DATA 201B Final Project
Run Yu & Yujia He
Heart Disease Prediction

INTRODUCTION

Heart disease is the leading cause of death in the United States. The frequency of this disease is increasing. Many factors will increase the risk for heart disease such as high cholesterol, high blood pressure, and diabetes. This study is going to use the heart disease data from UCI Machine Learning Repository to first provide an overview of current heart disease patterns and then use machine learning methods to make predictions of whether an individual has disease or not. We compare the methods will finally choose the best method which has highest prediction accuracy. This results of this study can help choose the suitable model to make prediction of the heart disease.

The dataset includes 303 samples. 13 attributes used in this dataset are the ones which will have effect on the heart disease. The attribute 'num' is the diagnosis of heart disease. The description of the attributes are as follows,

age: age in years
sex: sex(1 = male; 0 = female)
cp: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic
trestbps: resting blood pressure
chol: serum cholestoral
fbs: wheather fasting blood sugar > 120 mg/dl or not
-- Value 0: false
-- Value 1: true
restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality
-- Value 2: showing probable or definite left ventricular hypertrophy

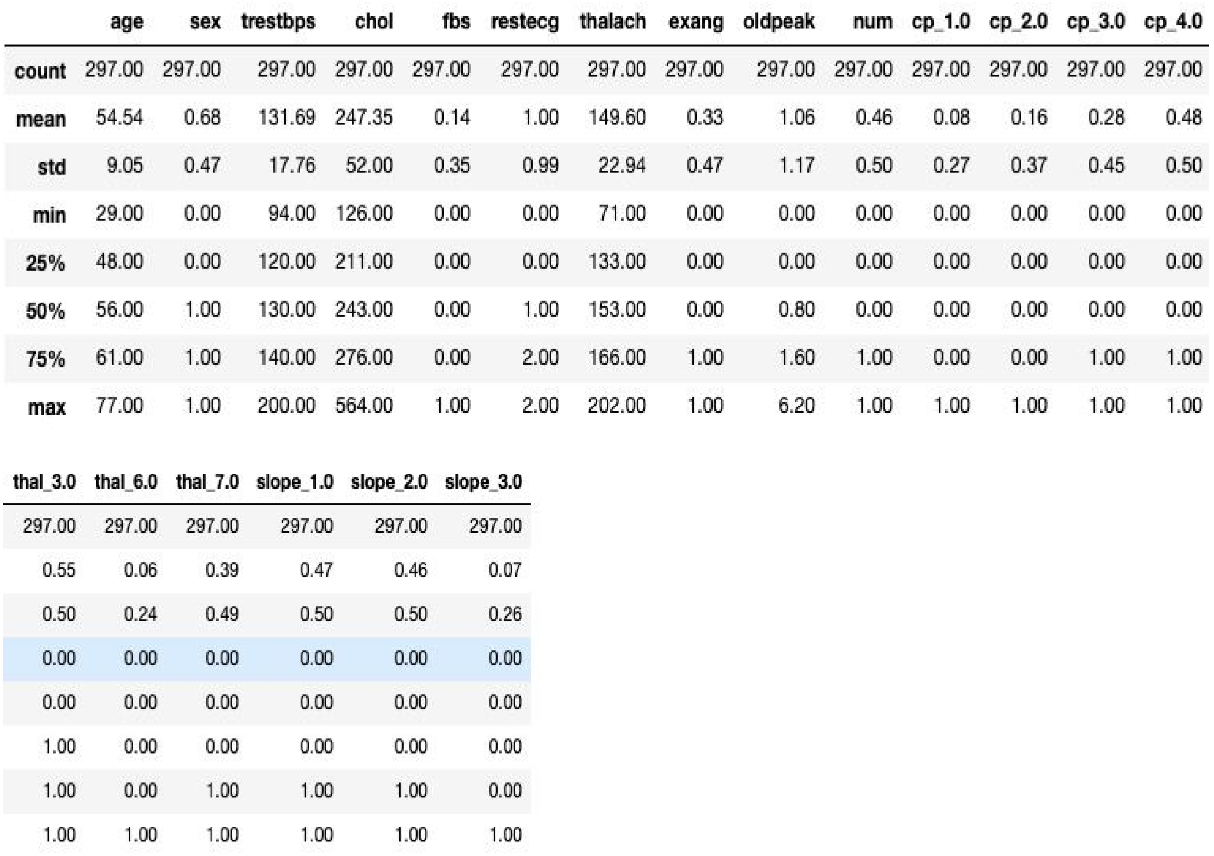
thalach: maximum heart rate achieved
exang: exercise induced angina (1 = yes; 0 = no)
oldpeak: ST depression induced by exercise relative to rest
slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping
ca: number of major vessels (0-3) colored by flourosopy
thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
num: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing

We assume that every value with 0 means heart is okay, and 1,2,3,4 means heart disease.

DATA CLEANING

When we retrieved data from UCI website, we checked the unique values for each variables and found some abnormal values. We masked those values to none values and drop them from our dataset dataframe. Instead of using all 303 samples, we eventually used 297 meaningful ones. Also, we re-defined the variable 'num' into a dummy variable by changing values of 2, 3, 4 into value of 1 since values of 1, 2, 3, 4 all mean having heart disease. Also, we created dummy variables for categorical variables including 'cp', 'thal' and 'slope' since we could use them to run our regression model.

