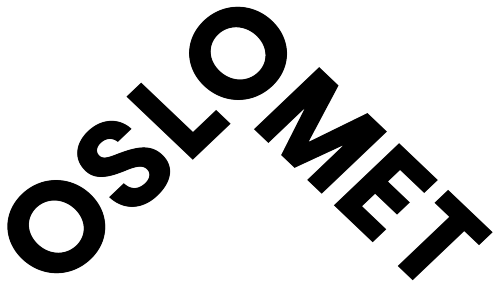


OsloMet- Oslo Metropolitan University

Department of Computer Science
Oslo Norway

ACIT4510-1 21H Statistical Learning

Master's Degree ACIT: Data Science

November 2021

Heart Disease Data set: Predicting the Probability of People Having Heart Disease

Shailendra Bhandari

s366261@oslomet.no

Contents

Chapter 1

Motivation and Introduction

1.1 Motivation

The main objective of healthcare facilities is to provide quality services at affordable cost. Effective diagnostic technology at affordable cost is an essential requirement for quality services. Most modern hospitals now use systematic data collection with appropriate computerized information systems. The application of such data mining techniques to health data can be found in reference [1]. In practice, a large amount of data is collected in the form of numbers, texts, tables and images during the treatment or diagnosis of a patient. There could be a wealth of information hidden in these data. So the question is: how can we decipher the patterns to extract the useful information from the collected data so that clinicians can make intelligent clinical decisions. So the motivation of this analysis is to turn the data into useful information.

1.2 Introduction

The term "heart disease" is often replaced by the term "cardiovascular disease". Cardiovascular disease generally refers to conditions involving narrowed or blocked blood vessels that can lead to heart attack, chest pain (angina), or stroke. Other heart diseases that affect the heart muscle, the heart valves or the heart rhythm are also called cardiac diseases. Heart disease is one of the leading causes of death in men and women of all ages worldwide. According to the World Health Organization, one in four people die from heart disease. Cardiovascular disease caused the most deaths in 2016 was 18.9 million, accounting for 44% of all deaths from non-communicable diseases [2]. A 30-year-old man had a higher risk of dying from coronary

heart disease before reaching age 70 than a 30-year-old woman. The increasing risk of heart disease may be due to excessive alcohol consumption, smoking, drugs, etc. Similarly, the major causes are high blood pressure, high blood sugar and high cholesterol.

Various studies show that most heart attacks are silent and about two-thirds of those affected die before reaching the hospital. Therefore, it is very important to recognize the early warning symptoms for prevention and further treatment. According to WHO, at least half of the world's population does not have full access to basic health services. Many of those who do have access to the services they need suffer from undue financial hardship. Therefore, the development of medical research is crucial for the cost-effective treatment of heart disease. Disease prediction can be considered as one of the challenges and is one of the most important subjects of investigation in the field of clinical data analysis. Medical diagnostic systems based on machine learning techniques can play an important role in this regard.

1.3 Objective of this Report

The main objective of this report is to select a problem, describe the data to be analyzed, propose a method for analyzing the data (appropriate algorithm for data analysis), and a methodology for interpreting the results. This includes the following methods:

- **Identification of Data:** Identifying relevant data includes specifying data sources, sampling data size, and developing a strategy for handling missing data and binning for continuous variables.

- **Data Understanding:** The understanding phase of the identified data uses the raw data to understand the data, identify its quality with preliminary findings, and identify subsets of interest to formulate hypotheses. This particularly done in Chapter [2].

- **Data Preparation**: In this phase, a final data set is created to be fed into a model. Three different computational intelligence technique (Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF)) methods are employed to calculate the accuracy.

- **Algorithm Specification:** The next step is to specify an appropriate algorithm. Depending on the type of dataset, any methodology can be used to show how it fits. Some examples are decision trees, neural networks, genetic algorithms, etc.

- **Deployment**: In this phase, the tasks required to use the model are defined. Details of this is shown in Chapter [3].

In this report, I will attempt to apply the above methods to model, evaluate, and deploy the dataset. The next objective is the binary classification of the structured data. Since each row contains information about a patient (a sample) and each column describes an attribute of the patient (a feature), these features are used to predict whether a patient has heart disease (binary classification). As mentioned earlier, disease prediction can be considered as one of the challenges and is one of the most important subjects of investigation in the field of clinical data analysis. Medical diagnostic systems based on machine learning techniques can play an important role in this regard. Nevertheless, a large volume of medical data can be easily collected and from the analysis of collected data, a wealth of hidden information could be extracted using various machine learning techniques.

Chapter 2

Data Source

2.1 Data Source

This data set is taken from the the data library UCI Machine Learning Repository for heart Disease dataset[1] with CSV source file[2]. Furthermore, the code was also taken from the same site and implemented for further analysis. This data was available in CSV file including both numerical and categorical features. The Keras preprocessing layers was used to normalized the numerical features and vectorize the categorical features. Table [2.1] shows the raw data set including 3 samples with 14 columns per sample. Altogether, there are 303 samples. The description of each features is shown in Table [2.2].

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	1	145	233	1	2	150	0	2.3	3	0	fixed	0
67	1	4	160	286	0	2	108	1	1.5	2	3	normal	1
67	1	4	160	286	0	2	108	1	1.5	2	3	normal	1

Table 2.1: Sample raw data

Creators:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

[1] http://archive.ics.uci.edu/ml/datasets/heart+disease
[2] http://storage.googleapis.com/download.tensorflow.org/data/heart.csv

Column	Description
Age	Age in years
Sex	Value 1: Male, Value 0: Female
CP	Chest pain type 1, 2, 3, 4
Trestbpd	Fasting blood pressure(in mm Hg on admission)
Chol	Serum Cholesteril in mg/dl
FBS	fasting blood sugar in 120 mg/dl (1 = true; 0 = false)
RestECG	Resting electrocardiogram results (0, 1, 2)
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina (1 = yes; 0 = no)
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment
CA	Number of major vessels (0-3) colored by fluoroscopy
Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
Target	Diagnosis of heart disease (1 = true; 0 = false)

Table 2.2: Features of the data set.

- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor: David W. Aha (aha@ics.uci.edu) (714) 856–8779

Figure [2.1] shows a correlation matrix displaying correlation coefficient for all the variables available in the data set. The correlation matrix shows the correlation between all the possible pairs of values in the table. This gives a visual summarizing of the data set to identify and visualize patterns in this data set. Figure [2.2] shows the correlation of different variables with the target. It is clearly seen from the figure that the variable fasting blood sugar in 120 mg/dl (fbs) and blood cholesterol (chol) are the least correlated with the target variables. Rest of all the variables have a significant correlation with the target variable.

