SC1015 Mini-Project

Robo-Medivisor

Heart Disease Prediction

SC19 Team 7

Koh Hao Sheng (U2122672K)
Lee An Ni (U2122370D)
Sim Guanyu (U2120328B)

Table of Contents

01

Our Motivation

Introduction

02

Exploring Dataset

Exploratory Data Analysis & Data-driven Insights

03

Core Analysis

Machine Learning Models

04

Our Outcome

Insights & Solution

01

Our Motivation

Introduction

Heart Disease

- Leading cause of death worldwide → 32% of all deaths
- In Singapore, about 19 people die from it everyday

Problem Definition

How can we assist doctors to speed up the diagnosis of heart disease to minimise further implications?

02

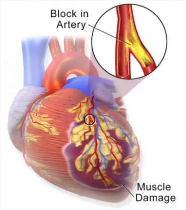
Exploring Dataset

Exploratory Data Analysis &
Data-driven Insights

Data Preparation

Dataset Used:

HEART DISEASE DATASET (COMPREHENSIVE)



4 ratings - Please login to submit your rating.

Citation Author(s): Manu Siddhartha *(Liverpool John Moore's University)*

Submitted by: MANU SIDDHARTHA

Last updated: Fri, 11/06/2020 - 04:17

DOI: 10.21227/dz4t-cm36

Data Format: *.csv

Links: A database for using machine learning and data mining techniques for coronary artery disease diagnosis

License: Creative Commons Attribution

ACCESS DATASET | CITE | SHARE/EMBED

16779 Views

Categories: Machine Learning, Health, Biomedical and Health Sciences

Keywords: Heart Disease, Coronary artery disease, Cardiovascular disease, heart disease dataset

Dataset Variables:

Heart Disease Dataset Attribute Description

S.No.	Attribute	Code given	Unit	Data type
1	age	Age	in years	Numeric
2	sex	Sex	1, 0	Binary
3	chest pain type	chest pain type	1,2,3,4	Nominal
4	resting blood pressure	resting bp s	in mm Hg	Numeric
5	serum cholesterol	cholesterol	in mg/dl	Numeric
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Binary
7	resting electrocardiogram results	resting ecg	0,1,2	Nominal
8	maximum heart rate achieved	max heart rate	71–202	Numeric
9	exercise induced angina	exercise angina	0,1	Binary
10	oldpeak =ST	oldpeak	depression	Numeric
11	the slope of the peak exercise ST segment	ST slope	0,1,2	Nominal
12	class	target	0,1	Binary

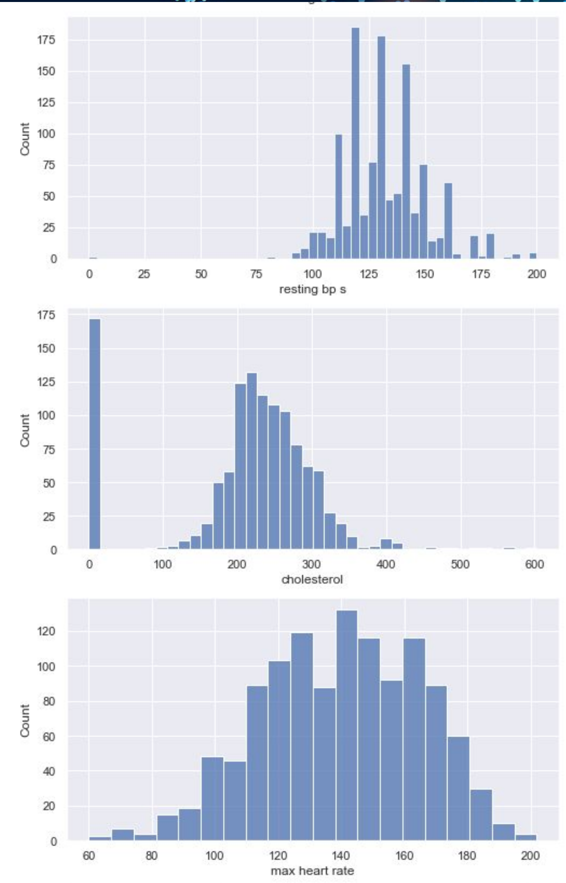
Data Preparation

Data Cleaning:

