Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Παράδειγμα – Κανόνων Συσχετίσεων

Πρόβλημα

Η Μαρία είναι διευθύντρια σε ένα σχολικό συγκρότημα το οποίο έχει πολλές και μεγάλες ανάγκες που δυστυχώς όμως δεν υπάρχουν οι απαραίτητοι πόροι για να ικανοποιηθούν.

Γνωρίζει όμως ότι πολλοί γονείς δραστηριοποιούνται σε διάφορους συλλόγους και οργανώσεις και θεωρεί πως αν καταφέρει να χρησιμοποιήσει τις γνωριμίες που έχουν οι γονείς ή πολύ περισσότερο αν καταφέρει να κάνει γονείς που ανήκουν σε διαφορετικούς συλλόγους να συνεργαστούν μεταξύ τους, τα οφέλη για το σχολικό συγκρότημα θα είναι πολύ μεγάλα.

Πρόβλημα

Οπότε αυτό που θα πρέπει να κάνει είναι να συλλέξει κάποια στοιχεία σχετικά με τους γονείς και τους συλλόγους στους οποίους συμμετέχουν και στη συνέχεια να διερευνήσει ενδεχόμενες συνδέσεις μεταξύ των συλλόγων.

Αφού συγκεντρώσει τα πρώτα αποτελέσματα τότε θα μπορέσει να έρθει σε επαφή με τους γονείς και να τους ζητήσει να χρησιμοποιήσουν τις γνωριμίες τους προκειμένου να ξεκινήσει η υλοποίηση κάποιων πολύ σημαντικών έργων για το σχολείο τα οποία όμως είχαν καθυστερήσει για πολλά χρόνια.

Τα Δεδομένα

Προκειμένου να μπορέσουν να συγκεντρώσουν τις κατάλληλες πληροφορίες δημιούργησε ερωτηματολόγια τα οποία περιλαμβάνουν τις παρακάτω ερωτήσεις.

Ελεύθερος Χρόνος {short-medium-long}, Ερώτηση σχετικά με τον χρόνο που μπορούν να αφιερώσουν

