Introduction

In this report, we propose two frameworks from analyzing a dataset containing features that our models will use to accurately predict heart disease. We will compare the usage of a Support Vector Machine model and using PCA to predict with as well. Both models used a 10-fold cross validation as part of one of the tuning parameters in the fitted models, and an 80/20 train-test split.

Data

A snippet of the features contributing to the prediction of heart disease are listed below:

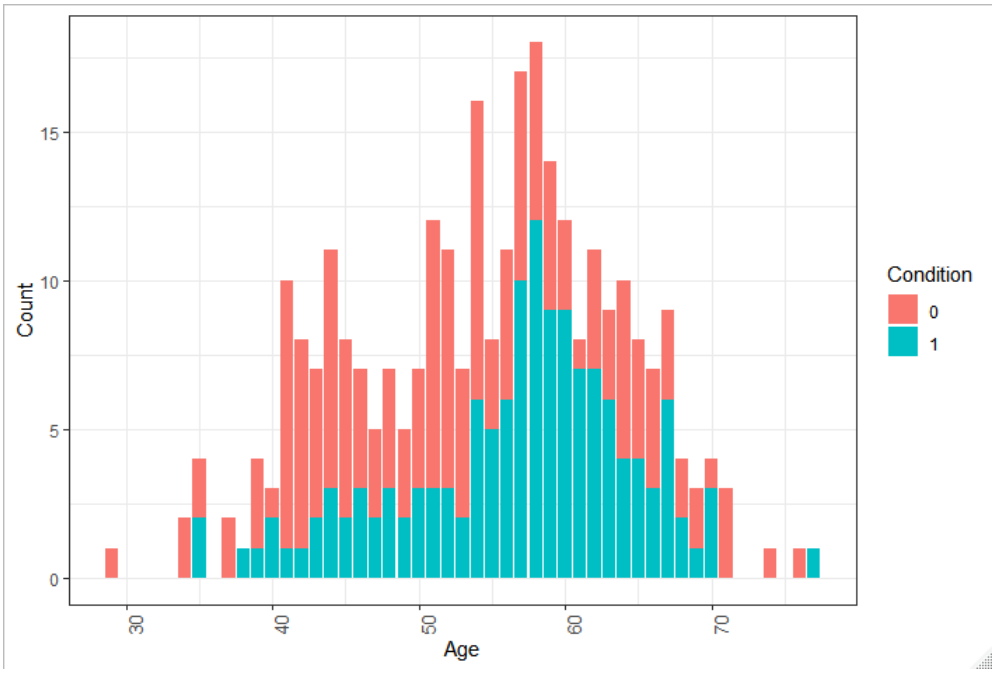
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	condition
69	1	0	160	234	1	2	131	0	0.1	1	1	0	0
69	0	0	140	239	0	0	151	0	1.8	0	2	0	0
66	0	0	150	226	0	0	114	0	2.6	2	0	0	0
65	1	0	138	282	1	2	174	0	1.4	1	1	0	1

Unlike age, some of these variables are less intuitive due to their abbreviations. Here is a list of the abbreviations and their meanings:

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 (in mm Hg on admission to the hospital)
chol: serum cholestoral in mg/dl
fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
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. Hemoglobin levels.

