Data Analysis on Heart Disease Data Set

CSE 5160, Summer 2022
Group 8 Project Presentation

Hirpara, Dhaval Chaturbhai
Inzunza, Kevin
Eddala, Likhitha Reddy

Introduction of Machine Learning

> What is Machine Learning?

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

> Why is machine learning important?

Data-driven decisions increasingly make the difference between keeping up with competition or falling further behind. Machine learning can be the key to unlocking the value of corporate and customer data and enacting decisions that keep a company ahead of the competition.

Background

> It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors.

> Among various life threatening diseases, heart disease has garnered a great deal of attention in medical research.

> The diagnosis of heart disease is a challenging task, which can offer automated prediction about the heart condition of patient so that further treatment can be made effective.

> The diagnosis of heart disease is usually based on signs, symptoms of the patient.

> There is one attribute we are particularly interested in and that’s exercise induced angina (Chest pain or tightness when exercising) as it a common symptom of Congenital Heart Disease (CHD) .

Background of Original Study

We are exploring a dataset donated by David W. Aha to the University of California Irvine Machine Learning Repository. This directory contains 4 databases concerning heart disease diagnosis. All attributes are numeric-valued. The data was collected from the four following locations:

1. Cleveland Clinic Foundation (cleveland.data)
2. Hungarian Institute of Cardiology, Budapest (hungarian.data)
3. V.A. Medical Center, Long Beach, CA (long-beach-va.data)
4. University Hospital, Zurich, Switzerland (switzerland.data)

Each database has the same instance format. While the databases have 76 raw attributes, only 14 of them are actually used.

Research Question

Among the 303 diagnosed in Cleveland with some degree of heart disease, is exercise induced angina an attribute that can be predicted? Which model is the most suitable, with a high enough accuracy to detect a patient with induced angina during exercise with non-fatal symptoms?

Solution

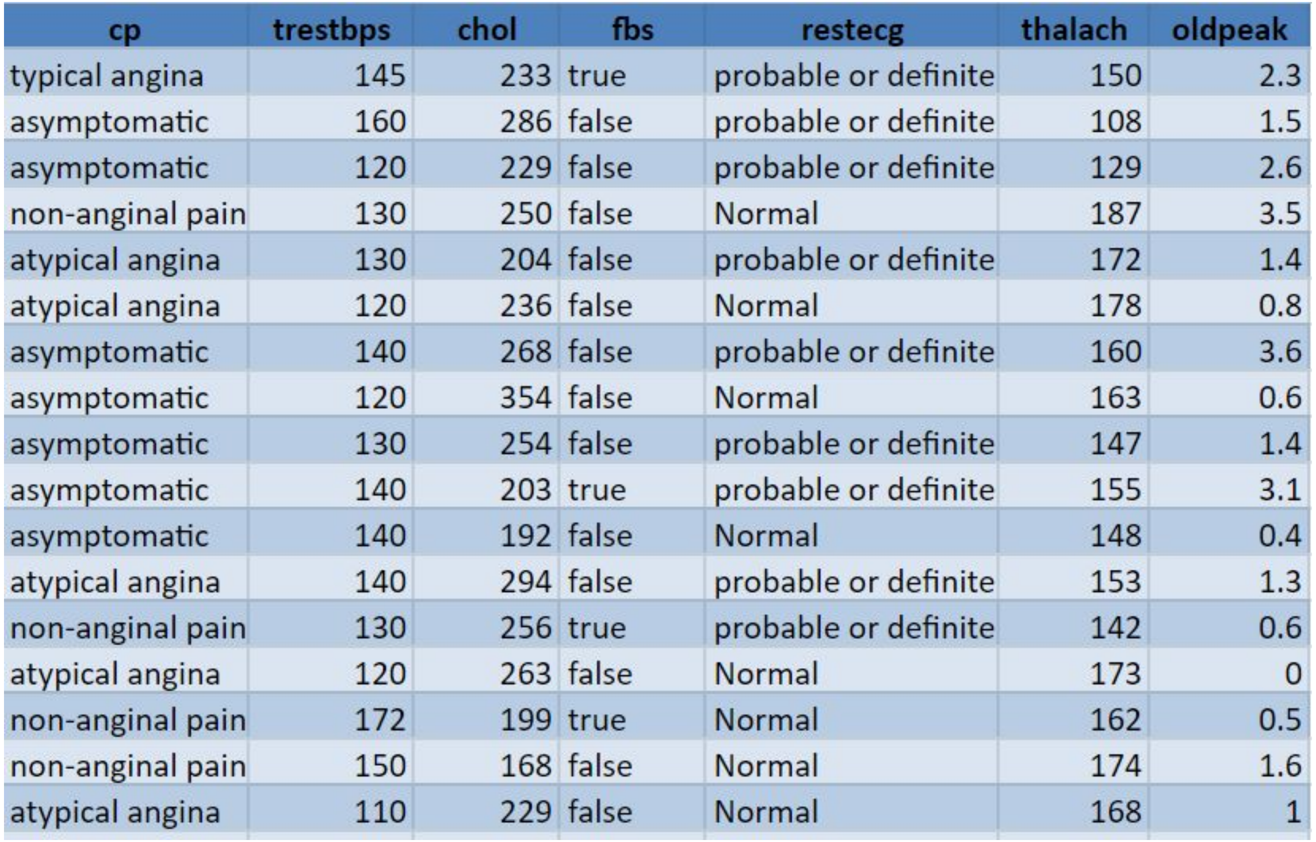
We are exercising everyday whether if it’s from walking to class or going to the gym. Having chest pain or tightness in the chest could be life threatening, and even more since it’s a symptom of CHD. If we could possibly predict who will get exercise induced angina from non-fatal attributes such as cholesterol and fasting blood sugar. We can use it as an early-stage detection for Congenital heart disease .

