A screenshot of a cell phone

Description automatically generatedEric Swanson

Cs303

Hw07

The goal of this assignment was to develop a model using machine learning to predict an unsupervised dataset of “user knowledge”. This was accomplished using the machine learning tool KNIME to explore the data relationships formed by a k-nearest neighbor training algorithm applied to the supervised User Knowledge training dataset, and Spider4 to process and output the findings. Using KNIME, I produced a visualization of the data in the form of a scatter matrix to get a better idea of the inherent relationships that existed. It appeared that the “STG”, “SCG”, and “STR” columns of the dataset had the least similarity, or largest distance between each of the data points, indicating a dissimilar relationship, while the “LPR” and “PEG” columns showed a tighter relationship with less distance on the plot between their data points, indicating a stronger relationship. Since our goal was to train a machine learning model to be able to predict the targets of another dataset of the same type based on this related data, I employed the K-Nearest-Neighbor algorithm in a cross validation loop using KNIME for the two most similar columns of data, and was able to achieve an accuracy score of 96%! This method seemed to be ideal for producing a high accuracy rating, as any less columns of training data would produce a less accurate score due to the lack of training information given to the machine learning algorithm. An example of this was performed using only the “PEG” column for training the model, which produced an inferior accuracy score of only 82.5%. Likewise, including any more columns for training the model produced an inaccurate result as well, due to the “noiseiness” of the other columns of data. As an example, a KNN classifier was trained including the addition of the “STR” on top of our initial run, and an accuracy score of 91% was achieved. It seemed that the addition of this extra training data only hurt the quality of the predictions because of the lack of similarity of the qualitative relationship between the data and the labels (as shown in the scatter matrix). This was an interesting find, and through this varied experimentation in KNIME I believed I had found the most Ideal training model for the dataset to be able to confidently predict the labels of the unsupervised dataset to be tested. Finally, I produced a table of these test results using my training model in Spider4.