TARC
TUNKU ABDUL RAHMAN UNIVERSITY COLLEGE
BEYOND EDUCATION

BACS2003 ARTIFICIAL INTELLIGENCE

202301 Session, Year 2022/23

Assignment Documentation

Full Name: Lee Jing Jet
Student ID: 2209855
Programme: Bachelor degree of software engineering
Tutorial Class: G4
Project Title: Heart Disease Prediction
Module In-Charged: Support Vector Machine Algorithm

Other team members' data

No	Student Name	Module In Charge
1	LIM MENG LEONG	Random Forest Algorithm
2	LOH WEI LUN	K-Nearest Neighbors Algorithm
3		

Lecturer: Goh Chin Pang	**Tutor : Ho Chuk Fong**

1. Introduction

1.1. Problem Background

Artificial intelligence adoption has been accelerating over the past few years, and there is no doubt that AI offers a large number of benefits and prospects. Providing relevant or useful information, knowledge, research, and prediction are a few examples. Artificial intelligence and the machine learning algorithms enable prediction tasks such as the prediction of heart disease, which is the subject of our assignment. Our team created a heart disease prediction software primarily because we firmly believe that excellent health is the best investment because it allows you to do anything.

However, Malaysians have a poor level of awareness of heart illnesses. Malaysians don't seem to care about their own health, which may be because they aren't exposed to enough health information. About 80% of Malaysia's elderly may experience chronic health issues like heart attacks, which may cause premature mortality, according to research by Murat N. Through regular medical examinations that help to identify any early indications, one of the greatest methods to prevent heart disease is to do so. However, CodeBlue has conducted research on this issue and from their findings it showed that nearly half of Malaysians are lacking health coverage beyond public care. It is very clear that the number of Malaysians who undergo monthly or yearly medical checkups are significantly lower than our neighboring countries. Our chief statistician Mohd Uzir Mahidin has stated that heart diseases have remained the principal cause of death in Malaysia with an increase of 5.4% from 11.6% in 2000 to 17% in 2020. Therefore, in Malaysia it is now more important than ever to prevent heart diseases.

The high cost of medical care is one of the main reasons Malaysians do not visit the doctor. Since, there are no obvious indicators of a significant health issue, they felt they are in good condition but it's vital to remember that not all medical issues will have readily observable early signs. Our team has developed a heart disease prediction tool that helps users determine whether they are susceptible to heart illnesses or not in the spirit of the adage "prevention is better than cure". Our team has employed a supervised machine learning algorithm in our heart disease prediction programme to determine whether a user was likely to develop heart disease. In order to ensure that people are more aware of this fatal disease, a good data driven approach helps to improve the entire research and prevention process.

A prediction system can be created by analyzing large and complex amounts of data with the aid of machine learning. Age, Sex, Type of chest discomfort, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, Result of a Resting ECG, Highest Heath Rate attained during exercise, and an older peak heart rate are the factors that be used to classify people who is being at risk for getting heart disease. The 3 different algorithms included Random Forest, K-nearest neighbor and Support Vector Machine(SVM).

1.2. Objectives/Aims

The main objective of this project is to increase the awareness of heart disease among Malaysians because heart disease is one of the principal causes of death in Malaysia. The project allows the users to check if they are suffering from the risk of getting heart disease or not. The notice message will be prompt on showing whether the users are prone to heart disease or not. There is a far greater possibility that they will either be able to reverse their heart condition or take the steps to stop it from worsening.

The heart disease prediction application helps to reduce the amount of cost used by the people to undergo monthly or yearly medical checkups as the users can get a grasp on how healthy their heart is. Although the application can only be taken with a grain of salt, it still lets users get an idea on the condition of their heart. Therefore, it helps the users to plan for their medical checkup to save costs by preventing excessive medical checkups.

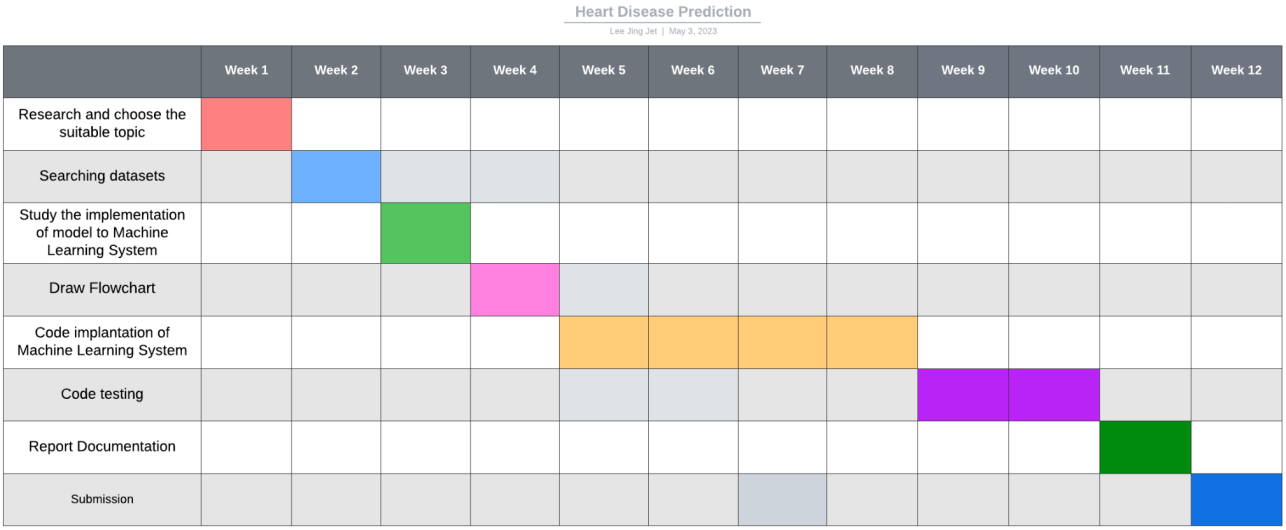
The adoption of machine learning for heart disease prediction helps users to save their time by knowing their heart condition without reading any relevant books, browse the internet to further find the symptoms of heart disease. The users can know whether they are prone to have heart disease or not by just need enter relevant details that are necessary into the system.

1.3. Motivation

There is a saying that goes, “no human being is perfect”. Doctors might make the mistakes when they are curing a patient but the mistakes they made may lead to a worse case which is a life lost. For instance, if a person who is healthy was said to have heart disease which was diagnosed wrongly by a doctor, this human error could cause serious complications to a person’s health. With the wrong prescription of medicines, it could take a major toll on the

person's health and might lead to other complications of diseases. The easiest and cheaper solution for this problem is to create a program for avoiding the problem. Therefore, our team has created this application that uses machine learning and natural language as the algorithm to make the tasks automated and provide a highly accurate and time efficient result.

2. Timeline/Milestone



3. Research Background

3.1. Background of the applications

In recent years, the world has started changing, everythings is connected to a data source and it is digitally recorded. Most of the tasks are being done automatically with the help of machine learning such as health prediction systems, stock price prediction systems and even financial prediction systems. With these prediction systems around the world, it cannot be denied that machine learning has given humans a lot of convenience in life. Machine learning is an algorithm that can automatically improve itself over time without requiring human programmers to feed in additional information. This is because after analyzing large amounts of data, Machine learning will start to modify itself in response to the data's quality and this helps to increase the precision and accuracy of the entire system overtime.

By adding and feeding Machine Learning with large amounts of medical history data, it will help to predict whether the person is prone to heart disease or not. This is because it will recognize

whether the individual is having any symptoms of heart disease such as high blood pressure, old age or different levels of chest pain. Before planning for a medical checkup, users can use our application to get a grasp on what is the condition of their heart. This helps the users to save money and time.

The application utilizes the machine learning algorithms to predict whether a user is prone to heart disease. Our team has chosen 3 types of algorithms which are Random Forest, K-Nearest Neighbor(KNN) and Stacking CV technique with the Extreme Gradient Boost algorithm as the basis of models for improving the performance of models. The reason for implementing multiple algorithms to use is because it helps to increase the accuracy and precision of the data being fed. Before any data is being sent to the models, we will be converting all the raw data into an intelligible format which is during the data preprocessing stage. During the data preprocessing stage, missing values, cleaning of data and also normalization will be done. Therefore, the accuracy and performance of our models can be easily evaluated through a variety of performance metrics.

3.2. Analysis of selected tool with any other relevant tools

Tools comparison	**Remark**	**Jupyter Notebook**
Type of license and open source license	State all types of license	BSD License Apache License 2.0
Year founded	When is this tool being introduced?	2014
Founding company	Owner	non-profit organization called the Jupyter Project
License Pricing	Compare the prices if the license is used for development and business/commercialization	Free
Supported features	What features that it offers?	• Interactive computing • Multiple programming languages • Data visualization • Extensibility • Collaboration
Common applications	In what areas this tool is usually used?	• Data exploration and preprocessing

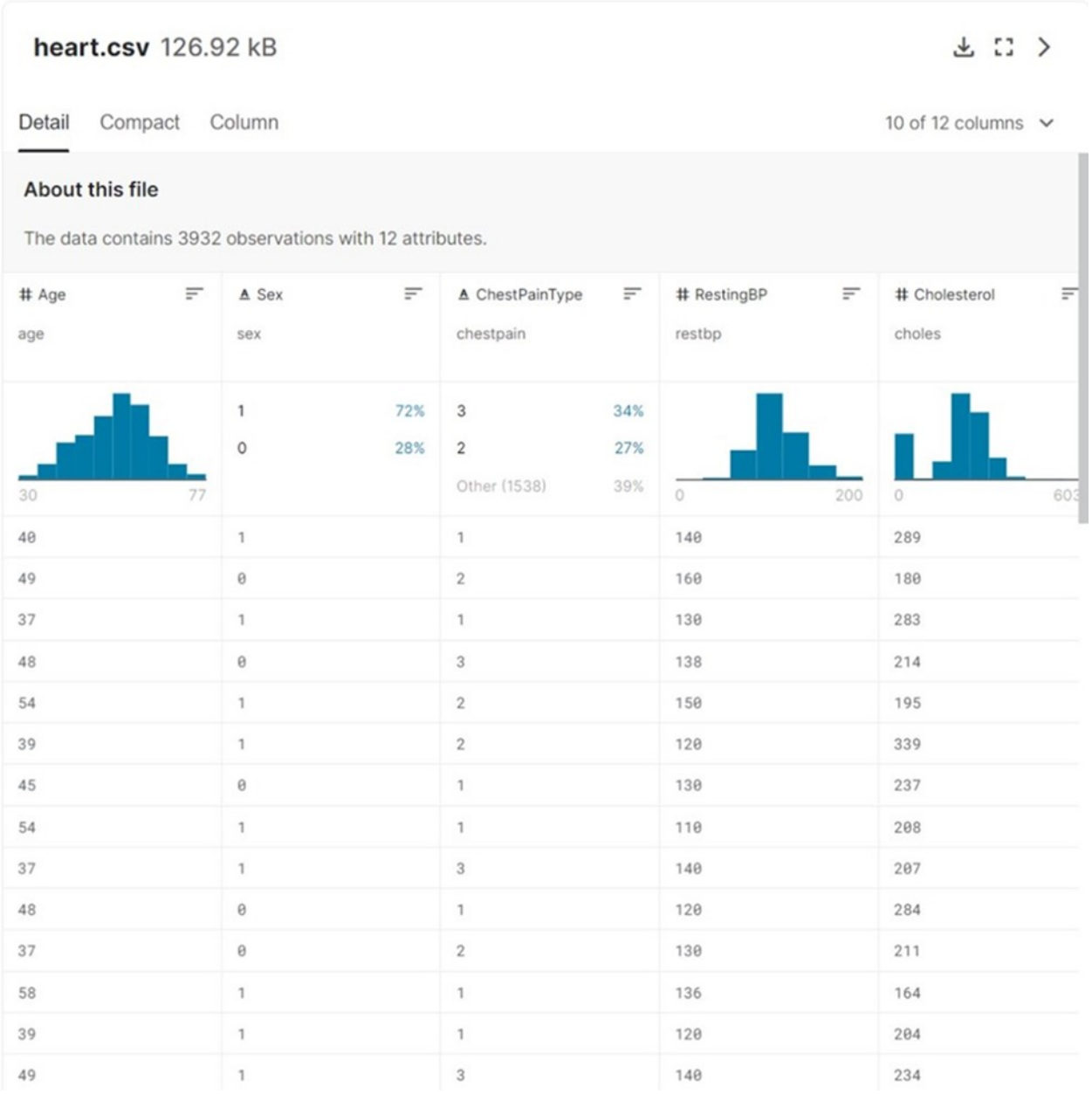
		• Model development and testing • Model visualization and interpretation • Model deployment
Customer support	How the customer support is given, e.g. proprietary, online community, etc.	• Online documentation • Community forums • GitHub repository • Third-party resources
Limitations	The drawbacks of the software	• Limited debugging capabilities • Steep learning curve

3.3. Justify why the selected tool is suitable

Tools	**Reason**
Jupyter Notebook	● **Interactive data exploration** : For data analysis, Jupyter Notebook offers an interactive and exploratory environment that makes it simple to visualise and modify data. For examining and comprehending the intricate correlations between various factors in heart disease data, this can be especially helpful. ● **Reproducibility** : Data cleansing, feature engineering, model training, and assessment are just a few of the steps of the data analysis process that may be documented using reproducible workflows created in Jupyter Notebook. For scientific research and clinical decision-making, it is crucial that the results are reliable and repeatable over time. ● **Collaboration** : Jupyter Notebook makes it simple to collaborate with coworkers, exchange code and results, and promote knowledge sharing by allowing for simple notebook and code sharing. ● **Visualization**: Jupyter Notebook makes it simple to produce educational and aesthetically pleasing plots and charts by supporting a range of data visualisation tools, including Matplotlib and Seaborn.

4. Methodology

4.1. Description of dataset



heart.csv 126.92 kB

Detail Compact Column 10 of 12 columns

About this file

The data contains 3932 observations with 12 attributes.

# Age	▲ Sex	▲ ChestPainType	# RestingBP	# Cholesterol
age	sex	chestpain	restbp	choles
30 77	1 72% 0 28%	3 34% 2 27% Other (1538) 39%	0 200	0 603
40	1	1	140	289
49	0	2	160	180
37	1	1	130	283
48	0	3	138	214
54	1	2	150	195
39	1	2	120	339
45	0	1	130	237
54	1	1	110	208
37	1	3	140	207
48	0	1	120	284
37	0	2	130	211
58	1	1	136	164
39	1	1	120	204
49	1	3	140	234

The heart.csv dataset was obtained through kaggle, This data is being used to build a heart disease prediction system. There are a total of 3932 rows and 12 columns in this dataset. The 12 columns are age, sex, chest pain, restbp, choles, fastbs, restecg, maxhr, exagina, oldpeak, stslope and target respectively.

Column's Name	Description
Age	Data under 28 to 77 years old
Sex	1: Male 0: Female
Chestpain	0: Typical Angina 1: Atypical Angina 2: Non-Aginal Pain 3: Asymptomatic
restbp	A range of up to 200
choles	A range of up to 600
fastbs	0: No 0: Yes
maxhr	A range from 60 to 200.
exagina	0: No 0: Yes
oldpeak	A range from 2.6 to 6.2
stslope	0: Upsloping = better heart rate with exercise(uncommon) 1: Flatsloping = minimal change(typical healthy heart) 2: Downsloping = signs of unhealthy heart
target	0: No Heart Disease 1: Heart Disease

4.2. Applications of the algorithm(s)

4.2.1 Data Representation

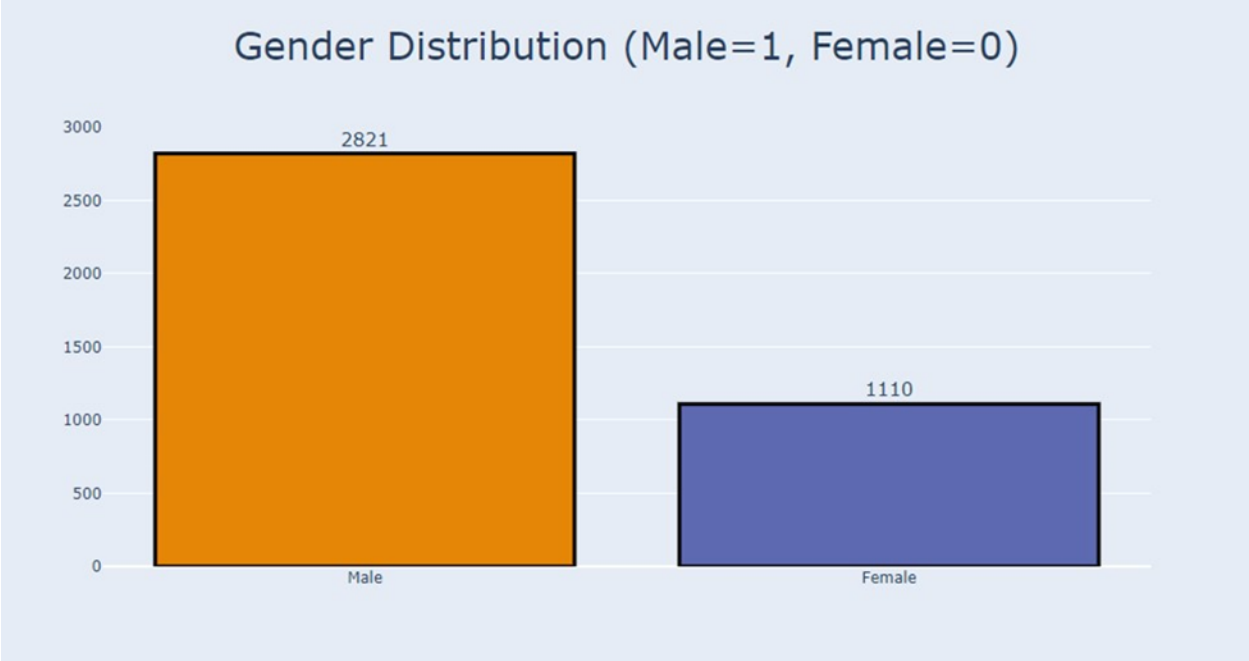


Figure 1: Bar Chart of the Gender Distribution

Figure above shows the bar chart of the gender distribution inside the dataset. The orange bar represents the number of males and the purple bar represents the number of females in the dataset.

