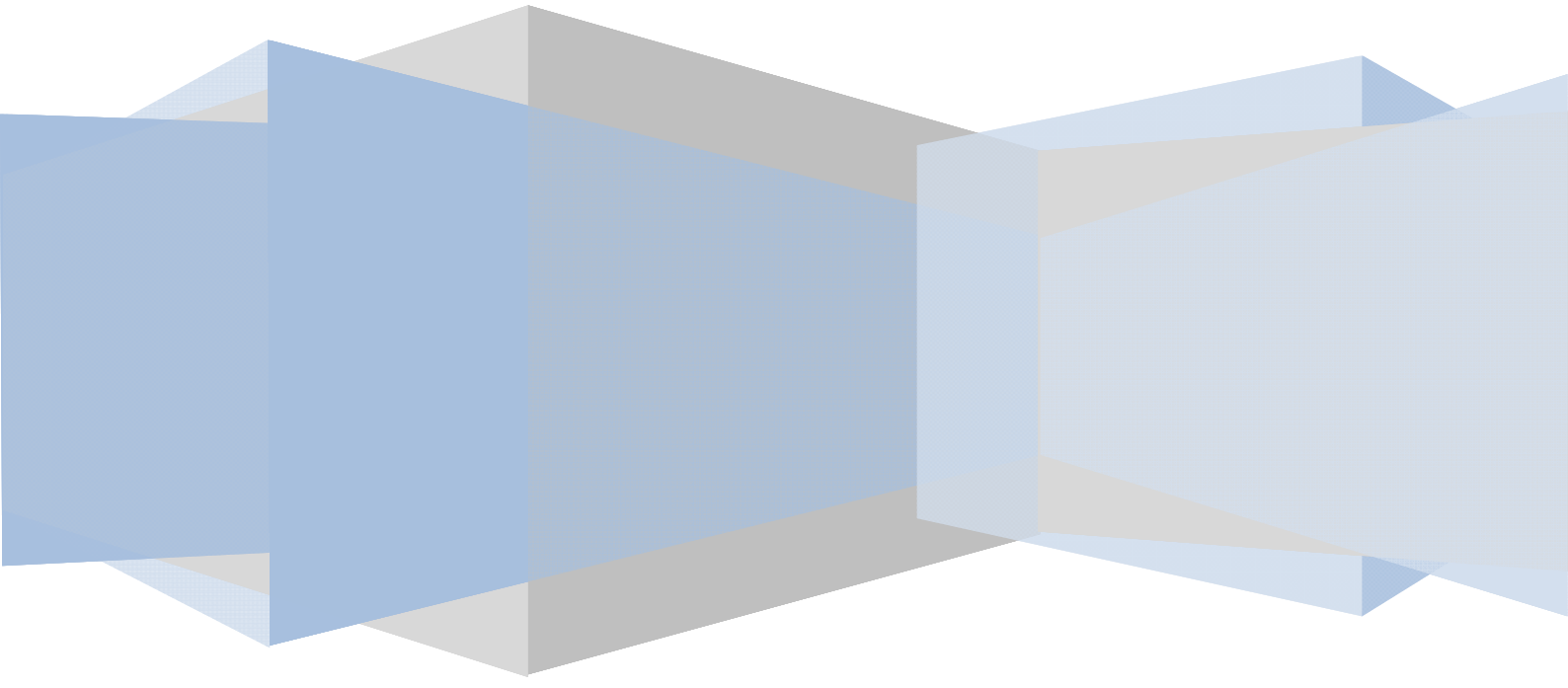
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2011-2012

Τεχνικές Εξόρυξης Δεδομένων (Data Mining)

Μελέτη Εφαρμογής σε Ελληνική Εταιρεία του κλάδου της Ένδυσης με τη Χρήση του Λογισμικού Weka

Παναγιωτάκος Θεόδωρος



Ευχαριστίες:

Θα ήθελα να ευχαριστήσω τόσο τον υπεύθυνο καθηγητή μου, κ. Σταύρο Πόνη, όσο και τον ιδιοκτήτη της Νότα Μασσέλος Α.Ε, κ. Βασίλειο Μασσέλο, και τον υπεύθυνο του τμήματος Πληροφορικής της εταιρείας, κ. Αναστάσιο Σωφρονά, για την αμέριστη βοήθεια που μου προσέφεραν για την ολοκλήρωση της εργασίας και την κατανόηση και την υπομονή που επέδειξαν κατά την περίοδο προσωπικών προβλημάτων.