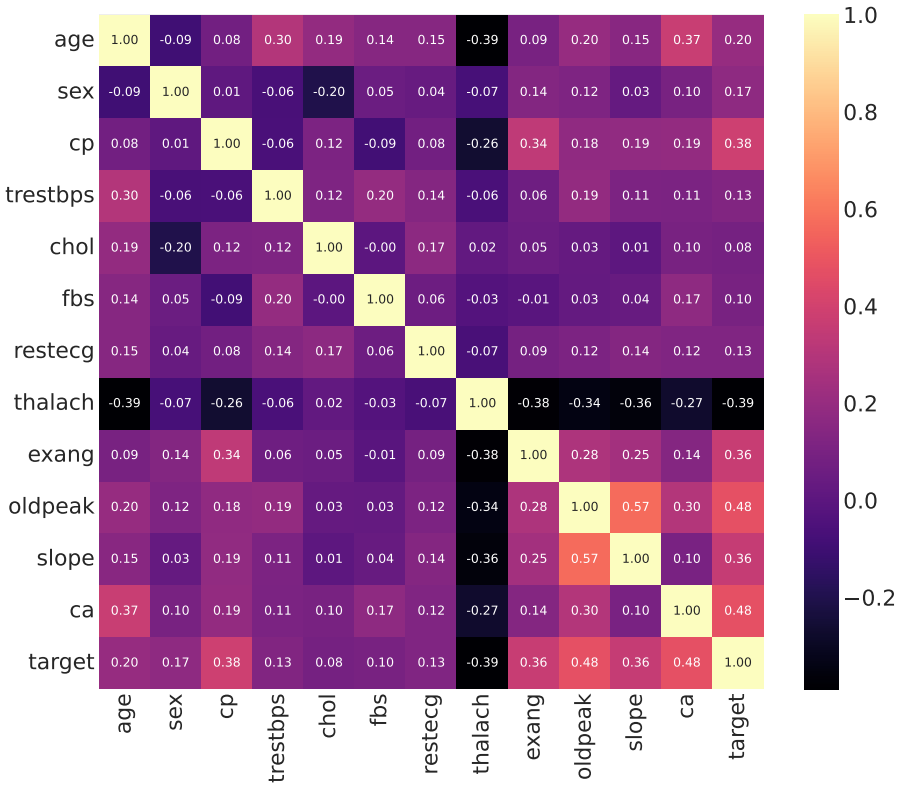


Figure 2.1: Correlation matrix

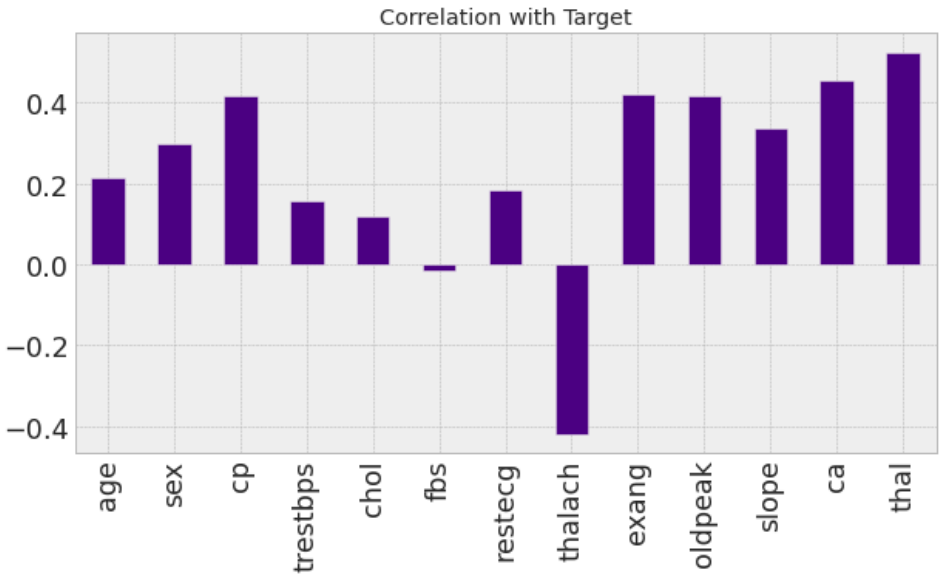


Figure 2.2: Correlation of the variables with target variables.

Chapter 3

Data Analysis

First, the data were divided into a training set (212 samples) and a validation set (91 samples). That is, the training set is 70% and the testing set is 30% of the total data. The validation set approach was used to estimate the test error rates resulting from fitting different linear models. The estimated test MSE for the linear regression fit is 72.56 and the test error for quadratic and cubic functions is 69.78 and 72.564, respectively. Since both numeric and categorical features are included in the dataset, the normalization layer for numeric, categorical and integer features was used to ensure that the mean is 0 and the standard deviation is 1.

3.1 Overview of the state of art

Classification of datasets is the most important and popular tool for decision making in the field of medical science. Some of the most popular models in this field are the following. Wang et al. [3] a Support Vector Machine (SVM) based learning algorithm for breast cancer diagnosis. By using this model, the accuracy was found to be increased by 33%. Subsequently, several researchers have adapted this SVM classifier with feature selection algorithms to predict coronary heart disease. Shao et al.[4] proposed another system to determine the accuracy of coronary heart disease prediction using machine learning logistic regression strategies. This system was able to achieve an accuracy of 82.14%.

3.1.1 Support Vector Machine

The Support Vector Machine (SVM) is an approach to classification and has proven to be useful in a variety of fields, including the medical field for predicting disease. The SVM is equivalent to support vector classifier with the polynomial kernel of degree $d = 1$. The purpose of the SVM algorithm is to create the line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category. This best decision boundary is called a hyperplane. The SVM computes the prediction of the form $\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p$ for each observation. We can classify the observation into the heart diseased of no heart disease categories depending on whether $\hat{f}(X) \leq t$ or $\hat{f}(X) \geq t$. The confusion matrix can be obtained by forming these predictions and computing the false positive and true positive rates for a range of values of t. SVM is highly preferred as it produces significant accuracy with less computation power for high dimensional data. The parameter used in SVM are: Kernel "rbf" i.e. to fit an SVM, a radial kernel is used, Penalty parameter C with error term =1. The analysis code below is adopted according to reference paper [5].

```
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler  # for data preprocessing
from sklearn.model_selection import train_test_split # for data splitting
from sklearn.metrics import confusion_matrix,accuracy_score,roc_curve,
    classification_report # for data modeling

model_for_support_vector_machine = 'Support Vector Classifier'
svc =  SVC(kernel='rbf', C=2) #kernel 'RBF' used, differeny kernel like '
    linear', 'poly', 'sigmoid'
                              #can also be used and check the accuracy ||
    Penalty parameter C=1 or 2..
svc.fit(X_train, y_train)
svc_predicted = svc.predict(X_test)
svc_conf_matrix = confusion_matrix(y_test, svc_predicted)
svc_acc_score = accuracy_score(y_test, svc_predicted)
print("confussion matrix")
print(svc_conf_matrix)
print("\n")
print("Accuracy of Support Vector Classifier:",svc_acc_score*100,'\n') #(
    TP + TN)/(TP + TN + FP + FN)
print(classification_report(y_test,svc_predicted))
```

```
Output:
confussion matrix
[[29 12]
 [10 40]]
Accuracy of Support Vector Classifier: 75.82417582417582
```

Code 3.1: Sample code for Support Vector Machine

3.1.2 Logistic Regression

The logistic model is a discriminative category approach where we must model a function $p(x)$ with a function that gives the outputs between 0 and 1 for all values of X. Many functions meet this description, however, in logistic regression we use the logistic function.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{3.1}$$

To fit the model (the above mention logistic regression function) a maximum likelihood method should have to be use. Equation 3.1 can be written as:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}. \tag{3.2}$$

The quantity $p(x)/[1 - p(X)]$ is called the odds and can take on any value between 0 and ∞. The major task of the logistic regression is to calculate the odds of an events. Mathematically, logistic regression estimates multiple linear regression function which is defined as follows:

$$log\frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X. \tag{3.3}$$

$$log\frac{p(X = 1)}{1 - p(X = 1)} = \beta_0 + \beta_1 X + \beta_2 X + ...\beta_k X_k. \tag{3.4}$$

```
import pandas_profiling as pp   # for data preprocessing
from sklearn.preprocessing import StandardScaler # for data splitting
from sklearn.model_selection import train_test_split # for data modeling
from sklearn.linear_model import LogisticRegression
model_for_Logistic_Regression = 'Logistic Regression'
lr = LogisticRegression()
model = lr.fit(X_train, y_train)
lr_predict = lr.predict(X_test)
lr_conf_matrix = confusion_matrix(y_test, lr_predict) #matrix
lr_acc_score = accuracy_score(y_test, lr_predict) # for validation set
print("confussion matrix")
print(lr_conf_matrix)
print("\n")
print("Accuracy of Logistic Regression:",lr_acc_score*100,'\n')
print(classification_report(y_test,lr_predict))
#Output
confussion matrix
[[30 11]
 [ 9 41]]


Accuracy of Logistic Regression: 78.02197802197803
```