Separated numerical and categorical variables
Renamed variable (sex) for exploratory data analysis
Removed anomalies for numerical data
Ensured dataset is balanced

Data Visualisation:

Importing pandas and NumPy to analyse data, seaborn to analyse relationship and several scikit-learn tools for regression and classification

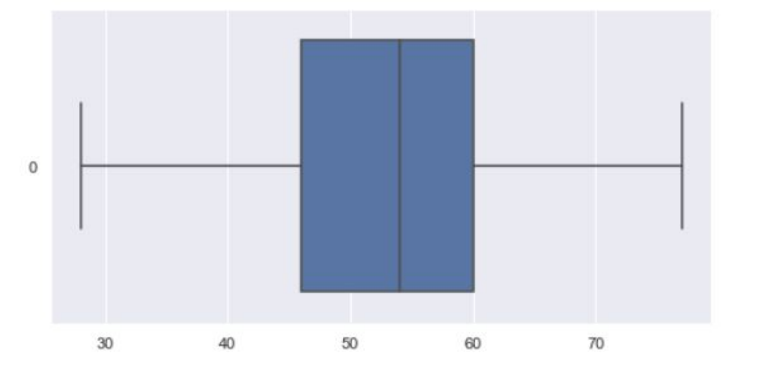


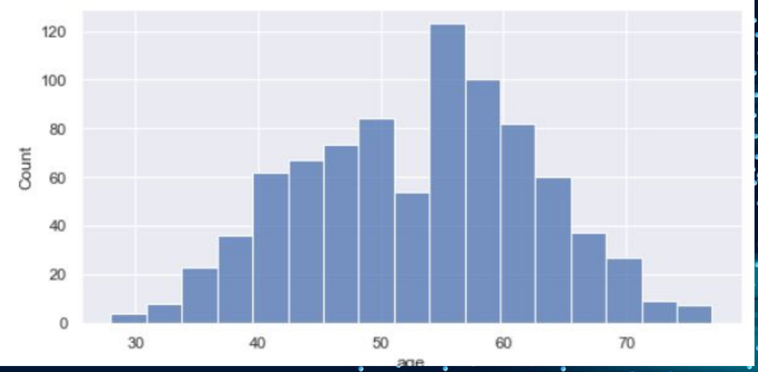
Exploratory Data Analysis

Numeric Variables (Uni-variate)

- Box Plot
- Histogram

	age	resting bp s	cholesterol	max heart rate	oldpeak
count	856.00	856.00	856.00	856.00	856.00
mean	53.10	130.99	243.72	137.97	0.99
std	9.47	15.67	56.13	22.40	1.09
min	28.00	92.00	85.00	69.00	-0.10
25%	46.00	120.00	208.00	122.00	0.00
50%	54.00	130.00	237.00	140.00	0.80
75%	60.00	140.00	274.00	155.00	1.70
max	77.00	170.00	603.00	185.00	6.20





Exploratory Data Analysis

Numeric Variables (Multi-variate)