Περιεχόμενα

Εικόνες

Πίνακες

Περίληψη της εργασίας

Η παρούσα εργασία έχει ως στόχο την μελέτη των μεθόδων Data Mining και την εφαρμογή τους για την επίλυση υπαρκτού προβλήματος και την εξαγωγή γνώσης από δεδομένα με τη χρήση του Weka. Στο περιεχόμενο της εργασίας παρουσιάζονται η σημασία και οι τεχνικές του Data Mining καθώς και τα πεδία στα οποία εφαρμόζεται. Στη συνέχεια περιγράφονται οι σημαντικότερες μέθοδοι του Data Mining (ταξινόμηση, συσχέτιση, ομαδοποίηση) και οι σπουδαιότεροι αλγόριθμοι που συναντώνται. Έπειτα παρουσιάζεται το λογισμικό που χρησιμοποιήθηκε για την εφαρμογή των μεθόδων αυτών, το Weka. Ακόμα παρουσιάζεται η εταιρεία (Νότα Μασσέλος Α.Ε.) με την οποία υπήρξε συνεργασία για την εκπόνηση της εργασίας και περιγράφονται οι στόχοι της από την εργασία αυτή, τόσο η δημιουργία μιας mailing list όσο και η ανάλυση του καλαθιού αγοράς. Στη συνέχεια παρουσιάζονται τα δεδομένα συναλλαγών που παραχωρήθηκαν για την εργασία και αναλύεται η μελέτη και η διαδικασία του Data Mining που ακολουθήθηκε: η προεπεξεργασία των δεδομένων, η εφαρμογή των μεθόδων της ομαδοποίησης, μέσω των αλγορίθμων EM και KMeans για τον διαχωρισμό των πελατών σε ομάδες, της ταξινόμησης, μέσω των αλγορίθμων Naïve Bayes αλλά και των JRip και Part για την ανακάλυψη της δομής των ομάδων αυτών αλλά και των κανόνων που τις διέπουν, και της συσχέτισης στα δεδομένα για την επιλογή της mailing list και την ανάλυση του καλαθιού αγοράς, μέσω των αλγορίθμων Apriori και FPGrowth. Τέλος παρουσιάζονται τα αποτελέσματα από την ανάλυση και τα συμπεράσματα που προκύπτουν από αυτή.

Summary of project

This project is about studying Data Mining methods and applying them to deal with an existing problem and extract knowledge out of data using Weka. We refer to the importance of Data Mining and the techniques it uses and also the fields they are used to. Next we describe the main methods of Data Mining, such as Classification, Association and Clustering, and the most important algorithms of it. Additionally, we present the software that was used in the project, Weka. Then we present the business case we worked on, the company we collaborated with, Nota Masselos SA, and which provided us with the data and the goals they set for this project, the creation of a mailing list for their new collection and a market basket analysis. Furthermore, we describe the transactional data that was provided for the project and analyze the procedure of Data Mining that took place: the preprocess of the data, the application of the methods of Clustering, through the EM and KMeans algorithms in order to create clusters of customers, Classification, through the use of the Naïve Bayes, JRip and Part algorithms to reveal the structure of the clusters and the rules that govern them, and Association, through the Apriori and FPGrowth algorithms in order to select the optimum mailing list and to do the market basket analysis. Finally we present the results that they produce and the conclusions that we came to.

1. Εισαγωγή

Ο αιώνας που διανύουμε έχει χαρακτηριστεί από πολλούς ως ο αιώνας της πληροφορίας. Σημαντικό εφόδιο για να πρωταγωνιστήσει κάποιος, σε όποιον τομέα και αν δραστηριοποιείται, είναι να έχει την ικανότητα να συγκεντρώνει πληροφορίες και δεδομένα και στα συνέχεια να τα αξιολογεί. Η αξιολόγηση και η αξιοποίηση της πληροφορίας και των δεδομένων είναι τα στοιχεία αυτά που θα επιτρέψουν στον κάθε ενδιαφερόμενο να αποκτήσει ανταγωνιστικό πλεονέκτημα στο χώρο που δραστηριοποιείται και να λάβει τις βέλτιστες αποφάσεις, δεδομένων των εκάστοτε συνθηκών, σε θέματα που τον αφορούν. Τέτοιου είδους αναλύσεις, που λαμβάνουν χώρα σε ποιοτικά αλλά και σε αριθμητικά δεδομένα πραγματοποιούνται, μεταξύ άλλων, με τη βοήθεια της επιστήμης της Τεχνητής Νοημοσύνης και πιο συγκεκριμένα με τη χρήση των τεχνικών Εξόρυξης Δεδομένων (*Data Mining*), οι οποίες δίνουν τη δυνατότητα εξαγωγής σχέσεων και κανόνων μέσω της χρήσης ηλεκτρονικών υπολογιστών [3].

1.1. Τεχνητή Νοημοσύνη

Η «Τεχνητή Νοημοσύνη» δεν είναι όρος που επιδέχεται έναν απλό και σαφή ορισμό. Ο όρος «Νοημοσύνη» από μόνος του επιδέχεται πολλές διαφορετικές προσεγγίσεις. Ο καθηγητής του MIT στον τομέα της Τεχνητής Νοημοσύνης, *Marvin Minsky*, υποστηρίζει ότι «Τεχνητή Νοημοσύνη είναι η επιστήμη που επιχειρεί να κάνει τις μηχανές να δημιουργούν πράγματα, τα οποία θα απαιτούσαν τον ανθρώπινο παράγοντα για να πραγματοποιηθούν» [2]. Ο *Elaine Rich*, λέκτορας του University of Texas at Austin, στον τομέα της Τεχνητής Νοημοσύνης, αναφέρει πως «Τεχνητή Νοημοσύνη είναι η μελέτη του πώς να κάνουμε τους ηλεκτρονικούς υπολογιστές να κάνουν πράγματα για τα οποία, προς το παρόν, οι άνθρωποι είναι καλύτεροι» [3]. Ωστόσο κοινά αποδεκτό από την επιστημονική κοινότητα είναι ότι η Τεχνητή Νοημοσύνη σε μια μηχανή είναι κάτι πολύ περισσότερο από απλά την εισαγωγή δεδομένων σε αυτή (input of data), καθώς συμπεριλαμβάνει επίσης την δυνατότητα της μηχανής να επεξεργαστεί και να χρησιμοποιήσει αυτά τα δεδομένα, να τα αξιολογήσει και να εξάγει πληροφορίες από αυτά και τέλος να «μάθει» και να βελτιωθεί μέσα από τις εμπειρίες της επαφής της με το περιβάλλον.