Φύλο {M,F},

Εργασία (NO,YES),

Ηλικία,

Παντρεμένος {0,1},

Hobbies {0,1},

Σύλλογος για τα κοινά {0,1},

Πολιτική Οργάνωση {0,1},

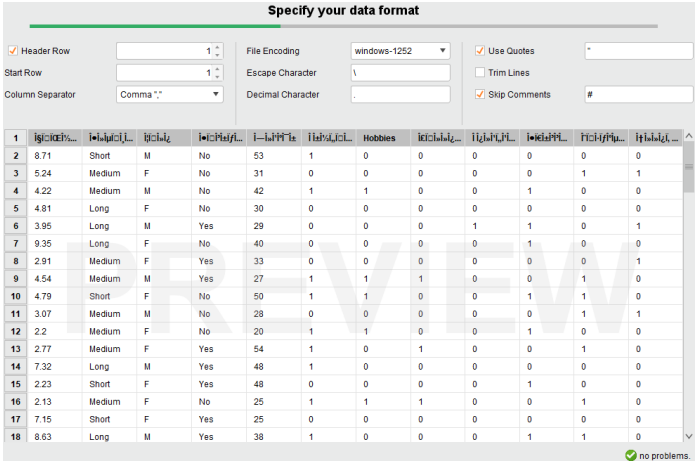
Επαγγελματικός Σύλλογος {0,1},

Θρησκευτικός Σύλλογος {0,1},

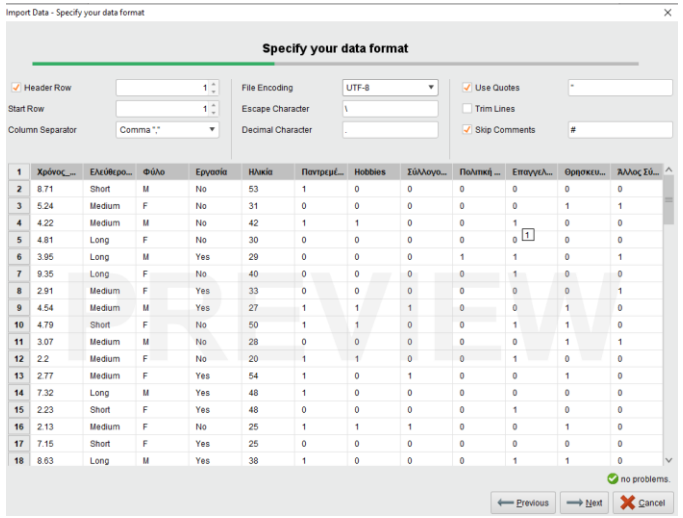
Άλλος Σύλλογος {0,1}

Προετοιμασία των Δεδομένων

- Το αρχείο με τα δεδομένα θα το βρείτε στον φάκελο dataset της ενότητας.
- Επειδή τα ονόματα των στηλών στο αρχείο είναι με Ελληνικούς χαρακτήρες, όταν θα εισάγεται τα δεδομένα ενδεχομένως να δείτε αυτή την οθόνη



- Για να διορθωθεί το πρόβλημα με την κωδικοποίηση θα πρέπει να επιλέξετε File Encoding: UTF-8



Προετοιμασία των Δεδομένων

Τα δεδομένα θα πρέπει να έχουν αυτή την μορφή

Row No.	Χρόνος_Συμ...	Ελεύθερος Χ...	Φύλο
1	8.710	Short	M
2	5.240	Medium	F
3	4.220	Medium	M
4	4.810	Long	F
5	3.950	Long	M
6	9.350	Long	F
7	2.910	Medium	F
8	4.540	Medium	M
9	4.790	Short	F
10	3.070	Medium	M
11	2.200	Medium	F
12	2.770	Medium	F
13	7.320	Long	M
14	2.230	Short	F
15	2.130	Medium	F
16	7.150	Short	F
17	8.630	Long	M

Result History | ExampleSet (Retrieve Chapter05DataSet)

Data | Statistics | Visualizations | Annotations

Filter (12 / 12 attributes): Search for Attributes

Name	Type	Missing	Statistics		
Χρόνος_Συμπλήρωσης	Real	0	Min 2.010	Max 10.150	Average 5.922
Ελεύθερος Χρόνος	Polynominal	0	Least Short (714)	Most Long (1465)	Values Long (1465), Medium (1304), ...[1 more]
Φύλο	Polynominal	0	Least M (1693)	Most F (1790)	Values F (1790), M (1693)
Εργασία	Polynominal	0	Least No (1739)	Most Yes (1744)	Values Yes (1744), No (1739)
Ηλικία	Integer	0	Min 17	Max 57	Average 36.731
Παντρεμένος	Integer	0	Min 0	Max 1	Average 0.390
Hobbies	Integer	0	Min 0	Max 1	Average 0.300