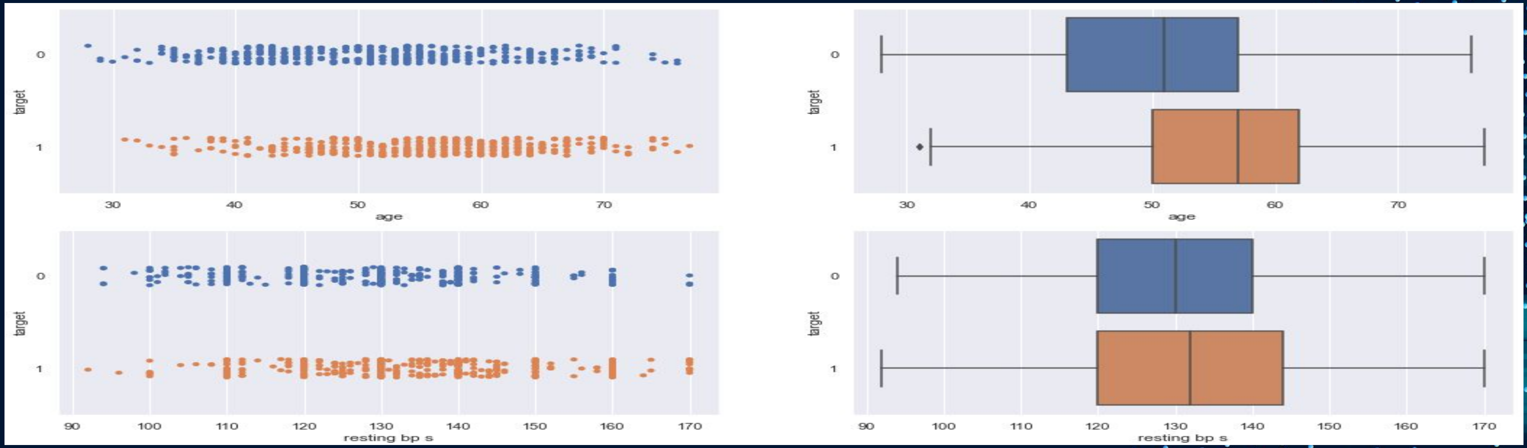
- Correlation Table and Heatmap

	age	resting bp s	cholesterol	max heart rate	oldpeak
age	1.000000	0.280102	0.047276	-0.443281	0.298883
resting bp s	0.280102	1.000000	0.099058	-0.155518	0.230970
cholesterol	0.047276	0.099058	1.000000	-0.020512	0.042241
max heart rate	-0.443281	-0.155518	-0.020512	1.000000	-0.233011
oldpeak	0.298883	0.230970	0.042241	-0.233011	1.000000

Exploratory Data Analysis

Numeric Variables (Predictors vs Target)

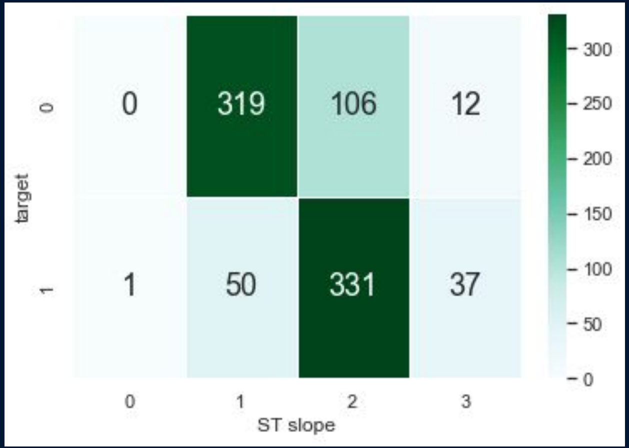
- Strip Plot
- Box Plot



Exploratory Data Analysis

Categorical Variables (Predictors vs Target)