The total number of males is 2821 and the total number of females is 1110.

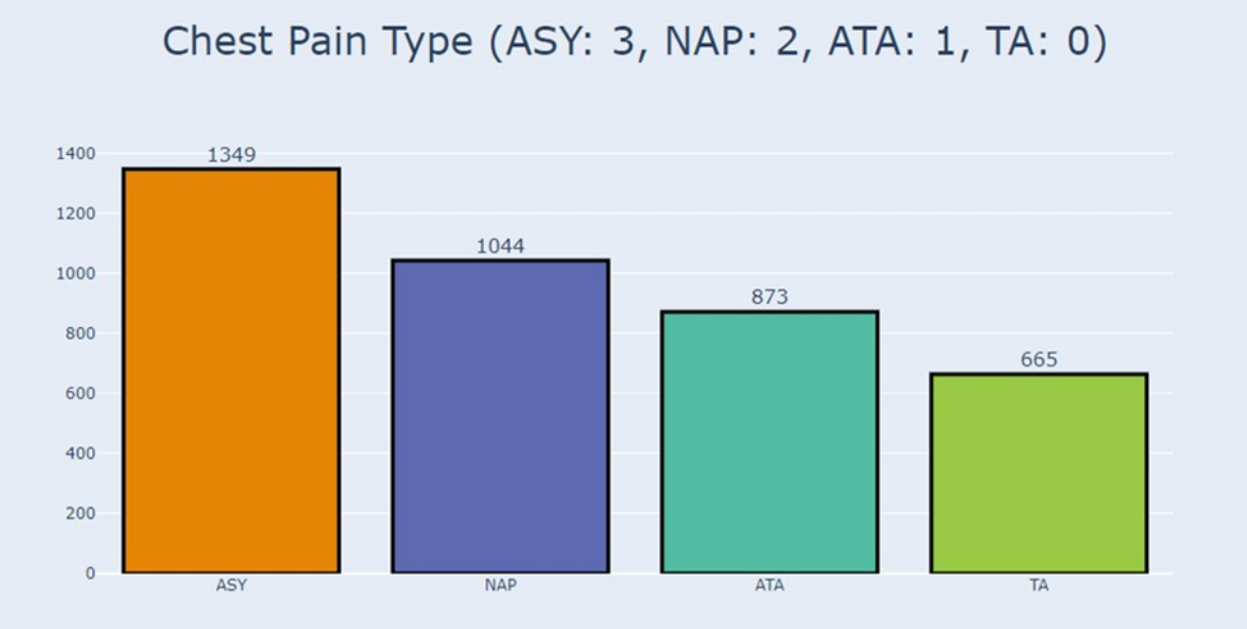


Figure 2: Bar Chart of the Chest Pain on different types

Above figure shows the bar chart of Chest Pain different types.

Orange Bar: Asymptomatic(ASY). Total = 1349

Purple Bar: Non-Anginal Pain(NAP). Total = 1044

Green Bar: Atypical Angina(ATA). Total = 873

LIght Green Bar: Typical Angina (TA). Total = 665

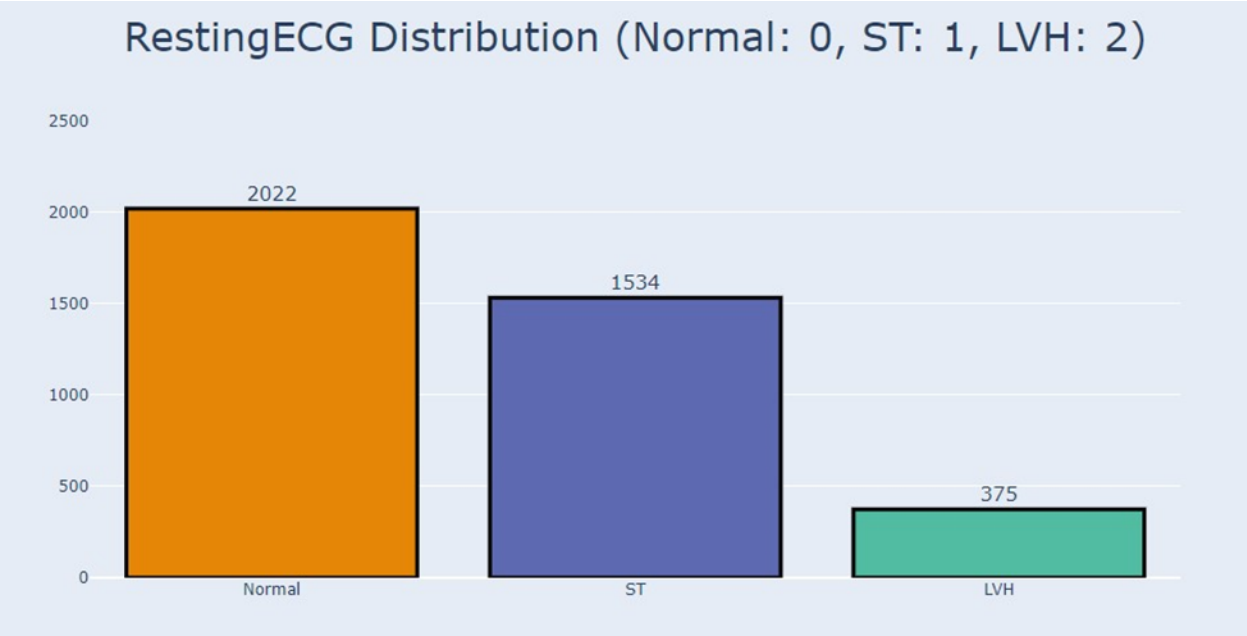


Figure 3: Bar Chart of the Resting ECG Distribution

Above figure shows the bar chart of the number of people who have the different types of resting electrocardiograms.

Orange Bar: Normal of resting electrocardiograms(Normal). Total = 2022

Purple Bar: ST-T Wave Abnormality(ST). Total = 1534

Green Bar: Left Ventricular Hypertrophy(LVH). Total = 375

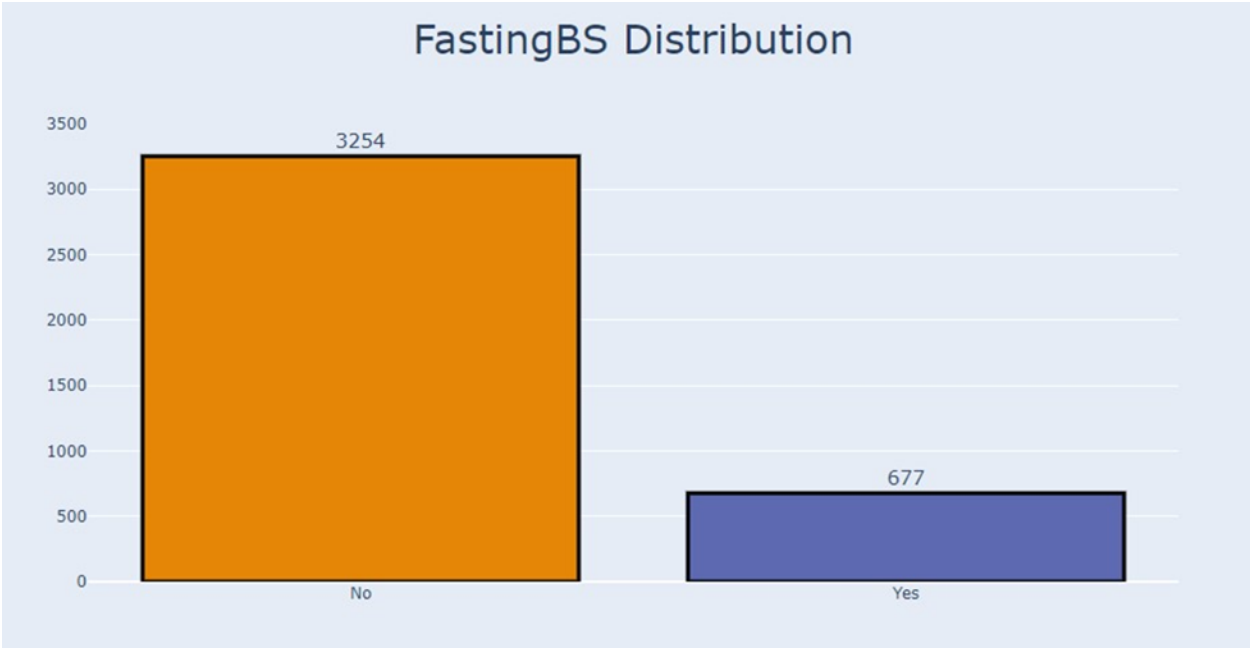


Figure 4: Bar Chart of the FastingBS Distribution

Above figure shows the bar chart of the fasting blood sugar distribution in the dataset.

Orange Bar: who have a blood sugar of less than 120 mg/dl. Total = 3254

Purple Bar: who have a blood sugar of more than 120 mg/dl. Total = 677

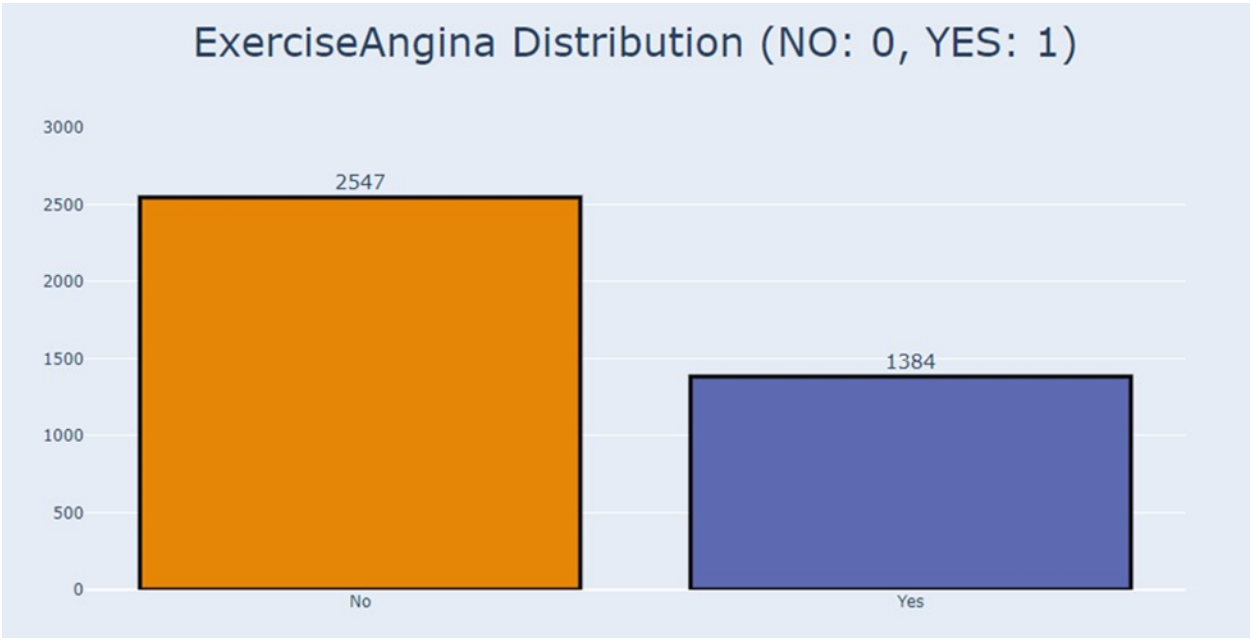


Figure 5: Bar Chart of the ExerciseAngina Distribution

Above figure shows the exercise induced angina distribution in the dataset.

Orange Bar: who do not have exercise induced angina. Total = 2574

Purple Bar: who have exercise induced angina. Total =1384

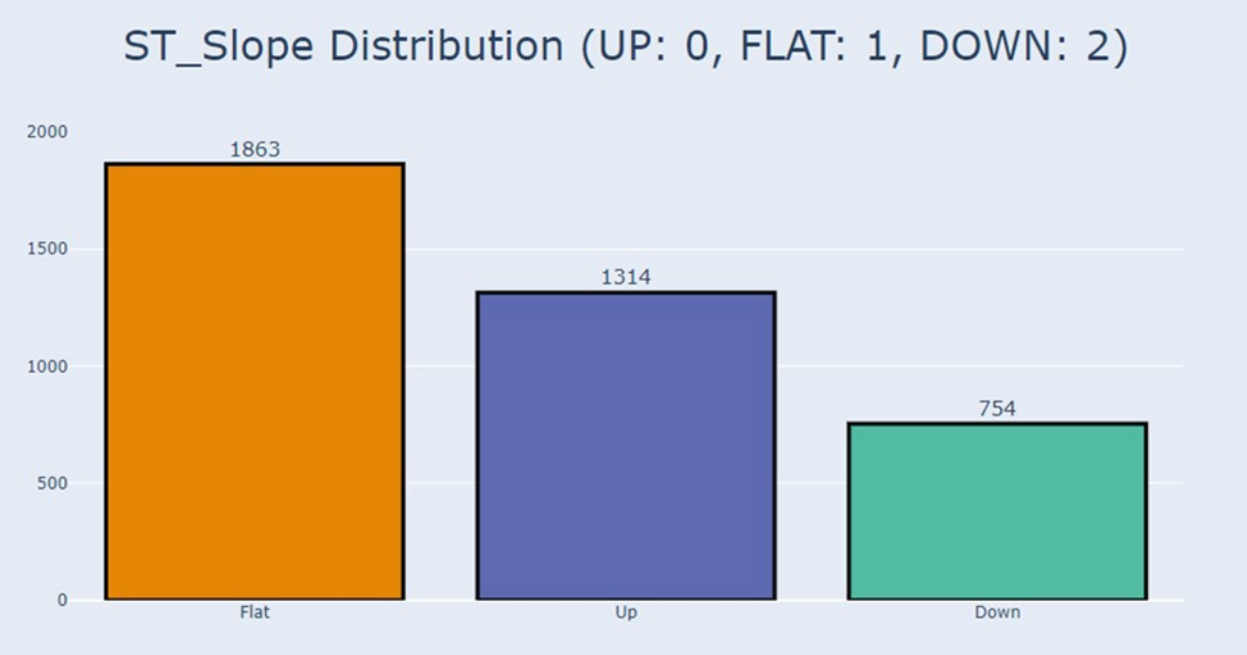


Figure 6: Bar Chart of the ST_Slop Distribution

Above figure shows the number of individuals who have different types of ST Slope in the dataset.

Orange Bar: who have a flat slope. Total = 1863

Purple Bar: who have an up slope. Total = 1314

Green Bar: who have a down slope. Total = 754

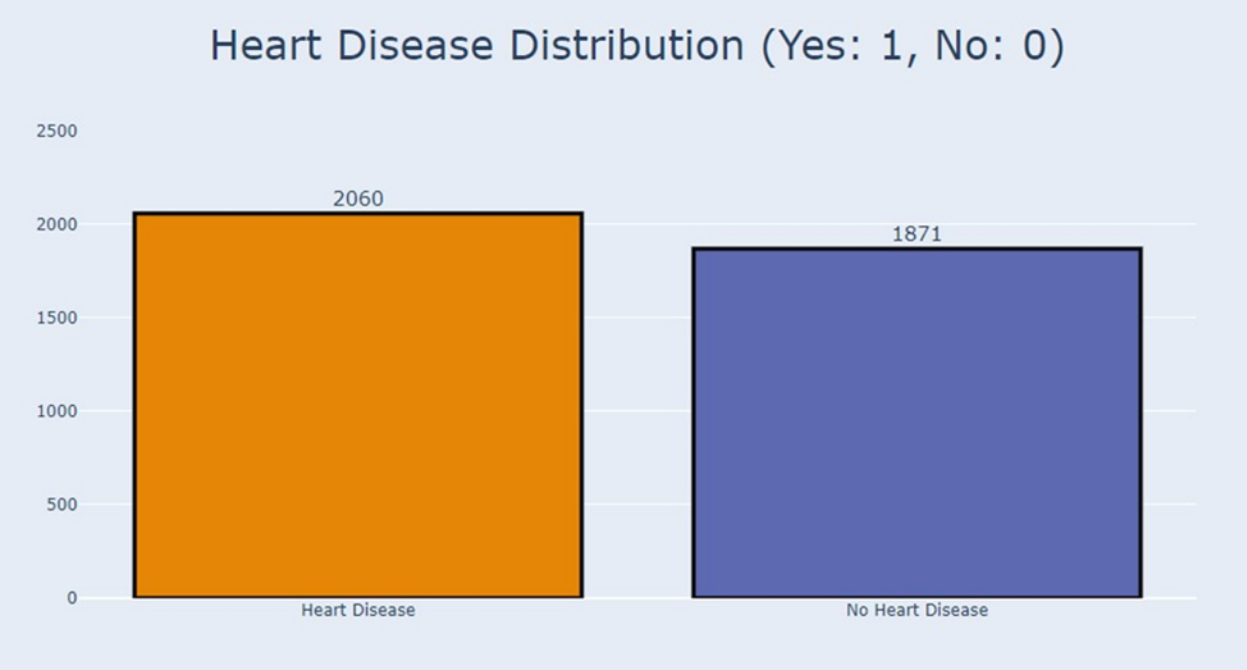


Figure 7: Bar Chart of the Heart Disease Distribution

Above figure shows the number of individuals who have different types of ST Slope in the dataset.

