CSE 5160, Summer 2022
Group 8 Project Report

Hirpara, Dhaval Chaturbhai
Inzunza, Kevin
Eddala, Likhitha Reddy

Data Analysis on Heart Disease Data Set

> Background:

The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. Over the last several years, the prevalence of cardiovascular disease has been rising quickly around the globe. Numerous studies have been carried out in an effort to identify the most important risk factors for heart disease and to precisely estimate the overall risk. Heart disease is also referred to as a silent killer because it can cause a person to pass away without any evident signs. In high-risk individuals, an early diagnosis of heart disease is crucial for helping them decide whether to change their lifestyle, which lowers consequences. By evaluating patient data that uses machine learning algorithms to categorize whether a patient has heart disease or not, this study hopes to predict future cases of heart disease.

The main difficulty with heart disease is detecting it. Although there are types of equipment that can forecast heart disease, they are either expensive or ineffective at calculating the likelihood of heart disease in humans. The mortality rate and total consequences can be reduced by early identification of heart disorders. Since it takes more intelligence, time, and knowledge, it is not always possible to accurately monitor patients every day, and a doctor cannot consult with a patient for a whole 24 hours. Since there is a lot of data available nowadays, we can use a variety of machine learning methods to search for hidden patterns. In medical data, the hidden patterns might be used for health diagnosis.

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

> Background of Original Study:

In order to answer these questions 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. Thus I've used the one with the 14 attributes, as 76 attributes weren't used among the machine learning researchers.

> 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?

> Potential Solution:

We are exercising every day whether it's by 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.

> Methods & Results:

The database for Cleveland contains 303 observations among those observations each patient has a number from 0-4 with 4 being the presence of heart disease. I used the Cleveland database that was processed, it doesn't contain all 76 attributes but rather the 14. Before training, data was preprocessed by normalization since the distribution of the data is not known and the algorithms that were used do not make assumptions about the data. Certain columns in the data were removed one of them was the diagnosis, the others were binary values. Since binary values don't work well with normalization. For training, I predicted the "exang" or induced exercise angina with two values "no" and "yes". When training, two methods were used: KNN and SVM to predict those patients with induced split in angina during exercise. For KNN, 5-fold cross-validation was used to resample data. The data was chunks according to its sample size. After having training data and testing data, the KNN function from the class library was used to predict values. For SVM, data was split in an 80:20 ratio. Training data was 80%, 20% of the data for testing. The data was then fed and tuned using SVM with a linear kernel and various cost parameters. The best-tuned model was saved and used for prediction. The same process was used for the radial kernel.

Sample of training attributes used:

