**Assignment #3: UCI Dataset Analysis and Model Development**

Student Name

Date

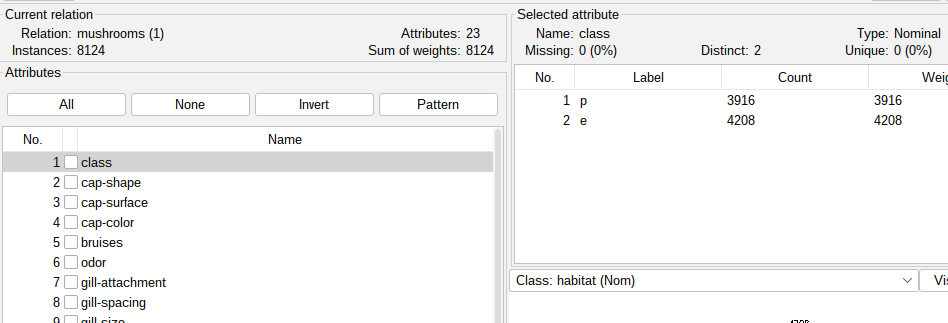
**Association Task**

In this assignment, two datasets were chosen from the UCI Machine Learning Repository to apply association rule mining with the support of the WEKA tool. The two selected datasets are the Mushroom dataset and the Groceries dataset. Both datasets were analyzed for structure as well as to find out if they are suitable for association activities.

**Dataset Analysis**

**Mushroom Dataset**

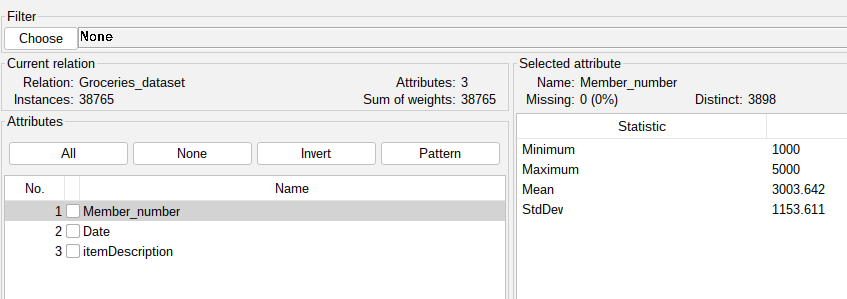
The Mushroom dataset, which was obtained from the Audubon Society Field Guide and is available in the UCI Machine Learning Repository, consists of 8124 instances and 22 categorical attributes describing physical characteristics of mushrooms, such as cap shape, odor, gill size, and habitat. The dataset is used for classification, in this instance, to classify whether a mushroom is edible or poisonous. The target attribute is "poisonous" with class labels poisonous or edible. All attributes are categorical, and while most of the attributes are complete, the attribute "stalk-root" has missing values. This data set is a real-world challenge for biology classification and is extremely popular for pattern extraction through data mining techniques like association rule mining and classification.



*Figure 1: Mushroom dataset*

**Groceries Dataset**

The Groceries Dataset is a 38,765-transaction record dataset of a grocery store that stores customers' shopping habits over time. Each transaction record includes customer ID, purchase date, and product description. The dataset is widely used in Market Basket Analysis and Association Rule Mining to reveal patterns and relations among items purchased together regularly. Using algorithms such as Apriori, analysts and researchers can derive descriptive association rules, such as butter, milk, and measures such as support, confidence, and lift, to gain a better understanding of customers' behaviors. The findings can be used to influence marketing efforts, promotions, and shelf merchandising to boost sales and customer satisfaction.



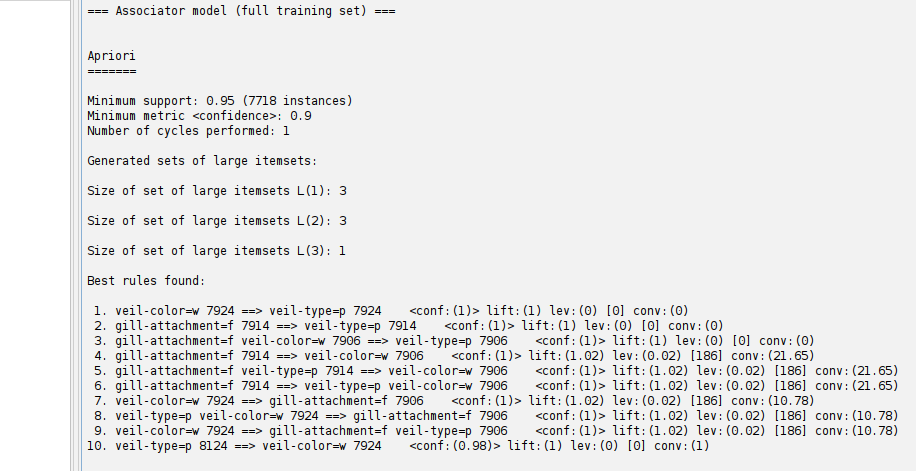
*Figure 2: Groceries dataset*

**Dataset Preparation**

For data preparation, both the Mushroom Dataset and the Groceries Dataset had to be cleaned and converted into a form compatible with association rule mining in WEKA. For the Groceries Dataset, customer and date were employed to partition transactions into creating itemsets corresponding to individual shopping baskets. Missing values in the Mushroom Dataset were handled, and categorical attributes were retained as they are suitable for association rule mining. Both datasets were then converted into the default input format of WEKA, which is the ARFF.

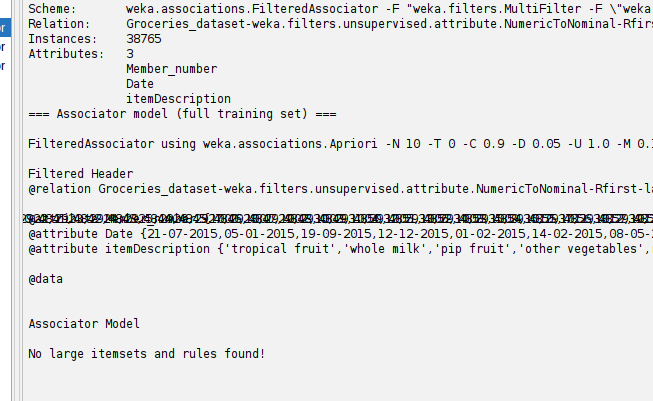
**Model Development**

The Apriori algorithm was run on the Mushroom dataset consisting of 8124 instances and 23 attributes, with minimum support = 0.95 and minimum confidence = 0.9. The model generated 10 robust association rules, and all of them reflect perfect confidence (1.0) except one rule that reflects marginally lower confidence (0.98). Rules primarily relate to the close association among features like veil-color=w, veil-type=p, and gill-attachment=f, which all suggest a highly consistent scheme throughout the data. For example, the rule veil-color=w → veil-type=p occurred 7924 times with a confidence level of 1.0, meaning this combination appears almost always together. All but the most common rules had a lift of 1 or just above that, which means that despite the strong associations, they may be capturing expected patterns given the dataset's structure.



*Figure 3: Apriori results*

The run was done with the Apriori algorithm through a FilteredAssociator in WEKA with a MultiFilter that entails missing value imputation. The data used for the task was the Groceries\_dataset, containing 38,765 instances and three attributes: Member\_number, Date, and itemDescription.



*Figure 4: Filtered associator results*

The Apriori parameters were minimum confidence = 0.9, minimum support = 0.1, and max rules = 10. Attributes were first converted from numeric to nominal where feasible, and all data was preprocessed entirely before association rule mining.