Machine Learning methods

Machine Learning Methods are used to make the system learn using methods like Supervised learning and Unsupervised Learning which are further classified in methods like Classification, Regression and Clustering. This selection of methods entirely depends on the type of dataset that is available to train the model, as the dataset can be labeled, unlabelled, large.

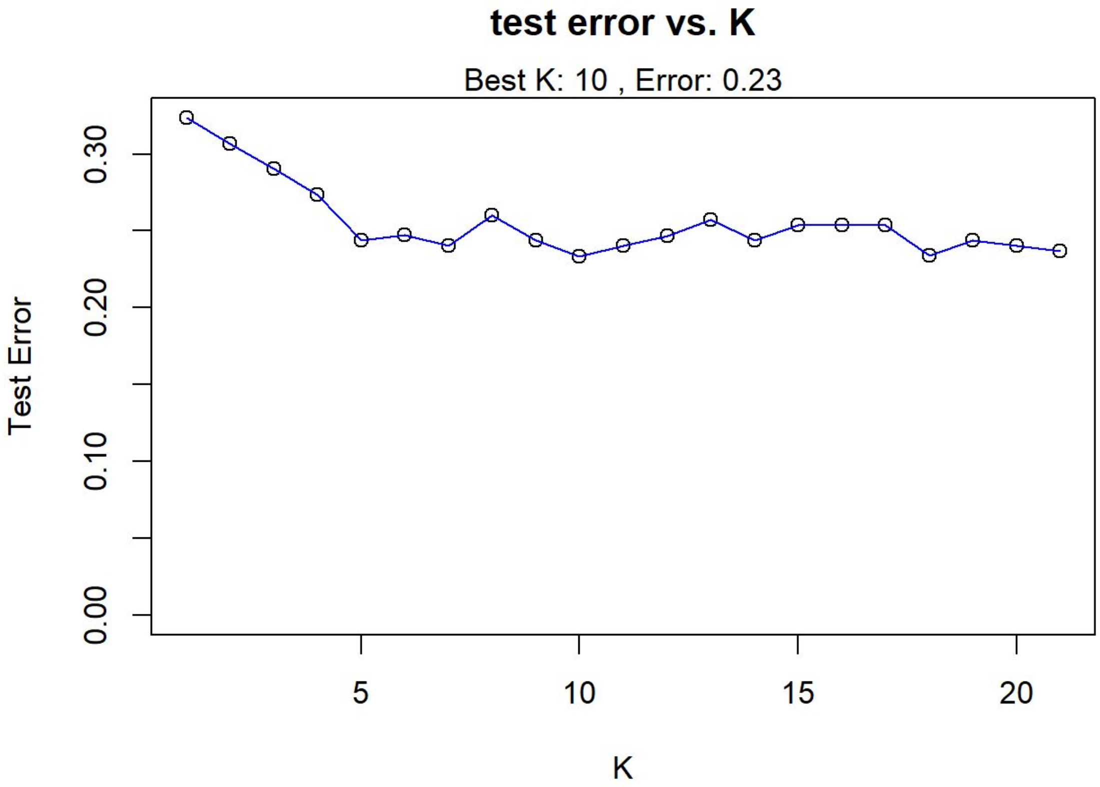
Here, we will be using two machine learning methods K-Nearest Neighbor Algorithm (KNN), SVM (Radial and Linear kernels)

