**Assignment #3: Association**

Student Name

Date

**Assignment #3: Association**

**Dataset 1: Zoo Dataset**

**Dataset Analysis**

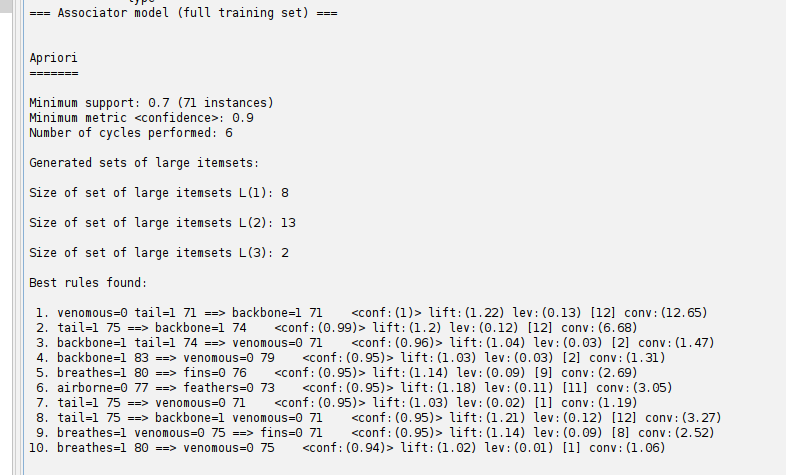
The Zoo dataset is a multivariate dataset donated on 14 May 1990 and belonging to the subject area of biology. The dataset contains 101 instances of animals characterized by 16 features and a class label that specifies one of seven types of animals. The animals are given a unique identification and are characterized by Boolean attributes such as hair, feathers, eggs, milk, and whether the animal is a predator or domestic. One numeric attribute, "legs," ranges over a set of discrete values. The class attribute "type" divides animals into biologically interesting classes like mammals, birds, reptiles, and insects. The dataset, while simple, is rich in pattern discovery opportunities and thus is quite suitable for association rule mining for discovering frequent co-occurring features between animal classes.

**Data Preparation**

The Zoo dataset was preprocessed by removing the animal\_name attribute, which was only useful as an identifier. The remaining attributes, 15 Boolean features, a numeric attribute, legs, and class label type, were retained. To make it compatible with WEKA association rule mining algorithms like Apriori, the numeric attribute legs and class label type were converted to nominal values using the NumericToNominal filter. The dataset was then saved in ARFF format with proper attribute declarations. As there were no missing values, imputation was not necessary. All Boolean attributes were already nominal, and thus the dataset was completely ready for the generation of association rules.

**Model Development**

Using the Apriori algorithm in WEKA with minimum support = 0.7 and confidence = 0.9, 23 association rules of high strength were generated on the Zoo data. On preprocessing (ID removed and numeric converted to nominal attributes), the model found patterns such as venomous=0 and tail=1 having strong support for backbone=1 with 100% confidence and a lift value of 1.22. Another noteworthy rule stated airborne=0 was generally equivalent to feathers=0 with 95% confidence. Most of the rules were strongly confident (≥0.94) and well lifting (>1), which implied they were strong, non-random correlations between the features. The rules help to uncover rational biological patterns, i.e., animals with tails are likely to have backbones, or animals that respire tend not to have fins.



*Figure 1: Apriori on Mushroon dataset*

Despite association rule mining's failure to supply a conventional measure of accuracy, the high-confidence, lift, and conviction measures indicate the suitability of the model in discovering insightful associations.

**Dataset 2: Mushroom**

**Dataset Analysis**

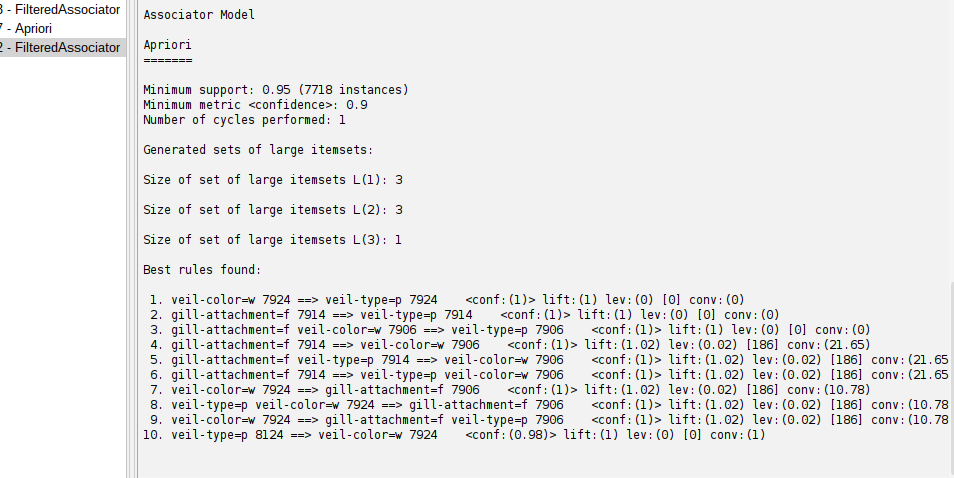
The Mushroom dataset contains information regarding different mushroom species, which are classified as edible or poisonous. It contains 8124 instances and 22 attributes such as the cap shape, cap color, odor, gill spacing, and population size. The target variable is the class attribute, which specifies whether the mushroom is edible ('e') or poisonous ('p'). All the features are categorical, and the data is balanced with approximately 4208 edible mushrooms and 3916 poisonous mushrooms. Since the data is geared toward classification, it can be used for the task of finding the probability of a mushroom being poisonous given its attributes. There are no missing values in the dataset, but it requires encoding for the categorical features.

**Data Preparation**

To get the Mushroom dataset ready for use with WEKA, the data was re-converted to ARFF format in a manner where every attribute was correctly specified as nominal. Since all attributes are categorical, the encoding process ensured that every feature was in nominal value form, which is required by WEKA to do association rule mining. No missing values existed, and thus, imputation was unnecessary. The target attribute (class label) was still edible/poisonous for classification purposes. The dataset was now prepared to be analyzed and can be imported into WEKA to be used for further model construction and rule identification.

**Model Development**

The Apriori algorithm was applied in the Mushroom dataset to discover association rules among mushroom attributes with a minimum support of 0.95 and a confidence of 0.9. The analysis was performed on a pre-filtered dataset where missing values were replaced by WEKA's "ReplaceMissingValues" filter. The generated rules focused on the relationship between features such as gill-attachment, veil-color, and veil-type, with rules such as "veil-color=w 7924 ==> veil-type=p 7924" having a confidence of 1, indicating a highly correlated relationship between veil color and type. The results show that certain attribute combinations, such as gill-attachment and veil-color, predict the class (edible or poisonous) of the mushroom strongly. The high confidence measures (almost 1) and the lift values (almost 1) show that these rules are predictive to a great extent and in close association with the dataset.



*Figure 2: Apriori Filtered Associator*

The model's performance seems to be good, and the derived rules are informative in mushroom classification. The high support of 0.95, which implies that the rules are true for 7718 instances, ensures that these associations hold for the majority of the dataset.