Code 3.2: Sample code for Logistic Regression

3.1.3 Random Forest

Random Forests have become widely used in medical diagnostics [6, 7]. In Random Forests, the trees are again drawn independently of random samples of the observations. The random forest algorithm can be applied to any tree at any split, using a random sequence of features, allowing for a more comprehensive exploration of the model space. RF consists of numerous decision trees. Each decision tree gives a vote indicating the decision on the class of the object. The parameters used are number of estimators = 20, random state =10 and maximum depth =10.

```

from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
model_random_forest_classifier = 'Random Forest Classfier'
rf = RandomForestClassifier(n_estimators=20, random_state=10,max_depth=10)
```

```
rf.fit(X_train,y_train)
rf_predicted = rf.predict(X_test)
rf_conf_matrix = confusion_matrix(y_test, rf_predicted)   #matrix
rf_acc_score = accuracy_score(y_test, rf_predicted)
print("confussion matrix")
print(rf_conf_matrix)
print("\n")
print("Accuracy of Random Forest:",rf_acc_score*100,'\n') #for validation
    set
print(classification_report(y_test,rf_predicted))

#Alternative approach
#print('Accuracy for training set for Random Forest = {}'.format((cm_train
    [0][0] + cm_train[1][1])/len(y_train)))
#print('Accuracy for test set for Random Forest = {}'.format((cm_test
    [0][0] + cm_test[1][1])/len(y_test)))

#Output:
confussion matrix
[[30 11]
 [12 38]]
Accuracy of Random Forest: 74.72527472527473
```

Code 3.3: Sample code for Random Forest Classifier

3.2 The performance Measurement

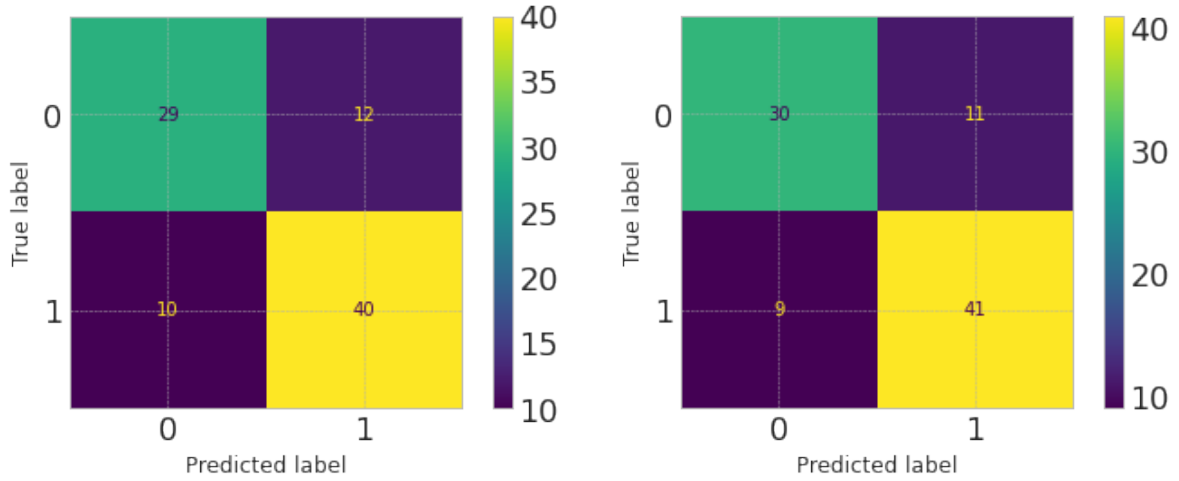
The confusion matrix, also known as the error matrix, is used to visualize or calculate the performance of computational intelligence techniques. The correctly and incorrectly predicted values are displayed in the confusion matrix. They are defined as follows:

- TP = True Positive(Correctly Identified)
- TN = True Negative(Incorrectly Identified)
- FP = False Positive(Correctly Rejected)
- FN = False Negative (Incorrectly Rejected)

The accuracy of the training and the validation sect can be calculated by the formula [5]:

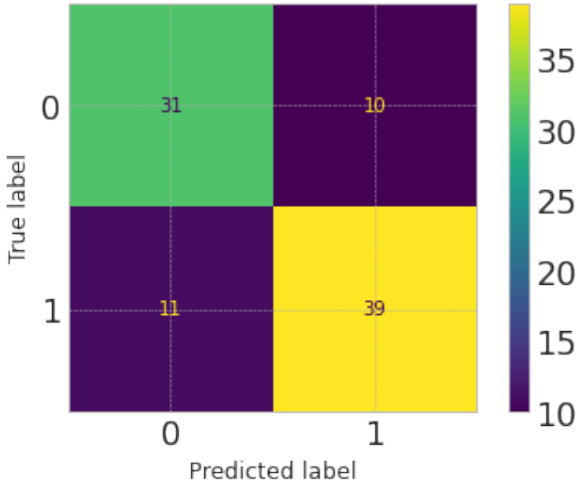
$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3.5}$$

Similarly, sensitivity, specificity, and precision can be calculated using the confusion matrix. The confusion matrix for SVM, LR, and Random Forest is shown in Figures [3.1a, b and c], respectively. Using the equation 3.5, the accuracy for LR for the validation set = $\frac{30+41}{30+11+9+41}$ = 0.7802, which corresponds to 78.02%. Similarly, the accuracy of SVM and Random Forest is 75.82% and 74.72%, respectively.



(a) SVM

(b) LR



(c) RF

Figure 3.1: Confusion matrix for Validation set of (a) SVM (b) LR and (c) RF.