Orange Bar: who has heart disease. Total = 2060

Purple Bar: who does not have heart disease. Total = 1871

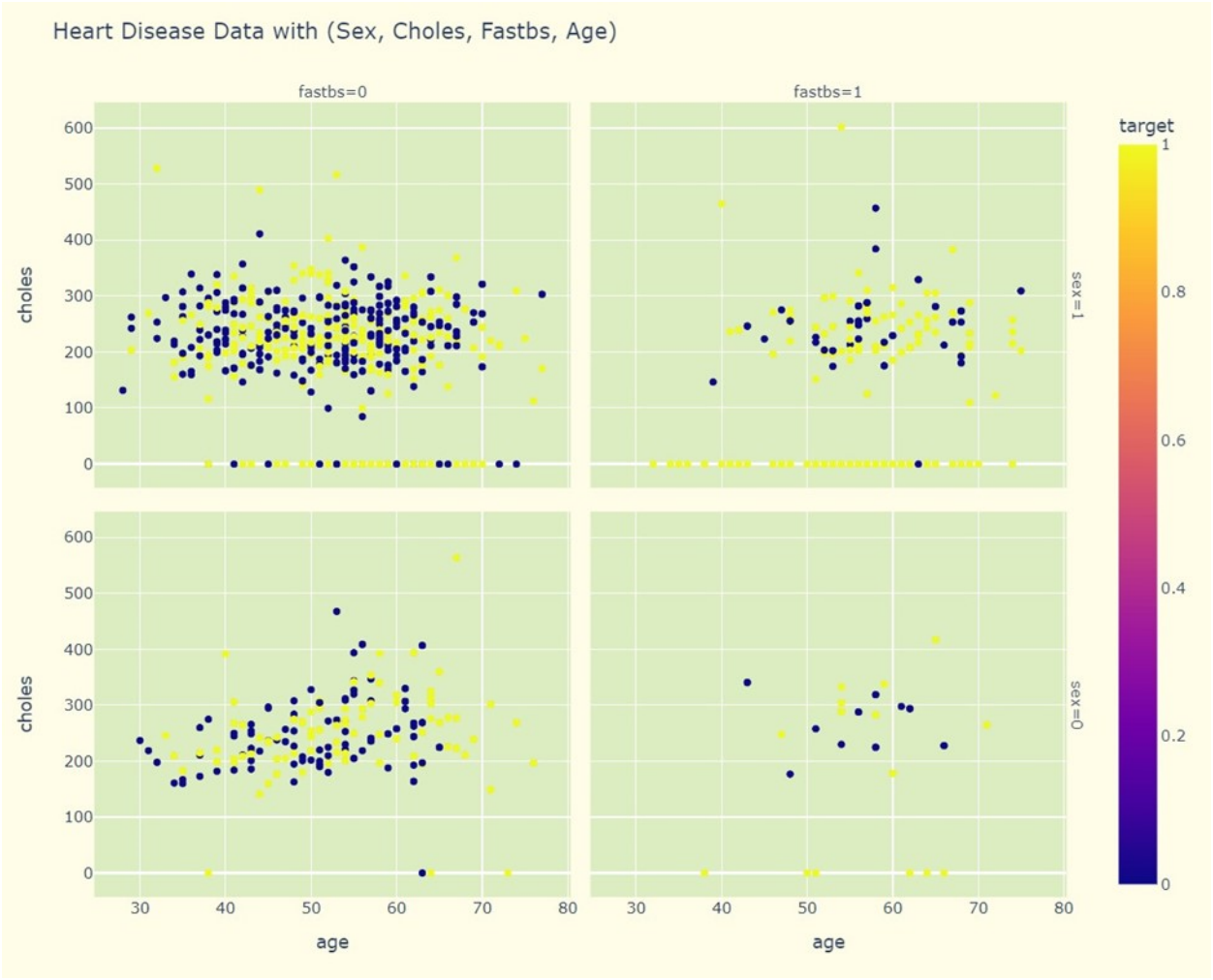


Figure 8: Scatter Plot Correlation Graph

Above figure shows the scatter plot graph which indicates the correlation between sex, choles, fastbs and age.

Blue Color Dots: who do not have heart disease

Yellow Color Dots: who have the heart disease

Top Left of the graph: who have a fasting blood sugar of less than or equal to 120 mg/dl. Top Right and Bottom Right of the graph: shows both sex with fasting blood sugar of more than 120 mg/dl.

Above graph shows that male who are in the age of 40 to 70 and have a cholesterol level between 150 to 300 mg/dl are more prone to heart disease.

The females who are in the age of 45 to 60 and cholesterol level between 150 to 300 mg/dl are more prone to heart disease.

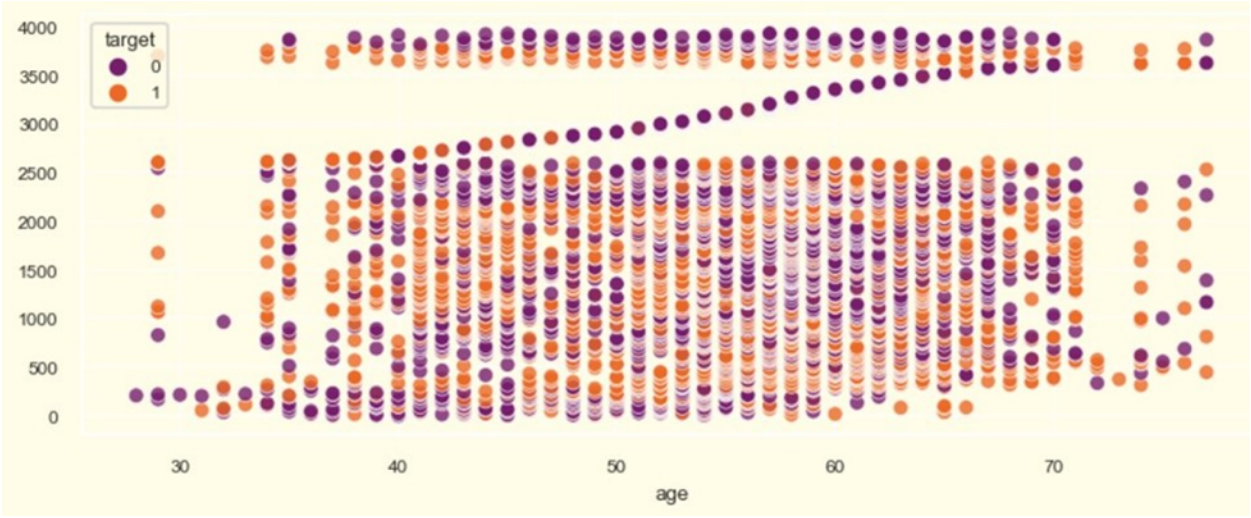


Figure 9: Scatter Plot of Age with target (heart disease)

Above figure shows the scatter plot of individuals whose age is between 28 to 77 and it shows which individual is more likely to have heart disease.

Those aged 30 to 70 are more prone to heart disease because the orange dots (1)representing those who have heart disease are frequently shown between age 30 to 70.

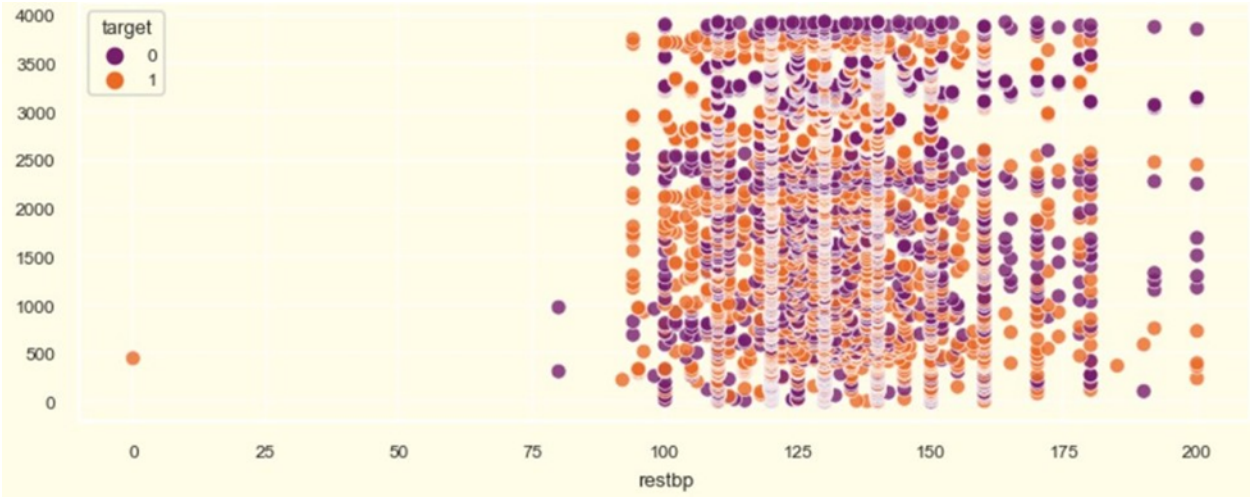


Figure 10: Scatter Plot of Resting Blood Pressure with target(heart disease)

Above figure shows the scatter plot of the resting blood pressure that determines whether an individual is prone to heart disease or not.

From the above figure, most of the individuals have a resting blood pressure between 100 to 180.

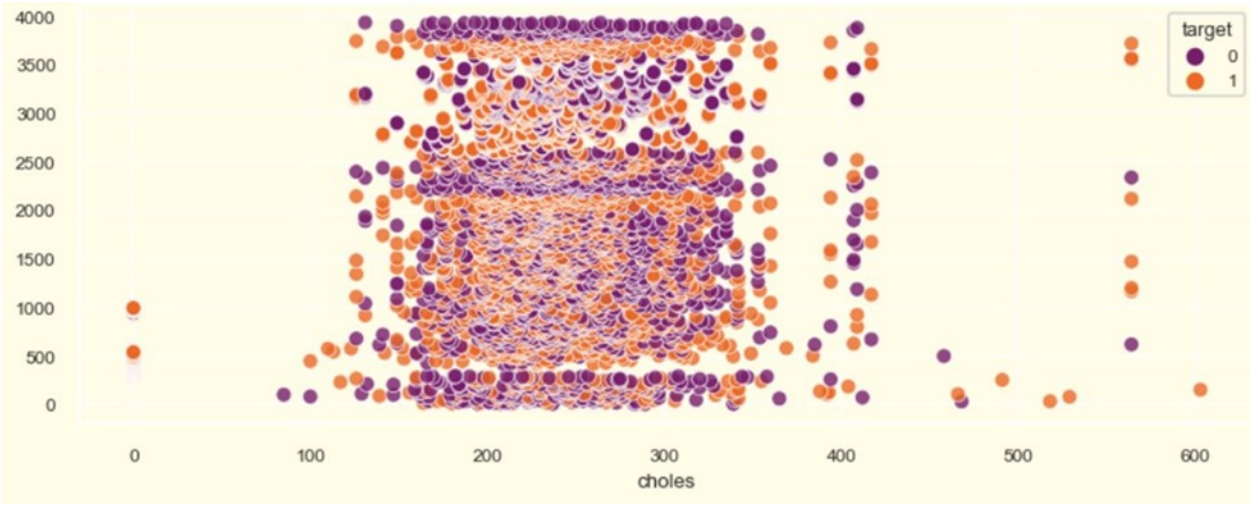


Figure 11: Scatter Plot of Cholesterol with target(heart disease)

Above figure shows the scatter plot of the cholesterol that determines whether an individual is prone to heart disease or not.

From the above figure, most of the individuals have a cholesterol between 150 to 350.

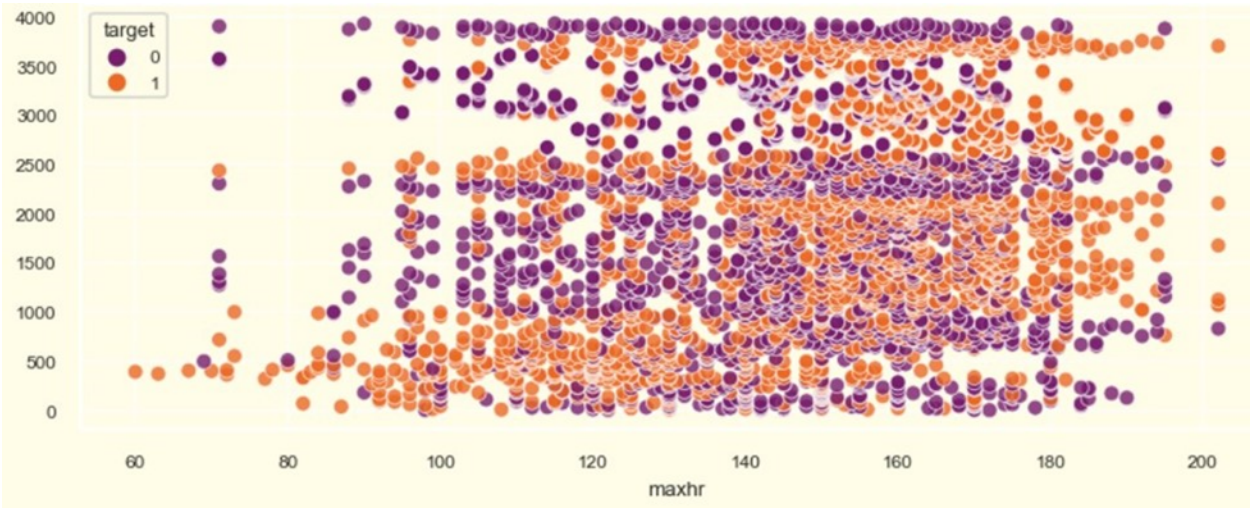


Figure 12: Scatter Plot of Maximum Heart Rate achieved with target(heart disease)

Above figure shows the scatter plot of the maximum heart rate achieved that determines whether an individual is prone to heart disease or not.

From the above figure, most of the individuals have a heart rate more than 80

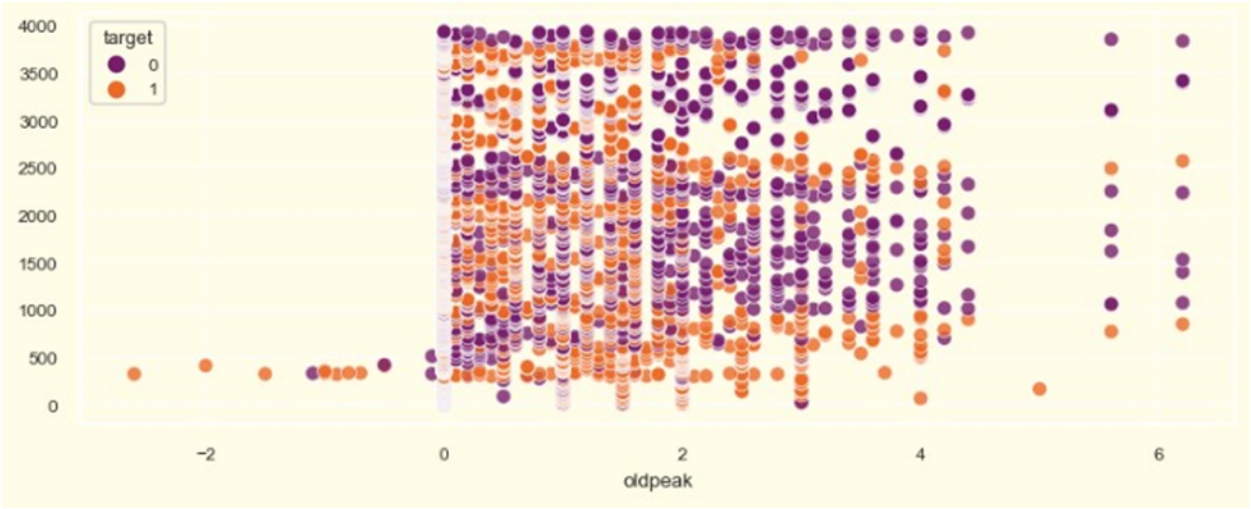


Figure 13: Scatter Plot of old peak with target(heart disease)

Above figure shows the scatter plot of the old peak with a target that determines whether an individual is prone to heart disease or not.

From the above figure, most of the old peaks are between 0 and 4.

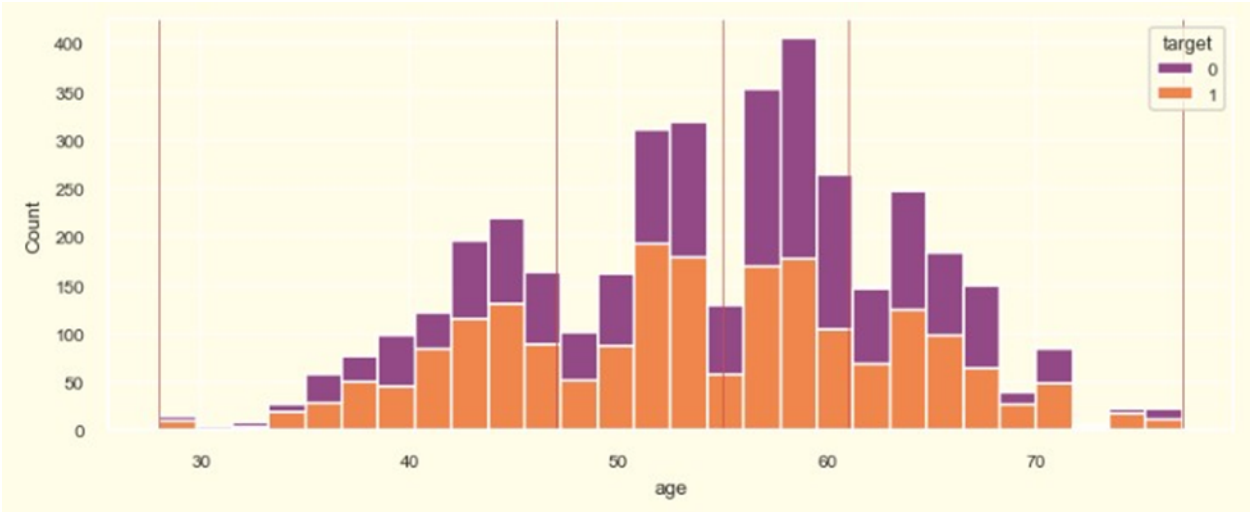


Figure 14: Histogram Plot of age with target(heart disease)

Above figure shows the histogram plot of the age with a target that determines whether an individual is prone to heart disease or not.

From the above figure, those who are aged 40 till aged 77 are at the risk of getting heart disease.