Attributes used in Training (Sample)



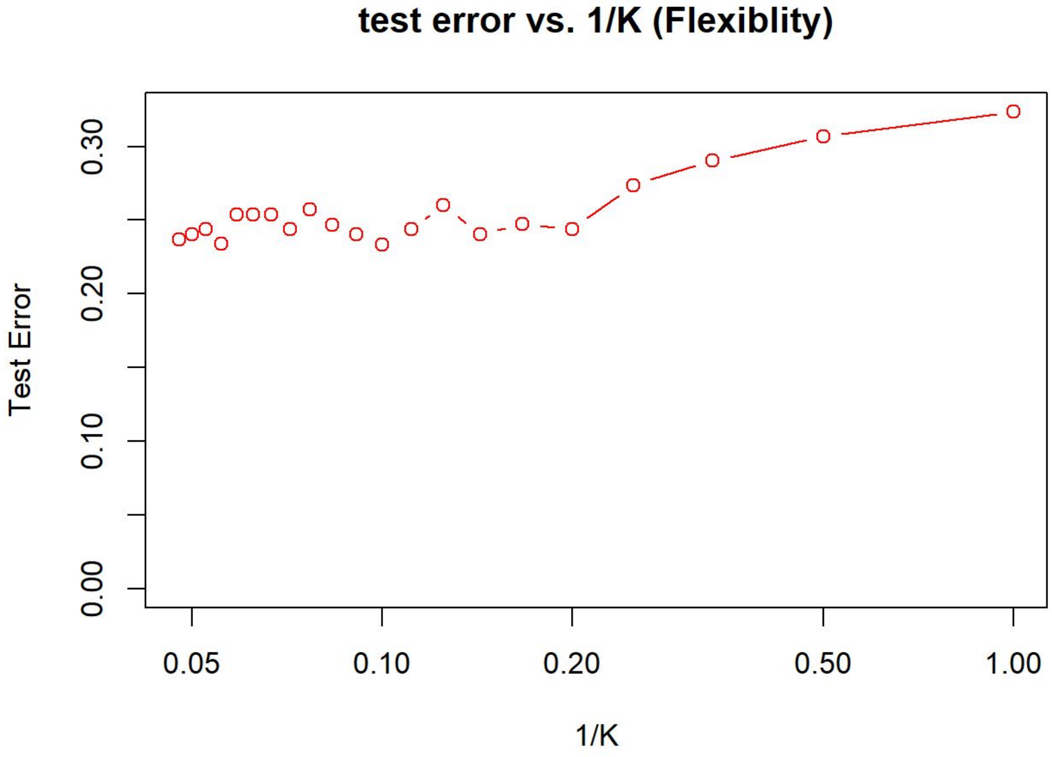
cp	trestbps	chol	fbs	restecg	thalach	oldpeak
typical angina	145	233	true	probable or definite	150	2.3
asymptomatic	160	286	false	probable or definite	108	1.5
asymptomatic	120	229	false	probable or definite	129	2.6
non-anginal pain	130	250	false	Normal	187	3.5
atypical angina	130	204	false	probable or definite	172	1.4
atypical angina	120	236	false	Normal	178	0.8
asymptomatic	140	268	false	probable or definite	160	3.6
asymptomatic	120	354	false	Normal	163	0.6
asymptomatic	130	254	false	probable or definite	147	1.4
asymptomatic	140	203	true	probable or definite	155	3.1
asymptomatic	140	192	false	Normal	148	0.4
atypical angina	140	294	false	probable or definite	153	1.3
non-anginal pain	130	256	true	probable or definite	142	0.6
atypical angina	120	263	false	Normal	173	0
non-anginal pain	172	199	true	Normal	162	0.5
non-anginal pain	150	168	false	Normal	174	1.6
atypical angina	110	229	false	Normal	168	1