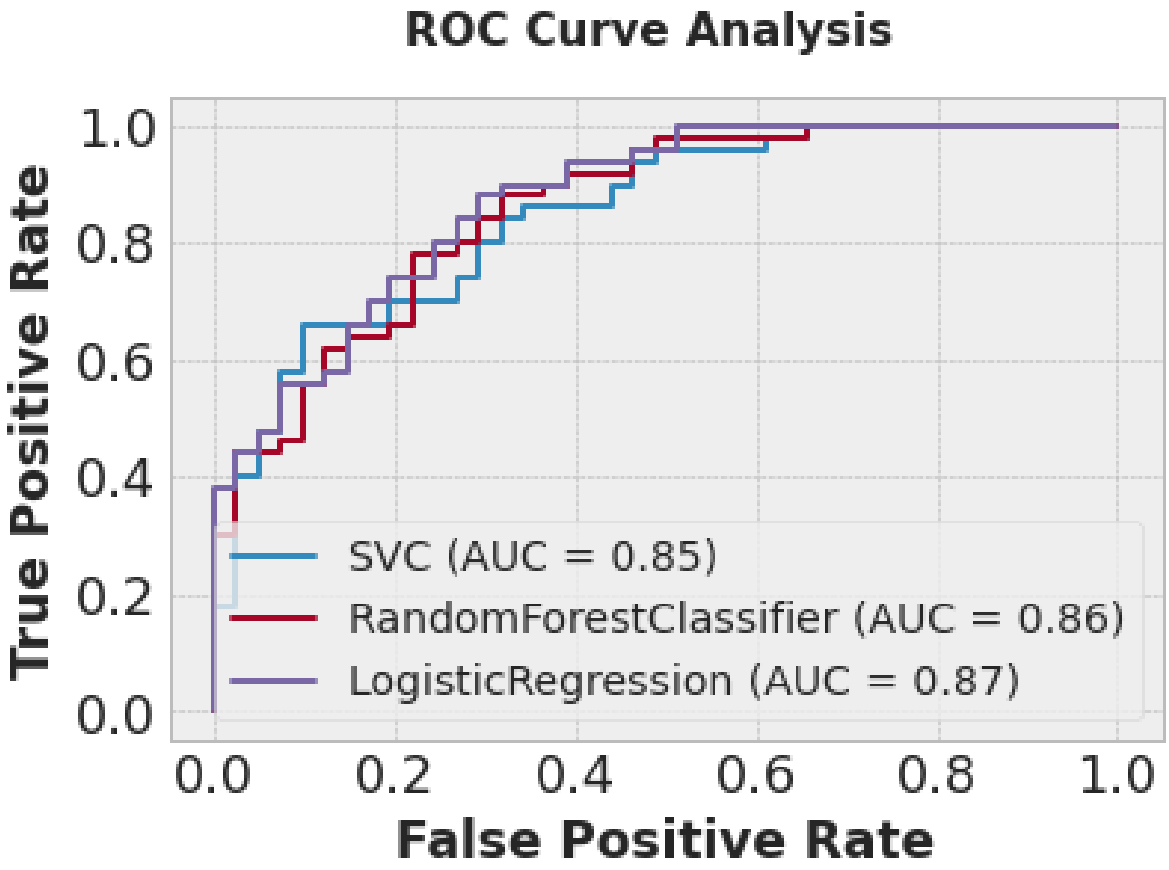


Figure 3.2: Receiver Operator Characteristics(ROC) curve analysis for SVC, LR and RF

Figure [3.2] shows the ROC curves for the heart data training set. The support vector classifier, Random Forest Classifier, and Logistic Regression Classifier are compared. Depending on the classifier scores $\hat{f}(X) < t$ or$\hat{f}(X) \geq t$ as mentioned in Section [3.1.1], ROC curve is obtained by forming these predictions and computing the false positive and true positive rates for a range of values of t. The ROC curves represent training error rates, which can be misleading in terms of performance on new test data.

In total, three different computational intelligence techniques are used to measure accuracy. Next, the probability of prediction based on each categorical variable is examined. The data consists of individuals who have heart disease and individuals who do not have the disease. Figure [3.3] shows the scatter plots for total number of individuals suffering from heart disease compared to those without heart disease. Heart disease in correlation to age and resting blood pressure, age and blood serum and age to maximum heart disease are shown in figure [3.4]. Both Resting Blood Pressure and Serum Cholesterol shows a bit positive correlation but not

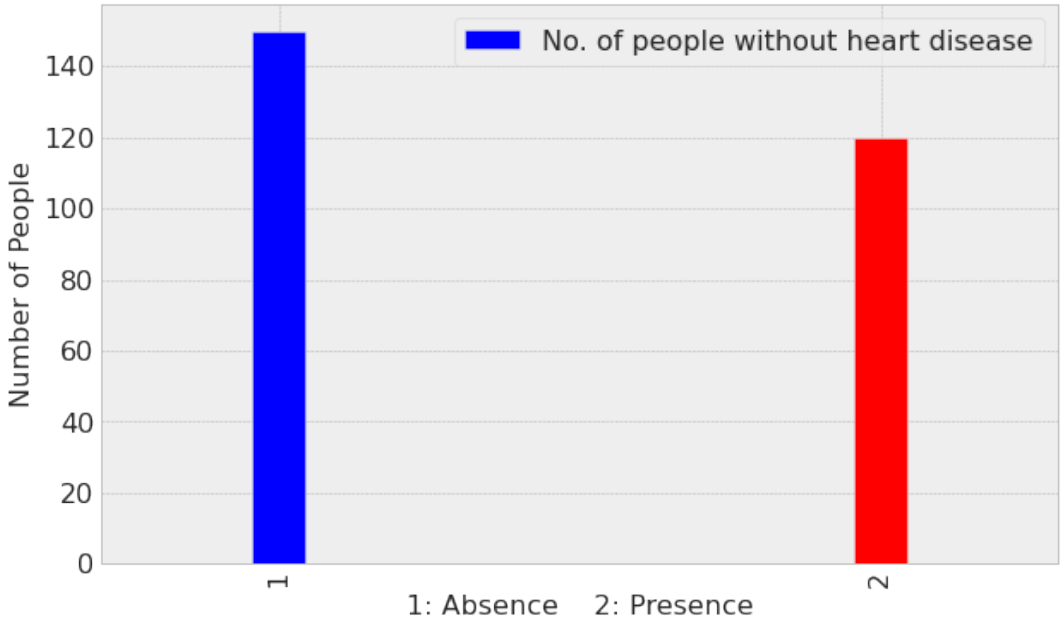


Figure 3.3: The total number of people who are suffering from heart disease versus without heart disease.

that much. However, observation from maximum heart rate doe not show any correlation.

In order to calculate the probability of having heart disease based on each variable (e.g. age, serum cholesterol level, maximum heart rate), a model was created and the model was trained. Validation of the model is found to be 85.71%. Since both numerical and categorical features are included in the data set, the normalization layer was used for numerical, categorical and integer features to make sure the mean is 0 and its standard deviation is 1.

3.3 The model and the Experiment

The model used to predict the probability of people having heart disease for this data set is described below:

- The data set is divided into validation and training set in the ratio 3:7. i.e. out of 303 samples, 212 samples for training and 91 for validation set.
- Batch size used is 64.

```
train_ds = train_ds.batch(64)
val_ds = val_ds.batch(64)
```

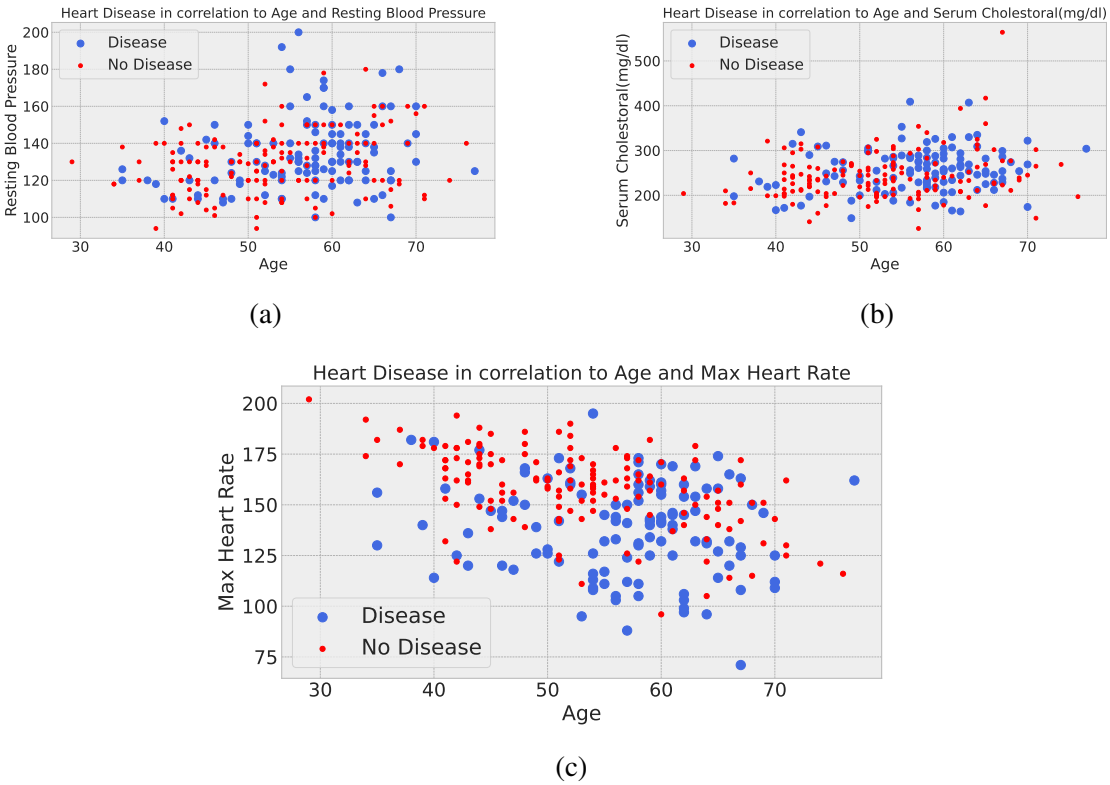


Figure 3.4: Scatter plot for (a) heart disease in correlation to the age and the resting blood pressure (b) heart disease in correlation to theAge and the serum Cholesterol and (c) heart disease in correlation to the age and the maximum heart rate.

