## **Machine Learning**

**Assignment 4**

## **Unsupervised Learning**

**NIM-BSCS-2020-03**

**Shamsa Batool**

## **Aim:**

The aim of this assignment is to predict if a person who applies for loan in bank, gets it or not. The reason for selecting this data set is my interest in application of Machine Learning in bank sector. Since bank involves various financial activities which produce a lot of data, this data can be used in Machine Learning to make interesting and useful models which can assist bank sector in various ways. I used **Apriori algorithm** which is basically used to extract Association rules. Using these rules, I classified the test data and find different evaluation measures.

## **Data:**

The data set I selected from kaggle.com involves 614 instances with 12 attributes. These aatributes are gender, married, dependents, education, self employed, applicant income, co-applicant income, loan amount, term, credit history, area and status. The status attribute represents class labels. Since in loan applications, there are two situations, the person gets loan or not. So, the output class is binary.

**Algorithm:**

The Apriori algorithm is a classic algorithm for finding frequent itemsets and generating association rules from a dataset. Firstly, the Apriori algorithm uses support and confidence of ruleset. Support is the measure of the frequency of an itemset in the dataset, while confidence is measure of the strength of the association between the antecedent and consequent in a rule. Secondly, the algorithm generates candidate itemsets until no more frequent itemsets can be found in mutiple iterations. Then the support for these candidate itemsets is calculated from the dataset. If the support of an itemset meets the minimum support threshold, it is considered a frequent itemset. In this way, frequent itemset is generated. These frequent itemsets are then used to extract association rules which can be used for classification. The generated rules are then evaluated using the support and confidence measures. Out of generated rules, those which pass the criterion of the programmer are selected. Following are the rules I got from my dataset through this algorithm.

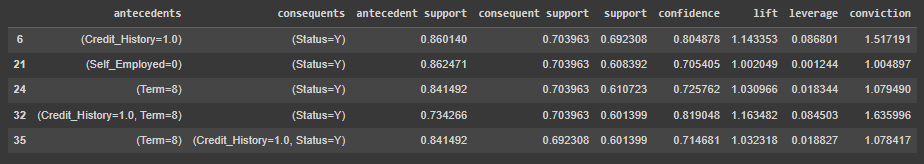


Image 1: Ruleset for Loan Prediction Dataset

### Findings:

Applying association rule procedure on classification dataset extracts those conditions from the features of the dataset which predict a class label. These conditions can then be checked on test instances and the results can be evaluated using different evaluation measures used in classification. The evaluation results I got in this task by applying apriori algorithm suggest that the association rule technique did not perform very well on the dataset. It has different reasons. First, the dataset was highly imbalanced, due to which I got rules for only one class label. As rules are extracted based on a threshold of support/confidence, if I decrease the threshold, the number of rules increases. However, these rules are not good as they get support from few data objects. Moreover, to get rules for the second class, I have to decrease the threshold to a very low value. However, these rules are poor for the reason I just discussed. Moreover, if we look at the classification report, the recall value is very good. It shows that the model is capable of correctly identifying the instances where the loan was actually approved. It means, the model is performing very well with respect to the positive class. However, the values for other evaluation measures are not that good. I got zero value of each measure for Status = No as the model could not retrieve any rule for this class label.

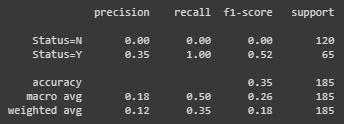


Image 2: Classification Report