KNN Results



test error vs. K
Best K: 10 , Error: 0.23
Test Error
0.00
0.10
0.20
0.30
5
10
15
20
K

KNN Results

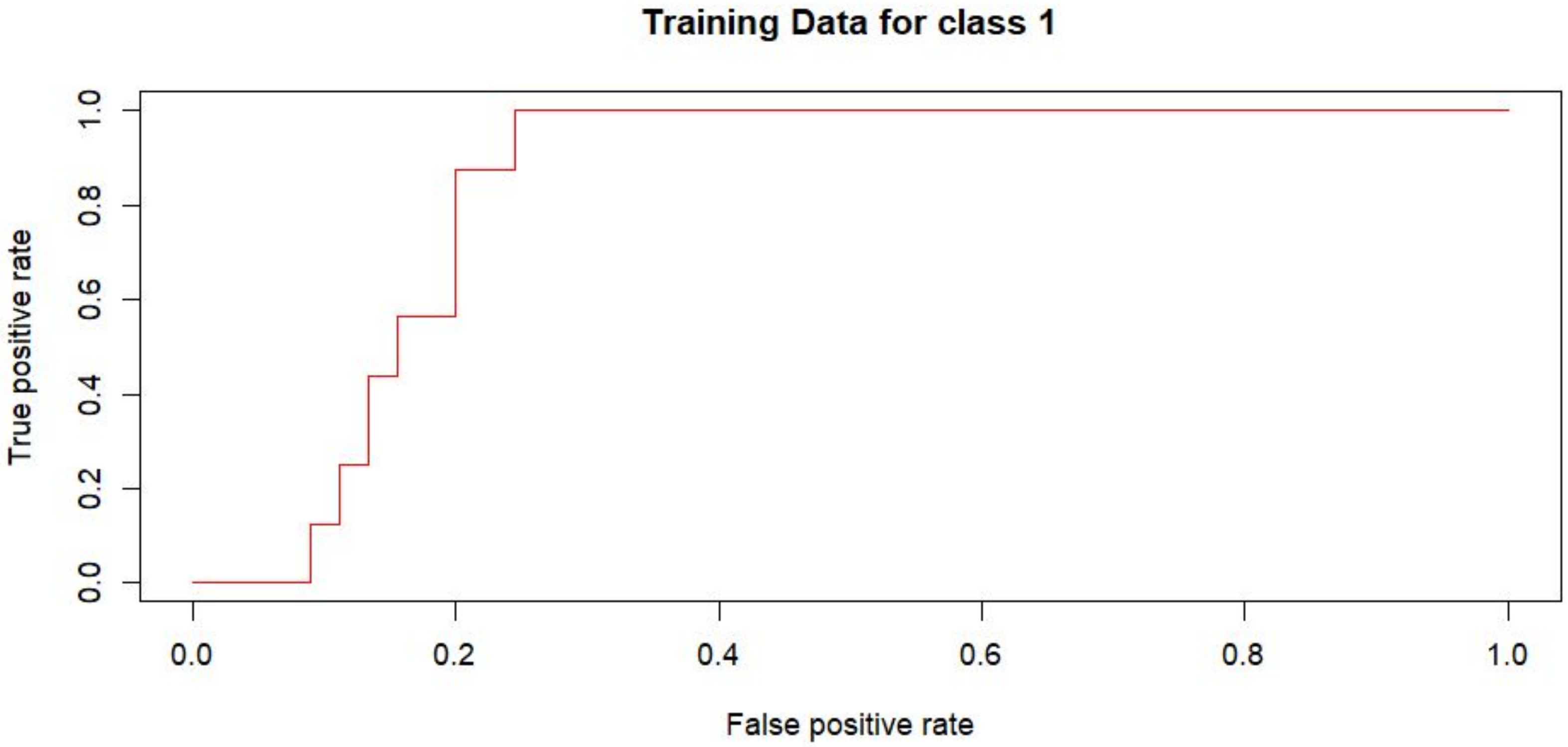

test error vs. 1/K (Flexiblity)
Test Error
0.00
0.10
0.20
0.30
0.05
0.10
0.20
0.50
1.00
1/K

KNN Confusion Matrix (K = 10)

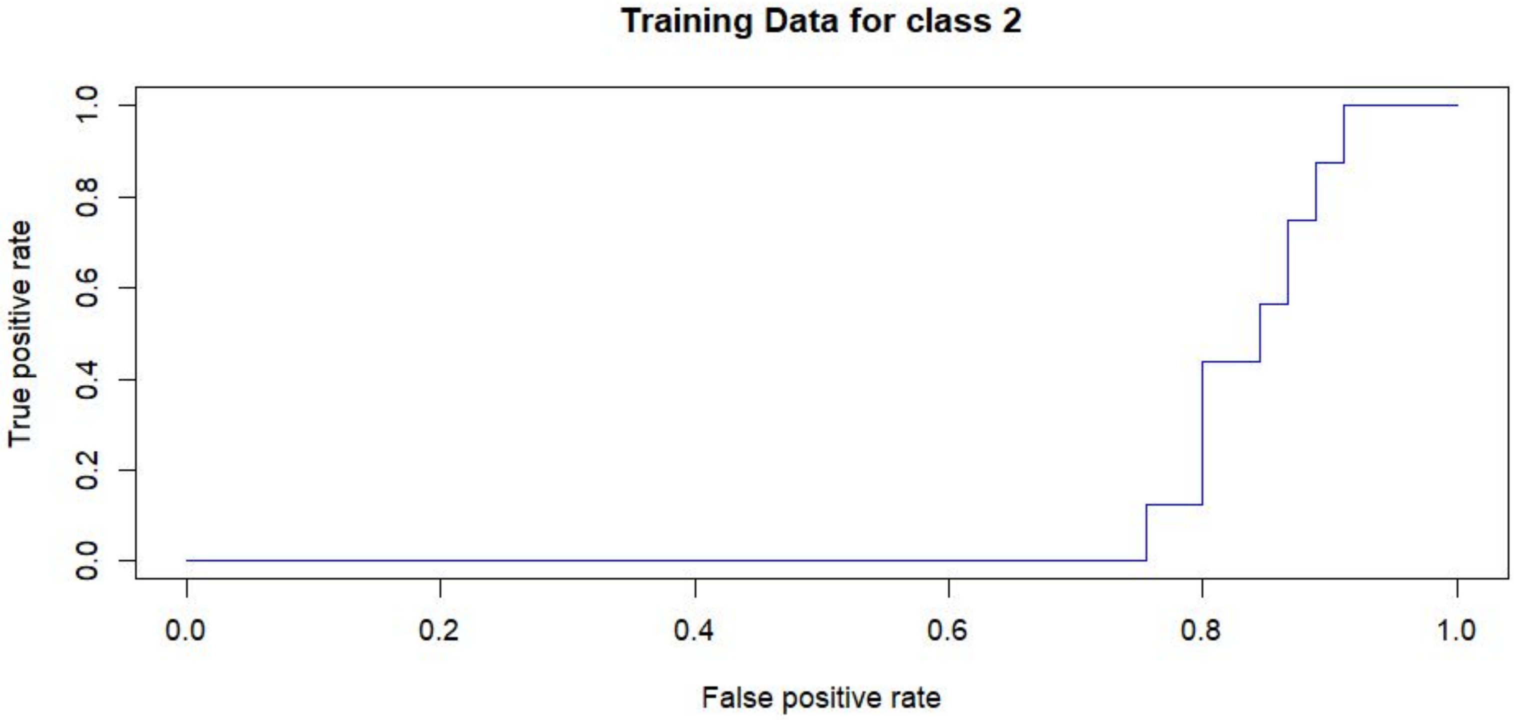
	no	yes
no	38	3
yes	5	13

Accuracy:
86.44%
Error Rate:
13.56%
Precision:
72.22%
Sensitivity:
81.25%
Specificity:
88.37%

SVM (Linear) Results



SVM (Linear) Results

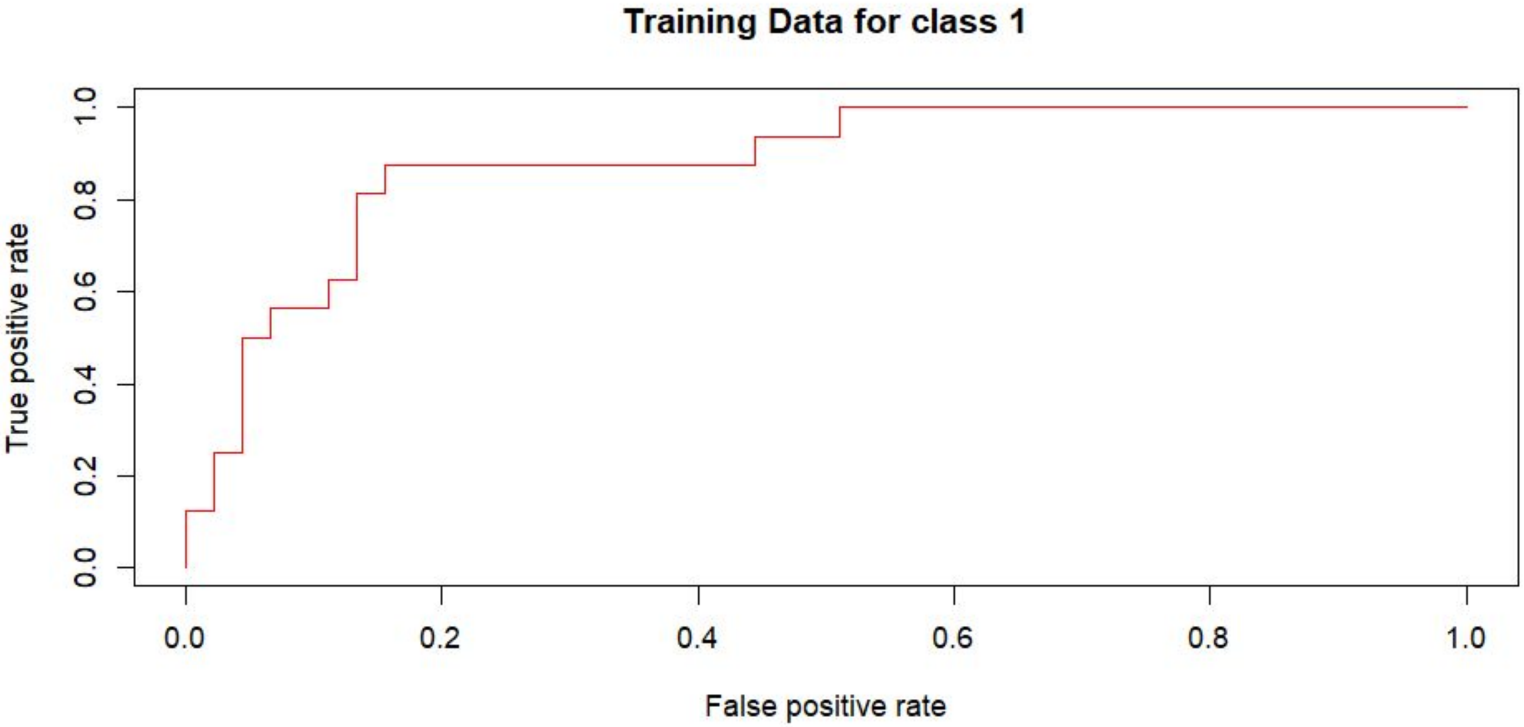


SVM (Linear) Confusion Matrix (Tuned w/ Cost 1)