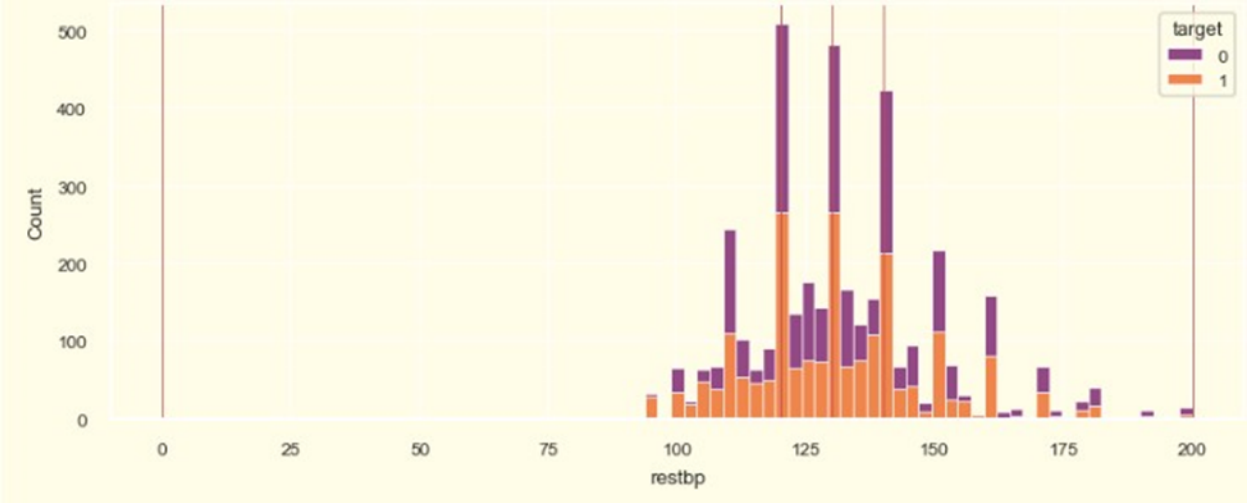


Figure 15: Histogram Plot of restbp with target(heart disease)

Above figure shows the histogram plot of the restbp with a target that determines whether an individual is prone to heart disease or not.

From the above figure, those who are have the resting blood pressure between 100 to 180 are at the risk of getting heart disease.

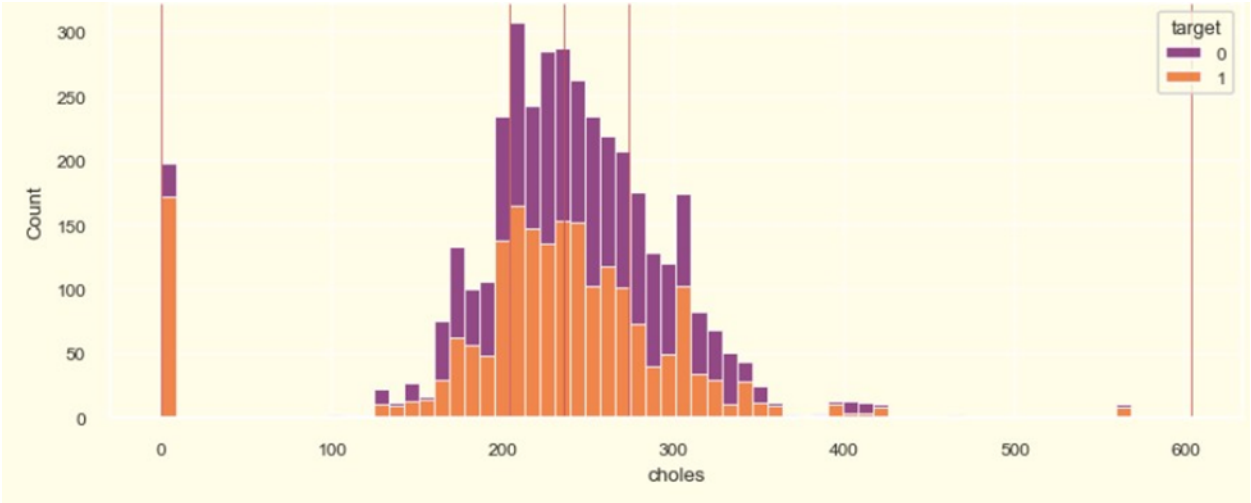


Figure 16: Histogram Plot of Cholesterol with target(heart disease)

Above figure shows the histogram plot of the cholesterol that determines whether an individual is prone to heart disease or not.

From the above figure, most of the individuals have a cholesterol between 150 to 350 s prone to heart disease.

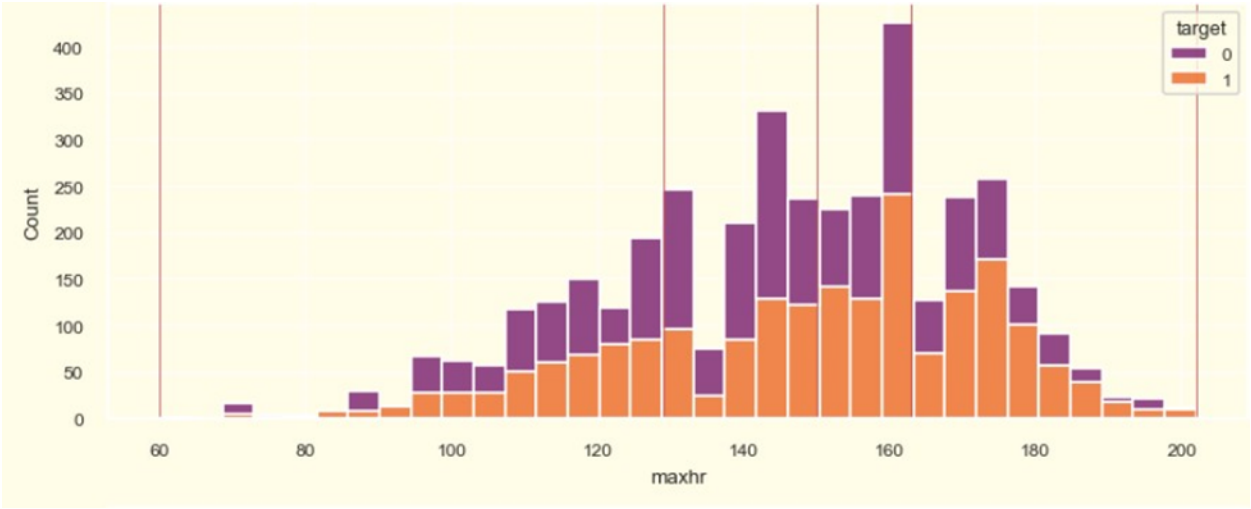


Figure 17: Histogram Plot of Maximum Heart Rate with target(heart disease)

Above figure shows the scatter plot of the maximum heart rate achieved that determines whether an individual is prone to heart disease or not.

From the above figure, most of the individuals have a heart rate more than 100 till 135 is prone to heart disease and start with 145 till 200 is prone to the heart disease.

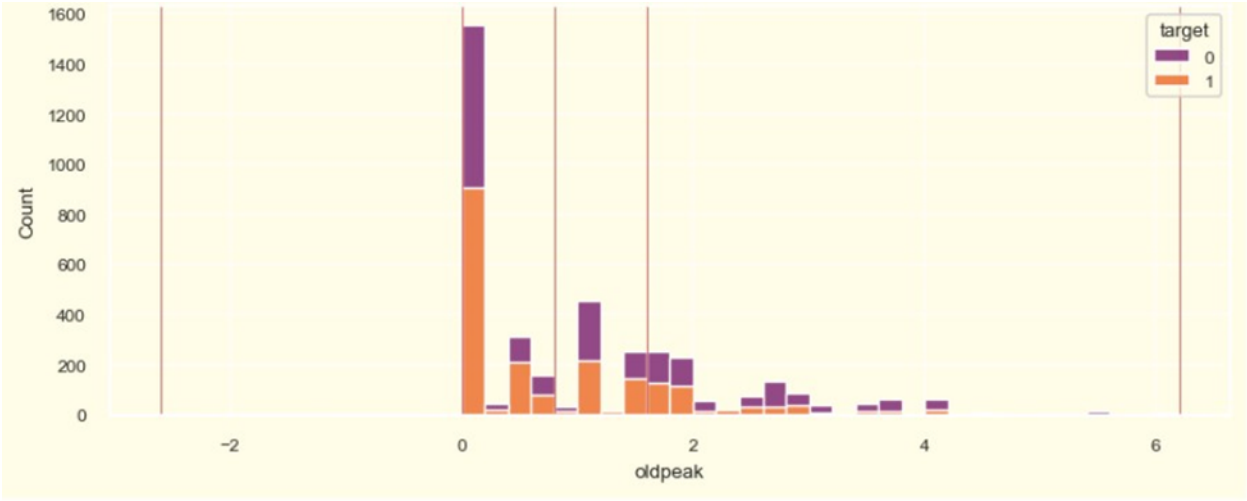


Figure 18: Histogram Plot of old peak with target(heart disease)

Above figure shows the scatter plot of the old peak with a target that determines whether an individual is prone to heart disease or not.

From the above figure, most of the old peaks are between 0 and 4 and the old peak at the 0 till 1 is prone to heart disease and starts with 1.7 to 2 also prone to heart disease.

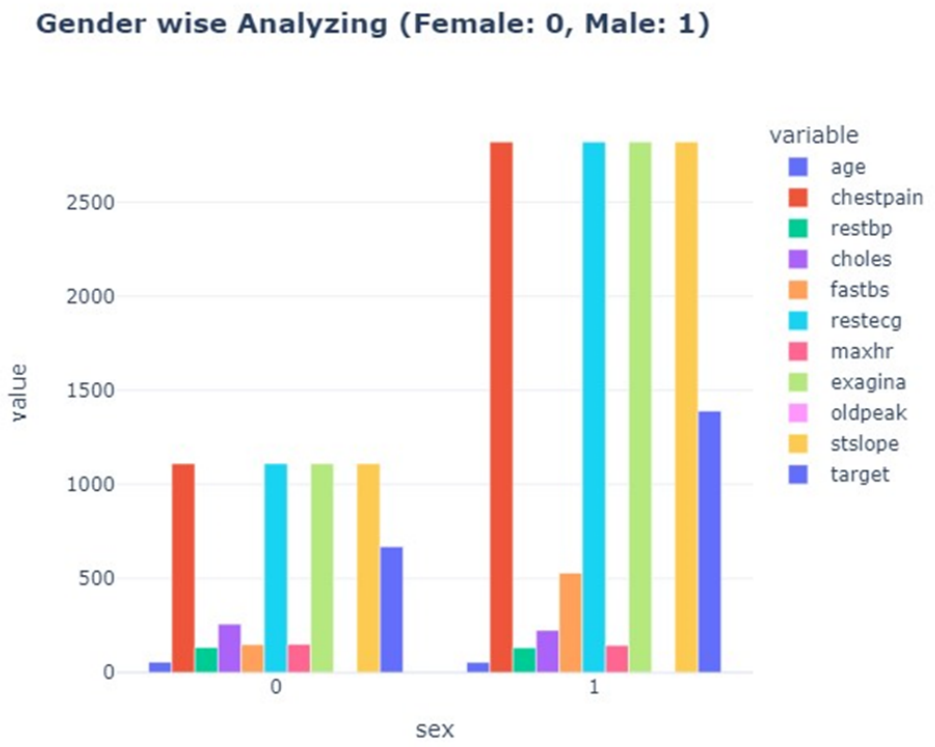


Figure 19: Bar Chart of showing the average gender wise analyzing with different variables

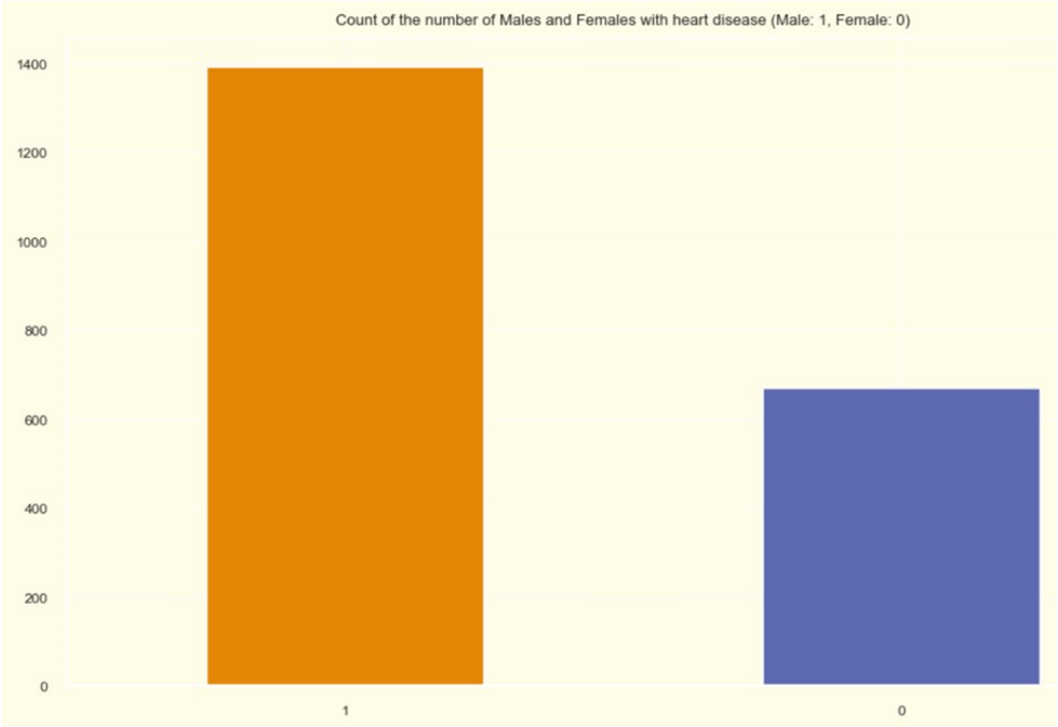


Figure 20: Bar Chart of Males and Females with Heart Disease

Above figure shows the bar chart of the genders who are prone to heart disease or not.

Orange Bar: Males who are prone to heart disease. Total = 1391

Purple Bar: Females who are prone to heart disease. Total = 669

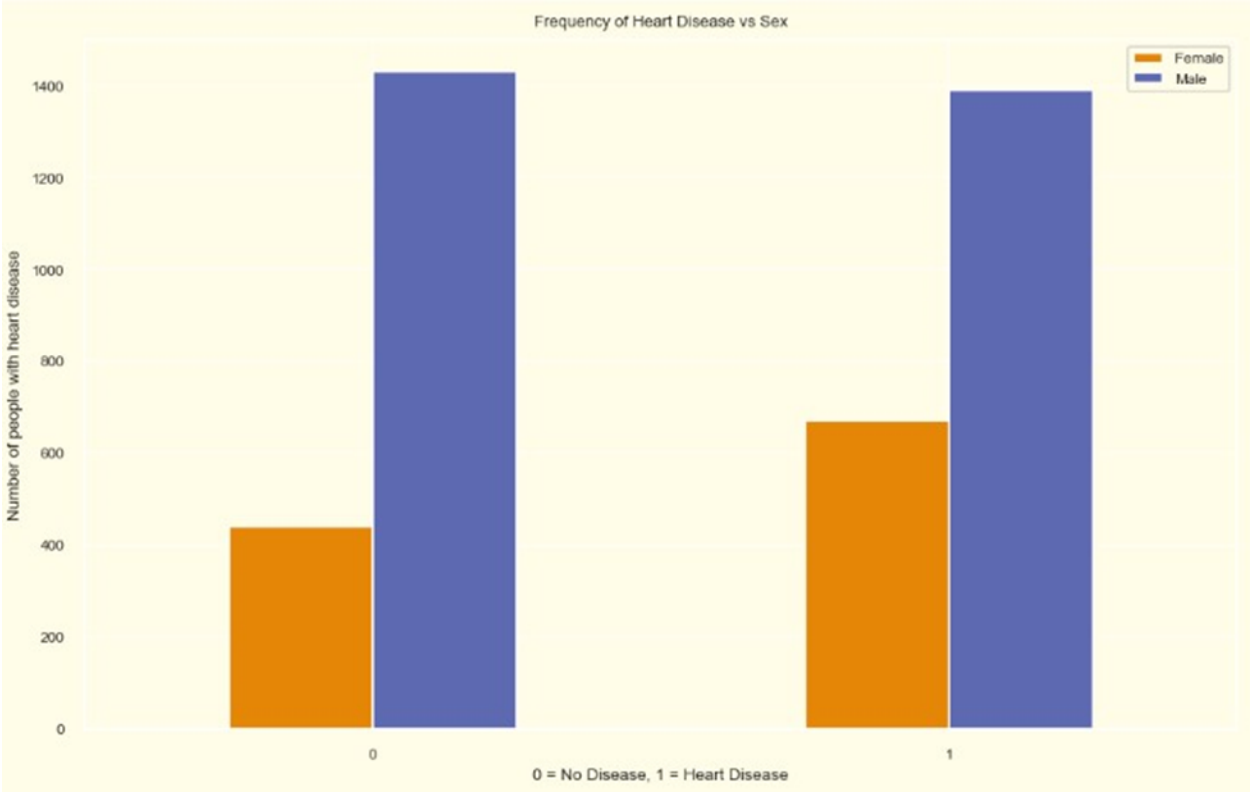


Figure 21: Bar Chart of Frequency of Heart Disease vs Sex

Above figure shows the bar chart of the Frequency of Heart Disease vs Sex.

Males have a higher risk of being prone to heart disease than Females.

4.2.2 Data Preprocessing

```
from sklearn.preprocessing import MinMaxScaler
scal=MinMaxScaler()
feat=['age', 'sex', 'chestpain', 'restbp', 'choles', 'fastbs', 'restecg', 'maxhr', 'exagina', 'oldpeak', 'stslope']
df[feat] = scal.fit_transform(df[feat])
df.head()
```

Code for performing the MinMaxScaler.

	age	sex	chestpain	restbp	choles	fastbs	restecg	maxhr	exagina	oldpeak	stslope	target
0	0.244898	1.0	0.333333	0.70	0.479270	0.0	0.0	0.788732	0.0	0.295455	0.0	0
1	0.183673	1.0	0.333333	0.65	0.469320	0.0	0.5	0.267606	0.0	0.295455	0.0	0
2	0.530612	1.0	0.666667	0.75	0.323383	0.0	0.0	0.436620	0.0	0.295455	0.0	0
3	0.224490	1.0	0.666667	0.60	0.562189	0.0	0.0	0.774648	0.0	0.295455	0.0	0
4	0.346939	0.0	0.333333	0.65	0.393035	0.0	0.0	0.774648	0.0	0.295455	0.0	0

Display the output after performing MinMaxScaler.