Προετοιμασία των Δεδομένων

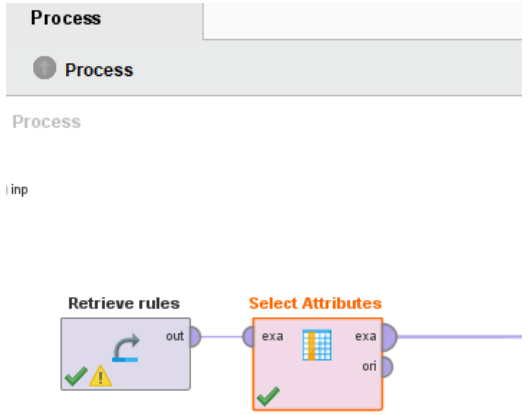
Όπως προκύπτει από τα δεδομένα

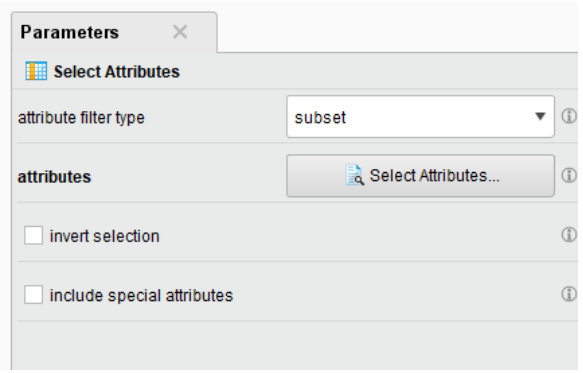
- Δεν υπάρχουν ελλιπείς τιμές
- Δεν χρειάζεται κανονικοποίηση.
- Το μόνο που παρατηρούμε είναι ότι εκτός από το πεδίο «Εργασία» που είναι Yes-No οι υπόλοιπες μεταβλητές είναι 0 ή 1. Επειδή όμως για την εξαγωγή των κανόνων οι τιμές θα πρέπει να είναι T-F θα πρέπει να γίνει μετατροπή από Numerical σε Binominal .
- Επίσης στο σετ των δεδομένων υπάρχουν οι μεταβλητές «Χρόνος Συμπλήρωσης», «Ηλικία», «Φύλο» οι οποίες δεν χρειάζεται σε αυτό το στάδιο, οπότε μπορούμε να εφαρμόσουμε κάποιο φίλτρο για να την απομακρύνουμε.

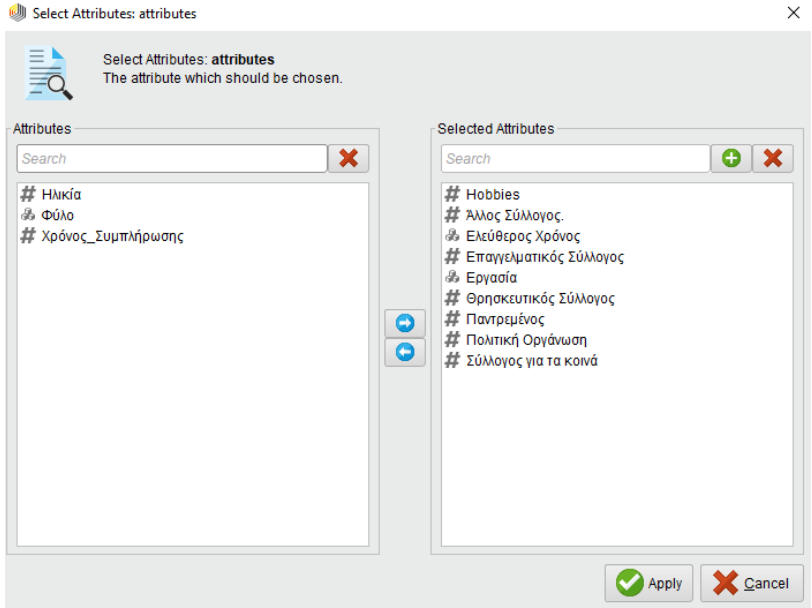
Προετοιμασία των Δεδομένων – Φίλτρο

Προκειμένου να απομακρύνουμε τις παραμέτρους θα επιλέξουμε το “Select Attributes” και στις παραμέτρους θα επιλέξουμε