- Input layers has activation function *relu*. The ReLU (Rectified Linear Unit) function is non-linear activation function. The main advantage of using the ReLU function is it does not activate all the neurons at the same time.

- Output layers has activation function *Sigmoid*. This function can be replaced by relu function but in this data sect the accuracy of model was found to be increased by around 10%, so the Sigmoid function is employed in the output layers. The mathematical expression for Sigmoid function is $f(x) = 1/(1 + e^{-x})$.

- The optimizer is *Adam* and loss functions used is *binary_crossentropy*. The categorical_crossentropy can also be used, but in this dataset the accuracy of the model was found to be higher with binary_crossentropy.

```
x = layers.Dense(64, activation="relu")(all_features)
x = layers.Dropout(0.5)(x)
```

```
output = layers.Dense(1, activation="sigmoid")(x)
model = keras.Model(all_inputs, output)
model.compile("adam", "binary_crossentropy", metrics=["accuracy"])
```

- The model is trained for 90 epochs.

```
model.fit(train_ds, epochs=90, validation_data=val_ds)
```

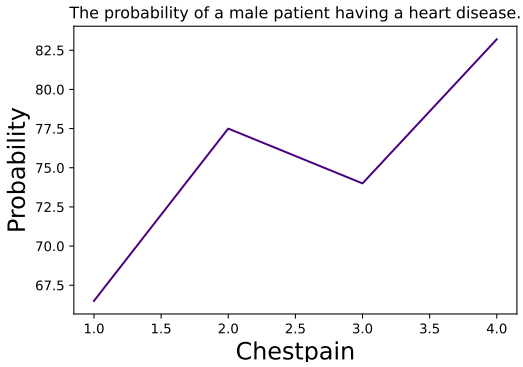
- Using the above mentioned condition, the accuracy of the model is 85.71%.

The next step is to get a prediction for a new sample. For this model.prediction() [8] is employed as shown in Code [3.4]. Since the models only processes batches of data, the scalars are wrapped into a list so as to have a batch dimensions.

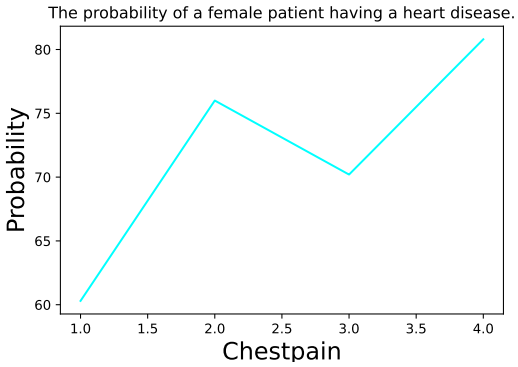
```
sample = { "age":60,  "sex": 1,   #1 for male and 0 female
    "cp": 4,  #Chest pain type 1, 2, 3, 4
    "trestbps": 120, # Fasting blood pressure(in mm Hg on admission)
    "chol": 308, # Serum Cholesteril in mg/dl
    "fbs": 1,  #fasting blood sugar in 120 mg/dl (1 = true; 0 = false)
    "restecg": 1, # Resting electrocardiogram results (0, 1, 2)
    "thalach": 150, #Maximum heart rate achieved
    "exang": 0,  #Exercise induced angina (1 = yes; 0 = no)
    "oldpeak": 2.3,"slope": 2,   #1 = upsloping 2 = flat 3 = downsloping
    "ca": 3,   #Number of major vessels (0-3)
    "thal": "fixed", # 0=fixed, 3 = normal; 6 = fixed defect; 7 =
   reversibl defect}
input_dict = {name: tf.convert_to_tensor([value]) for name, value in
   sample.items()}
predictions = model.predict(input_dict)
print(
    "A 60 year old male patient with cp=4,bps=120,chol=308, fbs 1 and so
   on had a %.1f percent probability "
    "of having a heart disease, as evaluated by the model." % (100 *
   predictions[0][0],))
    #Output:
    A 60 year old male patient with cp=4,bps=120,chol=308, fbs 1 and so on
     had a 78.6 percent probability of having a heart disease, as evaluated
     by the model.

```

Code 3.4: Sample code showing the model predicting the probability of people having heart disease.

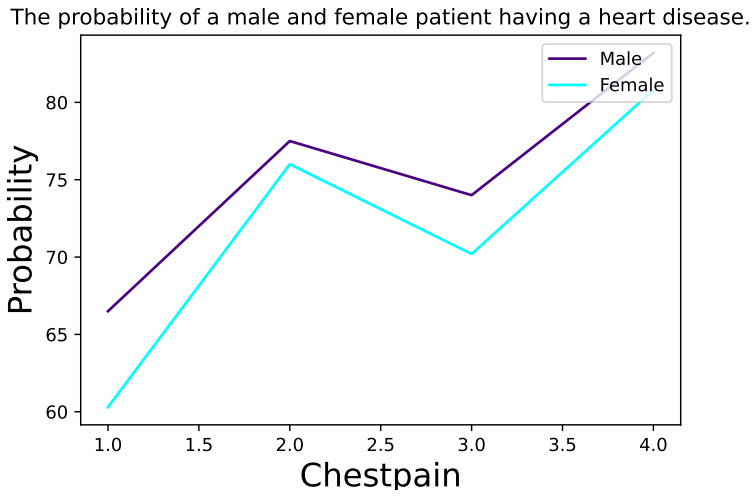
Keeping the other variables constant, I started calculating the probability of people suffering from heart disease and the results obtained are shown below. For example, a 60 year old male patient with cp=4,bps=120,chol=308, fbs 1, restecg =1, heart rate =150 had a 78.6 percent probability of having a heart disease, as evaluated by the model.



(a)



(b)



(c)

Figure 3.5: The probability of heart disease for (a) men (b) women and (c) the comparison between men and women as a function of chest pain type.

Figure [3.5] shows the prediction or probability of people having heart disease based on the type of chest pain. As mentioned in Chapter 2, chest pain is divided into four categories based on the type of pain experienced by the individual, using the following format: 1 = typical angina, 2 = atypical angina, 3 = nonanginal pain, and 4 = asymptomatic. The figure shows that people with typical angina have a low likelihood of heart disease, while people with asymptotic

type 4 pain have a higher risk of heart disease. However, the likelihood of type 3 chest pain (non-anginal pain) decreases compared to type 2 chest pain (atypical angina). Also, the figure shows that male have higher probability of having heart disease than female. In further analysis, chest pain type 4 is considered to calculate the likelihood of heart disease.