Η Τεχνητή Νοημοσύνη γεννήθηκε στις αρχές της δεκαετίας του 1940, με το ενδιαφέρον να περιστρέφεται στην κατασκευή προγραμμάτων για το παίξιμο παιγνίων αλλά και την απόδειξη θεωρημάτων. Η κυρίως έμφαση δόθηκε στην κατασκευή συστημάτων τα οποία είχαν κάποιο βαθμό γενικής ευφυΐας ή ικανότητας για την επίλυση προβλημάτων από διάφορα πεδία και χώρους [2]. Η γλώσσα LISP που δημιουργήθηκε εκείνη την περίοδο (1958), υιοθετήθηκε από τους ερευνητές της Τεχνητής Νοημοσύνης και έδωσε σημαντική ώθηση στην εξέλιξη της επιστήμης. Η βασική κατεύθυνση της Τεχνητής Νοημοσύνης και ο βασικός της σκοπός ήταν η δημιουργία δομών που παριστάνουν τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος αποθηκεύει δεδομένα, πληροφορία και γνώση και η

αντιστοιχία των δομών αυτών με τον ανθρώπινο συλλογισμό. Οι προσπάθειες όλες αφιερώθηκαν στην κατασκευή συστημάτων με περιορισμένο πεδίο εφαρμογής, που κατείχαν σημαντικό όγκο γνώσης για το συγκεκριμένο πεδίο (το πεδίο προσδιορίζει το ανάλογο αντικείμενο, επιστημονικό χώρο) [3] .

Οι σημαντικότεροι τομείς εφαρμογής της Τεχνητής Νοημοσύνης είναι:

- *Έξυπνα Ρομποτ – Intelligent Robotic*

Ρομποτ είναι μια ηλεκτρομηχανική συσκευή που μπορεί να προγραμματιστεί για να εκτελέσει συγκεκριμένες λειτουργίες όπως η μεταφορά αντικειμένων ή η χρήση συσκευών με ακρίβεια και αποδοτικά. Τα Έξυπνα Ρομποτ έχουν αισθητήρες διαφόρων ειδών που τους επιτρέπουν να αντιλαμβάνονται μεταβολές στο περιβάλλον τους και να αντιδρούν αναλόγως

- *Αυτοματοποιημένος Συλλογισμός – Automated Reasoning*

Ο Αυτοματοποιημένος Συλλογισμός περιλαμβάνει την κατανόηση των διαφόρων κομματιών της συλλογιστικής πορείας. Μέσα από αυτή τη διαδικασία είναι δυνατή η δημιουργία προγραμμάτων που να επιτρέπουν στους υπολογιστές να έχουν αυτόνομο λογισμό, σχεδόν ολοκληρωτικά. Οι πιο ανεπτυγμένοι κλάδοι του Αυτοματοποιημένου Συλλογισμού είναι η Αυτοματοποιημένη Απόδειξη Θεωρημάτων (*Automated Theorem Proving*) και ο Αυτοματοποιημένος Έλεγχος Αποδείξεων (*Automated Proof Checking*)

- *Έμπειρα Συστήματα – Expert System*

Τα Έμπειρα Συστήματα είναι μια προσπάθεια αυτοματοποίησης της συμβουλευτικής διαδικασίας από έναν εμπειρογνώμονα. Προγράμματα ηλεκτρονικών υπολογιστών αναλαμβάνουν τον ρόλο του συμβούλου, ενώ η γνώση τους για το εκάστοτε πεδίο βασίζεται στην εμπειρία που έχει αποκτηθεί.