Attribute filter type: subset

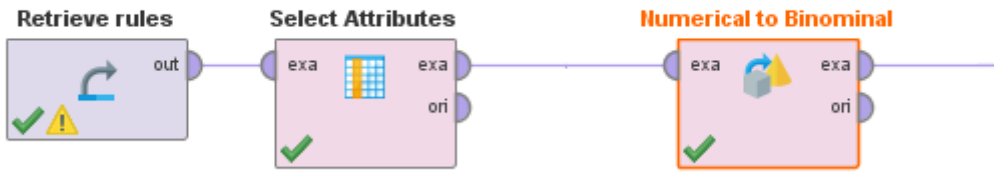


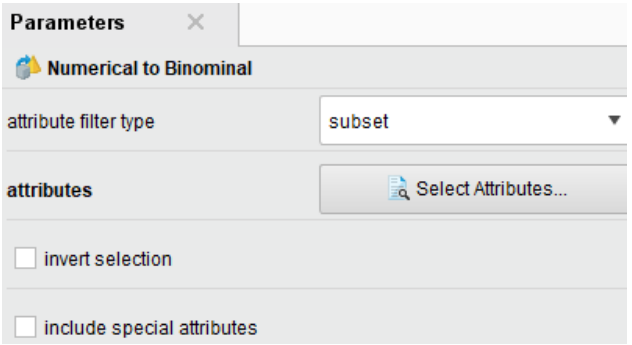


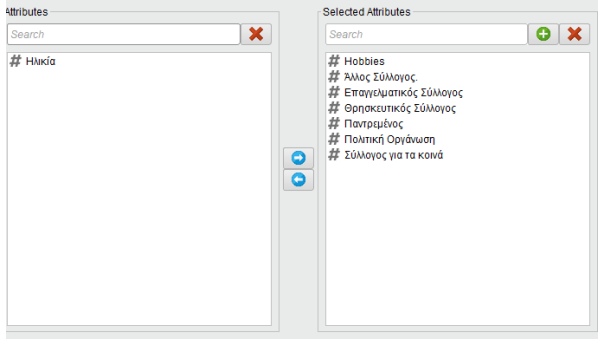


Προετοιμασία των Δεδομένων – Numerical σε Binominal

Προκειμένου να μετατρέψουμε τις μεταβλητές απο Numerical σε Binominal θα πρέπει να επιλέξουμε “Numerical to Binominal” και στις παραμέτρους να επιλέξουμε σαν attribute filter type: subset το οποίο θα έχει τιμές για τις οποίες θέλουμε να γίνει η μετατροπή