- Heatmap



03

Core Analysis

Machine Learning Models

Machine Learning Models

Decision Tree

Random Forest

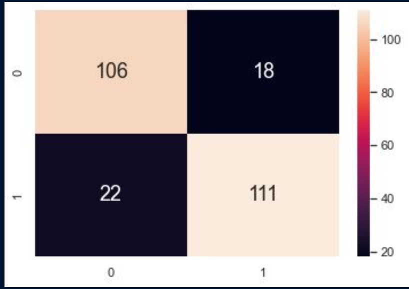
Logistic Regression

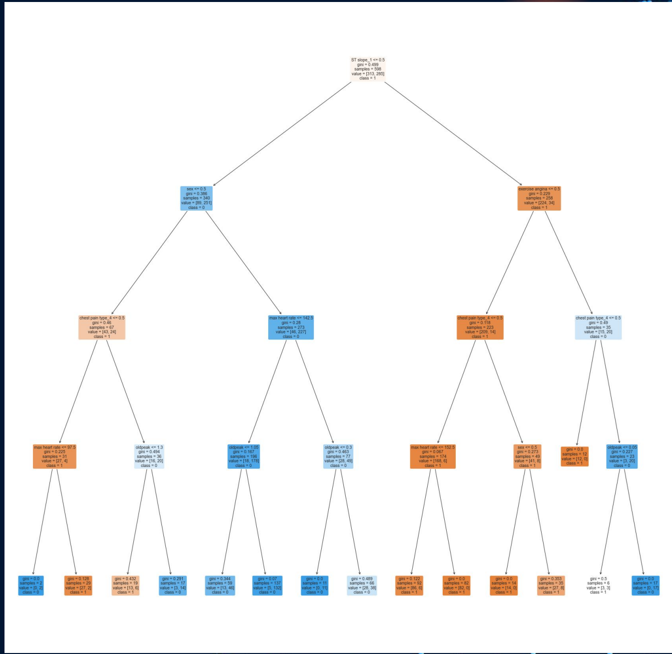
Decision Tree

Classification Accuracy

Train: ~87.63%

Test: ~84.44%





Random Forest

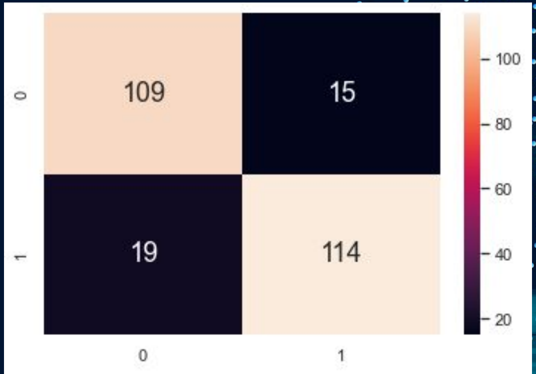
Classification Accuracy

Train: ~88.80%

Test: ~86.77%

Number of decision trees used: 100
Maximum depth of each tree: 4

Train



Test

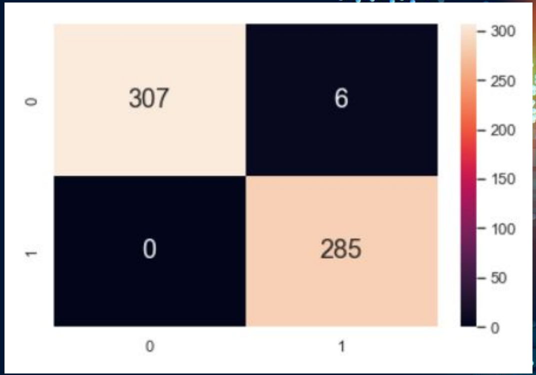
Random Forest

Classification Accuracy

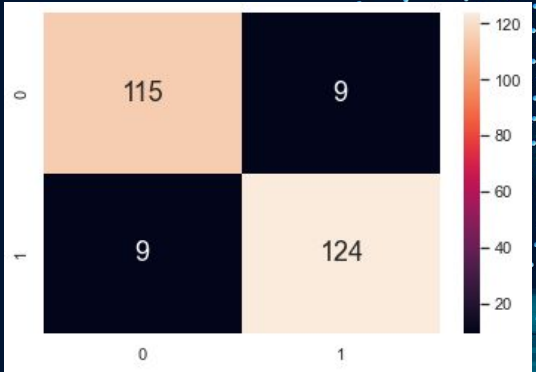
Train: ~99.00%

Test: ~93.00%

After adjusting 2 major hyper-parameters:
Number of decision trees used: 1000
Maximum depth of each tree: 10