- *Υπολογιστική Όραση – Computer Vision*

Ο στόχος της έρευνας στην Υπολογιστική Όραση είναι να δώσει στους υπολογιστές τη δυνατότητα να αντιλαμβάνονται το περιβάλλον τους. Η Τεχνητή Νοημοσύνη βοηθά στην ανάλυση των εικόνων που λαμβάνουν οι υπολογιστές μέσω καμερών, και στην αναγνώριση αντικειμένων και σχέσεων που πιθανώς υπάρχουν μεταξύ τους

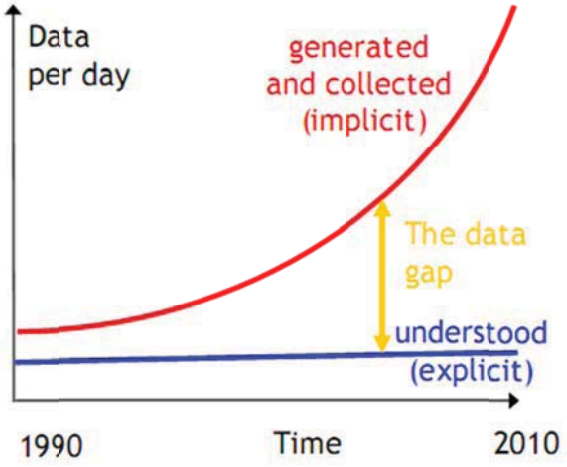
- *Μηχανική Μάθηση – Machine Learning*

Αναφέρεται στην ικανότητα ενός συστήματος για μάθηση από μόνο του και στην διερεύνηση των μηχανισμών και των υπολογιστικών διαδικασιών που επιτρέπουν την εξαγωγή και οργάνωση της γνώσης από την υπάρχουσα εμπειρία. Σε αυτόν τον τομέα βασίζονται οι τεχνικές *Data Mining*.

1.2. Από τα δεδομένα στη γνώση

Ο όγκος των αποθηκευμένων δεδομένων παγκοσμίως αυξάνεται συνεχώς. Παντοδύναμοι ηλεκτρονικοί υπολογιστές καθιστούν την αποθήκευση οποιουδήποτε είδους δεδομένων μια πολύ απλή υπόθεση. Φθηνοί δίσκοι με αποθηκευτική ικανότητα πολλών gigabyte συνεισφέρουν στην συνεχιζόμενη αύξηση των αποθηκευμένων δεδομένων. Ο Παγκόσμιος Ιστός μας κατακλύζει καθημερινά με δεδομένα. Σύμφωνα με το νόμο του Moore, η ταχύτητα των υπολογιστών διπλασιάζεται κάθε 18 μήνες ενώ τα δεδομένα που αποθηκεύονται διπλασιάζονται μόλις κάθε 9 μήνες [04]. Αν υποθέσει κανείς ότι ο χρόνος που δαπανάται για την επεξεργασία αυτών των δεδομένων παραμένει σταθερός, είναι εύκολα αντιληπτό το αυξανόμενο χάσμα μεταξύ της παραγωγής δεδομένων και της επεξεργασίας και εξαγωγής πληροφορίας από αυτά [5]. Από τα συλλεγόμενα δεδομένα μόνο ένα μικρό ποσοστό (5-10%) τυγχάνει ανάλυσης. Τη στιγμή λοιπόν που μια τυπική επιχειρησιακή βάση δεδομένων σήμερα περιέχει μεγάλο αριθμό εγγραφών ($10^8 - 10^{12}$) δεδομένων πολλών διαστάσεων ($10 - 10^4$), είναι πιο επίκαιρη από ποτέ η ρήση «Πνιγόμαστε στα δεδομένα αλλά λιμοκτονούμε για γνώση»

(«We are drowning in data but starving for knowledge» - John Naisbett) [05].



Εικόνα 1: ΑΝΕΚΜΕΤΑΛΕΥΤΑ ΔΕΔΟΜΕΝΑ[1]

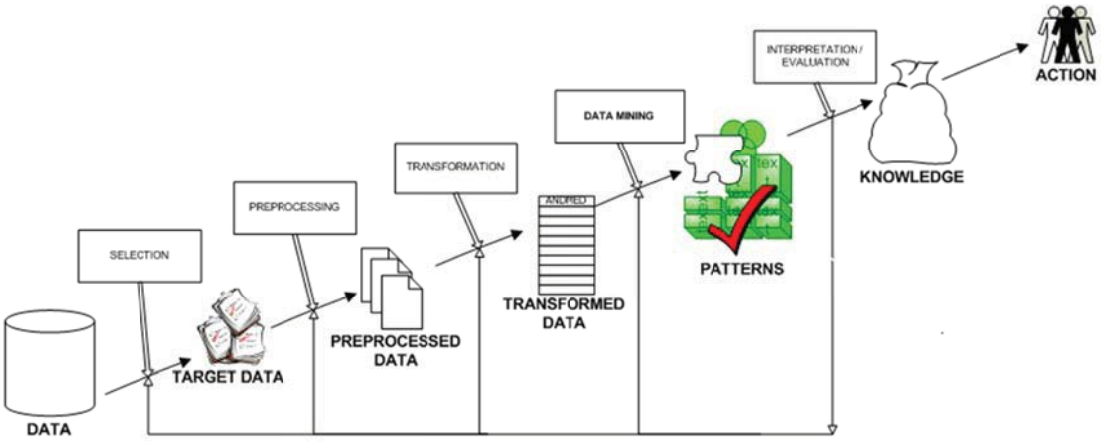
Για να εξερευνηθούν επομένως όλες αυτές οι εγγραφές πολλών μεταβλητών ώστε να εξαχθούν χρήσιμα συμπεράσματα από αυτές, απαιτείται μια εκ βαθέων μελέτη που οδηγεί στην ανακάλυψη γνώσης από τα δεδομένα (*Knowledge Discovery in Data – KDD*), γνώσης που θα δίνει αξία και νόημα στα δεδομένα [3]. Η KDD είναι μια μη τετριμμένη διαδικασία εύρεσης ***έγκυρων, πρωτότυπων, πιθανώς χρήσιμων*** και οπωσδήποτε ***κατανοητών*** προτύπων μέσα στα δεδομένα [04].