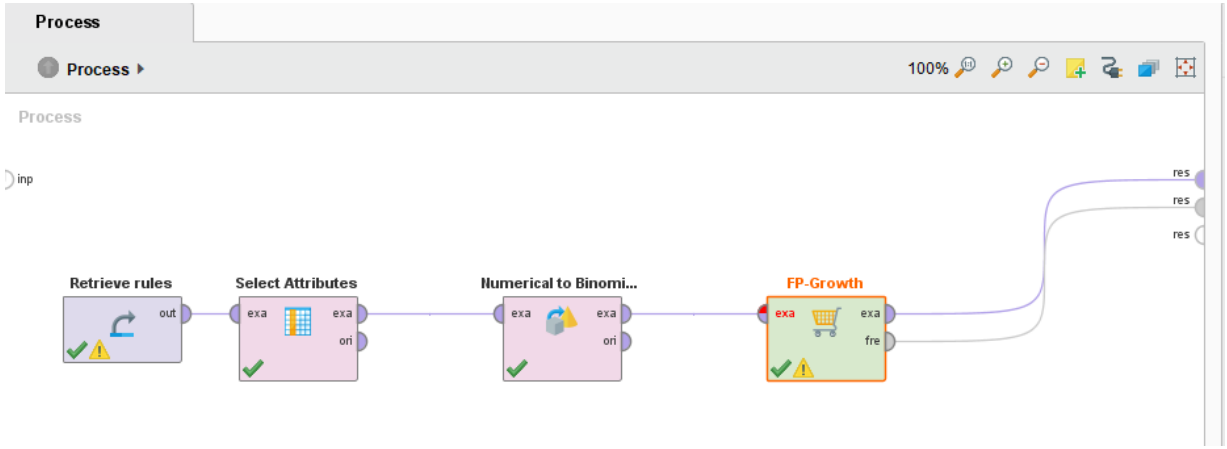


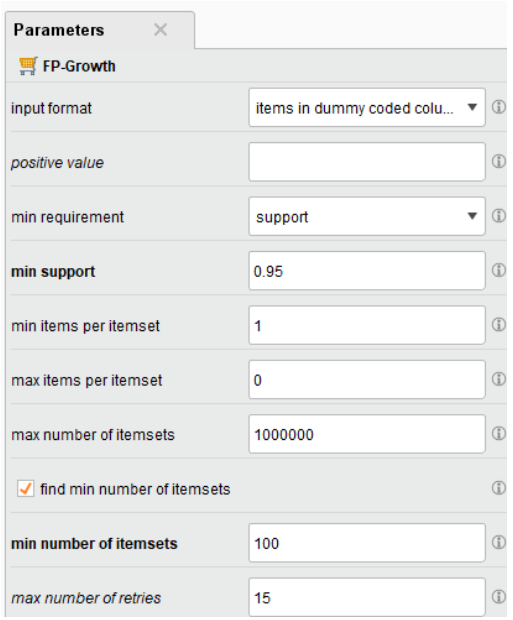
DATA PREPARATION – Τελική Μορφή

Row No.	Παντρεμένος	Hobbies	Σύλλογος γι...	Πολιτική Ορ...	Επαγγελματ...	Θρησκευτικ...	Άλλος Σύλλ...	Ελεύθερος Χ...	Φύλο	Εργασία	Ηλικία
1	true	false	false	false	false	false	false	Short	M	No	53
2	false	false	false	false	false	true	true	Medium	F	No	31
3	true	true	false	false	true	false	false	Medium	M	No	42
4	false	false	false	false	false	false	false	Long	F	No	30
5	false	false	false	true	true	false	true	Long	M	Yes	29
6	false	false	false	false	true	false	false	Long	F	No	40
7	false	false	false	false	false	false	true	Medium	F	Yes	33
8	true	true	true	false	false	true	false	Medium	M	Yes	27
9	true	true	false	false	true	true	false	Short	F	No	50
10	false	false	false	false	false	true	true	Medium	M	No	28
11	true	true	false	false	true	false	false	Medium	F	No	20
12	true	false	true	false	false	true	false	Medium	F	Yes	54
13	true	false	false	false	false	false	false	Long	M	Yes	48

Μοντελοποίηση

- Προκειμένου να βρούμε τους κανόνες συσχετίσεων θα πρέπει να χρησιμοποιήσουμε τον αλγόριθμο «FP-Growth» ο οποίος υπολογίζει τη συχνότητα που εμφανίζονται συνδυασμοί τιμών. Για αυτό στο παράδειγμα θα ενώσουμε και τις 2 εξόδους exa, fre με τις εισόδους res

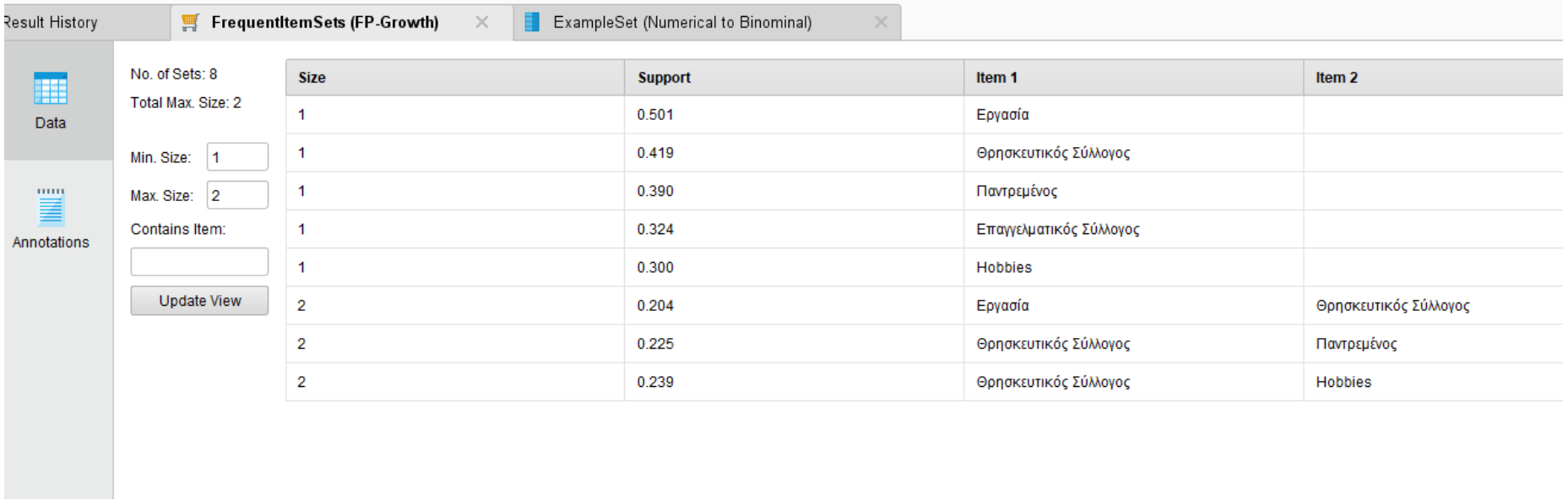




[Useful Link]

Μοντελοποίηση

Η exa έξοδος θα δημιουργήσει τα τελικά δεδομένα, ενώ η έξοδος fre θα δημιουργήσει ένα πίνακα με όλους τους συνδυασμούς που θα βρει καθώς επίσης και το βαθμό συχνότητας που εμφανίζονται.



