Ali Husain 2277224
Akshnoor Singh 2232759
Dinh Gam Phan 2281486
Anish Tottempudi 1950018
11/09/24
MATH 4323

MATH 4323, Project: Data Questions & Models.

1) We would like to see the main correlations between the primary causes of Cardiovascular disease (Heart disease), as it is one of the biggest health issues we are currently facing worldwide. We aim to highlight the most and least likely causes of this disease and shed light on them to create awareness and help reduce Heart disease in the population. Having the ability to predict the causes of this chronic disease that plagues around half of the U.S adult population is very meaningful and impactful for the overall health of the country.

2a) We will be using KNN for predictions to classify based on similarity to see which symptoms are close to each other when it comes to how they relate to CVD. We will be using SVM to find clear data lines with more complex data. The combination of the two will help us determine the strongest and weakest symptom correlations with Heart disease which is important when it comes to understanding the causes and effects. We use supervised learning models (KNN and SVM) here due to having the result given to us from the start (in the Classification column of the dataset). These 2 methods are also quite different, KNN utilizing a sort of voting system in order to determine the result of a certain observation while SVM utilizes hyperplanes, margins, and support vectors in order to obtain a result of an observation. Having these two differing methods allows us to have a diverse way of obtaining our results and predictions along with some general variety.
2b) SVM is a supervised learning algorithm that is used typically for classification that helps aid in finding hyperplanes and lines of best fit that can help separate data into different classes and also help us to use it as a predictor. SVM's typically have two classes separated by a hyperplane. Depending on the number of features, the hyperplane can change dimensions. For instance 4 input features can turn the hyperplane into a 3D plane. One of these margins that separate each class is known as the Maximal Margin Classifier that is determined from the observations. This is another crucial part of SVM's where the points near these margins serve as anchors and essentially set where they are. Another softer margin known as a support vector classifier is also determined by these anchor points. To conclude SVM uses multiple margins and classifiers in order to determine which class certain observations belong in whilst also removing noise from data sets.
KNN on the other hand is an algorithm that classifies based on how similar things are by looking at the closest neighbor. We use Euclidean distance and determine if an observation is more closely related to its other observations, whatever is the closest neighbor (or neighbors) determines the prediction of the observation. This is determined through a voting system where

the observation measures distance between a K amount of observations. Whichever class appears the most, that observation will then be assigned to that class. These neighbors are in classes due to their similarity in one another so, when an observation is placed and is close to that class it is considered part of that class.
2c) We plan on using cross-validation to see how each model performs, we want to use this because it can help us detect overfitting and prevent it which can help us see which model is the best. Also, we might use leave one out cross validation which can help use almost all of the data set which can reduce bias in our test estimate. Using these approaches, we can see which model is better to ensure accurate results.

3) Ali and Akshnoor will focus on implementing the SVM models and the related tasks, while Anthony and Anish will be implementing the KNN models and related tasks. In the end we will come together and analyze the models and come up with your thesis. In addition to this Ali and Akshnoor will conduct cross validation while Anish and Anthony will work on leave one out cross validation. At the end both of these groups will compare their findings and determine which one is better for our assignment.