[1] *ΠΗΓΗ:* http://sites.google.com/site/gtziralis/Lecture01_Introduction.pdf

Η διαδικασία ανακάλυψης γνώσης αποτελείται από την επαναλαμβανόμενη αλληλουχία των ακόλουθων βημάτων:

1. Ολοκλήρωση Δεδομένων (*Data Integration*) – όπου συνδυάζονται δεδομένα από διάφορες πηγές
2. Επιλογή Δεδομένων (*Data Selection*) – όπου ανακτούνται όσα δεδομένα είναι σχετικά με το αντικείμενο ανάλυσης
3. Καθαρισμός Δεδομένων (*Data Cleaning*) – για την απομάκρυνση θορύβου ή ασυνεπών δεδομένων
4. Μετασχηματισμός Δεδομένων (*Data Transformation*) – όπου τα δεδομένα μετασχηματίζονται ή ενώνονται σε μορφές κατάλληλες για εξόρυξη πληροφοριών από αυτά
5. Εξόρυξη Πληροφοριών από τα Δεδομένα (*Data Mining*) – Μια απαραίτητη διαδικασία όπου εφαρμόζονται έξυπνες μέθοδοι για την εξαγωγή προτύπων
6. Αξιολόγηση Προτύπων (*Pattern Evaluation*) – για την αναγνώριση των πραγματικά ενδιαφερόντων προτύπων που αναπαριστούν γνώση, βάσει κάποιων μέτρων σύγκρισης
7. Παρουσίαση γνώσης (*Knowledge representation*) – όπου χρησιμοποιούνται τεχνικές οπτικοποίησης και παρουσίασης συμπερασμάτων για να την γνώση που εξορύχτηκε στον χρήστη.

Τα βήματα 1 έως 4 είναι τα διάφορα στάδια της προεπεξεργασίας των δεδομένων, ώστε αυτά να ετοιμασθούν για το βήμα 5 (*Data Mining*). Παρ'ότι σύμφωνα με τα παραπάνω το *Data Mining* είναι μόνο ένα βήμα στην όλη διαδικασία, είναι το πιο σημαντικό και το πλέον απαραίτητο καθώς σε αυτό γίνεται η εξόρυξη των προτύπων και της γνώσης που κρύβεται μέσα στα δεδομένα. Λόγω αυτής της σημαντικότητας του κι επειδή έχει επικρατήσει και στο χώρο της βιομηχανίας και των επιχειρήσεων, ο όρος *Data Mining* χρησιμοποιείται πολλές φορές για να περιγράψει όλη τη διαδικασία ανακάλυψης προτύπων από δεδομένα. [4]



Εικόνα 2: Η ΠΟΡΕΙΑ ΠΡΟΣ ΤΗΝ ΕΞΑΓΩΓΗ ΓΝΩΣΗΣ[2]

1.3. Η εργασία

Είναι λοιπόν εμφανές ότι οι διαδικασίες εξόρυξης πληροφοριών από δεδομένα αποκτούν όλο και μεγαλύτερη αξία στη σύγχρονη κοινωνία. Η μελέτη τους και η εφαρμογή τους κερδίζουν συνεχώς έδαφος στον σύγχρονο κόσμο της πληροφορίας. Η χρήση τους είναι αυξανόμενη τόσο σε επιστημονικές όσο και σε εμπορικές εφαρμογές και ανάγκες. Με βάση αυτές τις τάσεις που περιγράψαμε επιλέχθηκε ως θέμα της εργασίας η μελέτη των τεχνικών *Data Mining* και η πρακτική εφαρμογή τους σε ένα υφιστάμενο πρόβλημα.

Η πρακτική εφαρμογή έλαβε χώρα με τη συνδρομή της Νότα Μασσέλος Α.Ε. Η Νότα είχε στόχο την μελέτη και ανάλυση της αγοραστικής συμπεριφοράς των διαφόρων πελατών λιανικής ώστε να εξάγει πρότυπα και πληροφορίες για τις αγορές που γίνονται από την εταιρεία αλλά και να διανείμει καταλόγους με τα προϊόντα της στοχεύοντας στους πελάτες εκείνους που ήταν πιο πιθανό να ανταποκριθούν. Απώτερος σκοπός μεταξύ άλλων ήταν η αύξηση των πωλήσεων στους υφιστάμενους πελάτες με στοχευμένη όμως διαφημιστική καμπάνια που θα είχε όσο το δυνατόν καλύτερο συντελεστή κόστους – απόκρισης από τους παραλήπτες. Η εξαγωγή ομάδων πελατών για στοχευμένη διαφήμιση είναι ένα από τα σημεία που χρησιμοποιούνται οι τεχνικές *Data Mining* ως επί το πλείστον, καθώς μπορούν να εξάγουν πρότυπα από πληθώρα δεδομένων τα οποία μπορεί να φαίνονται αχανή ή υπερβολικά πολύπλοκα. Έτσι ξεκίνησε η συνεργασία με τη Νότα, τα αποτελέσματα της οποίας περιγράφονται στην παρούσα εργασία και η οποία ήταν προσοδοφόρα τόσο για την εταιρεία όσο και για τον ίδιο τον συντάκτη του παρόντος κειμένου καθώς είχε την ευκαιρία να ασχοληθεί με το αντικείμενο και να εξασκηθεί πάνω σε πραγματικά δεδομένα.

