

Business Intelligence and Data Analysis

Fall 2021 Final Exam

Please download and save a local copy of the “Contraceptives.csv” data set¹. This medical data set has identifying characteristics of 1473 married couples across 8 attributes. Please refer to the Data Set Descriptions file (on Blackboard) for attribute definitions.

Use Weka to answer the following questions. (Always use “Use training set” option for testing).

1- Clustering

- Perform SimpleKMeans clustering with default parameters (2 clusters). How would you describe the two clusters based on the attribute characteristics? Interpret how the identified clusters are different based on average attribute values. Which attributes were more important to differentiate the clusters?
- Perform SimpleKMeans clustering with three clusters. How would you describe the three clusters based on the attribute characteristics? Discuss which subsets of the population each cluster represents.

2- Neural Networks

- Perform neural network analysis (MultilayerPerceptron) with two hidden layers (“hiddenLayers”=2). What is the overall prediction accuracy? Identify the attributes that significantly impact each of the two hidden nodes. How would you characterize these two hidden factors identified by the neural network analysis?
- Repeat the same analysis with three hidden layers. What is the new prediction accuracy? Interpret the confusion matrix. Why do you think the accuracy is different? Identify the attributes that significantly impact each of the three hidden nodes. How would you characterize these three hidden factors identified by the neural network analysis?

3- Association Rule Mining

Convert all numerical attributes to nominal by using the Unsupervised/Attribute/Discretize filter. Create 3 bins for each of the numerical attributes when converting to nominal, set “useEqualFrequency” parameter to True.

¹ Papers that cite this data set and relevant papers:

- Lim, T.-S., Loh, W.-Y. & Shih, Y.-S. (1999). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning.
- Earl Harris Jr. Information Gain Versus Gain Ratio: A Study of Split Method Biases. The MITRE Corporation/Washington C. 2001.
- Soumya Ray and David Page. Generalized Skewing for Functions with Continuous and Nominal Attributes. Department of Computer Sciences and Department of Biostatistics and Medical Informatics, University of Wis.
- Jos'e L. Balc'azar. Rules with Bounded Negations and the Coverage Inference Scheme. Dept. LSI, UPC.

- a. Perform general association rule mining with the default parameters.
 - i. Report the top 10 association rules.
 - ii. Interpret any 3 of these rules. What are the support and confidence values for these 3 rules?
- b. Next, perform targeted association rule mining by using the last attribute ("Contra") as the classification target. Set "car" parameter to True, and change "lowerBoundMinSupport" value from 0.1 to 0.02.
 - i. Report the top 10 association rules.
 - ii. Interpret any 3 of these rules. What are the support and confidence values for these 3 rules?