VISUALIZATION



	age	sex	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	num	cp_1.0	cp_2.0	cp_3.0	cp_4.0	thal_3.0	thal_6.0	thal_7.0	slope_1.0	slope_2.0	slope_3.0
count	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00
mean	54.54	0.68	131.69	247.35	0.14	1.00	149.60	0.33	1.06	0.46	0.08	0.16	0.28	0.48	0.55	0.06	0.39	0.47	0.46	0.07
std	9.05	0.47	17.76	52.00	0.35	0.99	22.94	0.47	1.17	0.50	0.27	0.37	0.45	0.50	0.50	0.24	0.49	0.50	0.50	0.26
min	29.00	0.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	48.00	0.00	120.00	211.00	0.00	0.00	133.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50%	56.00	1.00	130.00	243.00	0.00	1.00	153.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
75%	61.00	1.00	140.00	276.00	0.00	2.00	166.00	1.00	1.60	1.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.00
max	77.00	1.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure1: Summary Statistics

According to summary statistics, we know that after adding the dummy variables, we have 20 variables with ‘num’ as a binary dependent variable. There are 4 out of 20 independent variables are continuous which are age (‘age’) from 29 to 77, resting blood pressure (‘trestbps’) from 94 to 200, serum cholesterol (‘chol’) from 126 to 564 and maximum heart rate (‘thalach’) achieved from 71 to 202.

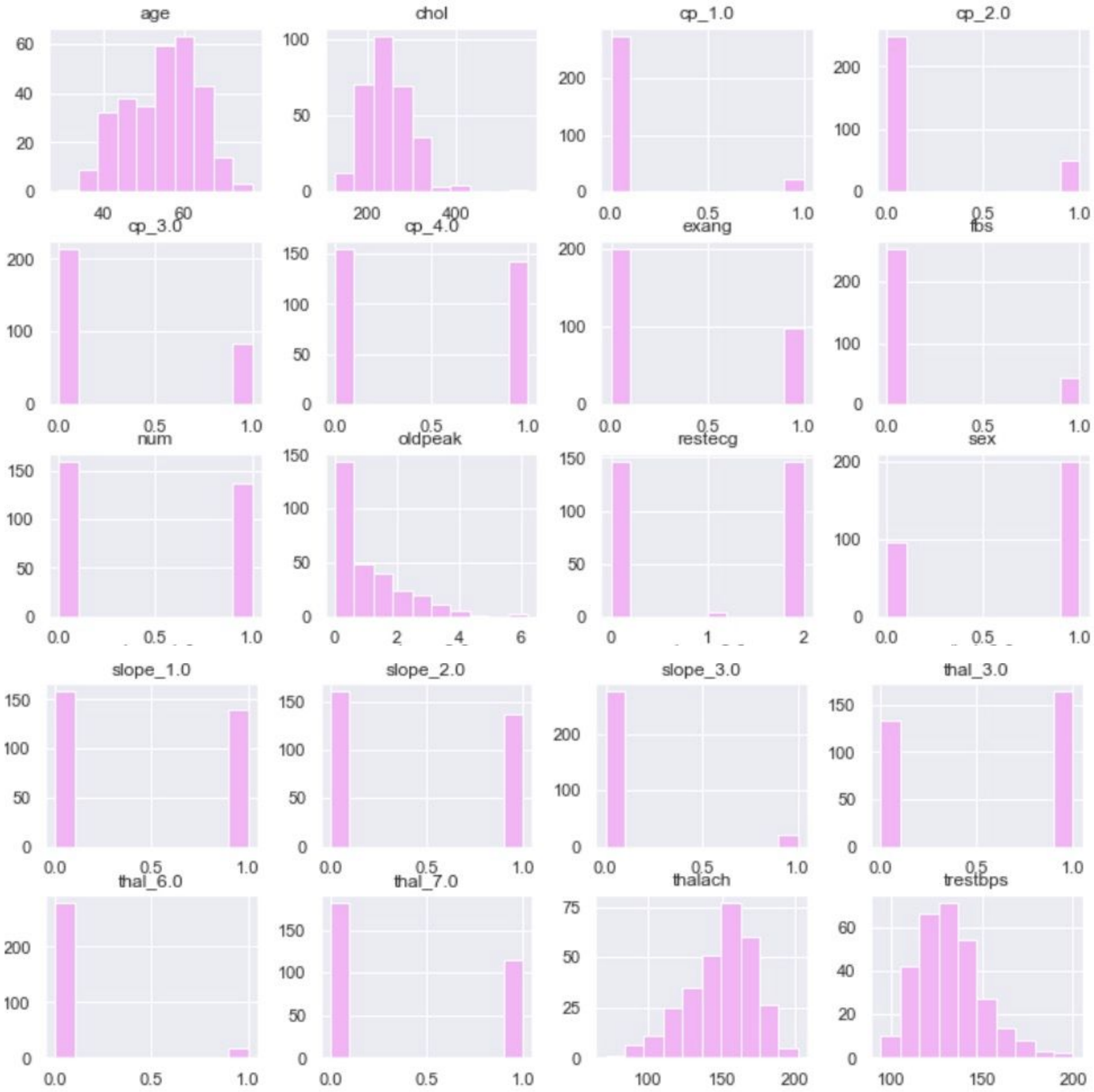
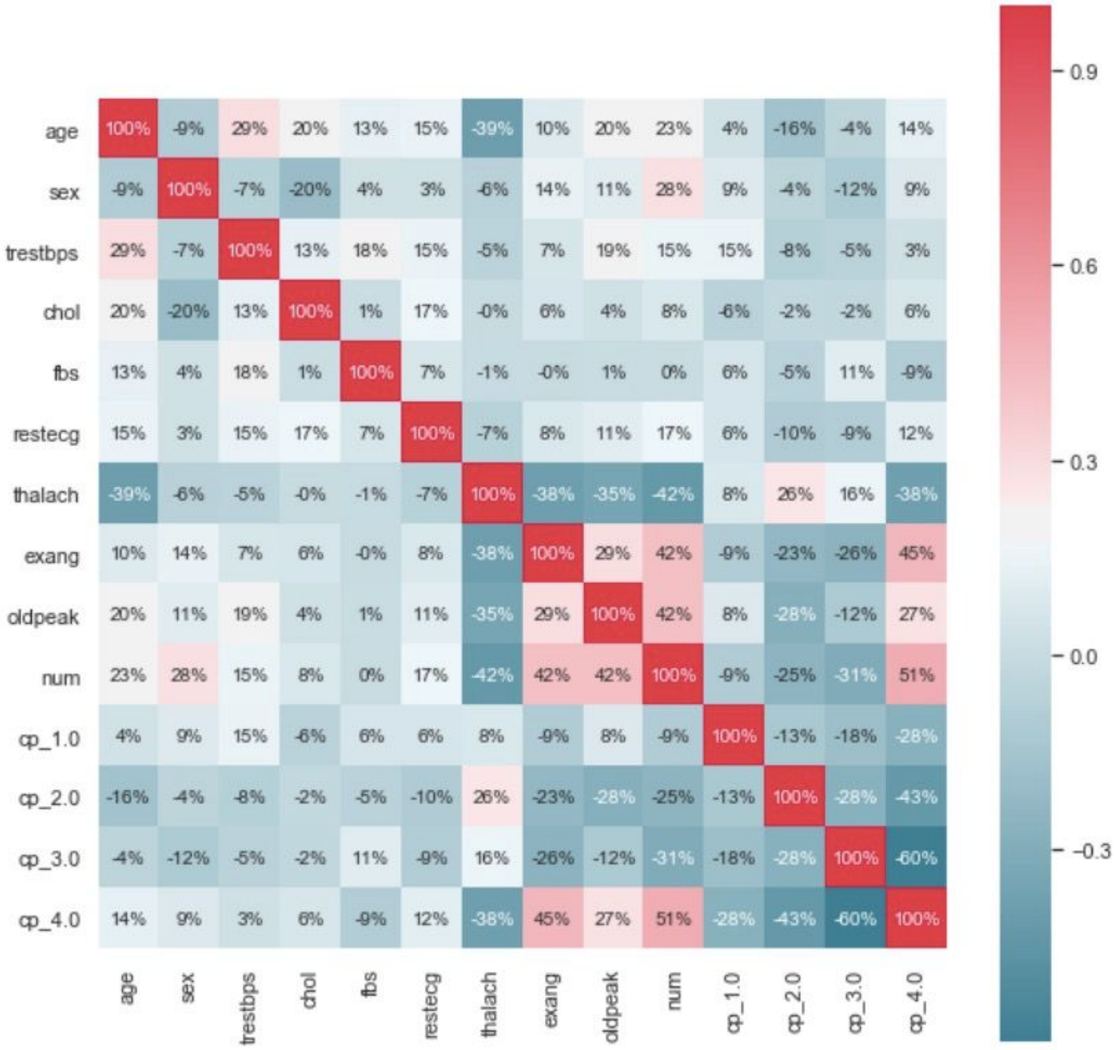