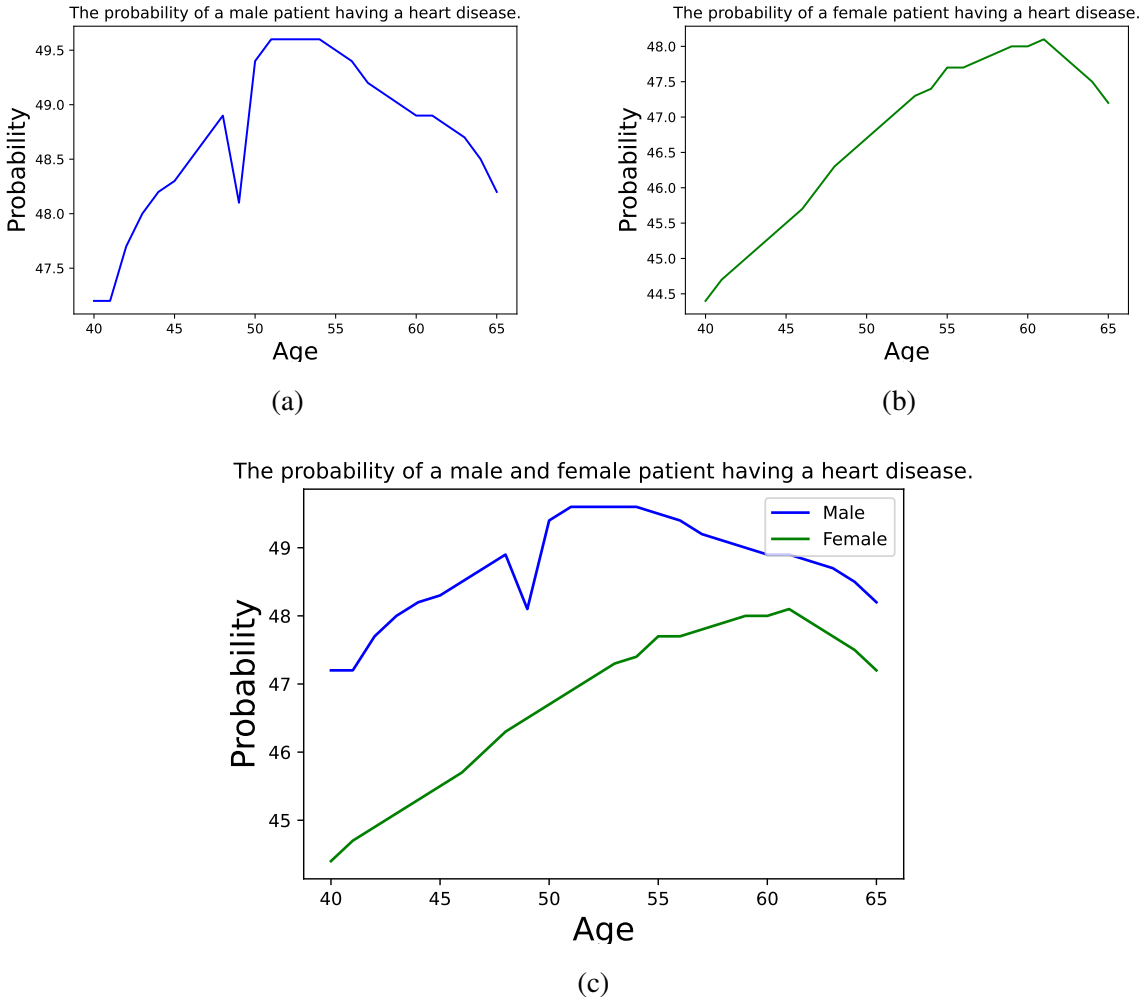


Figure 3.6: The probability of heart disease for (a) men (b) women and (c) the comparison between men and women as a function of age.

Figure [3.6] shows the probability of having heart disease to the patient of age between 40 to 65 years. It shows that male patient have higher probability of having heart disease than female. Similarly, maximum probability of having heart disease to male patient is at around 54 years whereas for female it is at around 61 years.

Figure [3.7] shows the probability of male and female people having heart disease depending

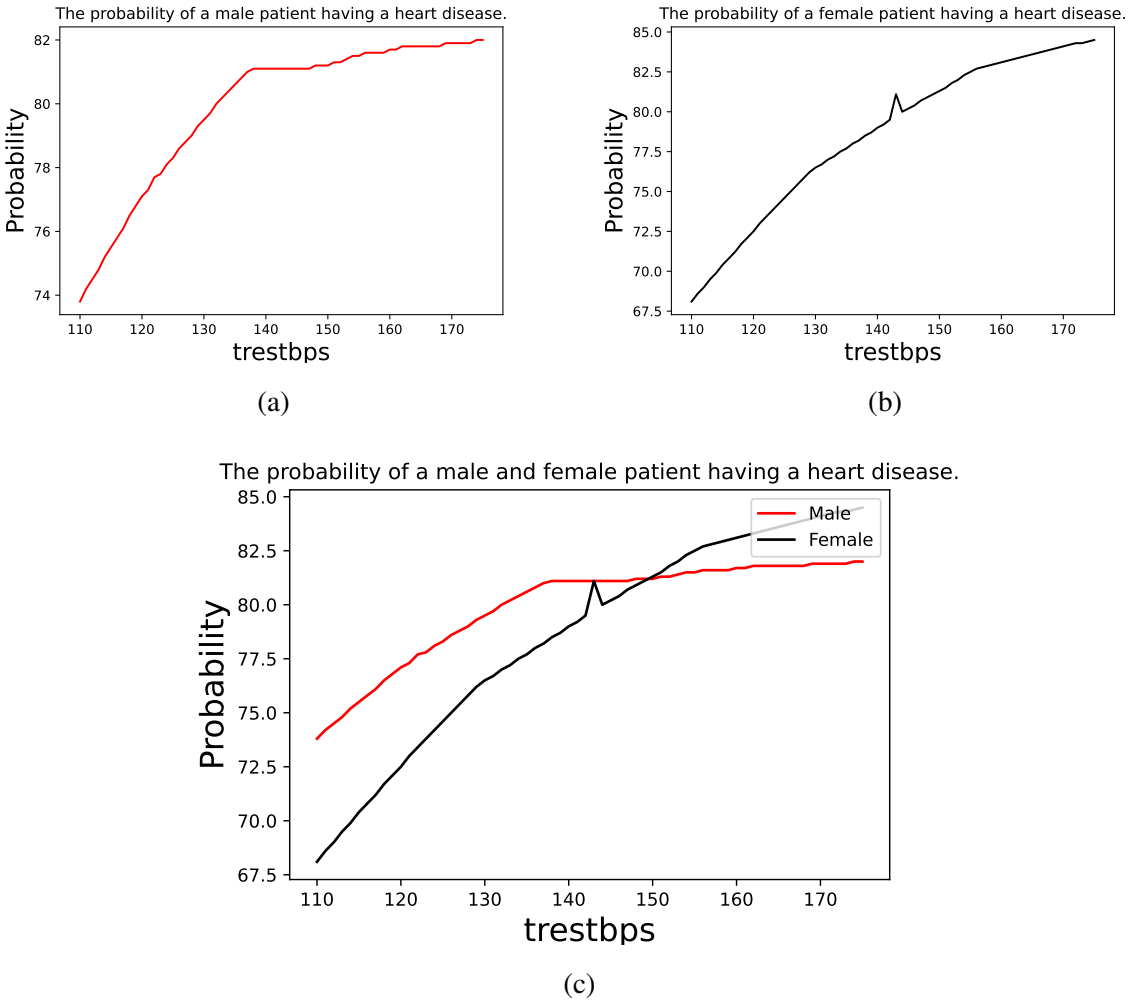
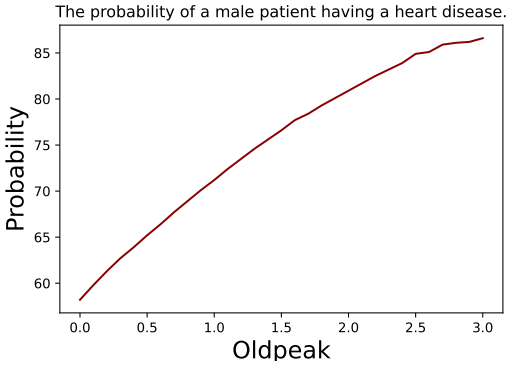


Figure 3.7: The probability of heart disease for (a) men (b) women and (c) the comparison between men and women as a function of trestbps (fasting blood pressure).