Size	Support	Item 1	Item 2
1	0.501	Εργασία	
1	0.419	Θρησκευτικός Σύλλογος	
1	0.390	Παντρεμένος	
1	0.324	Επαγγελματικός Σύλλογος	
1	0.300	Hobbies	
2	0.204	Εργασία	Θρησκευτικός Σύλλογος
2	0.225	Θρησκευτικός Σύλλογος	Παντρεμένος
2	0.239	Θρησκευτικός Σύλλογος	Hobbies

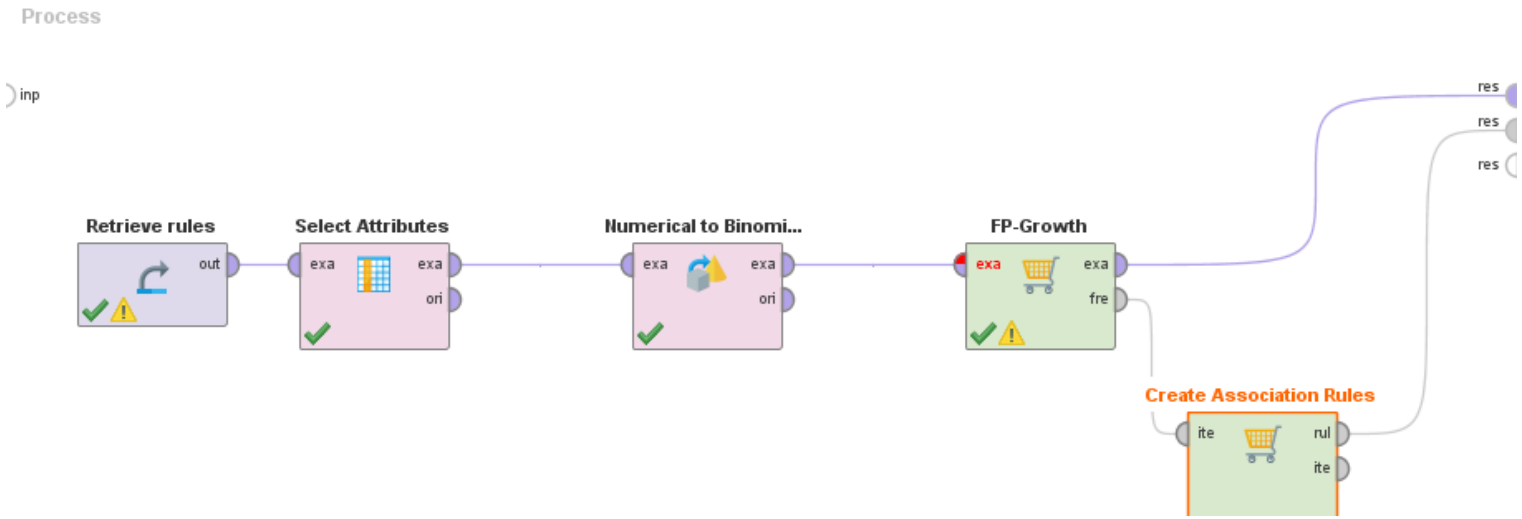
Μοντελοποίηση

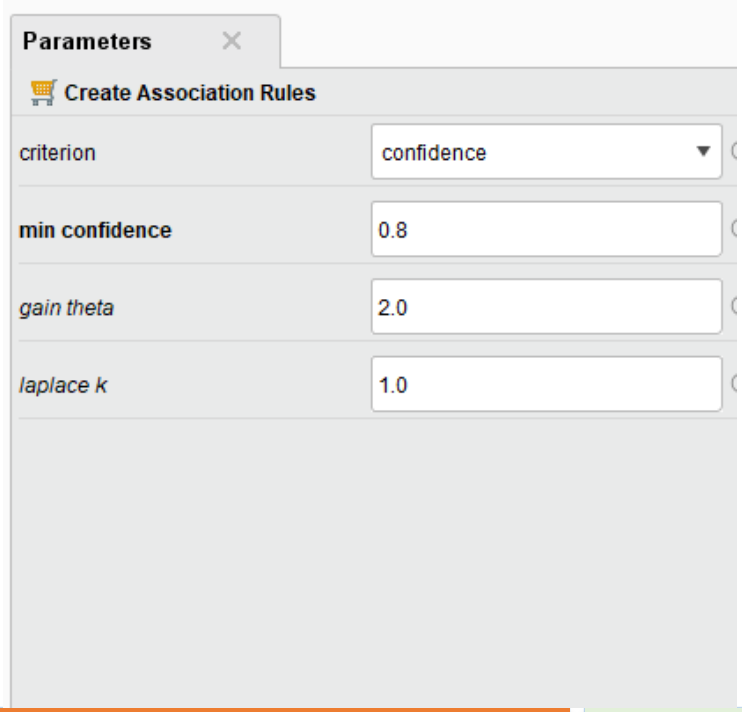
Αυτό που παρατηρούμε από τα αποτελέσματα είναι ότι κάποια μεταβλητές εμφανίζονται μαζί στο δείγμα μας. Για παράδειγμα όσοι είναι σε θρησκευτικούς συλλόγους έχουν Χόμπι και είναι παντρεμένοι.

Πάνω σε αυτά τα αποτελέσματα θα μπορούσαμε να δούμε αν υπάρχουν κανόνες συσχέτισης και με τι βαθμό εμπιστοσύνης αυτοί εμφανίζονται.

Μοντελοποίηση