“First Look” at the data

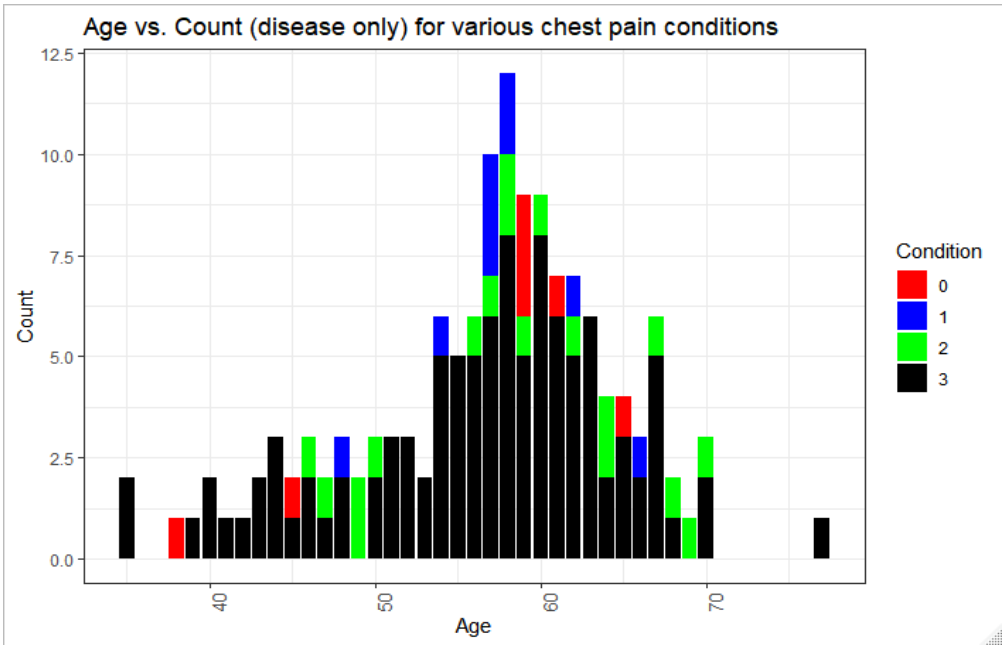
In order to give some insight to how our data is spread out, below is a graph displaying the heart disease distribution varied by age:



0’s -no disease
1’s - disease.

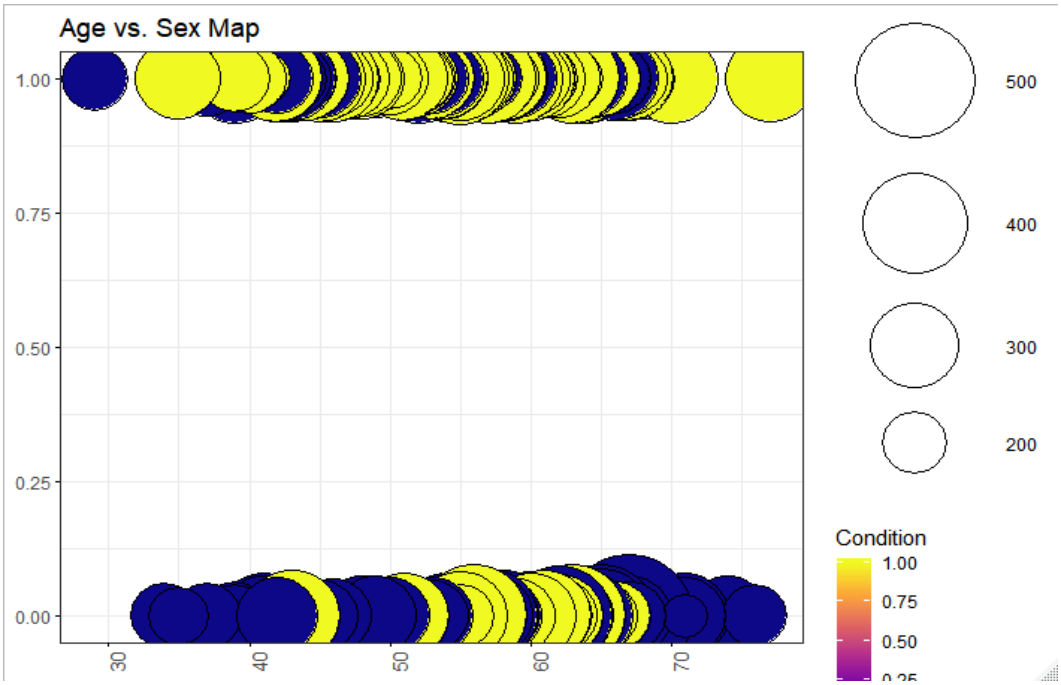
We see middle-age has the highest spike of patients with heart disease.

The plot below might give an indication of the type of chest pain(cp) as being an important factor to the prediction of heart disease amongst ages:



chest pain condition:
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic

The plot below is displaying age(x-axis) and sex(y-axis), size of circle is the cholesterol level, and condition is the color:



0 – female
1 – male
Yellow- heart disease
Blue- no disease

Seems like females have a higher variance in terms of cholesterol levels, but men have the majority of the patients with heart disease

Model Comparison

Each of the models had a training set, test set, and 10-fold cross validation was used to create the model, as well as a tuning parameter for the cost.

