**Machine Learning and Data Mining PMS 2024  
Εργασία 02 Association Rules**

**Γεώργιος Βώβος**

**Αριθμός μητρώου: Μ013124003**

*Να προσθέσετε το dataset mybasket.xlsx στο local repository του RapidMiner και στη συνέχεια να στήσετε μια διεργασία κατασκευής κανόνων συσχέτισης χρησιμοποιώντας ως τιμή support την μεγαλύτερη τιμή που παράγει 4-itemsets (βρείτε τη ποια είναι!) και confidence=0.8 (80%). Να καταγράψετε:*

*• Τα συχνά 4-itemsets και τα συχνά 3-itemsets.*

*• Τους κανόνες με την ισχυρότερη συσχέτιση καθώς και τα support, confidence και lift που αυτοί έχουν. Πώς τους κρίνατε ότι είναι αυτοί με την ισχυρότερη συσχέτιση?*

* Ξεκινάμε κάνοντας import to .xlsx **στο root folder** του Local Repository και φτιάχνοντας ένα νέο process το οπόιο απλώς κάνει Retrieve αυτό το αρχείο [1]
* Μιας και τα δεδομένα δεν είναι σε βολική μορφή όπως λέει και η εκφώνηση, e.g υπάρχουν πολλές εγγραφές για ένα product, κάνουμε έναν «group by» με τον τελεστή Aggregate. [2]  
  (Αυτό το βήμα φαίνεται να είναι προαιρετικό μιας και μπορούμε να πάρουμε τα ίδια αποτελέσματα που χρειαζόμαστε χρησιμοποιόντας μόνο το Pivot στο επόμενο βήμα)
* Βάζουμε ένα Pivot operator ώστε να φέρουμε τα δεδομένα σε μια τυπική δομή για association rules  
  (μια στήλη για το basketID και μια για κάθε product) [3]
* Κάνουνε κάποιες μικρές βελτιστοποιήσεις ώστα να φέρουμε τα δεδομένα απο το [3] στην τελική μορφή που χρειαζόμαστε. [4]
  + Μετονομάζουμε λίγο τις στήλες βγάζοντας το «count\(product\)\_»
  + Κάνουμε replace missing value με zero
  + Βάζουμε σαν Role το basketID για να μην το κάνουμε replace στο επόμενο βήμα
  + Χρησιμοποιούμε το «Numerical to Binominal» operator για να μετατρέψουμε τα 0 σε false και τα υπόλοιπα σε true
* Προσθέτουμε τον FP-Growth operator για να βρούμε το frequent item sets.  
  «Παίζοντας» με το min support, βλέπουμε ότι η μεγαλύτερη τιμή που παράγει 4-itemsets είναι περίπου **0.0574. (Παράγει δεκαπέντε 4-itemsets).[5]**Αν πάμε παρακατώ, πχ 0.0573,στα πάρουμε και 5-itemsets.  
  Υποθέτουμε ότι η άσκηση δεν ζητάει το μεγαλύτερο support που φτιάχνει έστω κι ένα 4-itemset, αυτό θα ήταν το 0.5175
* Τέλος, βάζουμε και ένα Create Association Rules με Min Confidence =0.8 για να πάρουμε τα αποτελέσματα που θέλουμε. [6]

**1)** **Συχνά 4-itemsets και τα συχνά 3-itemsets :** Εικόνες [7] και [8] αντίστοιχα **2)** **Τους κανόνες με την ισχυρότερη συσχέτιση καθώς και τα support, confidence και lift που αυτοί έχουν.   
 Πώς τους κρίνατε ότι είναι αυτοί με την ισχυρότερη συσχέτιση?**  
 Στις εικόνες [6] φαίνοται οι κανονες ταξινομημένοι ανά Lift/Confidence/Support. Ποιοί είναι οι καλύτεροι κανόνες, εξαρτάται από τι είναι σημαντικό για εμάς και δεν υπάρχει κάποιος τύπος με  
 απαντήσεις για όλες τις περιπτώσεις.  
 Συχνά, ένα μέτρο που χρησιμοποιέιτε είναι το Support × Confidence (που δεν φαίνεται εδώ)  
 Υπάχουν περιπτώσεις που θα θέλαμε μεγάλο confidence, π.χ για να βρούμε συσχετίσεις μεταξύ αντικειμένων   
 μέσα σετ ενα σύνολο. Στη δική μας περίπτωση ίσως το Lift να είναι καλό μέτρο μιας και θα δίξει «δυνατούς»   
 συσχετισμούς μεταξύ αντικειμένων.

[1]

A screenshot of a computer

Description automatically generated

[2]  
A screenshot of a computer

Description automatically generated

[3]  
A screenshot of a computer

Description automatically generated  
A screenshot of a computer

Description automatically generated

[4]  
A screenshot of a computer

Description automatically generated

[5]

A screenshot of a computer

Description automatically generated  
A screenshot of a computer

Description automatically generated

[6]

A screenshot of a computer

Description automatically generated  
A screenshot of a computer

Description automatically generated

**Sort by Lift Descending**  
A screenshot of a computer

Description automatically generated

**Sort by Confidence Descending**  
A screenshot of a computer

Description automatically generated

**Sort by Support descending**  
A screenshot of a computer

Description automatically generated

**[7] Frequent 4-itemsets**

A screenshot of a computer

Description automatically generated

**[8] Frequent 3-itemsets**  
A screenshot of a computer

Description automatically generated  
A screenshot of a computer

Description automatically generated