River Kelly & Kyler Gappa

CSCI-347: Data Mining

Final Project

April 14, 2022

Liver Disorder Data Mining

The problem that we are investigating is the association between developed liver disorders and the possible contributing factors that lead to such a diagnosis. It is no question that the consumption of alcoholic beverages increases the chances of developing a liver disorder, but the question still remains why some people are more at risk than others. In addition to alcohol consumption, other compounds found in the liver have been found to have confounding associations.

The data set that we have selected is the Liver Disorders Data Set (Forsyth). The data set includes a total of seven attributes. The first five attributes are numerical values pulled from the blood test of the patient. Since these values are thought to be sensitive to liver disease, they may show common traits so that early detection is more viable. The next variable is the number of drinks of alcohol per day. This could be a connecting factor for some of the test results. The final attribute is a categorical variable to help separate the data into test or training sets.

The data mining techniques that we will use to solve this problem include one-hot encoding, dimensionality reductions, and k-means. There exists one categorical attribute within our data set that will need to be converted using the one-hot encoding technique. We plan to use dimensionality reduction across all of the provided attributes in an attempt to identify which attributes have the greatest effect on the k-means. We also plan on utilizing the k-means

clustering method to assist in identifying attributes with the highest correlation (i.e. those that generate the clearest clusters.)

If we run out of time we plan to not explore the differences in the training and test categories to reduce the scope of the project from two separate datasets to a single dataset. If this were to happen we would totally ignore the categorical classification and run all testing over the entire dataset as a whole.

Work Cited

Forsyth, Richard S. “Liver Disorders Data Set.” *UCI Machine Learning Repository*, 15 May 1990, https://archive.ics.uci.edu/ml/datasets/Liver+Disorders. Accessed 14 April 2022.