Transforming all our features by scaling each of the features to a given range using the MInMaxScaler. For instance, chestpain can be categorized into 4 different scales which range from 0 1 2 3. In the first row 0, 0.33333 can be seen outputted after it is being processed through the minmaxscaler, this is calculated through the formula X-X(min) / X(max) – X(min). This formula is used throughout the whole dataset as shown in the picture above.

```
X=df.drop("target",axis=1).values
Y=df.target.values
```

Code to perform drop target column

Target column is being dropped before the dataset is written into the model.
Column from age to stslope is being written into variable X
Column target is being written into variable Y

```
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2,random_state=42)
```

Code to split data into train and test data

The data is split into train and test data and used 80/20 train/test split.
Train size is 80%
Test size is 20%

Source Code of Stacking CV Model (SCV)

```
from mlxtend.classifier import StackingCVClassifier
#Stack the the model
scv=StackingCVClassifier(classifiers=[Knn_clf,RF_clf],meta_classifier= Knn_clf)
scv.fit(X_train,Y_train)
scv_score=scv.score(X_test,Y_test)
scv_Y_pred=scv.predict(X_test)
evaluation(Y_test,scv_Y_pred)
```

```
{'accuracy': 0.978, 'recall': 0.983, 'F1 score': 0.98}
```

Above shows the source code of random forest. I have imported the SCV from mlxtend library. Then, define the name of the model known as scv. In order to stack the 2 algorithms together, we would have to include the two models that we have identified which are K-Nearest Neighbor (KNN) and Random Forest (RF). After stacking the models together, the model will be trained by fitting the training set of X_train and Y_train. The test data (X_test, Y_test) will be used to predict the score of the model. The accuracy score of the SCV model is 0.978.

4.2.3 Classification Method

Our team has a total of 3 different types of algorithm which are Random Forest, K-Nearest Neighbor(KNN) and Support Vector Machine(SVM). After we have identified all the accuracy scores of all the algorithms, our team will try to stack up 2 algorithms together to increase the accuracy of the algorithms.

Support Vector Machine Classifier Model

After preprocessing and splitting the data into training and testing sets, the training set of data is being fit to process the machining learning model. Then, the testing set of data will be used to estimate the model in order to get the accuracy of the result, The result of the testing set will be used for comparison.

The Support Vector Machine classifier was used to test the accuracy of the testing set of data.

I choose support vector machines because of their capacity to handle large amounts of data. As there may be many features or variables to take into account in medical applications, the ability of SVMs to handle high-dimensional datasets is crucial. Additionally, the classification is non-linear. SVMs can describe complicated connections between variables that may not be captured by linear models since they are capable of non-linear classification. SVMs are also less susceptible to outliers than other machine learning algorithms, which is crucial when working with medical data that may contain outliers as a result of measurement mistakes or other issues.

4.2.4 Comparing the result of different classification models

	Model	Accuracy
0	SCV	97.839898
1	Random Forest	97.331639
2	SVM	81.194409
3	KNN	93.646760

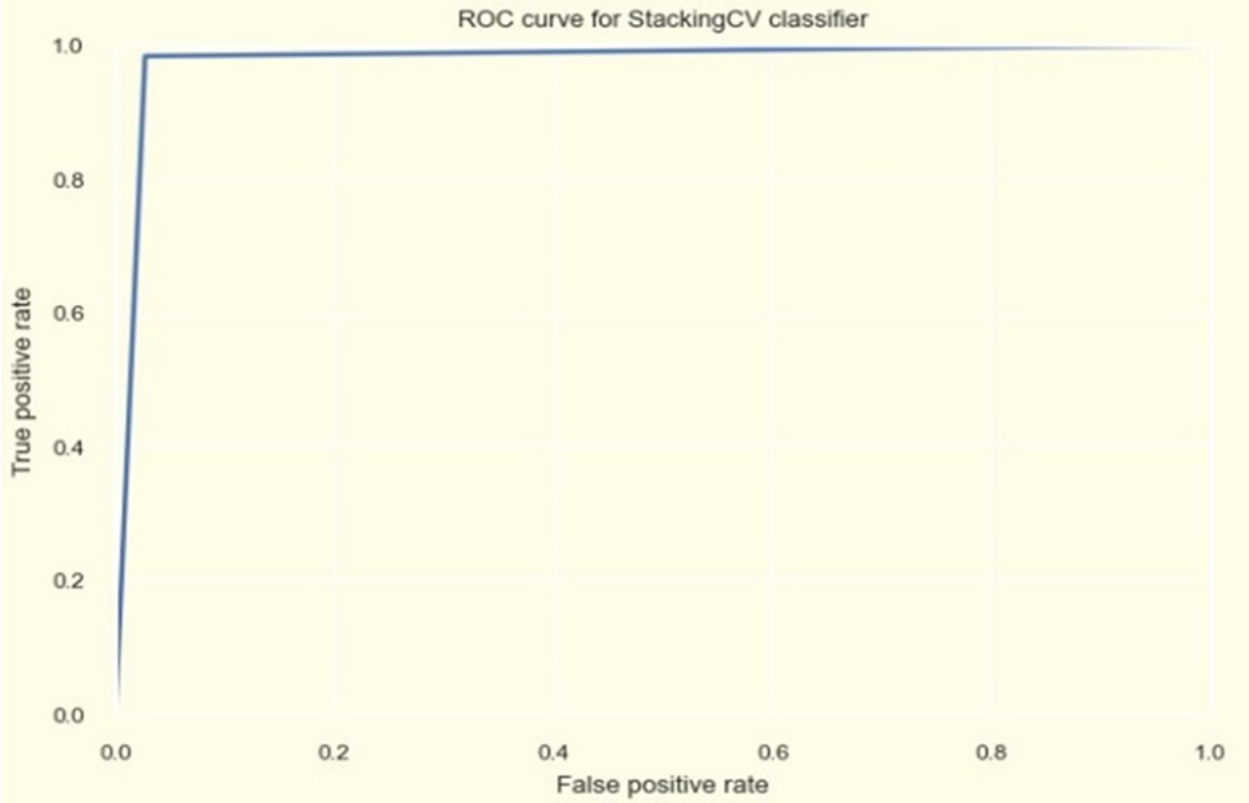
Accuracy result of different algorithms and Stacking CV(SCV) Model

The result of stacking up 2 algorithms is 97.83% which is higher than the Random Forest (RF), K-Nearest-Neighbours (KNN) and (Support Vector Machine (SVM) algorithms. The ranking of the accuracy of the model will be Stacking CV Classification which is 97.83%, followed by Random Forest which is 97.33% and the K-Nearest Neighbor will be the last which is 93.64%.then SVM which is 81.19%. In short, the SCV is the most accurate model to be used for the heart disease prediction application and the Random Forest can be said is the most accurate algorithm compared to other algorithms such as SVM and KNN that are being used in the application.

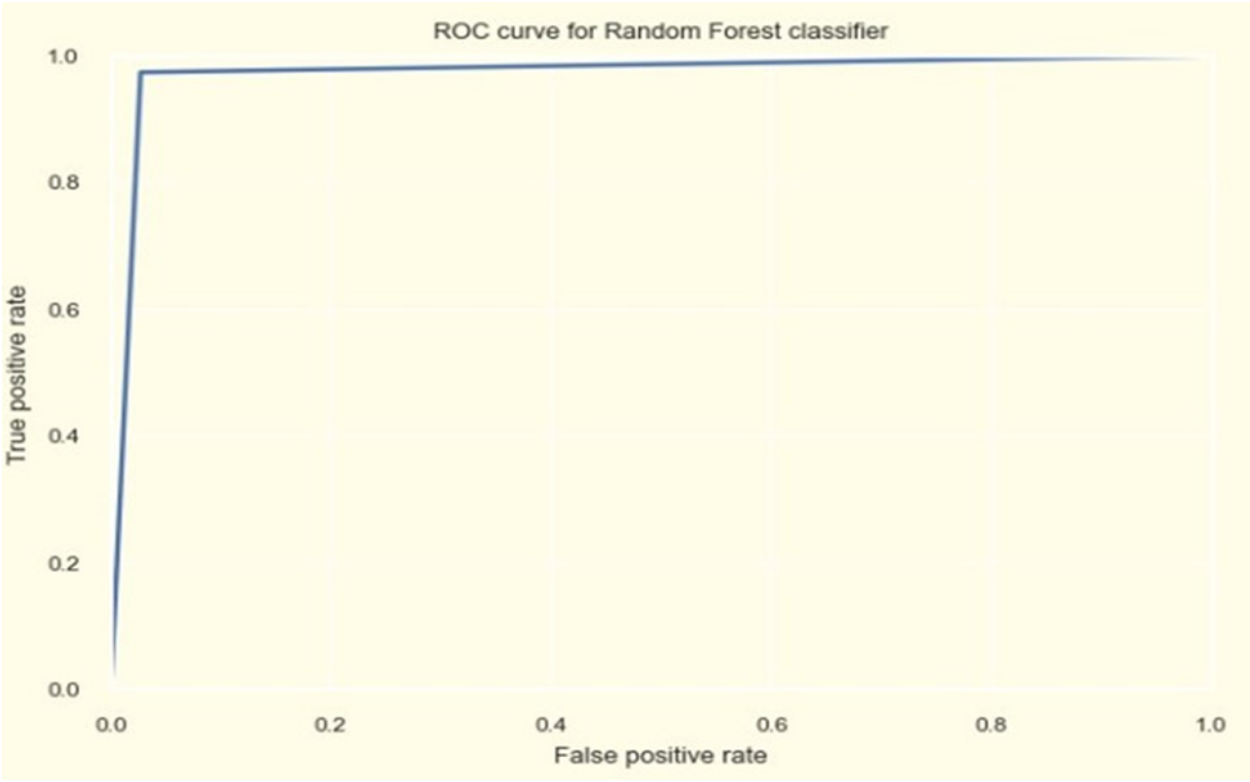
Roc Curve

Roc Curve is used to summarize the trade off between the true positive rate and false positive rate for the predictive model using different probability thresholds. All the roc curve shows below can be seen that all the models are slanting towards the y-axis. The accuracy for the Roc Curve of all the models are followed by 97.81% and 97.33%, 93.59%, 81.19%. Hence, stacking cv will be the highest accuracy score and if compared to algorithms Random Forest will be the highest compared to KNN and SVM.

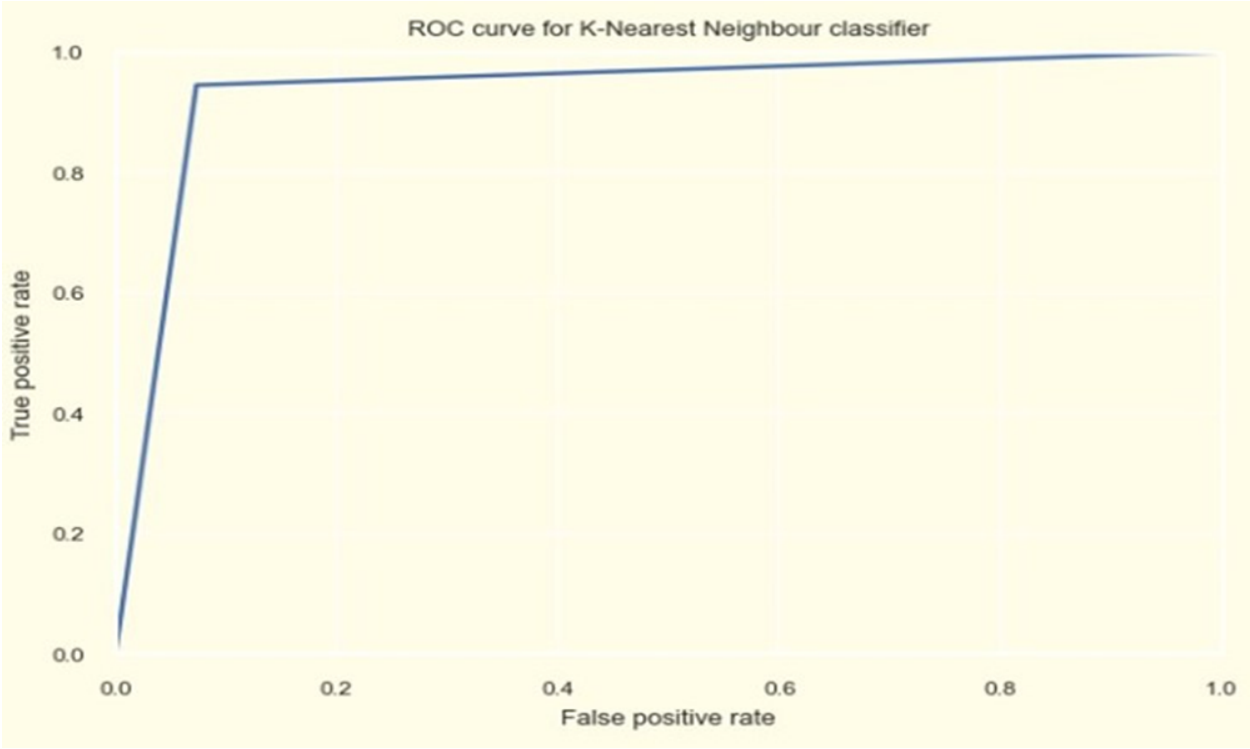
Stacking CV(SCV)



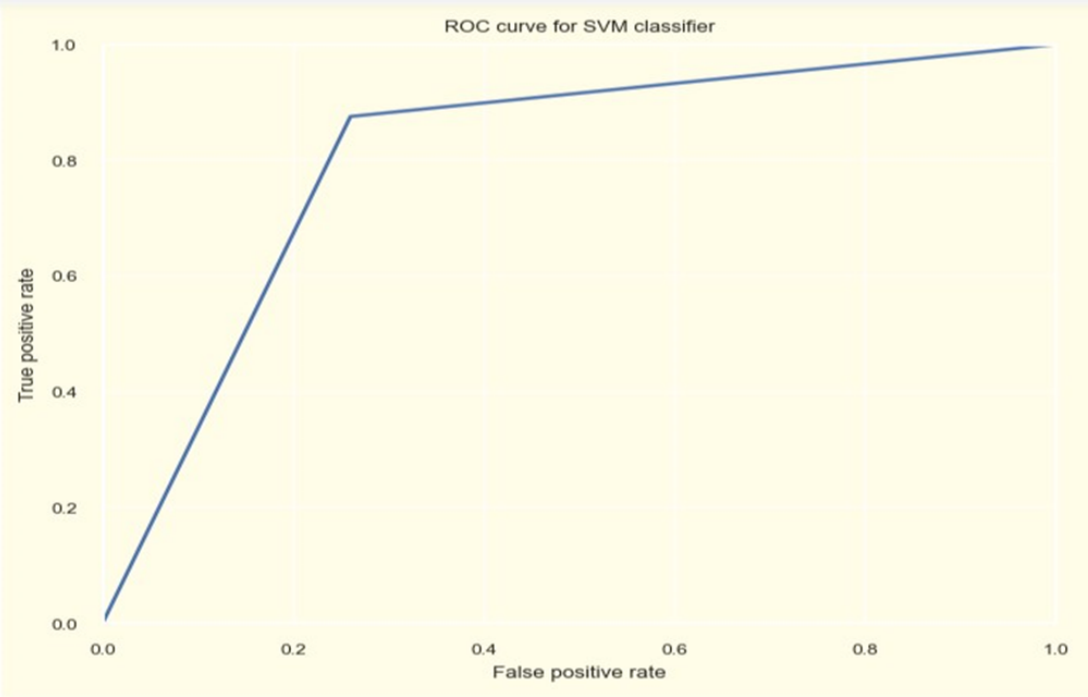
Random Forest



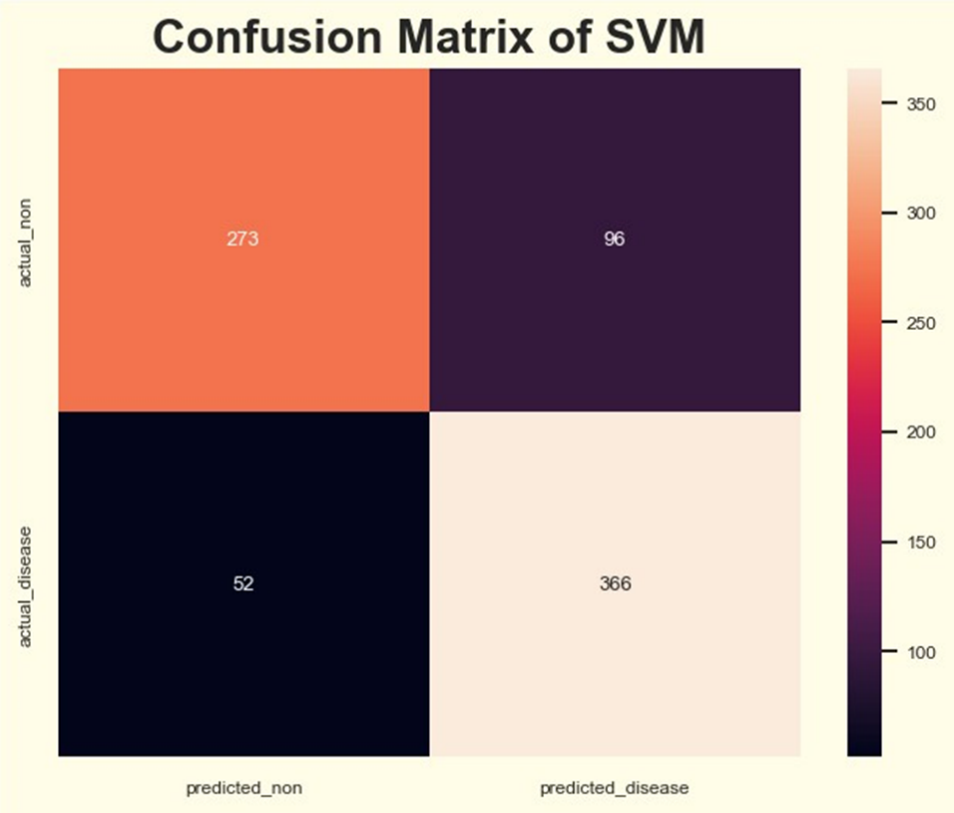
K-Nearest Neighbors(KNN)