[2] *ΠΗΓΗ:* http://student.dcu.ie/~aldeirm2/data-mining.gif

Πιο συγκεκριμένα στην παρούσα εργασία συναντάμε:

Στο πρώτο κεφάλαιο μια εισαγωγή την οποία ήδη διαβάσαμε και αποτελεί μια εισαγωγή στην διαδικασία εύρεσης προτύπων και εξαγωγής πληροφοριών στην σύγχρονη κοινωνία της πληροφορίας.

Στο δεύτερο κεφάλαιο γίνεται μια αναφορά στην επιστήμη του *Data Mining*, τις διαδικασίες που ακολουθεί κανείς όταν την εφαρμόζει και τα εργαλεία που μπορεί να χρησιμοποιήσει, καθώς και τις μεθόδους που συναντά κανείς στην ασχολία του με το αντικείμενο. Επίσης υπάρχει ένα κομμάτι αφιερωμένο στα ηθικά ζητήματα που μπορεί να προκύψουν εφαρμόζοντας την επιστήμη αυτή

Στο τρίτο κεφάλαιο παρουσιάζονται οι σπουδαιότεροι αλγόριθμοι *Data Mining* όπως αυτοί αναγνωρίστηκαν από τη «*Διεθνής Διάσκεψη για το Data Mining», (IEEE International Conference on Data Mining (IEEE ICDM))*, που έλαβε χώρα στο Hong Kong το 2006 καθώς και μια σύντομη περιγραφή του καθενός από αυτούς.

Στο τέταρτο κεφάλαιο γίνεται μια παρουσίαση του Weka, του λογισμικού που χρησιμοποιήθηκε στην παρούσα εργασία, και μια αναφορά στα διάφορα περιβάλλοντα που μπορεί κάποιος να το χρησιμοποιήσει. Επίσης υπάρχει και μια αναφορά για την μετατροπή των διαφόρων τύπων αρχείων σε αρχεία τύπου *.arff* που είναι αυτά που χρησιμοποιεί το Weka.

Στο πέμπτο κεφάλαιο παρουσιάζεται η Νότα Μασσέλος Α.Ε. και γίνεται μια περιγραφή τόσο της εταιρείας όσο και του αντικειμένου που μας αφορά και των στόχων που αυτή έχει μέσα από την παρούσα εργασία.

Στο έκτο κεφάλαιο αναφερόμαστε στην προετοιμασία των δεδομένων για την εφαρμογή των τεχνικών *Data Mining* και πιο συγκεκριμένα ασχολούμαστε με την αρχική εξερεύνηση των δεδομένων, τον καθαρισμό και τους πιθανούς μετασχηματισμούς που χρειάζονται αλλά και την αντιμετώπιση τεχνικών ζητημάτων που ενδεχομένως να συναντώνται.

Στο έβδομο κεφάλαιο παρουσιάζεται η πρακτική εφαρμογή των τεχνικών που μελετάμε και αναφέρεται στην επεξεργασία των δεδομένων μέσω διαφόρων μεθόδων *Data Mining* και τα αποτελέσματα αυτών παρουσιαζόμενων σε κατάλληλους πίνακες.

Τέλος στο όγδοο κεφάλαιο γίνεται μια αποτίμηση της εργασίας και των αποτελεσμάτων της και παρατίθενται κάποια συμπεράσματα και κάποιες προτάσεις που προέκυψαν κατά την ενασχόληση με το αντικείμενο.

2. Data Mining

Με τον όρο *Data Mining* περιγράφουμε την αυτόματη ή ημιαυτόματη μη τετριμμένη διαδικασία, μέσω χρήσης ηλεκτρονικού υπολογιστή, εξαγωγής χρήσιμων πληροφοριών και προτύπων από μεγάλες βάσεις δεδομένων [06].

Εικόνα 3: ΑΝΑΚΑΛΥΨΗ ΠΡΟΤΥΠΩΝ ΣΕ ΔΕΔΟΜΕΝΑ[3]

Γενικά ο όρος *Data Mining* αναφέρεται σε υψηλού επιπέδου εφαρμογές και μεθόδους που χρησιμοποιούνται για να παρουσιάσουν και να αναλύσουν δεδομένα σε πεδία λήψης αποφάσεων [3]. Η βασική ιδέα που κρύβεται πίσω από τον όρο *Data Mining* είναι η εύρεση εκείνης της μη μηδενικής λύσης, η οποία δίνει τη δυνατότητα εξαγωγής ουσιαστικών κανόνων από δεδομένα [2]. Η όλη διαδικασία βασίζεται στην χρησιμοποίηση αλγορίθμων οι οποίοι αναζητούν κανόνες μεταξύ των μεταβλητών των δεδομένων και στην συνέχεια καταχωρούν τα δεδομένα σε νέες βάσεις δεδομένων [6]. Το *Data Mining* αποτελεί σύγχρονη και πειραματική επιστήμη, με το πρώτο σχετικό συνέδριο να πραγματοποιείται το 1995 [05].

2.1. Στόχοι

Οι κυριότεροι στόχοι της τεχνικής *Data Mining* είναι η Πρόβλεψη (*Prediction*) και η Περιγραφή (*Description*) [7]. Η Πρόβλεψη χρησιμοποιεί τις υπάρχουσες μεταβλητές σε μια βάση δεδομένων, ώστε να προβλέπει άγνωστες ή μελλοντικές αξίες ενδιαφέροντος. Η Περιγραφή από την άλλη μεριά επικεντρώνεται στην εύρεση προτύπων και κανόνων, περιγράφοντας τις γενικές ιδιότητες των δεδομένων [4].

[3] *ΠΗΓΗ:* http://techpubs.sgi.com/library/dynaweb_docs/books/MineSet_T/sgi_html/figures/dataprocess.gif

2.2. Συστατικά Στοιχεία Αλγορίθμων

Οι αλγόριθμοι που χρησιμοποιούνται στο *Data Mining* αποτελούνται από τρία συστατικά στοιχεία [7]:

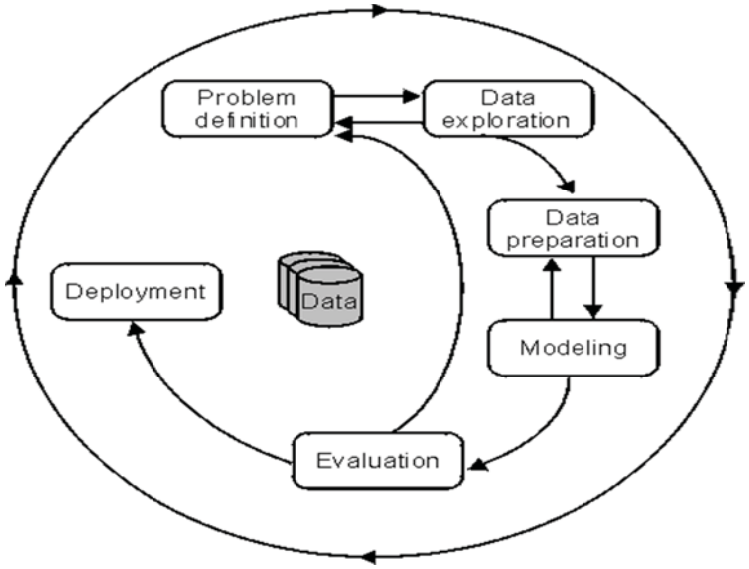
1. Model Representation – Αναπαράσταση Μοντέλου
   Είναι η γλώσσα L, η οποία περιγράφει ανακαλυφθείσες μεθόδους. Αν η αναπαράσταση είναι μικρή και περιορισμένη αρκετά, τότε τίποτα από παραδείγματα ή από χρόνο εκπαίδευσης δεν είναι αρκετό για να παράγει ένα ικανοποιητικό μοντέλο για τα δεδομένα

2. Model Evalutation – Αξιολόγηση Μοντέλου
   Υπολογίζει κατά πόσο μια συγκεκριμένη μέθοδος, ένα μοντέλο με τις παραμέτρους του, πληρεί τα κριτήρια της διαδικασίας KDD (*Knowledge Discovery in Data*)

3. Search Method – Μέθοδος Αναζήτησης
   Αποτελείται από δύο επιμέρους στοιχεία. Το πρώτο είναι η parameter search όπου ο αλγόριθμος αναζητεί παραμέτρους που συνοψίζουν τα μοντέλα αποτίμησης που έλαβαν δεδομένα, ενώ το δεύτερο είναι η model research όπου γίνεται μια διορθωτική έρευνα όλου του μοντέλου [3]

2.3. Η Διαδικασία του Data Mining

Η διαδικασία του *Data Mining* αποτελείται από τρία κύρια στάδια: (1) το στάδιο της αρχικής εξερεύνησης των δεδομένων (*the initial exploration*), (2) το στάδιο της κατασκευής του κατάλληλου μοντέλου πρόβλεψης και της αξιολόγησής του (*model building and evaluation*) και (3) το στάδιο της εφαρμογής του μοντέλου (*deployment*)



Εικόνα 4: ΤΑ ΣΤΑΔΙΑ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ Data Mining[4]

