UTFPR – CPGEI – Campus Curitiba

Mineração de Dados e Descoberta de Conhecimento

Profs. Heitor S. Lopes e Thiago H. Silva (UTFPR, 2025) - Exercise #1

A) Decision Trees

1. **Objective**: to explore data classification using decision trees and the Orange software.
2. **Procedure**:
    a. The *Soybean* dataset is regarding the diagnosis of 19 common soybean diseases. The dataset has 35 attributes and 683 instances. Do a preprocessing as necessary for uploading the dataset into Orange.

    b. Using data visualization tools, answer what is possible to infer, preliminarly, about the attributes of this dataset?

    c. Select column "*class*" as the classification target, and all the remaining columns as predicting attributes. Use 5-folds stratified cross-validation for training a Decision Tree using its default parameters. Take notes regarding the size of the obtained tree (number of nodes, depth and number of leaf nodes), as well as the quality measures (accuracy, precision recall and F1-score). Amongst these measures, justify which one is the most adequate for this dataset.

    d. Show the confusion matrix generated by training/testing the decision tree. Identify in this tree for which classes the classification achieved 100% and 0%, respectivelly. Justify this behavior, with special interest to the classes that achieved the worst performance (0% of correct classification).

B) Classification rules

3. **Objective**: to explore data classification using decision trees and classification rules using the Orange software.
4. **Procedure**:
    a. Use the *Contraceptive Method Choice Dataset*, available at the *Machine Learning Repository*. This objective here is to predict the most suitable contraceptive method to be used by Indonesian women, based on demographic and socioeconomic data.This dataset has 1473 instances, 9 predicting attributes, and 3 unbalanced classes (*No-use, Long-term e Short-term*).

b. Using data visualization tools, answer what is possible to infer, preliminarly, about the attributes of this dataset?

c. Using the discretization tool available in the Orange software, preprocess the numerical attributes *wifes_age* and *number_chd_born*, so that data must be constrained in **equal frequency** bands, respectively as: {young, adult, mature, middle-age} and {1, 2, 3-4, 5+}.

d. Use the baseline algorithms (ZeroRule and OneRule) to establish references for further comparison regarding the classification quality, by annotating accuracy, precision, recall and F1-score.

e. Use the CN2 algorithm to obtain classification rules that must be **compreensibles** and **interesting** (but, **not obvious**). Based on these rules, contextualize the findings and clarify the socio-economic and cultural profile of the indonesian women who have opted to use the short-term and long-term (more important) contraceptive methods.

f. Compare the predictive qualit of the thres methods (ZeroRule, OneRule and CN2).

g. Create a decision-tree of such a depth that allows human comprehension. Then, compare it with the results previously obtained by the classification rules, and discuss the positive and negative points of both approaches (trees X rules) regarding classification quality and comprehensibility.