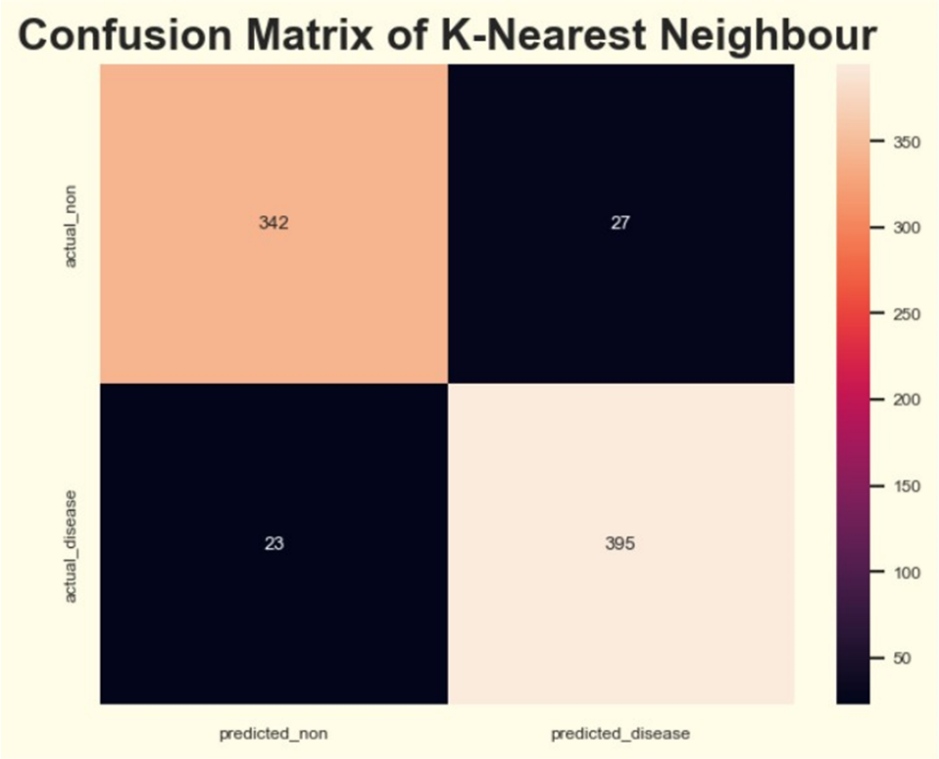
Support Vector Machine(SVM)



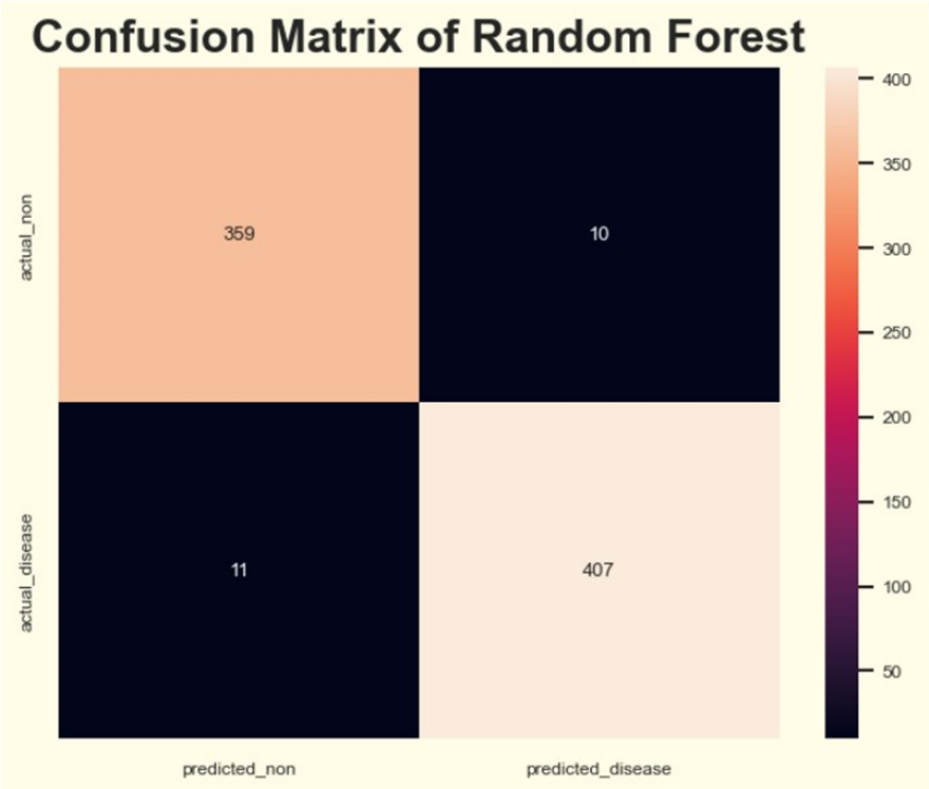
Confusion Matrix Support Vector Machine (SVM)



K-Nearest Neighbors (KNN)



Random Forest (RF)



Stacking SV (SVC)

Confusion matrix is used to help the summarization on how each model performed based on the testing data. Below are the descriptions of the confusion matrix.

	FP = False Positive FN = False Negative		
Actual	**actual_non**	TP: An individual that is not having heart disease and correctly identified by the algorithm.	FN: An individual that is not having heart disease but the algorithm said is diagnosed with heart disease.
	actual_disease	FP: An individual that is diagnosed heart disease but the algorithm said not having heart disease.	TN: An individual that is diagnosed with heart disease and correctly identified by the algorithm.
		predicted_non	**predicted_disease**
	Predicted		

The TP and TN indicate how many times the algorithm has been correctly classified. According to the confusion matrix above, we can see that the number of correctly classified by Support Vector Machine(SVM) is 273+366= 642, K-Nearest Neighbor(KNN) is 342+395=737, Random Forest(RF) is 359+407=766 and Stacking CV (SCV) is 359+411=770. With these numbers, we can conclude that the StackingCV classifier is the most suitable classification model to be used for our heart disease prediction application and Random Forest will be the second choice.

Classification Report: Support Vector Machine(SVM), K-Nearest Neighbor(KNN), Random Forest(RF) and Stacking CV (SCV)

Suppot Vector Machine Classification Report

	precision	recall	f1-score	support
0	0.84	0.74	0.79	369
1	0.79	0.88	0.83	418
accuracy			0.81	787
macro avg	0.82	0.81	0.81	787
weighted avg	0.81	0.81	0.81	787

K-Nearest Neighbour Classification Report

	precision	recall	f1-score	support
0	0.94	0.93	0.93	369
1	0.94	0.94	0.94	418
accuracy			0.94	787
macro avg	0.94	0.94	0.94	787
weighted avg	0.94	0.94	0.94	787

Random Forest Classification Report

	precision	recall	f1-score	support
0	0.97	0.97	0.97	369
1	0.98	0.97	0.97	418
accuracy			0.97	787
macro avg	0.97	0.97	0.97	787
weighted avg	0.97	0.97	0.97	787

Stacking CV Classification Report

	precision	recall	f1-score	support
0	0.98	0.97	0.98	369
1	0.98	0.98	0.98	418
accuracy			0.98	787
macro avg	0.98	0.98	0.98	787
weighted avg	0.98	0.98	0.98	787

Precision

The proportion of properly predicted positive observations to the total anticipated positive observations.

Recall

The ratio of properly predicted observations to the total number of observations in the actual class.

F1 Score

Showing us the weighted average of Precision and Recall.

The formula for precision, recall and F1 score is shown below:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

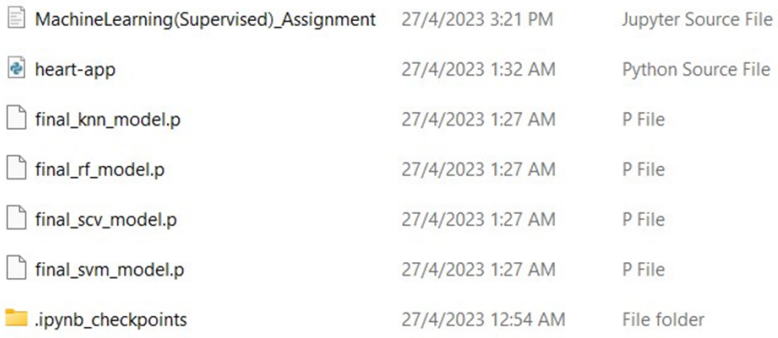
Support Machine Vector(SVM) has the worst precision, recall and F1 score and K-Nearest Neighbor (KNN) will be the second last even though it is higher than the SVM. Therefore, SVM and KNN will be excluded from being used in our heart disease prediction application . In contrast, both models of Random Forest (RF) and Stacking CV (SCV) have a high value when it comes to the precision, recall and F1 value. The SCV model has a slight advantage when compared to RF with a slight increase by 0.01 in terms of precision, recall and F1 value. With this, it is able to conclude that the SCV model has the highest accuracy out of all the models therefore it is the most suitable model to be used for our heart disease prediction application and Random Forest will be the second choice.

4.2.5 User Interface of Heart Disease Prediction

System Source Code of Writing pkl file:

```
import pickle as pkl

#Save Model

# StackingCV
pkl.dump(scv,open("final_scv_model.p","wb"))

#Random Forest
pkl.dump(RF_clf, open("final_rf_model.p","wb"))

#SVM
pkl.dump(Knn_clf,open("final_knn_model.p","wb"))

#KNN
pkl.dump(svm_model,open("final_svm_model.p","wb"))
```

In this step we will store all the models that will be displayed in the User Interface (UI). All the models will be included which are the StackingCV classifier model, the Random Forest classifier model Support Vector Machine Model(SVM) and K-Nearest-Neighbor (KNN) to make the comparison easily. The models are being stored in a pkl format which enables us to load the model into the heart-app.py. After storing the model into pkl file:

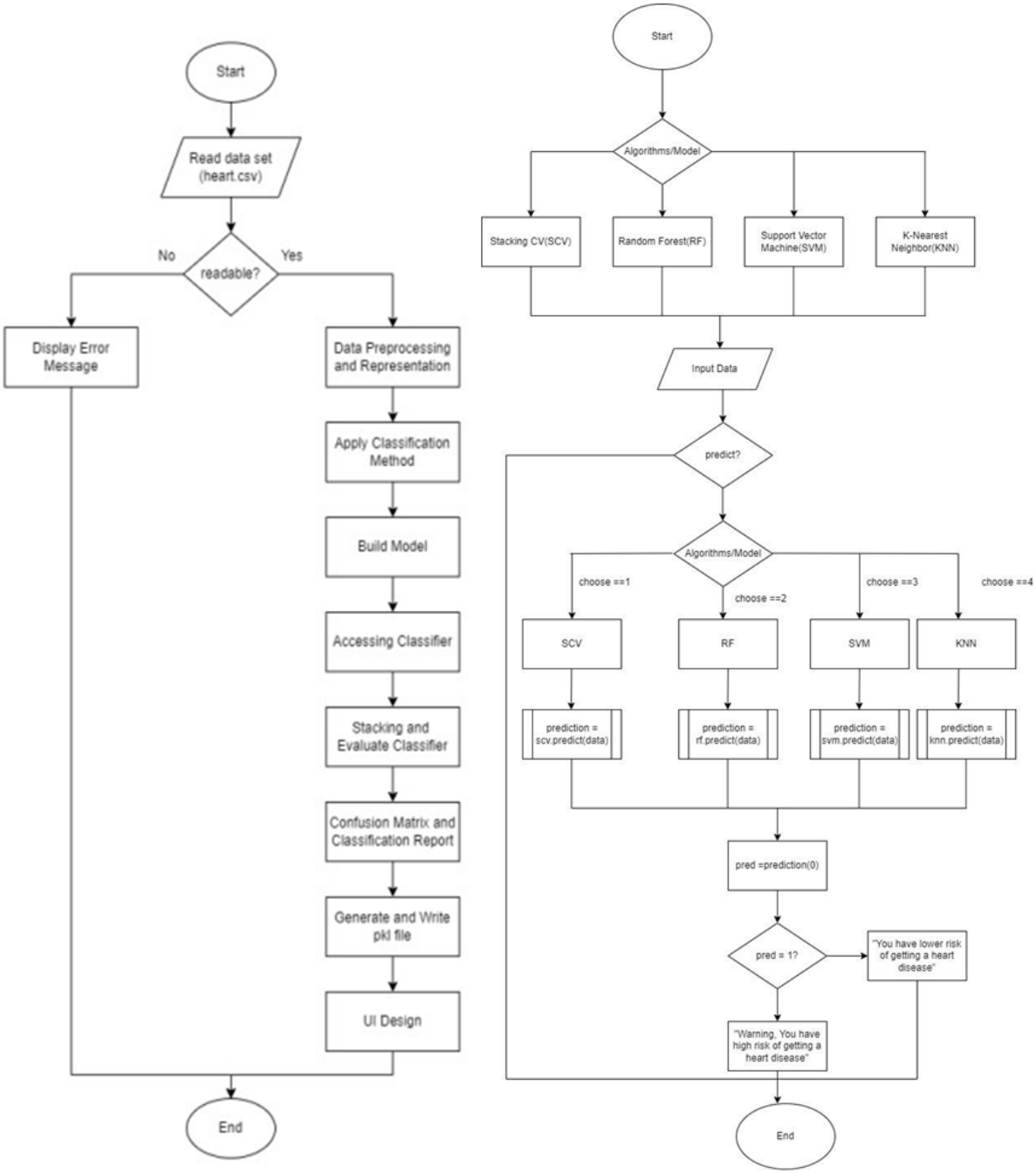


The Random Forest Classifier Model pkl file named as final_rf_model.p ,StackingCV Classifier Model pkl file named as final_scv_model.p, K-Nearest Neighbor (KNN) pkl file named as final_knn_model.p and Support Vector Machine (SVM) pkl file names as final_svm_model.p.

User Interface of Heart Disease Prediction Application:

In the heart disease prediction application, it allows users to choose which algorithm to use for predicting whether they are prone to heart diseases or not.

4.3. System flowchart/activity diagram



4.4. Proposed test plan/hypothesis

Step 1: Select the data from the dataset heart.csv that I chose to input and test in the project.

S	Age	Sex	Chest Pain	RestBp	Choles	FastBs	RestEeg	MaxHr	Exagina	Oldpeak	Stslope
S1	58	Male	Atypical Angina	136	164	No	ST-T Wave abnormality	99	Yes	2.00	Flatsloping
S2	70	Male	Asymptomatic	170	192	No	ST-T Wave abnormality	129	Yes	3.00	Downsloping
S3	52	Male	Atypical Angina	140	100	No	Nothing to note	138	Yes	0.00	Upsloping
S4	42	Male	Non-Anginal Pain	160	147	No	Nothing to note	146	No	0.00	Upsloping

S5	48	Female	Asympt omatic	138	214	No	Nothin g to note	108	Yes	1.50	Flatslo ping
S6	59	Female	Asympt omatic	130	338	Yes	ST-T Wave abnorm ality	130	Yes	1.50	Flatslo ping
S7	45	Female	Atypica l Angina	130	237	No	Nothin g to note	170	No	0.00	Upslopi ng
S8	48	Female	Atypica l Angina	120	284	No	Nothin g to note	120	No	0.00	Upslopi ng

Step 2: State the hypothesis for these selected input data

Predicting based on the SVM Model

H1)	*S1 will be diagnosed as a heart disease patient*
H2)	*S2 will be diagnosed as a heart disease patient*
H3)	*S3 will be diagnosed as a non-heart disease patient*

H1)	*S1 will be diagnosed as a heart disease patient*
H4)	*S4 will be diagnosed as a non-heart disease patient*
H5)	*S5 will be diagnosed as a heart disease patient*
H6)	*S6 will be diagnosed as a heart disease patient*
H7)	*S7 will be diagnosed as a non-heart disease patient*
H8)	*S8 will be diagnosed as a non-heart disease patient*

5. Result

5.1. Results

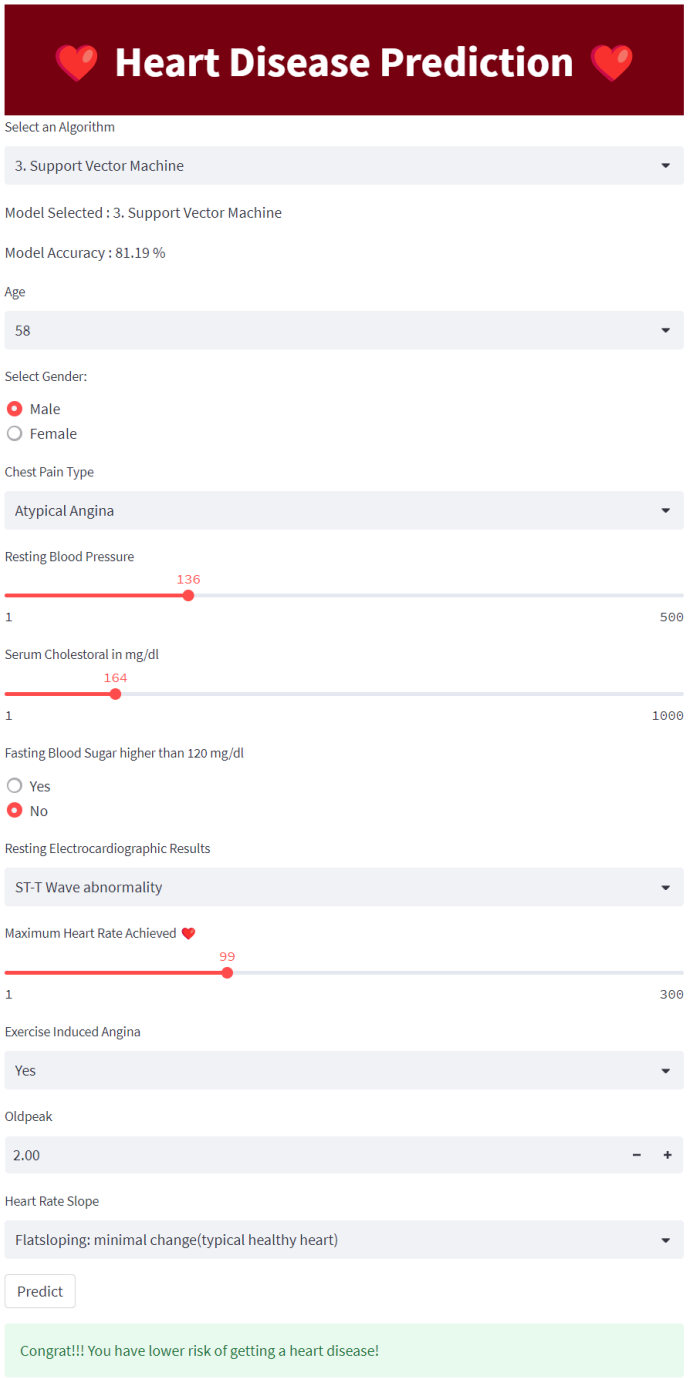


Figure 1: Predict S1 (SVM)