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

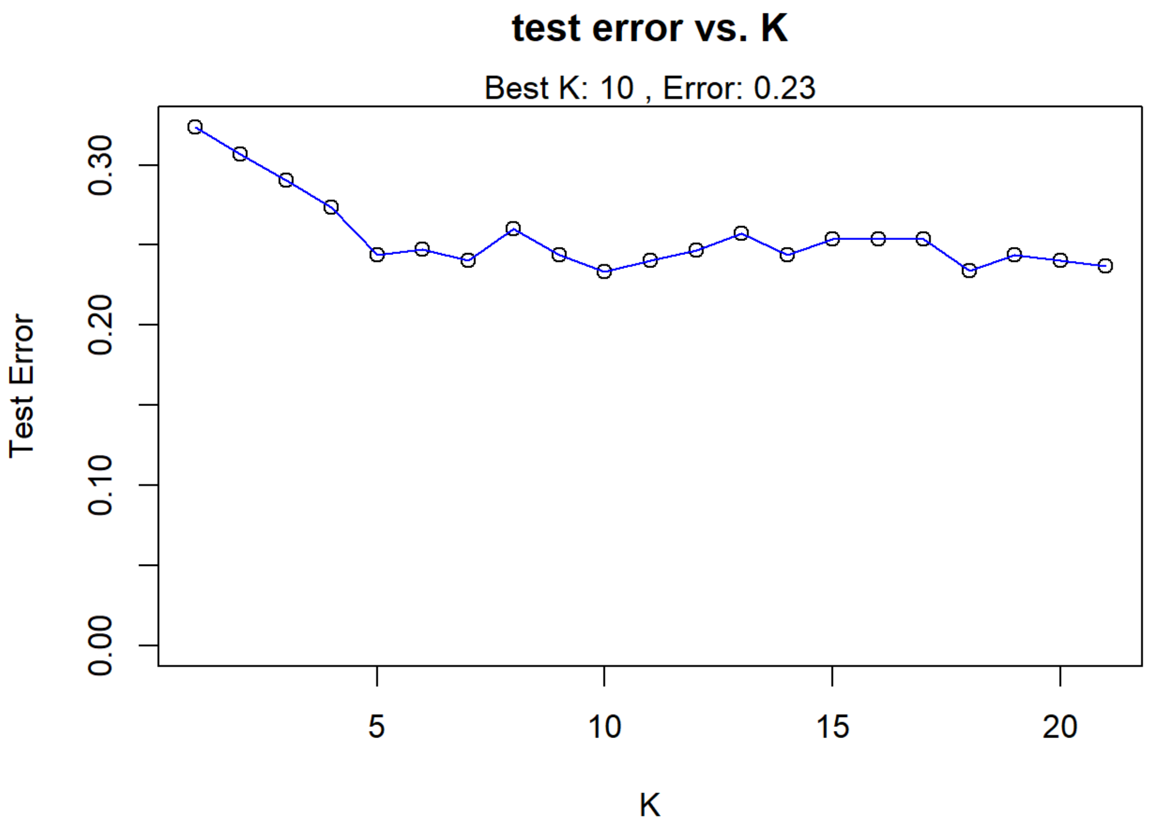
asymptomatic	120	229	false	probable or definite	129	2.6
non-anginal pain	130	250	false	Normal	187	3.5
atypical angina	130	204	false	probable or definite	172	1.4
atypical angina	120	236	false	Normal	178	0.8
asymptomatic	140	268	false	probable or definite	160	3.6
asymptomatic	120	354	false	Normal	163	0.6
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