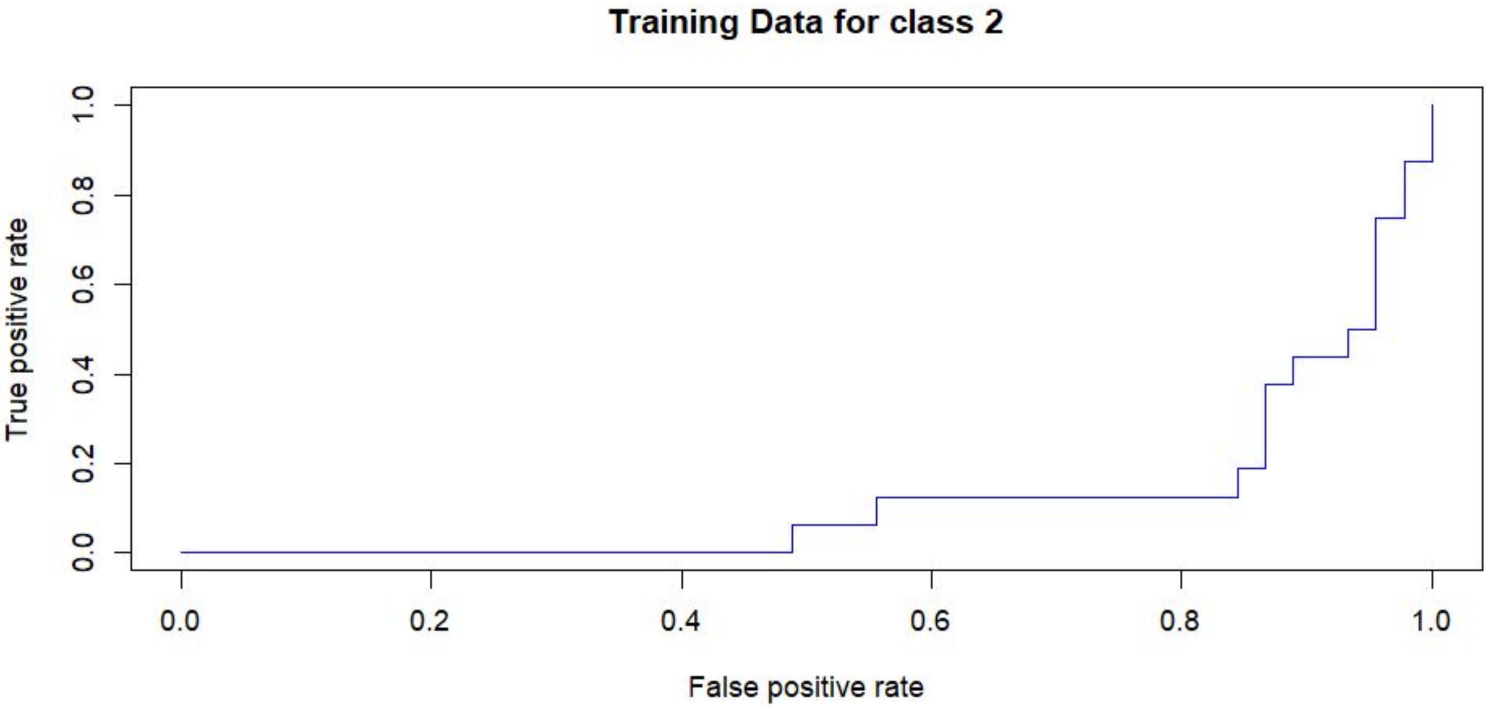
	no	yes
no	34	1
yes	11	15

Accuracy:
80.33%
Error Rate:
19.67%
Precision:
57.69%
Sensitivity:
93.75%
Specificity:
75.56%

SVM (Radial) Results



SVM (Radial) Results



SVM (Radial) Confusion Matrix (Tuned w/ cost 0.01)

	no	yes
no	39	5
yes	6	11

Accuracy:
81.97%
Error Rate:
18.03%
Precision:
64.71%
Sensitivity:
68.75%
Specificity:
86.67%

Comparison of Confusion Matrix

Model	Accuracy	Error Rate	Precision	Sensitivity	Specificity
KNN	86.44%	13.56%	72.22%	81.25%	88.37%
SVM - Linear	80.33%	19.67%	57.69%	93.75%	75.56%
SVM - Radial	81.97%	18.03%	64.71%	68.75%	86.67%

The most optimal model in this case is KNN. The accuracy of KNN is higher than the SVM's and has the lowest error rate. The precision value is also higher. However, it seems to have the second-best Sensitivity rate, the "SVM – Linear model" has the highest sensitivity rate but the lowest accuracy of them all. The Specificity is also the highest for KNN. Therefore, it makes KNN the most optimal.

Discussion

Exercise-induced angina is a symptom of an underlying heart problem, usually coronary heart disease. If we could predict patients with exercise-induced angina among other non-fatal attributes, then we could possibly predict CHD. The following models resulted in success in predicting exercise-induced angina. Among the models, KNN seems to have the best metrics to predict exercise-induced angina. KNN has both high accuracy and a low error rate as well as high precision. One drawback with KNN is the sensitivity (the performance to predict true positive rates). The “SVM – Linear model” performs best in sensitivity but worst in all other areas, making it less viable. KNN also performs the best in specificity (detecting true negatives). Thus, the most optimal model to predict exercise-induced angina is KNN.

Limitations

Observing the metrics within each model shows that one of the worst performances was in precision. Precision consists of correctly classified patients over the total of patients that have the symptom. Considering that the lowest count in most of our confusion matrix was our false positives. This probably has made our precision go down. This is more apparent when observing the ROC curves, they all have a moment where they are constant. To improve precision, we will need to increase the number of true positives but as a result, there will be more false positives. The limitation here is the threshold.

References

https://archive.ics.uci.edu/ml/datasets/heart+disease
https://bradleyboehmke.github.io/HOML/svm.html
https://bradleyboehmke.github.io/HOML/knn.html
https://www.rdocumentation.org/packages/ROSE/versions/0.0-3/topics/ovun.sample
https://shiring.github.io/machine_learning/2017/04/02/unbalanced
https://quantdev.ssri.psu.edu/sites/qdev/files/kNN_tutorial.html