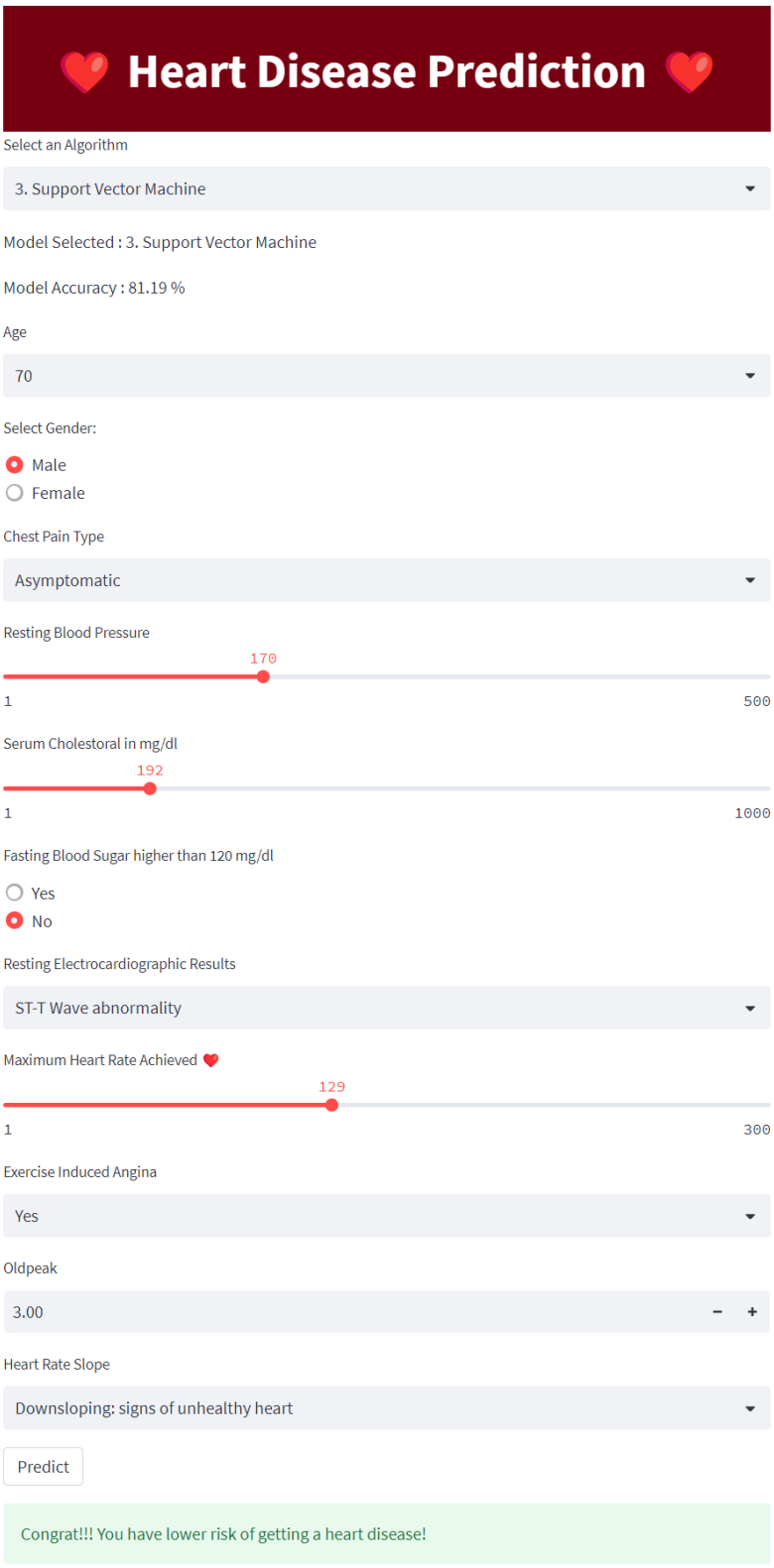


Figure 2: Predict S2 (SVM)

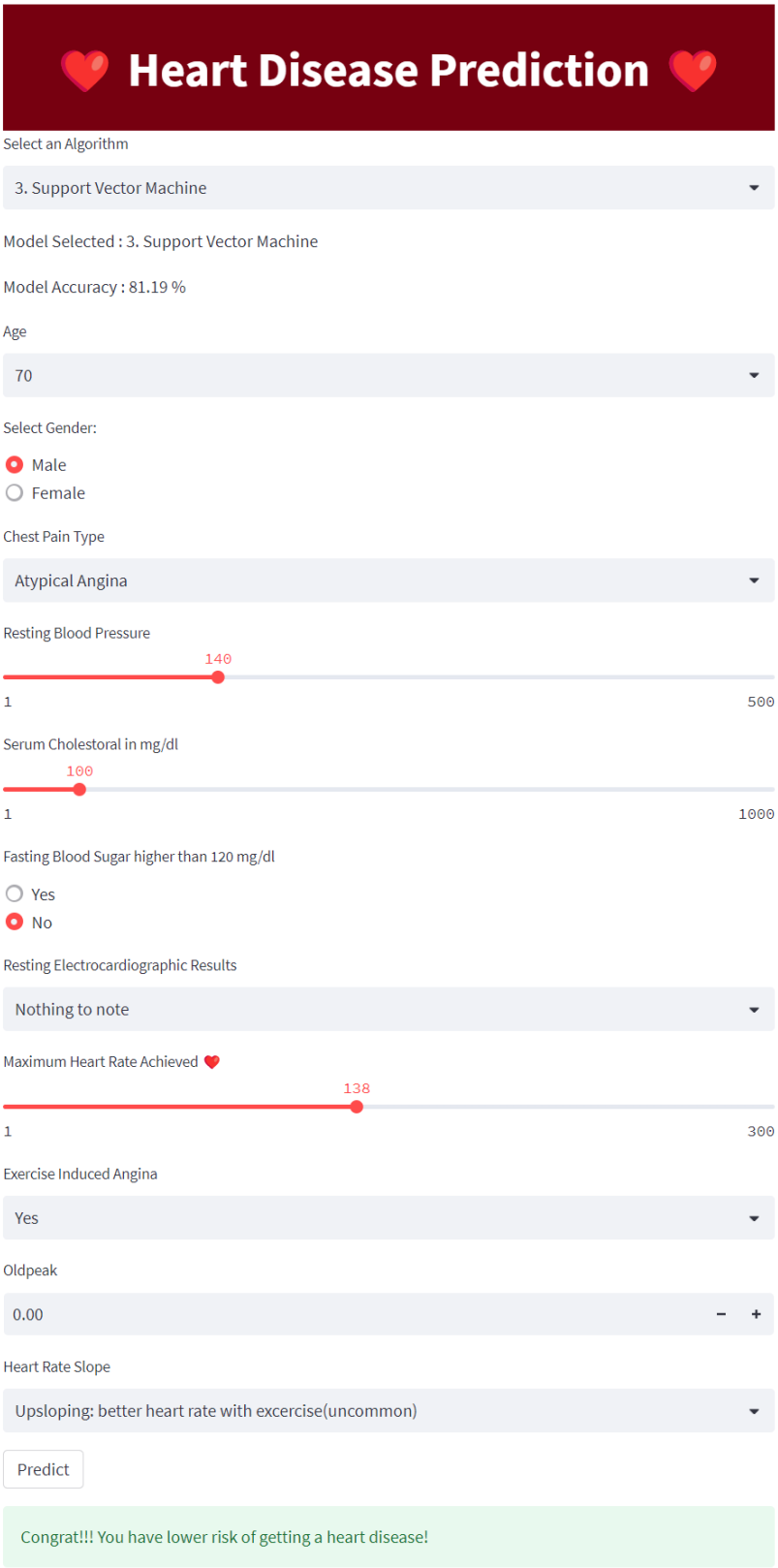


Figure 3: Predict S3 (SVM)

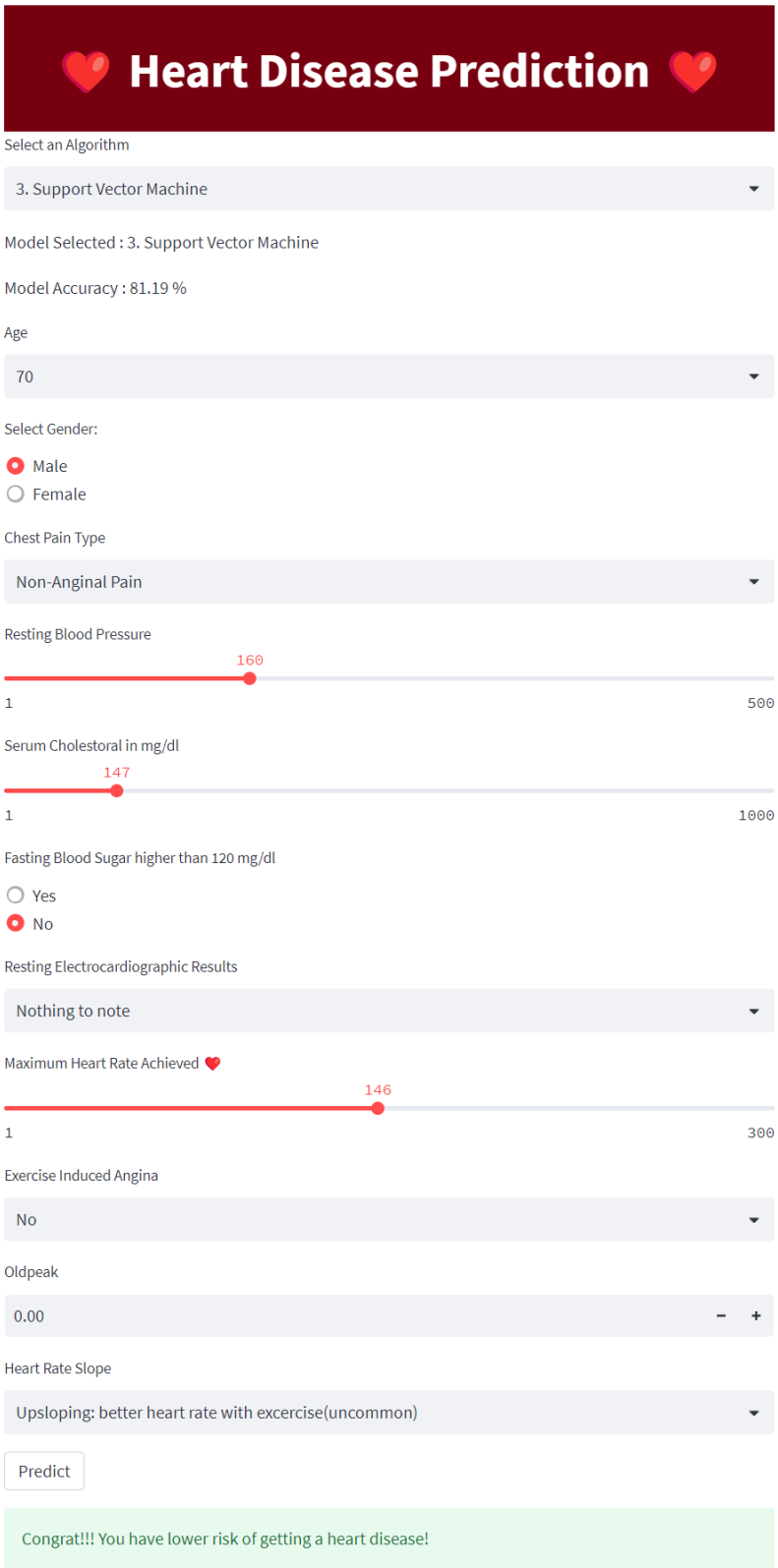


Figure 4: Predict S4 (SVM)

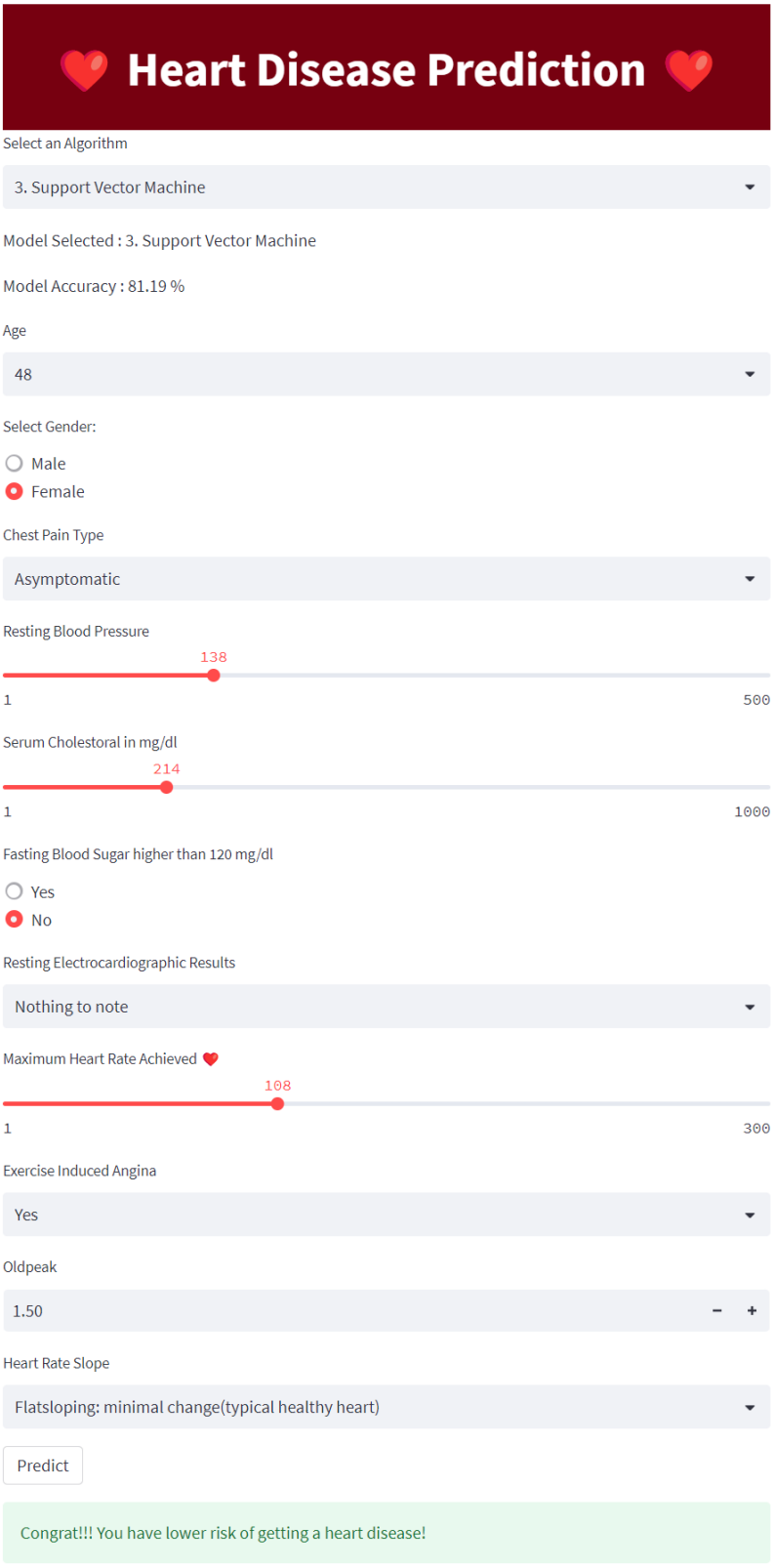


Figure 5: Predict S5 (SVM)