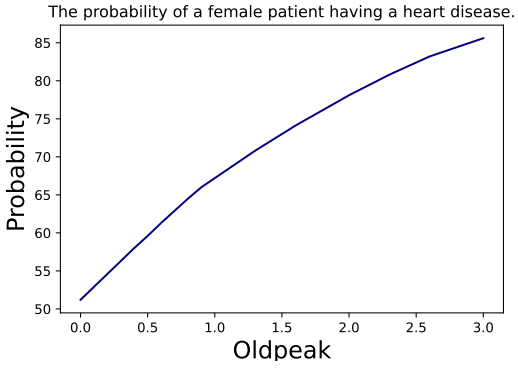
on the fasting blood pressure. The probability increases with increasing the blood pressure. The probability is higher for male people at lower blood pressure however the probability of female exceeds the probability of male with increasing blood pressure.

Figure [3.8] shows the probability of heart disease in men and women as a function of ST depression induced by physical activity compared to rest. Again, the probability increases with increasing integer value of the oldpeak. Moreover, the probability of heart disease is higher in men than in women. In the further analysis, the integer value of oldpeak 3 is used to calculate the probability of heart disease.

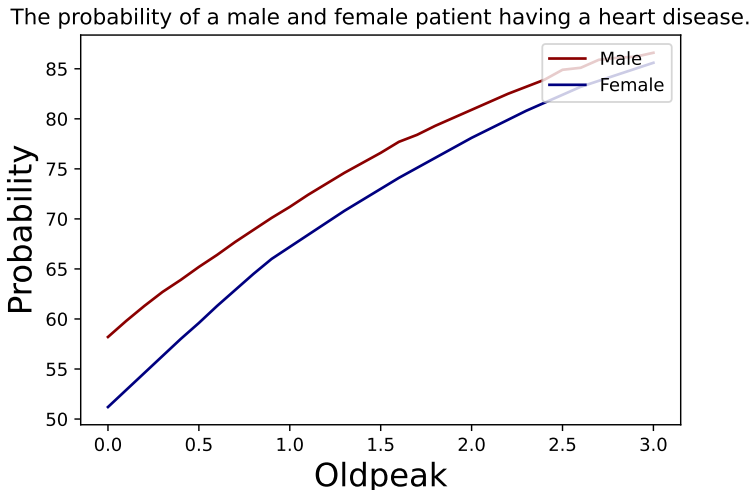
> “The pain or discomfort associated with angina usually feels tight, gripping, or pressing, and can vary from mild to severe. Angina is usually felt in the center of the chest, but can spread to one or both shoulders, the back, the neck, the jaw, or the arm. It may even be felt in the hands. Types of Angina a. Stable Angina / Angina Pectoris b. Unstable angina c. Variant (Prinzmetal) angina d. Microvascular angina. [9].”



(a)



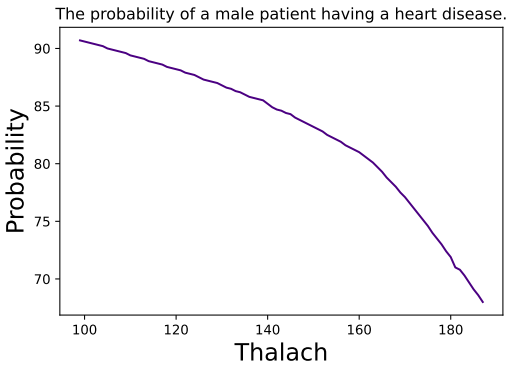
(b)



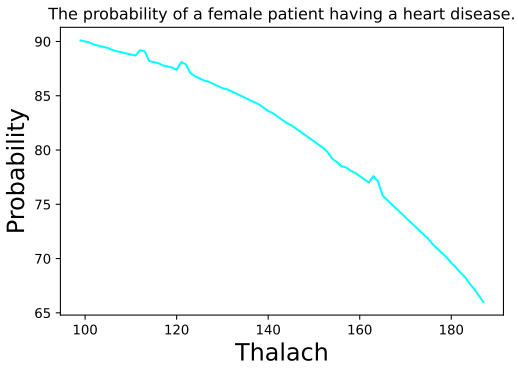
(c)

Figure 3.8: The probability of heart disease for (a) men (b) women and (c) the comparison between men and women as a function of oldpeak.

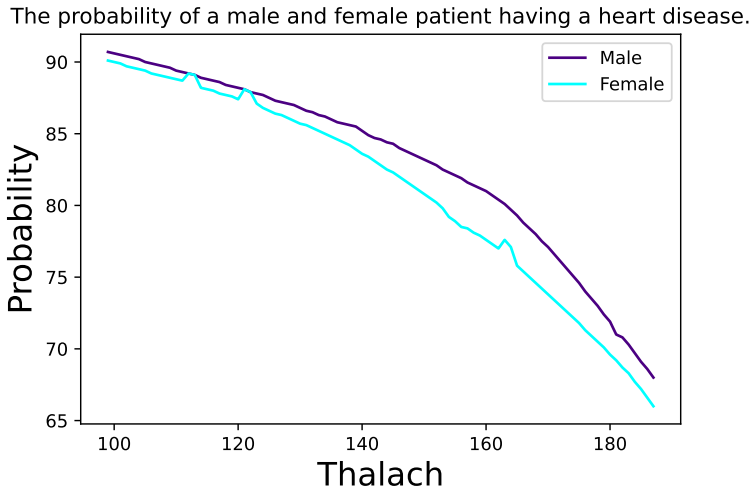
Figure [3.9] shows the predictive probability of heart disease for males and females and the comparison between the two in terms of maximum heart rate achieved. In this case, the prediction probability decreases with increasing maximum heart rate for both men and women. I have no idea why the probability of heart disease decreases as heart rate increases. It is expected to increase, but in this case the opposite is true.



(a)



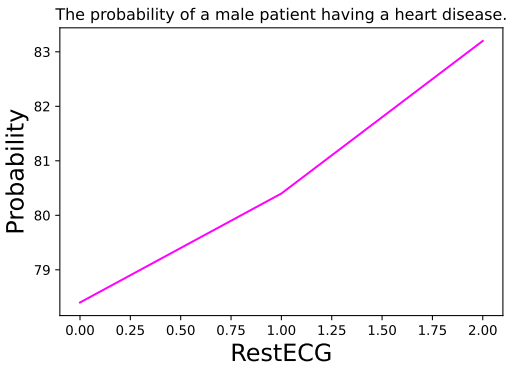
(b)



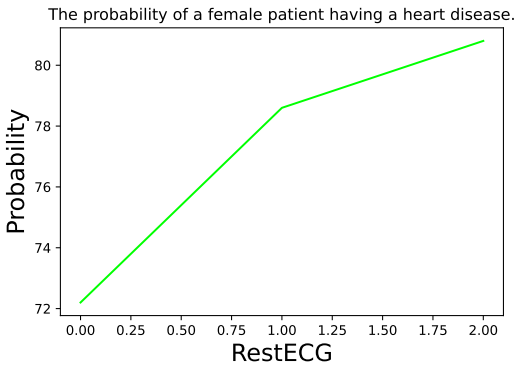
(c)

Figure 3.9: The probability of heart disease for (a) men (b) women and (c) the comparison between men and women as a function of thalach (maximum heart rate achieved).