SVM:

SVM-Type: C-classification
SVM-Kernel: linear
Cost: 0.1

Number of Support Vectors: 102
After running the SVM model there was a balanced accuracy of 85.97%

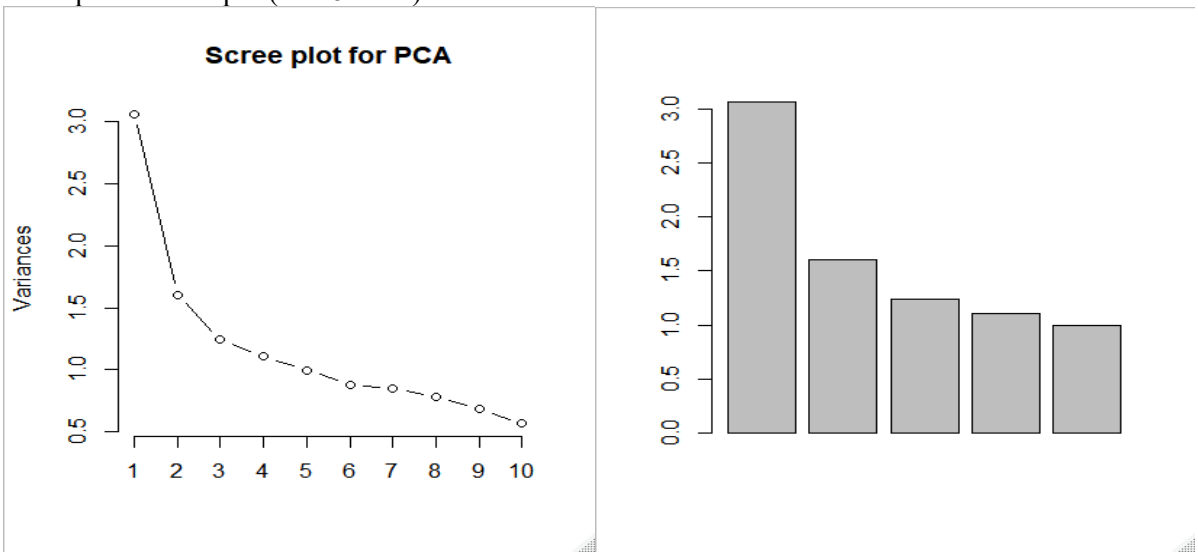
	Actual – 0	Actual – 1
Prediction – 0	31	5
Prediction - 1	3	21

Accuracy : 0.8667
95% CI : (0.7541, 0.9406)
P-Value [Acc > NIR] : 5.858e-07
Mcnemar's Test P-Value : 0.7237

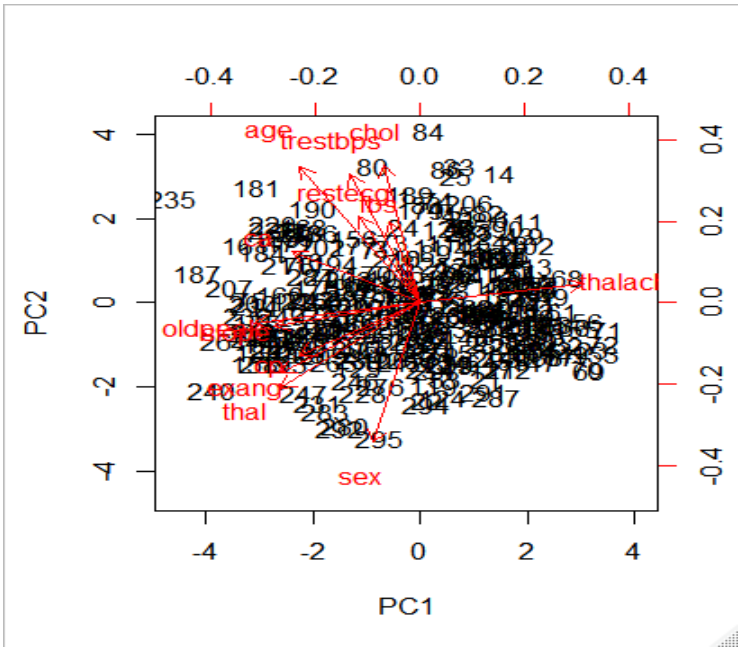
Sensitivity : 0.9118
Specificity : 0.8077
Pos Pred Value : 0.8611
Neg Pred Value : 0.8750
Prevalence : 0.5667
Detection Rate : 0.5167
Detection Prevalence : 0.6000
Balanced Accuracy : 0.8597