asymptomatic	140	239	false	Normal	160	1.2
non-anginal pain	130	275	false	Normal	139	0.2
atypical angina	130	266	false	Normal	171	0.6
typical angina	110	211	false	probable or definite	144	1.8
typical angina	150	283	true	probable or definite	162	1
atypical angina	120	284	false	probable or definite	160	1.8
non-anginal pain	132	224	false	probable or definite	173	3.2

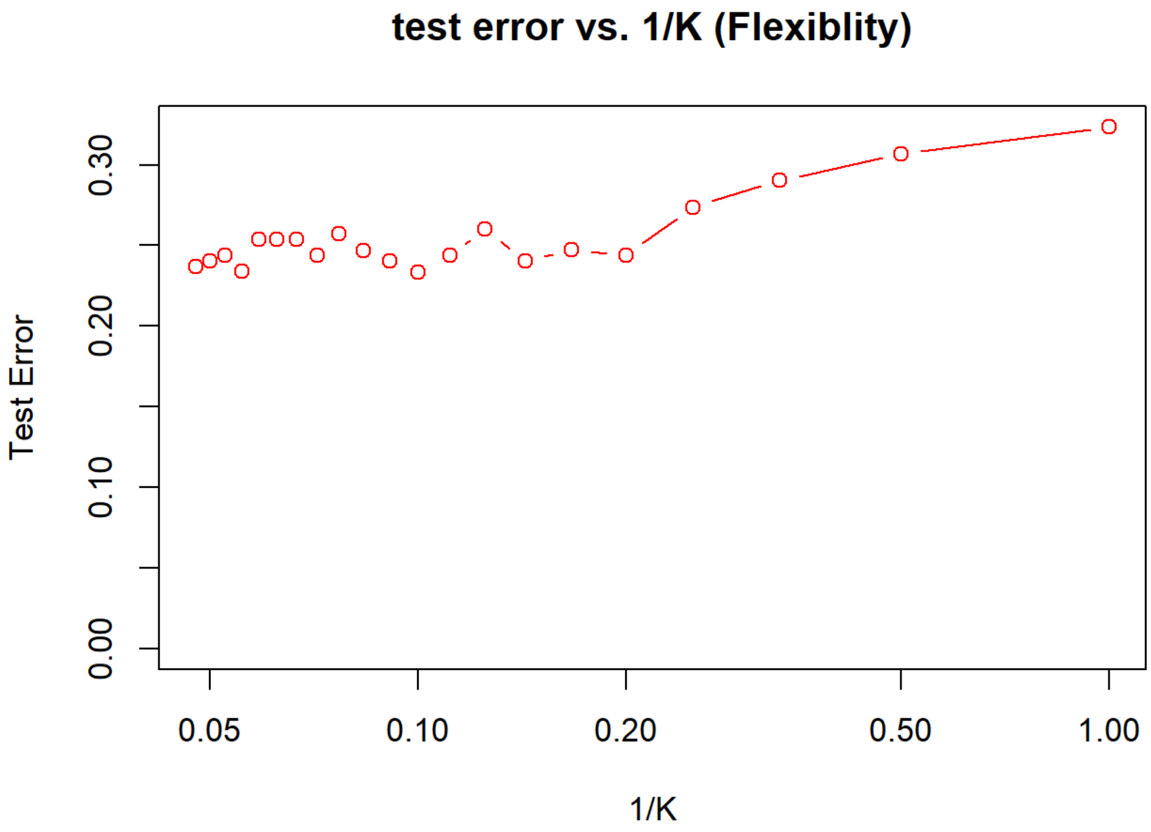
Where:

- CP : Chest Pain
- Trestbps : Resting Blood Pressure
- Chol : Cholestrol
- RestEcg : resting electrocardiographic results

- Thalach : maximum heart rate achieved
- Oldpeak : ST depression induced by exercise relative to rest

<u>KNN Results:</u>





Confusion Matrix (K = 10)

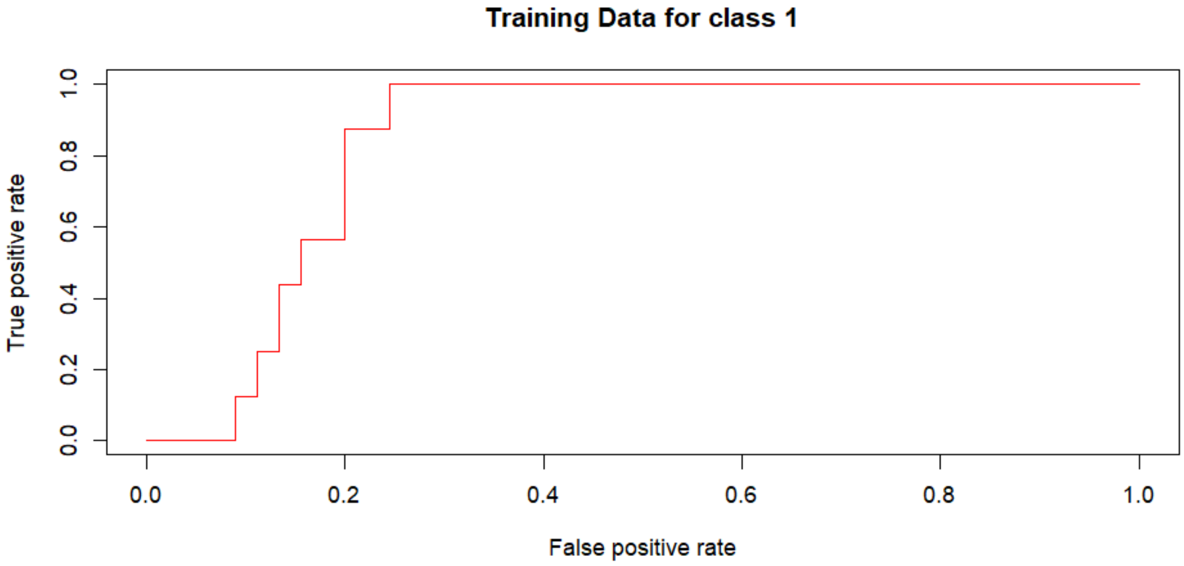
	no	yes
no	38	3
yes	5	13

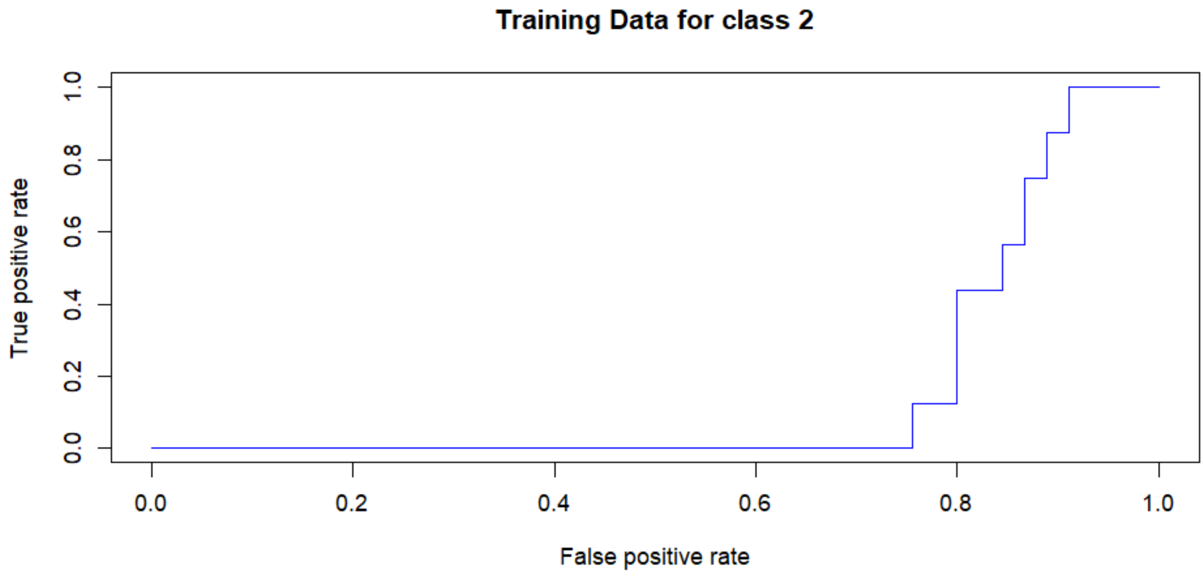
Accuracy:
86.44%
Error Rate:
13.56%
Precision:
72.22%
Sensitivity:
81.25%
Specificity:
88.37%

Analysis:

The lowest test error for 5 k-fold for KNN is 0.23, with the best K being 10. When the flexibility at 0.1 seems to be around 0.20. Overall, the whole graph for flexibility seems to be somewhat squished together. It doesn't seem to have any extreme points that have less flexibility, or more flexibility. The same with the test error vs K graph. There is a nice accuracy of 86%, a low error rate of 13% and the precision is high. Sensitivity is high and specificity is high, meaning that both true negatives and true positives are correctly detected. Nothing abnormally wrong with the model itself.

SVM Results:
Linear Kernel:



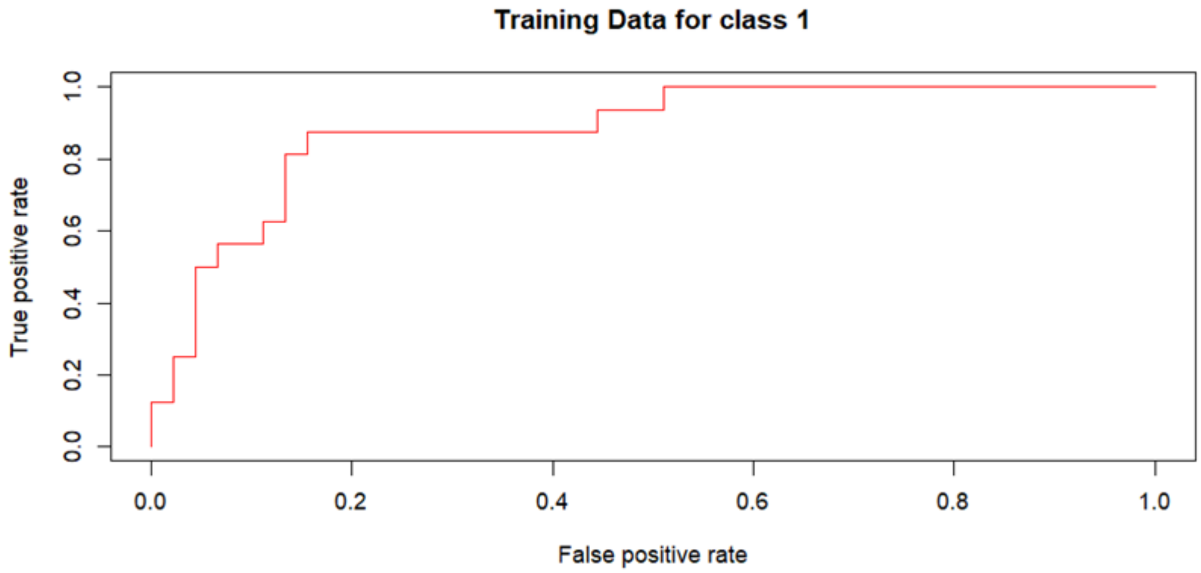


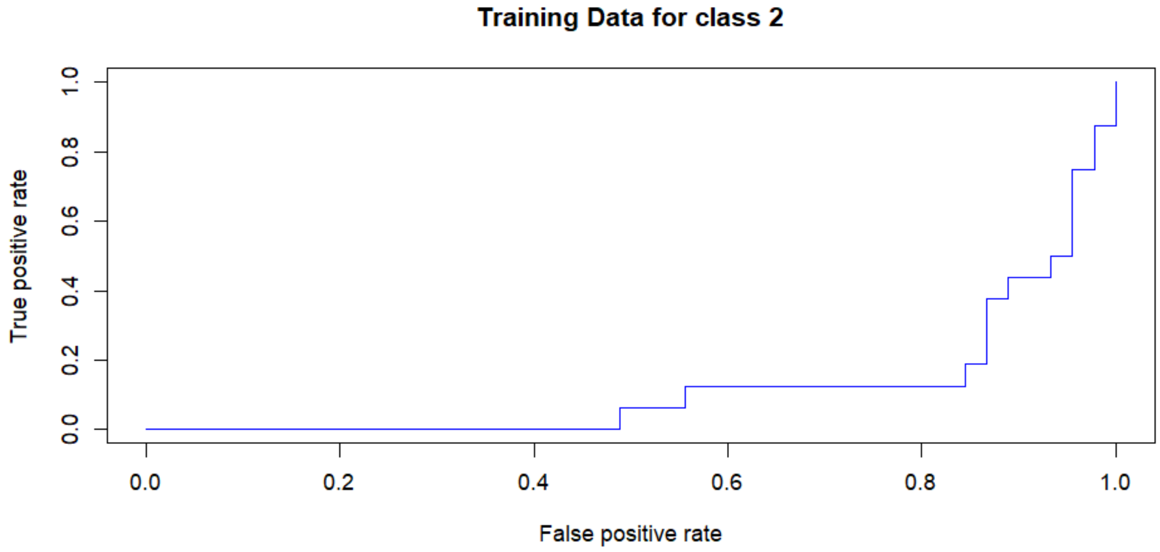
Confusion Matrix (Tuned w/ Cost 1)

	no	yes
no	34	1
yes	11	15

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

Radial Kernel





Confusion Matrix (Radial and Tuned)

	no	yes
no	39	5
yes	6	11

Accuracy:
81.97%
Error Rate:
18.03%
Precision:

64.71%
Sensitivity:
68.75%
Specificity:
86.67%

<u>Analysis:</u>

The ROC curve for the linear for Class 1 (no's), seems to be constant for a while. This is worrying because it means there is significant number of False Positives than True positives. For Class 2 (Yes'), seems to have a much higher rate compared to class 1 of false positives. The precision seems to be somewhat low 57%, a little more than half of the data is being classified correctly. For the radial kernel, the training data for Class 1 seems to be alright, the curve seems to hug the left corner. However, for Class 2 it is constant for a while, this again has a high false postive rate which as result a low true postive rate. The accuracy seems to be an improvement over the linear kernel and error rate, but worse on sensitivity. The specificity seems to have increase too.

> Model Comparision:

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.

> Reference:

- 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