Figure2: Distribution of all variables

From the histogram showing the distribution of all variables, we can see all categorical variables are dummy variables. The distribution of variable 'thalach' and 'trestbps' are inverse. For dependent variable 'num' which represent the diagnostic result, the number of 0 (no heart disease) and 1 (heart disease) are similar, so we can confirm the data is balanced which would make the result more accurate.



	age	sex	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	num	cp_1.0	cp_2.0	cp_3.0	cp_4.0
age	100%	-9%	29%	20%	13%	15%	-39%	10%	20%	23%	4%	-16%	-4%	14%
sex	-9%	100%	-7%	-20%	4%	3%	-6%	14%	11%	28%	9%	-4%	-12%	9%
trestbps	29%	-7%	100%	13%	18%	15%	-5%	7%	19%	15%	15%	-8%	-5%	3%
chol	20%	-20%	13%	100%	1%	17%	-0%	6%	4%	8%	-6%	-2%	-2%	6%
fbs	13%	4%	18%	1%	100%	7%	-1%	-0%	1%	0%	6%	-5%	11%	-9%
restecg	15%	3%	15%	17%	7%	100%	-7%	8%	11%	17%	6%	-10%	-9%	12%
thalach	-39%	-6%	-5%	-0%	-1%	-7%	100%	-38%	-35%	-42%	8%	26%	16%	-38%
exang	10%	14%	7%	6%	-0%	8%	-38%	100%	29%	42%	-9%	-23%	-26%	45%
oldpeak	20%	11%	19%	4%	1%	11%	-35%	29%	100%	42%	8%	-28%	-12%	27%
num	23%	28%	15%	8%	0%	17%	-42%	42%	42%	100%	-9%	-25%	-31%	51%
cp_1.0	4%	9%	15%	-6%	6%	6%	8%	-9%	8%	-9%	100%	-13%	-18%	-28%
cp_2.0	-16%	-4%	-8%	-2%	-5%	-10%	26%	-23%	-28%	-25%	-13%	100%	-28%	-43%
cp_3.0	-4%	-12%	-5%	-2%	11%	-9%	16%	-26%	-12%	-31%	-18%	-28%	100%	-60%
cp_4.0	14%	9%	3%	6%	-9%	12%	-38%	45%	27%	51%	-28%	-43%	-60%	100%

Figure3: Correlation matrix

As the correlation matrix,Value 4: asymptomatic of cp has largest correlation value with diagnostic ‘num’ , the next are ‘exang’ and ‘ddpeak’. And only two absolute correlation values between feature ‘num’ and ‘cp_4.0’(0.51),’cp_3.0 and ’cp_4.0’(-0.6) are higher than 0.5, so multicollinearity is not a problem for these data.

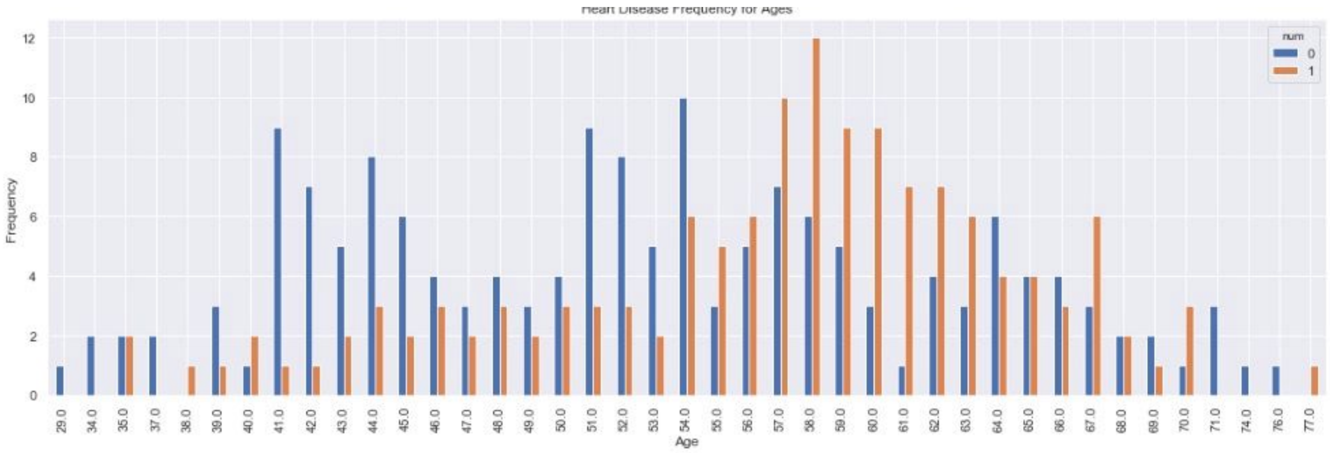


Figure4: Heart disease VS age population proportion

In real life, people may think that the older people has the higher risk of heart disease. According to diagnostic result for different age people, we can see the heart disease population proportion(1-‘heart disease’) appears large for age between 57-63,67,70,77. For the rest of age period, the heart disease proportion is low. Amount all age levels, age of 61 group has the highest heart disease proportion than other age’s. In general, from age 57-60 has the highest frequency of heart disease diagnostics.

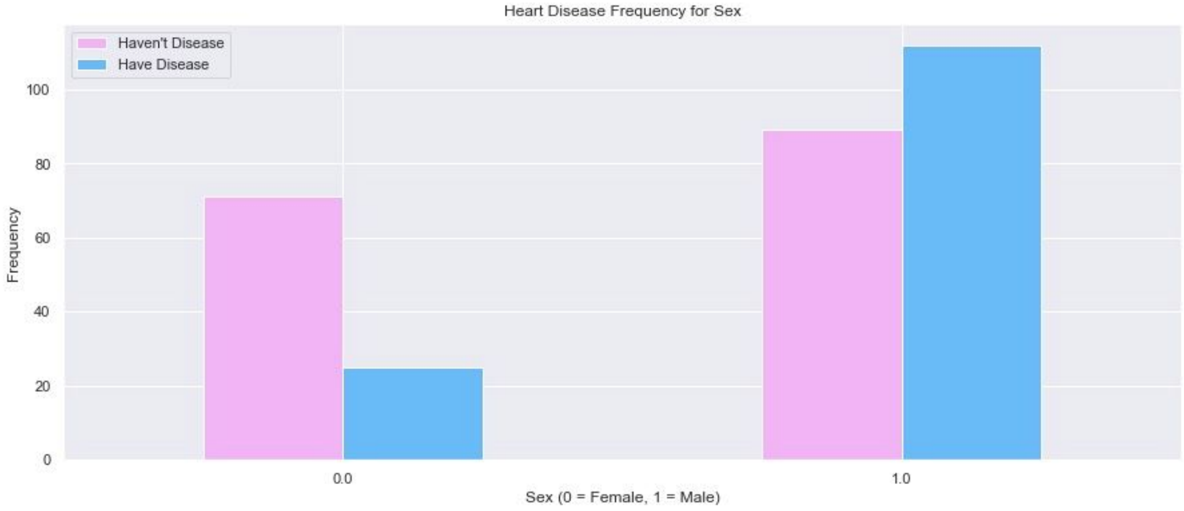


Figure5: Heart disease VS sex population proportion

As for gender feature, apparently, male has the highest the frequency to be diagnostics while female has much lower frequency of heart disease. Also the number of no heart disease men and women are relevantly similar.

METHODS

Method 1: Logistic Regression

The logistic regression is a supervised classification algorithm and is used for binary classification. It models the probability of the default class. A binary logistic regression describes the relationship between the dependent binary variable and other independent variables. In our study, ‘num’ is our dependent binary variable indicating whether having heart disease or not.

The key representation in logistic regression are the coefficients. The coefficients of logistic regression is estimated from training data using maximum-likelihood estimation. The better the coefficients, the higher the value for the default class will be predicted.

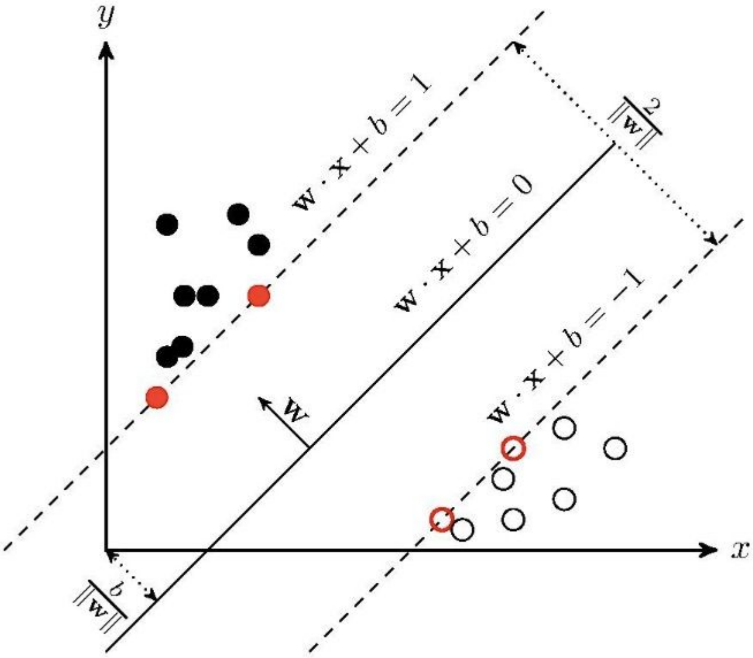
Result:

Test Accuracy of logistic regression algorithm is 81.67%.

The accuracy using logistic regression model is kind of high. We assume the reason to be that logistic regression assumes no error in output variables. Also, we do not have highly-correlated inputs in our case. We then perform other models to see whether there is other model which can have higher accuracy.

Method 2: SVM

Firstly we performed linear SVM which is a classification method, the goal is to find the optimal hyperplane in the N-dimensional space where N is the number of features to maximizes the distance between the classes and distinctively classifies the dataset. Where support vector are the red point which is closest to hyperplane and can influence the direction and position of hyperplane.



y
x
$\mathbf{w}\cdot\mathbf{x}+b=1$
$\mathbf{w}\cdot\mathbf{x}+b=0$
$\mathbf{w}\cdot\mathbf{x}+b=-1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$

Figure6

As this picture shows that Wx+b=0 is the hyperplane, we want to maximize the distance 2/||w||. In other word to minimize 1/2 ||w||^2. By utilizing SVM we fit the SVM model using train data then extract the features from the test data and predict values. We get predict accuracy by extract the parameter in SVM model: svm.score.

Secondly, in order to optimize the accuracy of SVM model, we perform PCA for dimension reduction, combining 20 features to 7 principle components. The split train and test data based on new x variables.

<u>Result</u>:

Figure shows the scatter plot of dependent variables y and 2D projection of 20 independent variables x. The green point represents diagnostic result which is having heart disease and red point is the result of no heart disease.

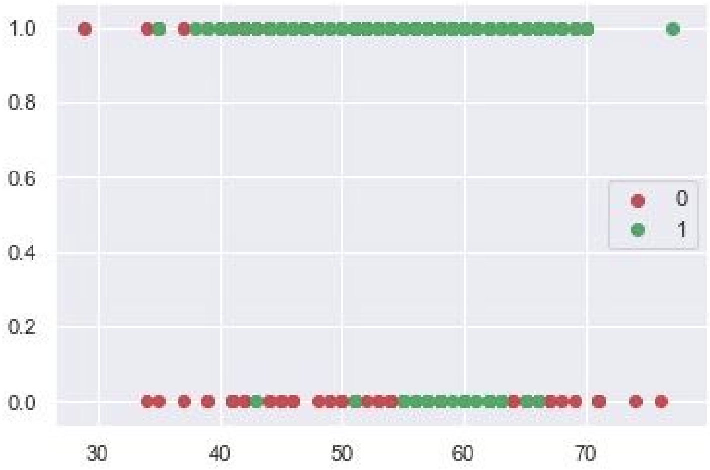


Figure7:Scatter plot of dependent variables y

Model 1: Test Accuracy of RBF SVM Algorithm: 50.00%

Model2: Test Accuracy of Linear SVM Algorithm: 81.67%

As we can see from the SVM model results, the test accuracy of SVM using radius basis function is low. We try to find method to optimize the accuracy, perform PCA to reduce variable dimensions to 7. Also these 7 dimensions are projected to 2D flat. The following is the scatter plot of PCA components and y.

Method 3: PCA

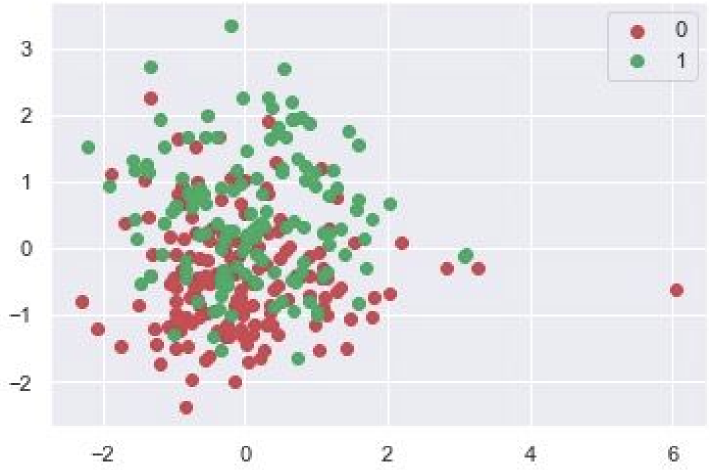


Figure8: Scatter plot of PCA components and y

Result:

PCA-SVM model1: Test Accuracy of linear SVM-PCA model: 85.00%

PCA-SVM model2: Test Accuracy of rbf SVM-PCA Algorithm: 85.00%

After utilizing the dimension reduction method PCA, the accuracy of both linear SVM and RBF SVM are improved to 85%.

Method 4: Decision Tree

Decision tree is a non-parametric supervised learning method used for classification which cannot affected by probability distribution assumption. The classes for heart disease dataset is 'heart disease' and 'no heart disease' .Decision trees can handle high dimensional data with good accuracy. The goal of this method is to create a model that can predict the diagnostic result. The reason why we use decision tree is that : Decision tree mimic how human thinking so it's simple for us to understand the data and interpret data easily. Furthermore, decision trees actually show you the logic for the data to interpret, not like SVM a kind of black box algorithms.For heart disease example, if we classify the diagnostic result of heart disease and the decision tree graph display how to make the decision clearly.

As we can see the decision tree graph, where each node represents an attribute and branch or link represents a decision and each leaf is an outcome of a categorical value. The idea of decision is to create a tree for dataset and minimize the error in every leaf. The process to utilize decision tree is as following:

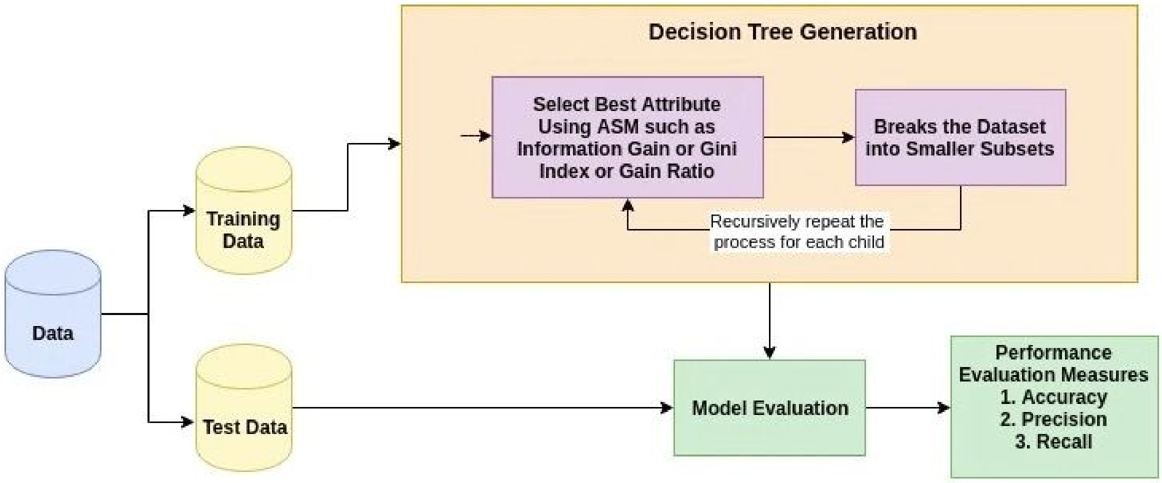

Decision Tree Generation
Select Best Attribute Using ASM such as Information Gain or Gini Index or Gain Ratio
Breaks the Dataset into Smaller Subsets
Recursively repeat the process for each child
Training Data
Data
Test Data
Model Evaluation
Performance Evaluation Measures
1. Accuracy
2. Precision
3. Recall

Figure 9

- Select the best attribute according to Attribute Selection Measures(ASM) to split records
- Make the best attribute a decision node and then breaks data into subsets
- Start repeat above step until one of conditions matches:
    1) no remaining features
    2) all tuples are the same attribute value

<u>Result</u>:

Model1: Decision Tree Test Accuracy 70.00%

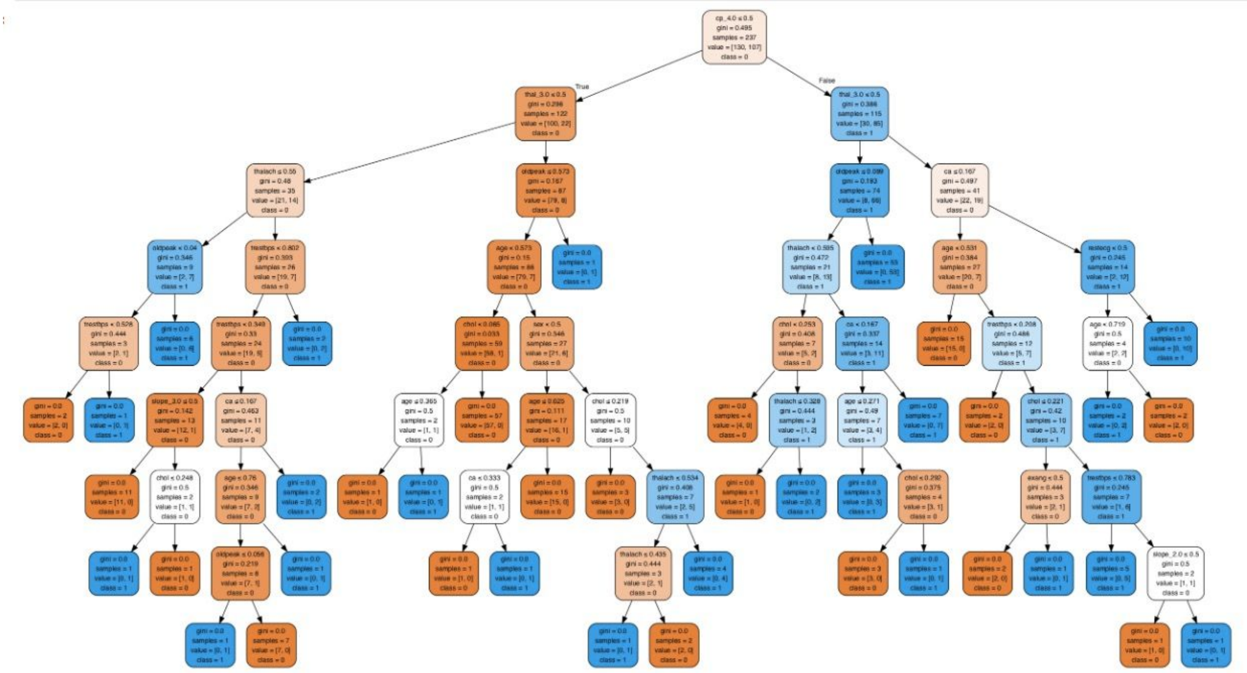


Figure10: Unpruned tree

Here, this tree is unpruned. This unpruned tree is unexplainable and have many branches. Furthermore the accuracy rate of non-pruning tree is low so we then optimized it by using pre-pruning trees.The deeper the tree, the model is more overfitted.