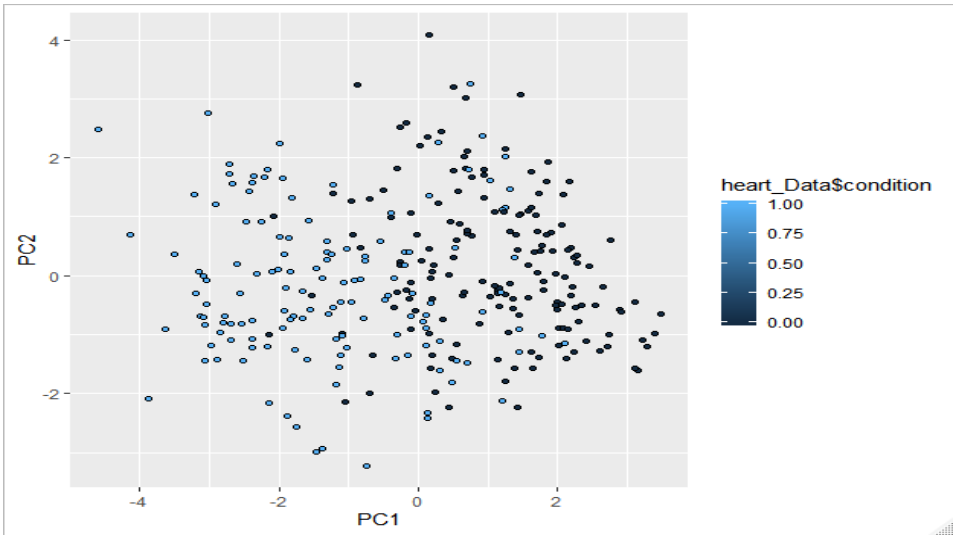
PCA:

Scree plot and bar plot(first 5 PC's)



The plot below is showing a biplot of the first 2 PC's. Majority of the variables are negatively correlated to PC1 except for the thalach(maximum heart rate achieved). For example, as an observation has a higher value for PC1, it has an extremely low oldpeak, exercise induced angina(exang), and hemoglobin levels(thal), but as an obersation has a higher value for PC2, it has a very high age and cholesterol level.





The plot above is a clearer pictured of the first two PC's. It is showing the same thing, but as well as the condition where a patient had heart disease or not(0 – no disease, 1 – heart disease). As we recall from the first plot, maximum heart rate achieved(thalach) was the only positively correlated feature as PC1 increases, and it interesting to point out that that is where majority of the non-heart diseased patients land for these PC's.

PCA Parameters:

SVM-Type: C-classification
SVM-Kernel: linear
Cost: 0.1
Number of Support Vectors: 114

Overall accuracy for the PCA model was 87.44%

	Actual – 0	Actual – 1
Prediction – 0	32	5
Prediction - 1	2	21

Accuracy : 0.8833
P-Value [Acc > NIR] : 1.119e-07
Mcnemar's Test P-Value : 0.4497

Sensitivity : 0.9412
Specificity : 0.8077
Pos Pred Value : 0.8649
Neg Pred Value : 0.9130
Prevalence : 0.5667
Detection Rate : 0.5333
Detection Prevalence : 0.6167
Balanced Accuracy : 0.8744

Conclusion

The best model for this experiment was the SVM with PCA with a balanced accuracy of 87.44% compared to 85.97%. So, it showed a slight improvement to the model after implementing the first 5 PC's into the Support Vector Machine.
Below is a plot of the SVM to show its decision boundaries for the first 2 PC's: