Asuka Li

Dr. Zhao

CS 470 Data Mining

3/25/22

CS 470 Homework 3

Decision Tree Classifier for Heart Disease Prediction

Model Evaluation

I utilized hold-out method to evaluate my Decision Tree Classifier with 30 randomly selected combinations of training and testing sets. I chose hold-out method to evaluate my model as I understood it the most and felt that the random selection would give a fair evaluation of the model from the metrics derived from it. As a result, my model showed an 83.16% precision and 79.47% accuracy. The recall, F-measure, specificity, and sensitivity scores were 0.7987, 0.8146, 0.7987, and 0.7894. The recall metric allows us to understand how good this model is when there is a high cost associated with false negative, which in this case, detecting a person’s heart disease. 83.16% of the tuples my classifier labeled as positive were actually positive, while 79.87% of positive tuples were marked as positive by my classifier. With the harmonic mean of precision and recall (F-score) being 0.8146, I think my classifier is not the most accurate but predicts the results in a somewhat accurate manner.

Through this project, I was able to understand the decision tree classifier along with impurity indices in a more holistic way to successfully build it. I was also able to gain knowledge of some of the optimizations that could improve the algorithm and think in a strategic way to limit the height and minimum samples split for the classifier for the classifier to run relatively faster than without it. I found it very interesting to utilize already existing data to train the classifier and then apply it to the testing sets. I would love to think more about what I can do to improve the classifier and improve the precision, accuracy, and recall scores of the model in order to better predict the data if I get a chance.