Optimal Model2: Pre_Pruning Decision Tree Test Accuracy 76.67%

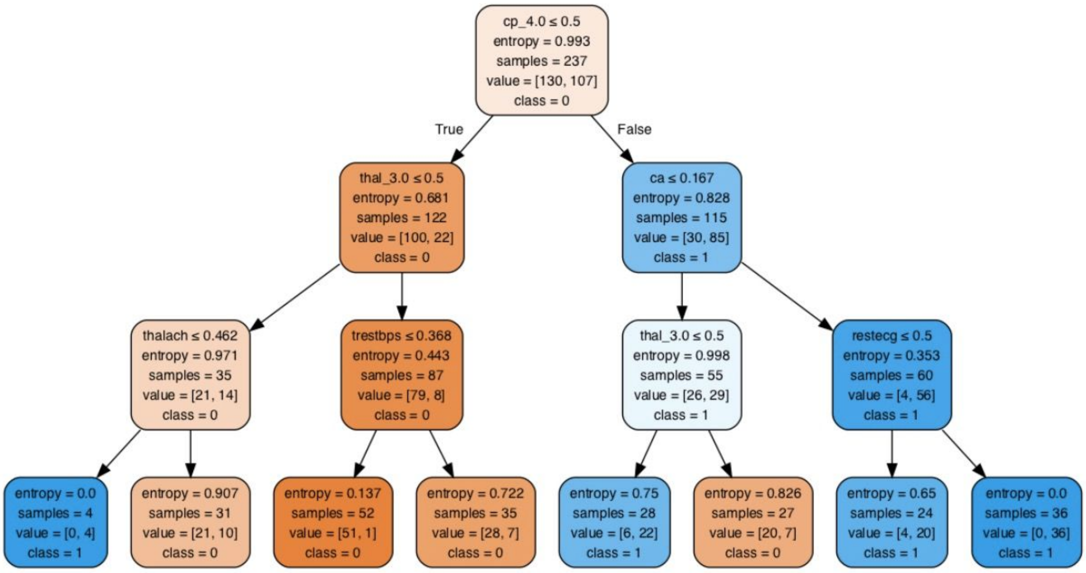


Figure11: Pre_Pruning Decision Tree

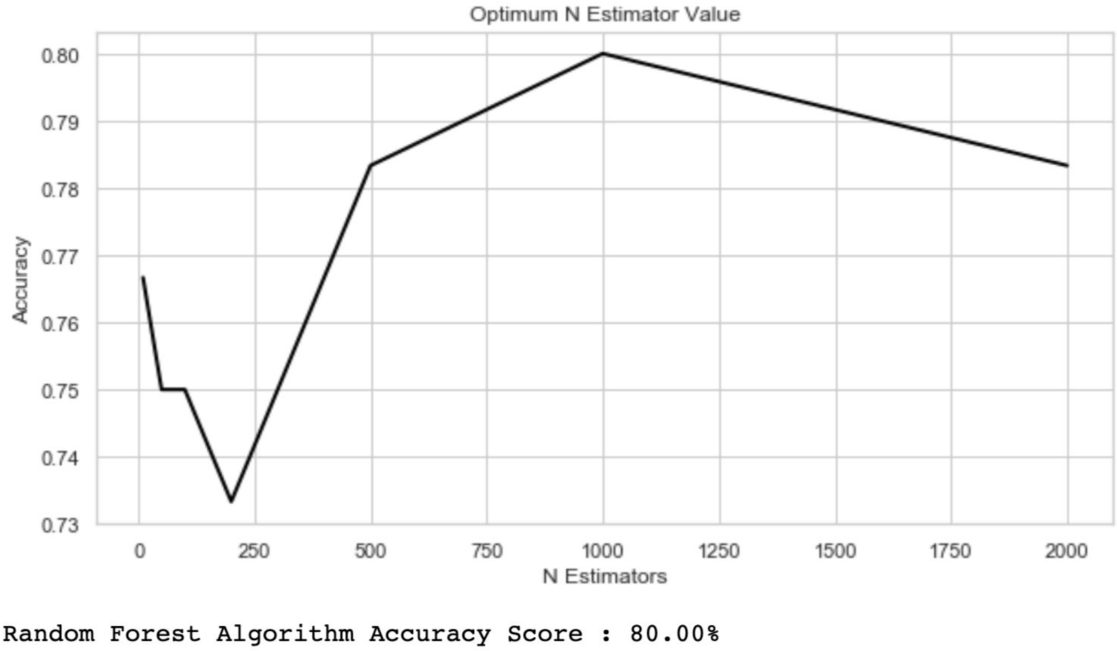
Also decision trees are biased with imbalance dataset, but fortunately our data is balanced. Decision tree is sensitive to noisy data thus, it can overfit noisy data. And the small variance in data can result in the different decision tree. This can be reduced by random forest.So we then consider to use random forest.

Method 5: Random Forest

Due to the overfitting problem of the decision tree, we combine many decision trees into a single model which is random forest. In this way, the depth of the tree which will reduce the variance and increase the bias is limited. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

At the training time, each tree learns from a random sample of dataset and is trained on different samples. Also, the samples are drawn with replacement which is bootstrapping.

At the test time, final predictions are made by averaging the predictions of each decision tree.



```
Random Forest Algorithm Accuracy Score : 80.00%
```

Figure11:Number of estimators using for loop

Result:

Random Forest Test Accuracy 80.00%

We first look for the optimal number of estimators using for loop. From the graph, we can see the accuracy reaches to the highest when the number of estimators is 1000. In our study, we then choose 1000 estimators in order to have the highest accuracy. The accuracy we get from decision tree is 70% which is lower than the accuracy using random forest. It thus been proved that the random forest method can improve the accuracy in our case.

MODEL COMPARISION

We performed model comparison to compare the predictive ability of all 8 models we get. According to figure(12), RBF SVM model has the least accuracy while the combination of PCA and linear/RBF model have the highest accuracy (85%), Logistic regression model, linear SVM model, random forest model have about 80% accuracy. As Figure(13) shows the confusion matrix of 8 models, PCA+SVM RBF models show that there are 28 out of 30 cases are predicted no heart disease, while 27 out of 30 cases are predicted heart disease. And PCA+SVM linear models successfully predicted 29 out of 30 no heart disease, can 22 out of 30 heart disease result. As to pre-pruning decision tree predicted 27 out of 20 no heart disease which is more than non-pruning decision tree, however, pre-pruning and non-pruning decision tree predicted similar number of heart disease diagnostic result.

