 1
SMOKER: 0
ALCOHOL_INTAKE: 0
ACTIVE: 1
CARDIO_DISEASE: 0
BMI: 22.0
WEIGHT_STATUS: normal
OBESITY_STATUS: no

❤️ Click Me

Dashboard :
Web Application to Predict Cardiovascular Disease



Age (in years)

41

Height (in cm)

165

Weight (in lbs)

150

Systolic BP (higher number)

120

Diastolic BP (lower number)

80

High Cholesterol (Y/N)

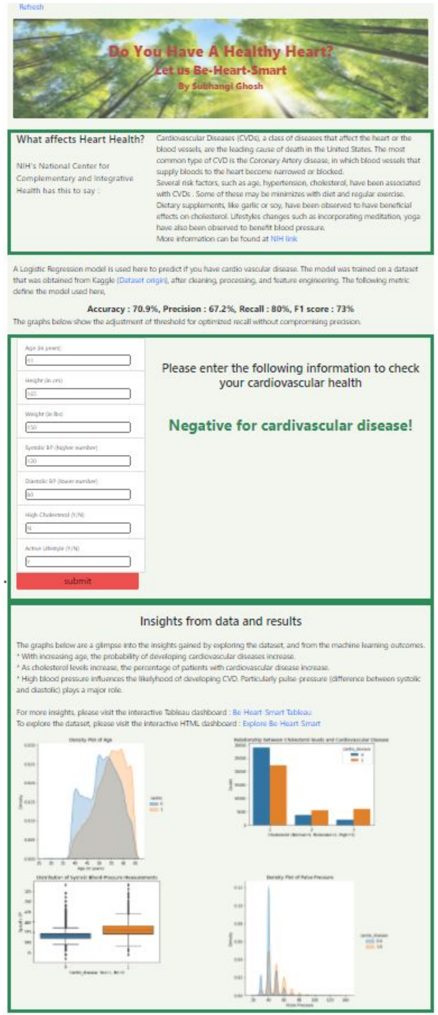
N

Active Lifestyle (Y/N)

Y

Please enter the following information to check your cardiovascular health

Negative for cardivascular disease!



Do You Have A Healthy Heart?

Let us Be-Heart-Smart

By Subhangi Ghosh

What affects Heart Health?

NIH's National Center for Complementary and Integrative Health has this to say :

Cardiovascular Diseases (CVDs), a class of diseases that affect the heart or the blood vessels, are the leading cause of death in the United States. The most common type of CVD is the Coronary Artery disease, in which blood vessels that supply bloods to the heart become narrowed or blocked.

Several risk factors, such as age, hypertension, cholesterol, have been associated with CVDs . Some of these may be minimizes with diet and regular exercise. Dietary supplements, like garlic or soy, have been observed to have beneficial effects on cholesterol. Lifestyles changes such as incorporating meditation, yoga have also been observed to benefit blood pressure.

More information can be found at NIH link

A Logistic Regression model is used here to predict if you have cardio vascular disease. The model was trained on a dataset that was obtained from Kaggle (Dataset origin), after cleaning, processing, and feature engineering. The following metric define the model used here,

Accuracy : 70.9%, Precision : 67.2%, Recall : 80%, F1 score : 73%

The graphs below show the adjustment of threshold for optimized recall without compromising precision.

Age (in years)

Height (in cm)

Weight (in lbs)

Systolic BP (higher number)

Diastolic BP (lower number)

High Cholesterol (Y/N)

N

Active Lifestyle (Y/N)

Please enter the following information to check your cardiovascular health

Negative for cardivascular disease!

Insights from data and results

The graphs below are a glimpse into the insights gained by exploring the dataset, and from the machine learning outcomes.

* With increasing age, the probability of developing cardiovascular diseases increase.
* As cholesterol levels increase, the percentage of patients with cardiovascular disease increase.
* High blood pressure influences the likelyhood of developing CVD. Particularly pulse pressure (difference between systolic and diastolic) plays a major role.

For more insights, please visit the interactive Tableau dashboard : Be Heart Smart Tableau

To explore the dataset, please visit the interactive HTML dashboard : Explore Be Heart Smart

Density Plot of Age
Relationship between Cholesterol levels and Cardiovascular Disease
cardio_disease
0
1
Distribution of Systolic Blood Pressure Measurements
Density Plot of Pulse Pressure