Train



Test

Logistic Regression

```
Optimization terminated successfully.
         Current function value: 0.366927
         Iterations 7
Table 2:
                          Results: Logit
===================================================================
Model:              Logit            Pseudo R-squared: 0.470
Dependent Variable: target           AIC:              460.8447
Date:               2022-04-23 01:54 BIC:              509.1742
No. Observations:   598              Log-Likelihood:   -219.42
Df Model:           10               LL-Null:          -413.85
Df Residuals:       587              LLR p-value:      2.2210e-77
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-------------------------------------------------------------------
                     Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
-------------------------------------------------------------------
age                 -0.0218   0.0137 -1.5942 0.1109 -0.0486  0.0050
sex                  1.6003   0.3071  5.2103 0.0000  0.9983  2.2022
chest pain type      0.4838   0.1330  3.6376 0.0003  0.2231  0.7445
resting bp s        -0.0054   0.0075 -0.7298 0.4655 -0.0201  0.0092
cholesterol          0.0019   0.0022  0.8373 0.4024 -0.0025  0.0063
fasting blood sugar  0.2432   0.3392  0.7170 0.4734 -0.4217  0.9081
resting ecg          0.1319   0.1376  0.9583 0.3379 -0.1378  0.4016
max heart rate      -0.0330   0.0049 -6.7075 0.0000 -0.0426 -0.0234
exercise angina      1.0641   0.2682  3.9682 0.0001  0.5385  1.5897
oldpeak              0.6580   0.1459  4.5094 0.0000  0.3720  0.9439
ST slope             1.1627   0.2506  4.6389 0.0000  0.6714  1.6539
===================================================================
```

Model 1

```
Optimization terminated successfully.
         Current function value: 0.334024
         Iterations 7
Table 1:
                          Results: Logit
====================================================================
Model:              Logit            Pseudo R-squared: 0.517
Dependent Variable: target           AIC:              421.4930
Date:               2022-04-23 01:54 BIC:              469.8225
No. Observations:   598              Log-Likelihood:   -199.75
Df Model:           10               LL-Null:          -413.85
Df Residuals:       587              LLR p-value:      9.2910e-86
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
--------------------------------------------------------------------
                   Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
--------------------------------------------------------------------
age                0.0248   0.0162  1.5324 0.1254 -0.0069  0.0565
max heart rate    -0.0086   0.0071 -1.2103 0.2262 -0.0225  0.0053
oldpeak            0.5110   0.1510  3.3832 0.0007  0.2150  0.8070
sex                2.0473   0.3401  6.0190 0.0000  1.3806  2.7139
exercise angina    0.9549   0.2844  3.3576 0.0008  0.3975  1.5124
chest pain type_2 -0.3124   0.6158 -0.5073 0.6119 -1.5193  0.8946
chest pain type_3  0.0718   0.5495  0.1306 0.8961 -1.0052  1.1487
chest pain type_4  1.4195   0.5295  2.6807 0.0073  0.3816  2.4574
ST slope_1        -4.7011   1.7692 -2.6571 0.0079 -8.1687 -1.2335
ST slope_2        -2.6127   1.7210 -1.5182 0.1290 -5.9858  0.7603
ST slope_3        -3.6212   1.8168 -1.9931 0.0462 -7.1821 -0.0603
====================================================================
```

Model 2

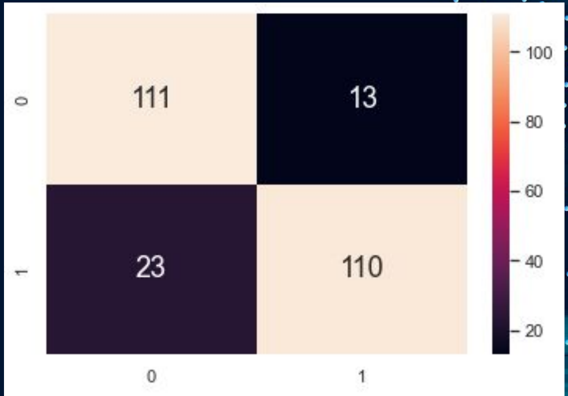
Logistic Regression

Classification Accuracy (Model 1)

Train: ~86.79%

Test: ~85.99%

Train



Test

Learning Points

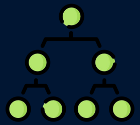
Random Forest	Handles numeric variables (regression) and categorical variables (classification)
Logistic Regression	Binary classification Categorical target

04

Our Outcome

Insights & Solution

Evaluation



Decision Tree

Pros:
Faster computation time compared to Random Forest

Cons:
Relatively less accurate since only one tree is used in the prediction, overfitting without control

Random Forest

Pros:
Constructs multiple decision trees to improve predictions, making it more stable and accurate

Cons:
Slower computation time as compared to Decision Tree

Random Forest returns a higher accuracy for our dataset

Evaluation

Random Forest

Pros:
Offers higher accuracy than Logistic Regression

Cons:
Slower computation time and harder to interpret as compared to Logistic Regression

Logistic Regression

Pros:
Easier to interpret and shorter computation time as compared to Random Forest

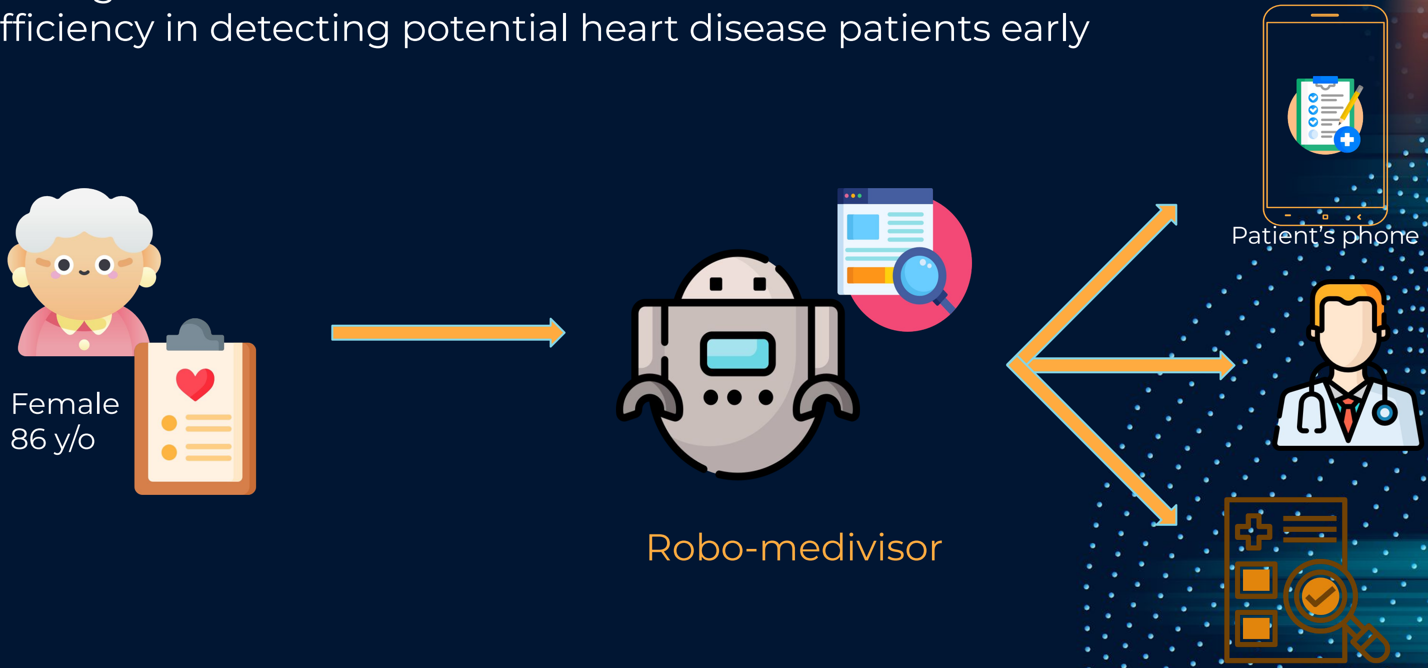
Cons:
Lower accuracy

Random Forest returns the highest accuracy for our dataset

Solution: *Robo-Medivisor*

Chosen model: **Random Forest**

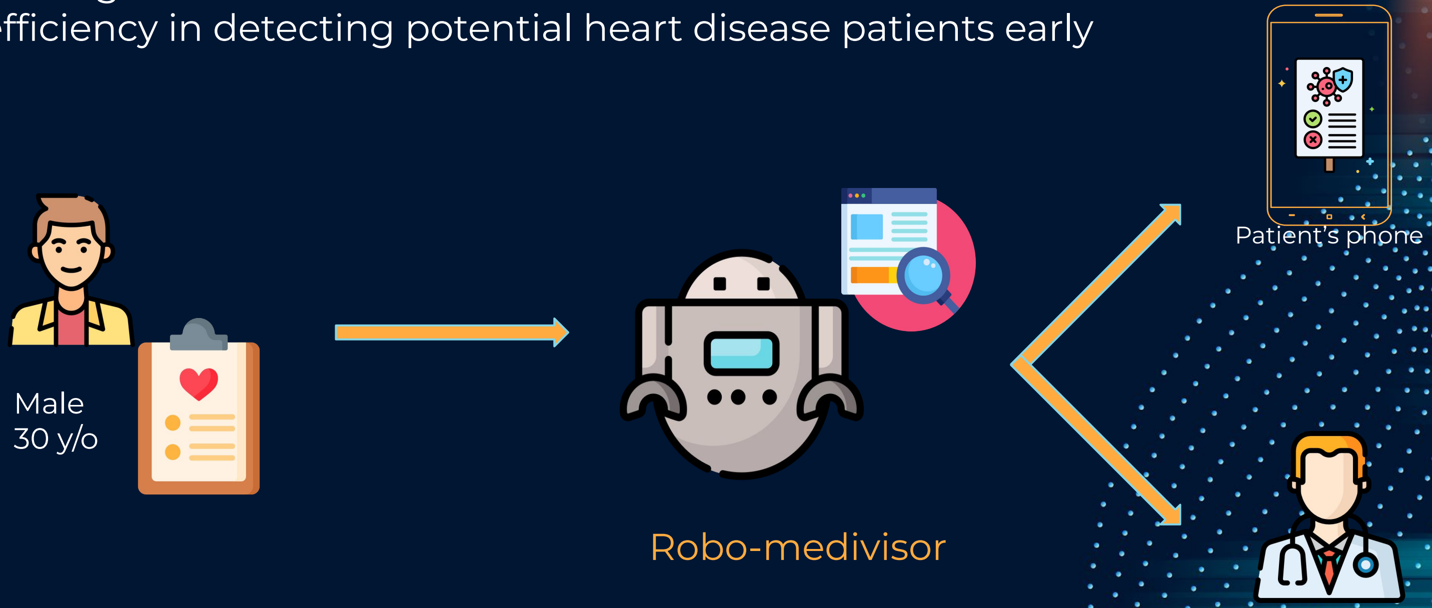
Aim: lighten the workload of doctors and increase doctors' efficiency in detecting potential heart disease patients early



Solution: *Robo-Medivisor*

Chosen model: **Random Forest**

Aim: lighten the workload of doctors and increase doctors' efficiency in detecting potential heart disease patients early



THANK YOU