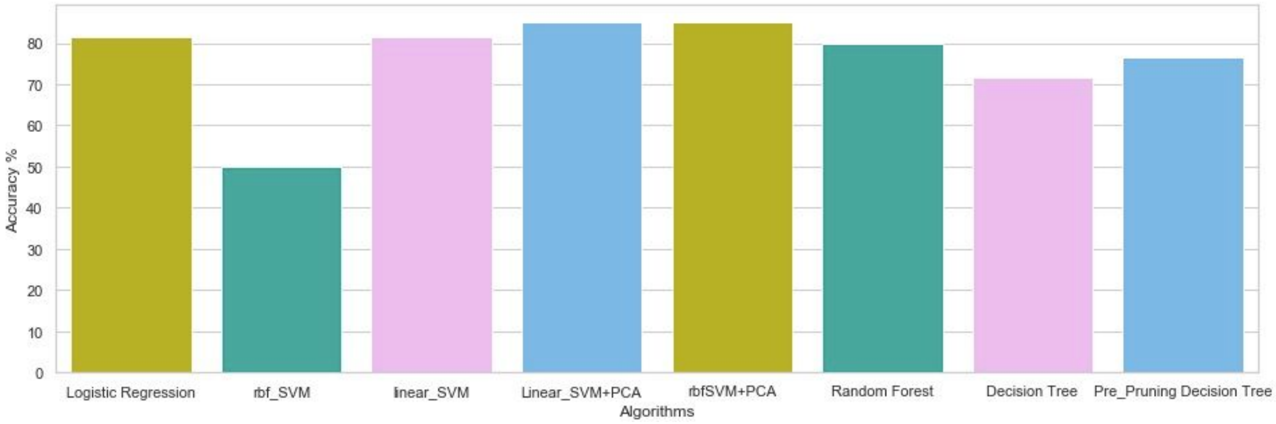
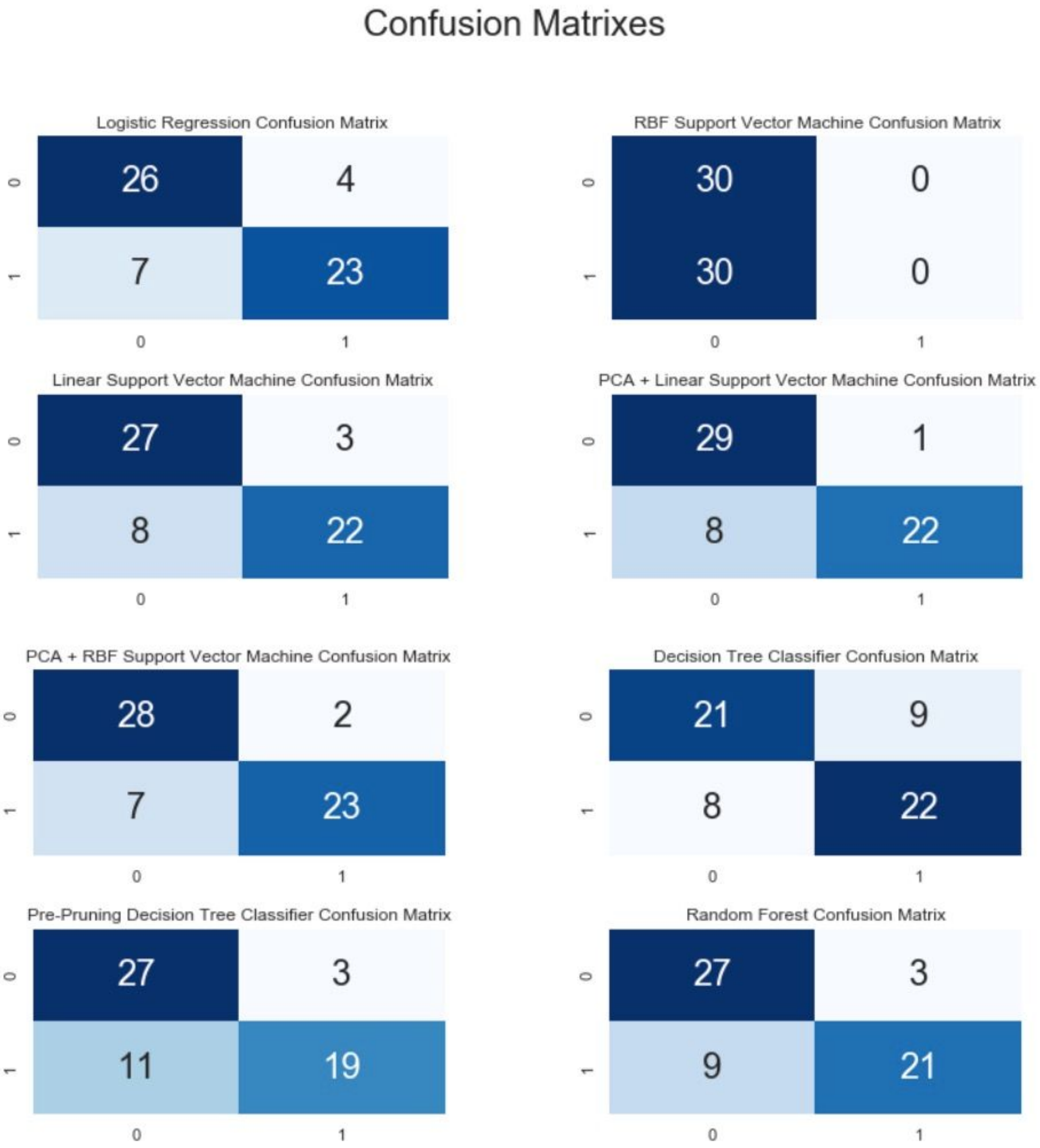


Figure12: Accuracy Histogram



Confusion Matrixes

Logistic Regression Confusion Matrix

	0	1
0	26	4
1	7	23

RBF Support Vector Machine Confusion Matrix

	0	1
0	30	0
1	30	0

Linear Support Vector Machine Confusion Matrix

	0	1
0	27	3
1	8	22

PCA + Linear Support Vector Machine Confusion Matrix

	0	1
0	29	1
1	8	22

PCA + RBF Support Vector Machine Confusion Matrix

	0	1
0	28	2
1	7	23

Decision Tree Classifier Confusion Matrix

	0	1
0	21	9
1	8	22

Pre-Pruning Decision Tree Classifier Confusion Matrix

	0	1
0	27	3
1	11	19

Random Forest Confusion Matrix

	0	1
0	27	3
1	9	21

Figure 13:Confusion Matrix