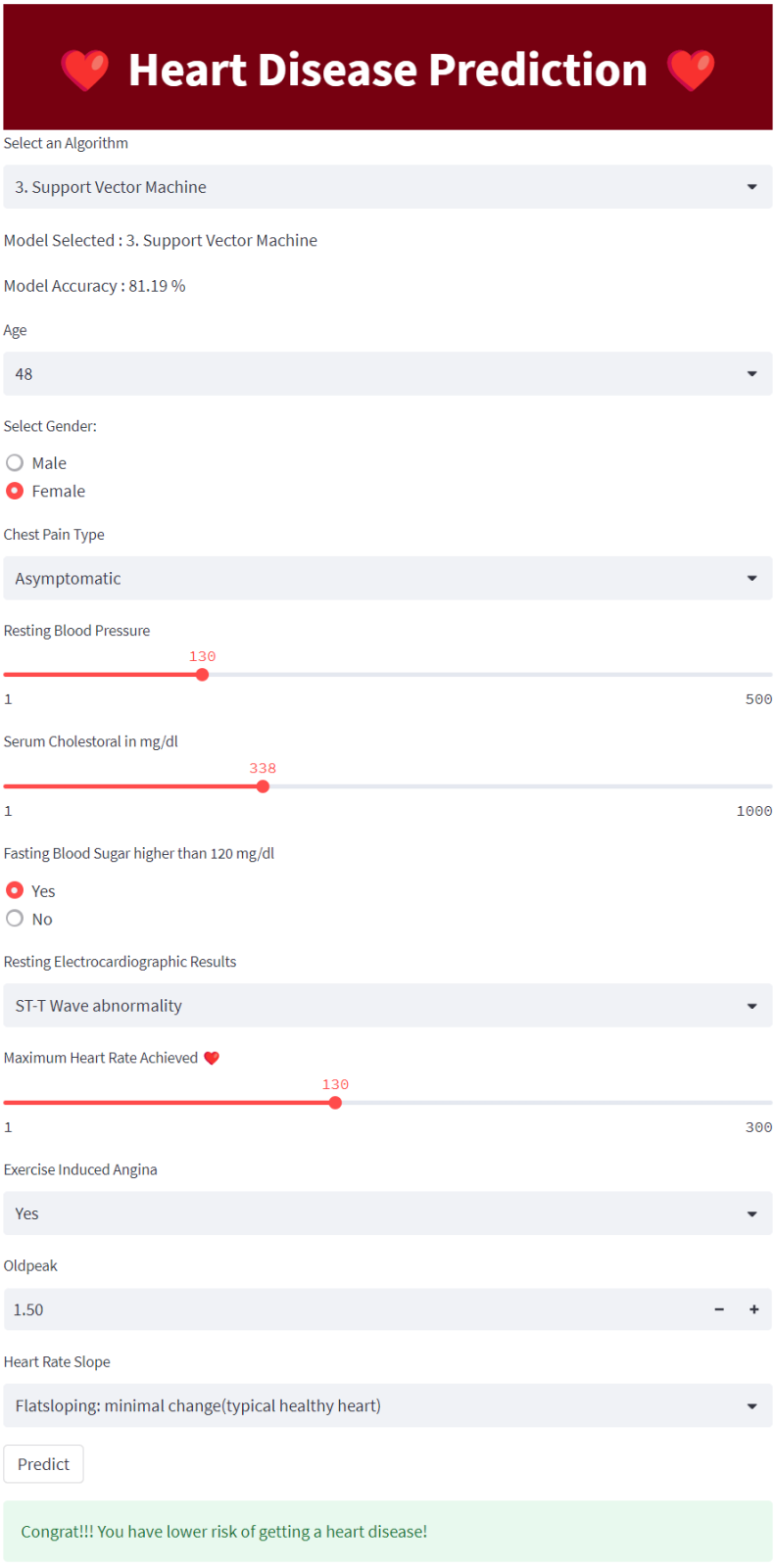


Figure 6: Predict S6 (SVM)

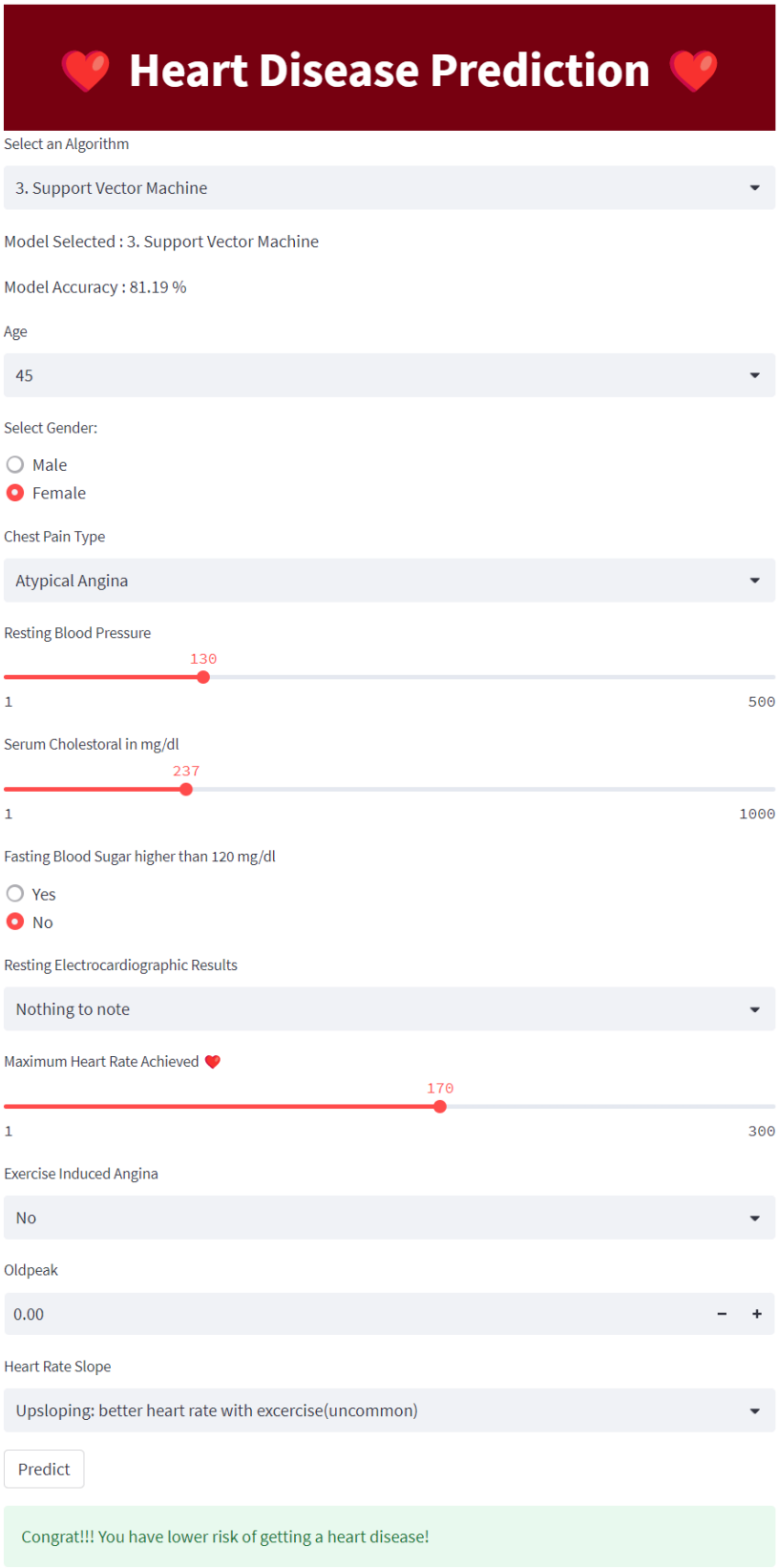


Figure 7: Predict S7 (SVM)

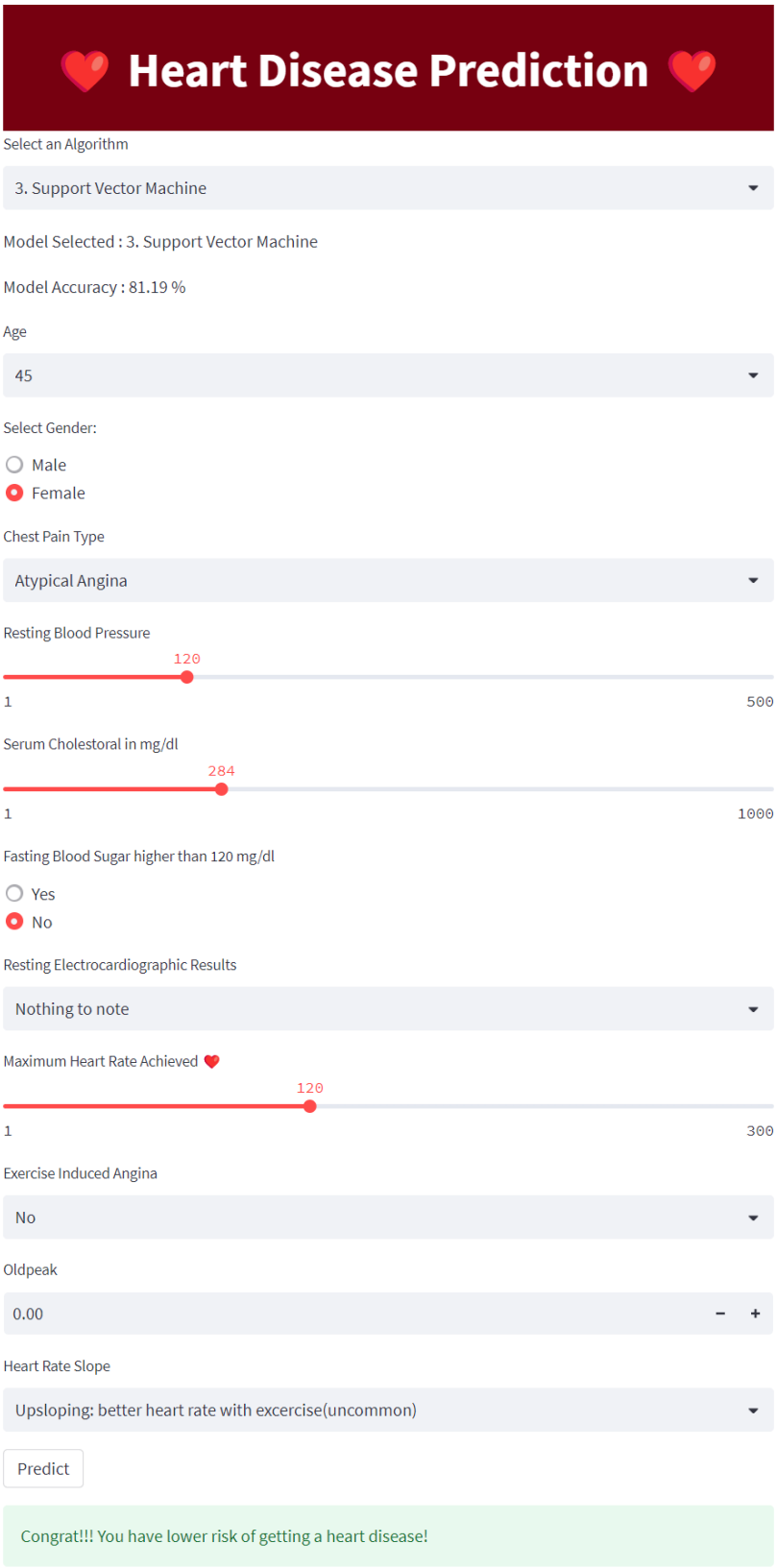


Figure 8: Predict S8 (SVM)

5.2. Discussion/Interpretation

Predicting results based on the SVM Model.

Hypothesis:

Hypothesis	Prediction Result based on SVM
H1	S1 will be diagnosed as non-heart disease patient
H2	S2 will be diagnosed as non-heart disease patient
H3	S3 will be diagnosed as non-heart disease patient
H4	S4 will be diagnosed as non-heart disease patient
H5	S5 will be diagnosed as non-heart disease patient
H6	S6 will be diagnosed as non-heart disease patient
H7	S7 will be diagnosed as non-heart disease patient
H8	S8 will be diagnosed as non-heart disease patient

Input Data:

"Congrat!!! You have lower risk of getting a heart disease!"

H1 is not rejecting and concludes that S3 is not prone to heart disease after predicting using SVM

S1	58	Male	Atypical Angina	136	164	No	ST-T Wave abnormality	99	Yes	2.00	Flatsloping

"Congrat!!! You have lower risk of getting a heart disease!"

H2 is not rejecting and concludes that S3 is not prone to heart disease after predicting using SVM

S2	70	Male	Asymptomatic	170	192	No	ST-T Wave abnormality	129	Yes	3.00	Downsloping

“Congrat!!! You have lower risk of getting a heart disease!”

H3 is not rejecting and concludes that S3 is not prone to heart disease after predicting using SVM

S3	52	Male	Atypical Angina	140	100	No	Nothing to note	138	Yes	0.00	Upsloping

“Congrat!!! You have lower risk of getting a heart disease!”

H4 is not rejecting and concludes that S3 is not prone to heart disease after predicting using SVM

S4	42	Male	Non-Anginal Pain	160	147	No	Nothing to note	146	No	0.00	Upsloping

“Congrat!!! You have lower risk of getting a heart disease!”

H5 is not rejecting and concludes that S3 is not prone to heart disease after predicting using SVM

S5	48	Female	Asymptomatic	138	214	No	Nothing to note	108	Yes	1.50	Flatsloping

“Congrat!!! You have lower risk of getting a heart disease!”

H6 is not rejecting and concludes that S3 is not prone to heart disease after predicting using SVM

S6	59	Female	Asymptomatic	130	338	Yes	ST-T Wave abnormality	130	Yes	1.50	Flatsloping

“Congrat!!! You have lower risk of getting a heart disease!”

H7 is not rejecting and concludes that S3 is not prone to heart disease after predicting using SVM

S7	45	Female	Atypical Angina	130	237	No	Nothing to note	170	No	0.00	Upsloping

“Congrat!!! You have lower risk of getting a heart disease!”

H8 is not rejecting and concludes that S3 is not prone to heart disease after predicting using SVM

S8	48	Female	Atypical Angina	120	284	No	Nothing to note	120	No	0.00	Upsloping

In conclusion, the result come out with SVM model is not very accurate because all the data we have selected is predict as non-heart disease patient and different with the hypothesis that have been stated early. So, SVM is not a very suitable algorithm for predicting this heart disease due to its accuracy is not high enough.

6. Discussion and Conclusion

6.1. Achievements

Throughout the heart disease prediction project, I gained knowledge and expertise in utilizing various machine learning algorithms, including Random Forest, Support Vector Machine, K-nearest neighbor, and Stacking cross-validation. By experimenting with each algorithm, I was able to identify their benefits and drawbacks, and differentiate between them. Additionally, I applied Stacking CV, an ensemble learning technique that combines multiple models to enhance the overall performance, which helped to minimize overfitting and improve generalization performance.

By using a dataset consisting of 3932 observations from Kaggle, I evaluated the algorithms and techniques based on precision and accuracy, ultimately determining that Random Forest was the most reliable algorithm, with Stacking CV performing even better. On the other hand, Support Vector Machine was found to be the least reliable due to the accuracy of svm is the lowest and the prediction result of heart disease is will not be accurate enough.

In conclusion, the project has met our objectives fulfilling its main objective of providing a reliable and accurate prediction machine for users to assess their risk of developing heart disease. Due to the accuracy of the prediction is quite high and more accurate prediction of heart disease risk.We also have a user interface with simple operation, so that users can easily use this system. Therefore, the heart disease prediction project has successfully achieved its objectives by providing a simple, reliable, and accurate prediction machine that can help users make informed decisions about their health.

6.2. Limitations and Future Works

The first limitation that we may encounter when using machine learning algorithms for heart disease diagnosis and prediction is the lack of domain-specific knowledge. Without sufficient knowledge of cardiovascular physiology and heart disease, students may not be able to identify the most relevant features or variables to include in their models, which can lead to inaccurate predictions. To address this limitation, we can collaborate with experts in cardiovascular physiology and heart disease, such as clinicians or researchers. Additionally, we can stay up to date with the latest developments in the field by reading relevant research articles and attending conferences or workshops.

The second limitation we will face is data quality is crucial in ensuring the accuracy and reliability of machine learning models in heart disease diagnosis and prediction. The quality of data used to train the models can have a significant impact on their performance. Inadequate data or data bias can lead to inaccurate results, hindering the effectiveness of the models. To address this limitation, future works should focus on gathering high-quality data that is free of biases and accurately represents diverse populations. Additionally, standardization of data collection and preprocessing protocols can improve the consistency and reproducibility of results.

Not only that, generalization is essential to ensure the effectiveness of machine learning models in heart disease diagnosis and prediction. Models trained on a specific dataset may not perform well on new datasets, leading to poor generalization. To address this limitation, future works should focus on evaluating and validating the performance of machine learning models on diverse datasets that accurately represent different populations. This can increase the robustness of the models and improve their effectiveness in clinical practice.

The next limitation is integration of machine learning models into clinical practice is challenging due to the lack of infrastructure and regulatory approval. The successful integration of these models requires a collaborative effort between healthcare providers, policymakers, and researchers. Future works should focus on developing strategies that address these challenges, such as creating guidelines for the ethical use of machine learning models in clinical practice and improving the infrastructure necessary for their integration.

Lastly, limited availability of data is a significant limitation of machine learning models in heart disease diagnosis and prediction. The amount of data available for research purposes is often limited due to privacy concerns and difficulty in collecting data from diverse populations. This can lead to overfitting, where the model performs well on the training data but poorly on new data. To address this limitation, future works should aim to increase the availability of diverse and high-quality data, including data from underrepresented populations. This can improve the effectiveness of machine learning models in heart disease diagnosis and prediction and ensure their generalizability to different populations.

Reference & Source

1. Brownlee, J. (2014). A Gentle Introduction to Scikit-Learn. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machinelearning-library/#:~:text=Scikit%2Dlearn%20was%20initially%20developed.

2. Kaggle (2022). Kaggle: Your Home for Data Science. [online] Kaggle.com. Available at: https://www.kaggle.com/.

3. Donges, N. (2021). A Complete Guide to the Random Forest Algorithm. [online] Built in. Available at: https://builtin.com/data-science/random-forest-algorithm.

4. JavaTpoint (2021). K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint. [online] www.javatpoint.com. Available at: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

5. Khan Academy. (n.d.). Machine learning algorithms (article). [online] Available at: https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101/x2d2f703b37b450a3:machine-learning-and-bias/a/machine-learning-algorithms#:~:te [Accessed 27 Apr. 2023].

6. Sunil, R. (2019). Understanding Support Vector Machine algorithm from examples (along with code). [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

7. Jindal, H. et al. (2021) “Heart disease prediction using machine learning algorithms,” in IOP Conference Series: Materials Science and Engineering. IOP Publishing Ltd. doi:10.1088/1757-899X/1022/1/012072.

8. FTM. 2021. [online] Available at: <https://www.freemalaysiatoday.com/category/nation/2021/11/16/heart-diseases-remain-thetop-killer-in-malaysia/>

9. CodeBlue (2020). Nearly Half Of Malaysians Lack Health Coverage Beyond Public Care. [online] CodeBlue. Available at: https://codeblue.galencentre.org/2020/06/02/nearly-half-of-malaysians-lack-health-coverag-beyond-public-care/.

10. Ray, S. (2023) Learn how to use support vector machines (SVM) for Data Science, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (Accessed: May 4, 2023).

11. *Support Vector Machines (SVM) algorithm explained* (2017) *MonkeyLearn Blog*. Available at: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/ (Accessed: May 4, 2023).