Για να εξάγουμε τους κανόνες θα χρησιμοποιήσουμε τη λειτουργία “Create Association Rules”. Όπου στο πεδίο min confidence συμπληρώνουμε τον ελάχιστο βαθμό εμπιστοσύνης που θα πρέπει να έχουν οι κανόνες που θα μας υπολογίσει.

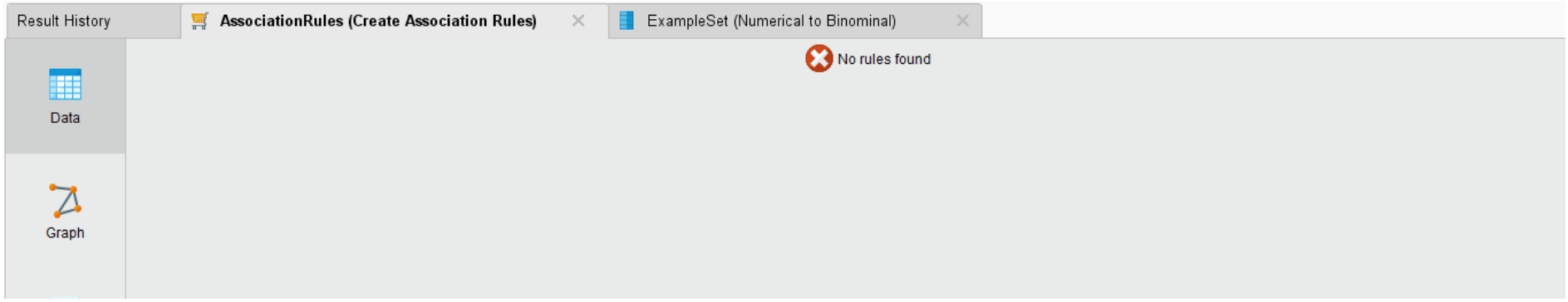




Useful Link

Αποτελέσματα

Όπως βλέπουμε από το σύνολο των δεδομένων που έχουμε δεν βρέθηκε κανένας κανόνας.



Αποτελέσματα

Όπως προκύπτει από την προηγούμενη εικόνα δεν βρέθηκε κανενας κανόνας με βαθμό εμπιστοσύνης 0.8. Αν κατεβάσουμε λίγο το βαθμό εμπιστοσύνης σε 0.7 τότε εμφανίζεται ο πρώτος κανόνας

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s
2	Θρησκευτικός Σύλλογος	Hobbies	0.239	0.571	0.873	-0.598	0.113
3	Παντρεμένος	Θρησκευτικός Σύλλογος	0.225	0.576	0.881	-0.555	0.061
4	Hobbies	Θρησκευτικός Σύλλογος	0.239	0.796	0.953	-0.361	0.113

ΤΕΛΟΣ ΕΝΟΤΗΤΑΣ