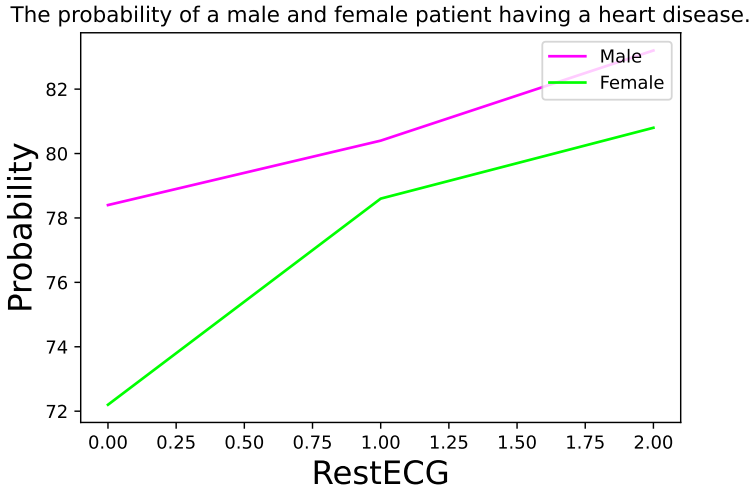
And finally, Figure [3.10] shows the probability of heart disease relative to the resting ECG. The resting ECG shows the results of the resting electrocardiogram: 0 = normal, 1 = with ST-T-wave abnormality, and 2 = left ventricular hypertrophy. The probability is lowest with a normal ECG and increases to left ventricular hypertrophy with ST-T-wave abnormality in both men and women. Also in this case, the probability of heart disease is higher in men than in women under different ECG conditions.



(a)



(b)



(c)

Figure 3.10: The probability of heart disease for (a) men (b) women and (c) the comparison between men and women as a function of RestECG.

Chapter 4

Results and Discussion

Three different computational intelligence techniques (SVM, LR, Random Forest) are used to measure accuracy. The dataset consisted of 303 samples. The dataset is divided into two categories, the training set and the testing or validation set. The accuracy of the validation set for SVM, LR and Random Forest are 75.82%, 78.02% and 74.72% respectively as shown in figure [4.1]. LR achieves 78.2% accuracy which is comparatively better than the other two. These accuracies or performance measures are calculated using the confusion matrix or error matrix. These types of computational intelligence techniques play an important role in medical diagnosis.

On the other hand, the probability of heart disease is calculated with respect to each categorical variable for men and women. From the analysis of each categorical variable, it is found that the probability of heart disease is always higher for males than females. This means that men are more likely to have cardiovascular disease after analyzing this data set. Going through the results one by one, the prediction probability of heart disease increases in men up to age 55 and in women up to age 60. After that, the prediction probability decreases with age. This result seems somewhat strange for application in daily life, but further analysis should be done with large data sets to determine the true prediction. Similarly, the predictive probability of heart disease in both sexes increases with increases in fasting blood pressure, exercise-induced depression compared to resting (Oldpeak) ST and resting ECG. In contrast to all of the above predictions, the predictive probability for the patient with heart disease decreases as heart rate increases. "However, an increase in heart rate of 10 beats per minute has been shown to be associated with an increase in the risk of cardiac death of at least 20%, and this increase in risk is similar to that observed with an increase in systolic blood pressure of 10 mm Hg" [10]. It has also been shown that the maximum heart rate measured in elderly men has

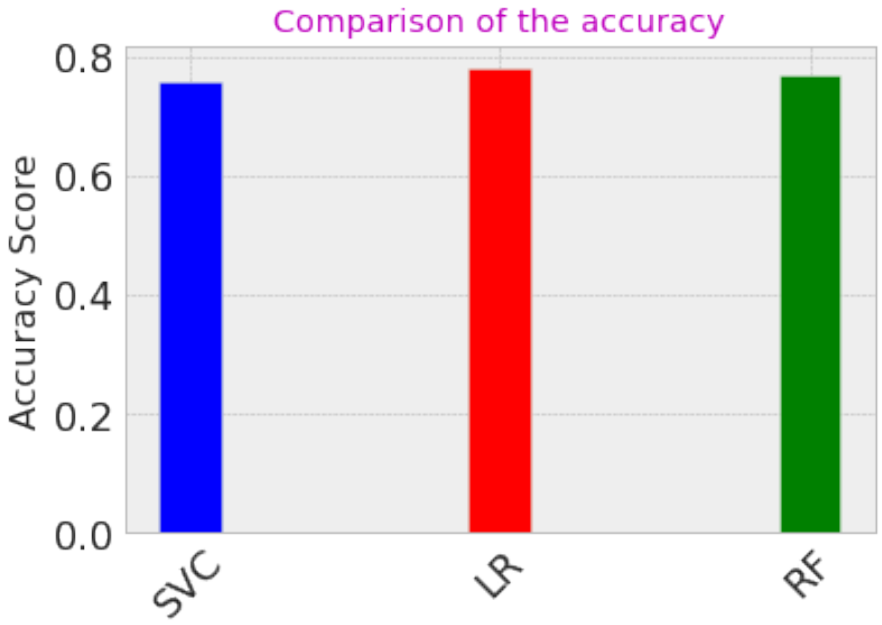


Figure 4.1: Accuracy for SVC, LR and RF

a strong predictive value in survival to a very old age.

Since cardiovascular diseases are complicated and thousands of people lose their lives every year. The early symptoms of heart disease should be treated in time to avoid drastic consequences. These computational intelligence techniques could be applicable and make the work of doctors easier and simpler. The accuracy of the model can be increased and also with some intelligence techniques coronary heart disease could be predicted. Since there is a variety of data and ways of data collection in the health sector, making the various sources available will help in better performance in gaining knowledge and clear understanding of the problem.

Bibliography

[1] Obenshain, M. K. Application of data mining techniques to healthcare data. *Infection Control and Hospital Epidemiology* **25**, 690–695 (2004). URL http://www.jstor.org/stable/10.1086/502460.

[2] Organization, W. H. *World health statistics 2018: monitoring health for the SDGs, sustainable development goals* (World Health Organization, 2018).

[3] Wang, H., Zheng, B., Yoon, S. W. & Ko, H. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research* **267**, 687–699 (2017).

[4] Shao, Y., Hou, C.-D. & Chiu, C.-C. Hybrid intelligent modeling schemes for heart disease classification. *Applied Soft Computing* **14**, 47–52 (2014).

[5] Ayon, S. I., Islam, M. M. & Hossain, M. R. Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. *IETE Journal of Research* **0**, 1–20 (2020). URL https://doi.org/10.1080/03772063.2020.1713916.

[6] Campanella, G. *et al.* Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. *Computerized Medical Imaging and Graphics* **65** (2017).

[7] Nguyen, H. & Bui, X.-N. Predicting blast-induced air overpressure: A robust artificial intelligence system based on artificial neural networks and random forest. *Natural Resources Research* 1–15 (2018).

[8] Chris. Using model.predict() with your TensorFlow / Keras model – (2021). URL https://www.machinecurve.com/index.php/2020/02/21/how-to-predict-new-samples-with-your-keras-model/.

[9] Heart disease prediction (rawat, s. 2021, june 28). https://towardsdatascience.com/heart-disease-prediction-73468d630cfc. Accessed: 2021-11-10.

[10] Perret-Guillaume, C., Joly, L. & Benetos, A. Heart Rate as a Risk Factor for Cardiovascular Disease. *Progress in Cardiovascular Diseases* **52**, 6–10 (2009). URL https://www.sciencedirect.com/science/article/pii/S0033062009000322.