HEART DISEASE PREDICTION

By Python Clown

TABLE OF CONTENTS

Overview

Background:

Heart Disease is a leading cause of death in the developed world. The government has decided to ramp up its efforts to support people who may be at high risk of Heart Disease. However, the population is way too large to screen and provide support to everyone. As such, you have been tasked with building an ML algorithm that can quickly identify such people.

Target:

Develop an ML algorithm to identify people with high risk of Heart Disease

INTRODUCTION TO DATA SET

This data set contains data about the general and health information about 270 records, including age, sex, chest pain type, etc., as well as whether the patient is having heart disease.

- Age
- Sex
- Chest pain type (1-4, 1 = typical angina, 4 = no symptoms)
- BP - blood pressure
- Cholesterol
- FBS over 120 - Fasting Blood Sugar (1 = True, 0 = False)
- EKG results (0-2, 0 = normal, 2 = abnormal)
- Max Heart Rate(HR)
- Exercise angina (1 = True, 0 = False)
- ST depression
- Slope of ST(1-3)
- Number of vessels fluro
- Thallium(3 values)
- Heart Disease - Presence/Absence

METHODOLOGY

Conducted Train-Test split - 80:20

↓

Fit 3 models , Logistic Regression, Random Forest and Support Vector Machine with the training set

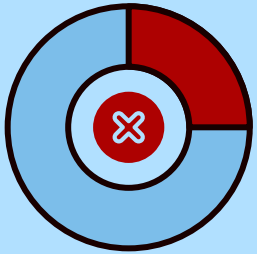
↓

Evaluated the model using test set, including the metrics: accuracy score, confusion matrix

↓

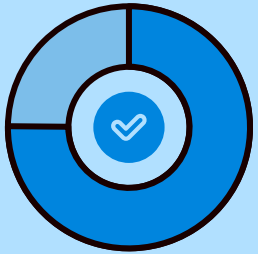
Interpret the results and select the best classifier

Test Results



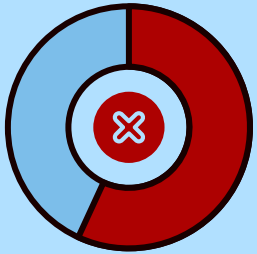
83.3%

Logistic Regression



85.1%

Random Forest



84.4%

SVM

Confusion Matrix

Logistic Regression

	Present	Absent
Present	19	**4**
Absent	5	26

Recall = TPR = 19/(19+4) = 82.6%

Random Forest

	Present	Absent
Present	19	**3**
Absent	5	27

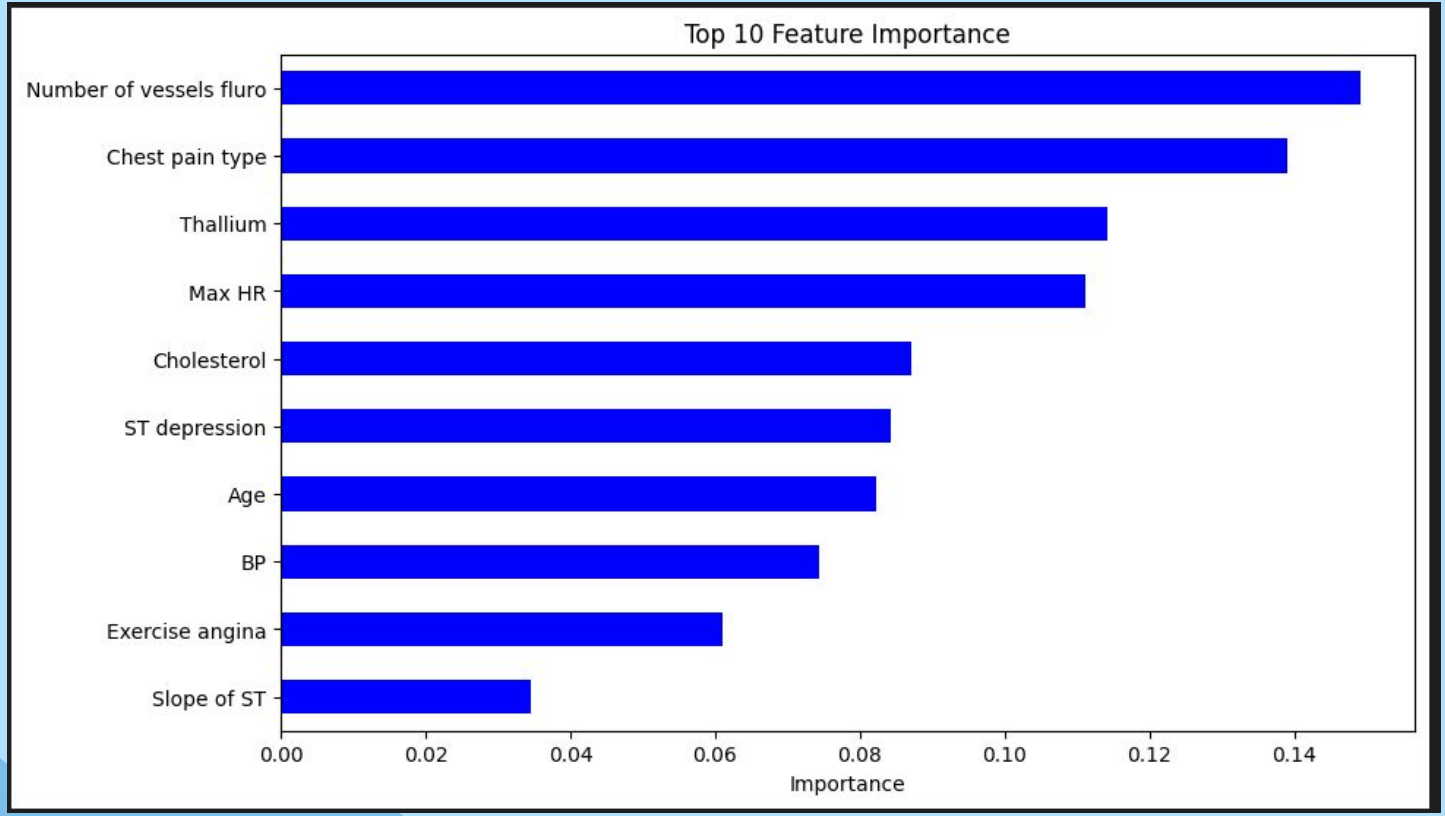
Recall = TPR = 19/(19+3) = 86.4%

SVM

	Present	Absent
Present	95	**25**
Absent	18	135

Recall = TPR = 95/(95+25) = 79.1%

Important features



Justification and Conclusion

Among the 3 models, the Random Forest Classifier has the highest accuracy and true positive rate.

Insights:

1. The government can utilise Random Forest Algorithm to predict people with high risk of Heart Disease
    a. Identified with “present” can be quickly sent to doctors for consultation while those identified with “absence” can be allocated to doctors for initial medical examination. (Resource Optimisation)
2. In their further research and treatment, the government and hospitals can focus more on the top 10 features related to Heart Diseases. E.g. number of vessel fluros and Chest pain type