2.3.1. Εξερεύνηση των Δεδομένων

Το στάδιο αυτό ξεκινά με την προετοιμασία των δεδομένων, κάτι που περιλαμβάνει τον καθαρισμό των δεδομένων από πιθανούς θορύβους, το μετασχηματισμό τους σε κατάλληλες μορφές, την επιλογή υποσυνόλου εγγραφών και στην περίπτωση που αφορά σετ δεδομένων με μεγάλο αριθμό μεταβλητών την εφαρμογή κάποιων προκαταρκτικών διαδικασιών επιλογής χαρακτηριστικών ώστε να μειωθεί ο αριθμός των μεταβλητών σε ένα μέγεθος κατάλληλο για την εκάστοτε μέθοδο. Έπειτα, αναλόγως τη φύση του προβλήματος, αυτό το πρώτο στάδιο μπορεί να περιλαμβάνει μια απλή επιλογή από μεθόδους πρόβλεψης για ένα μοντέλο παλινδρόμησης, χρησιμοποιώντας μια ποικιλία από γραφικές και στατιστικές μεθόδους , ώστε να προσδιορίσει τις πιο σχετικές μεταβλητές και να καθορίσει την πολυπλοκότητα και την γενική φύση των μοντέλων που μπορούν να χρησιμοποιηθούν σε επόμενα βήματα. [011]

2.3.2. Κατασκευή Μοντέλου Πρόβλεψης

Το στάδιο αυτό περιλαμβάνει την δοκιμή διαφόρων μοντέλων και την επιλογή του πιο κατάλληλου, με κριτήριο την απόδοση των προβλέψεων του. Ίσως ακούγεται σαν μια απλή διαδικασία αλλά πολλές φορές πρόκειται για μια πολύπλοκη και δύσκολη διαδικασία. Έχει αναπτυχθεί πλήθος τεχνικών για να επιτύχουν αυτό τον σκοπό, με τις περισσότερες από αυτές να βασίζονται στο *competitive evaluation of models,* δηλαδή στην εφαρμογή διαφορετικών μοντέλων στο ίδιο σύνολο δεδομένων και στη συνέχεια στην σύγκριση των αποδόσεων τους για την επιλογή του πιο κατάλληλου. Τέτοιες τεχνικές, που συχνά θεωρούνται ο πυρήνας του

[4] *ΠΗΓΗ:* http://publib.boulder.ibm.com/infocenter/db2luw/v9/topic/com.ibm.im.easy.doc/idmk0s01.gif

predictive data mining είναι: *Bagging (voting, averaging), Boosting, Stacking (stacked generalizations)* και *Meta-learning.* [011]

2.3.3. Εφαρμογή του Μοντέλου

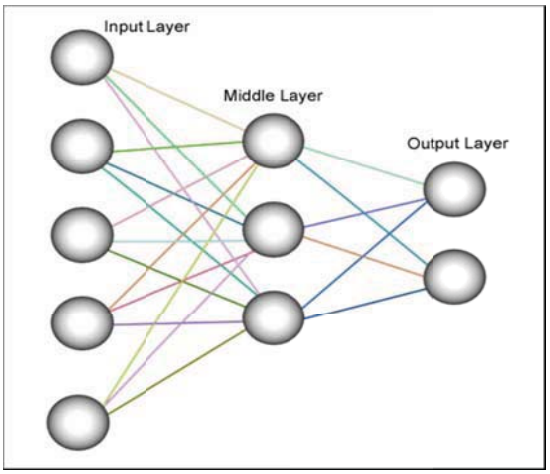
Το τελικό στάδιο αναφέρεται στην χρήση του μοντέλου που επιλέχθηκε από προηγούμενα στάδια ως το βέλτιστο και την εφαρμογή του σε νέα δεδομένα ώστε να παράγει προβλέψεις ή να εκτιμήσει το αποτέλεσμα. [011]

2.4. Χρησιμοποιούμενα Εργαλεία

Τα εργαλεία εξόρυξης (*Mining Tools or Search Engines*) που χρησιμοποιούνται είναι συνήθως «έξυπνα», από το πεδίο της Τεχνητής Νοημοσύνης, και σχετικά με βάσεις δεδομένων. Τα κυριότερα είδη αυτών των εργαλείων αναφέρονται στη συνέχεια [2].

- *Τεχνητά Νευρωνικά Δίκτυα – Artificial Neural Networks*

  Τα Νευρωνικά Δίκτυα είναι μη γραμμικά μοντέλα πρόβλεψης για λήψη αποφάσεων, τα οποία χρησιμοποιούν υπάρχοντα δεδομένα που έχουν γνωστά αποτελέσματα (outcomes), για να εκπαιδεύσουν ένα μοντέλο το οποίο να μπορεί μετά να χρησιμοποιηθεί για να κάνει προβλέψεις [3]. Η τυπική τους μορφή αποτελείται από ένα δίκτυο από παράλληλες μονάδες επεξεργασίας, οργανωμένες σε μια σειρά επιπέδων, και συνδεδεμένες μεταξύ τους μέσω του κατάλληλου βάρους (*weight*), που χαρακτηρίζει την ισχύ της σύνδεσης [2]. Βάση της μορφής αυτής είναι η προσπάθεια εξομοίωσης του δικτύου με τον τρόπο λειτουργίας των νευρώνων στον ανθρώπινο εγκέφαλο κατά την επεξεργασία των σημάτων που λαμβάνει από το περιβάλλον.

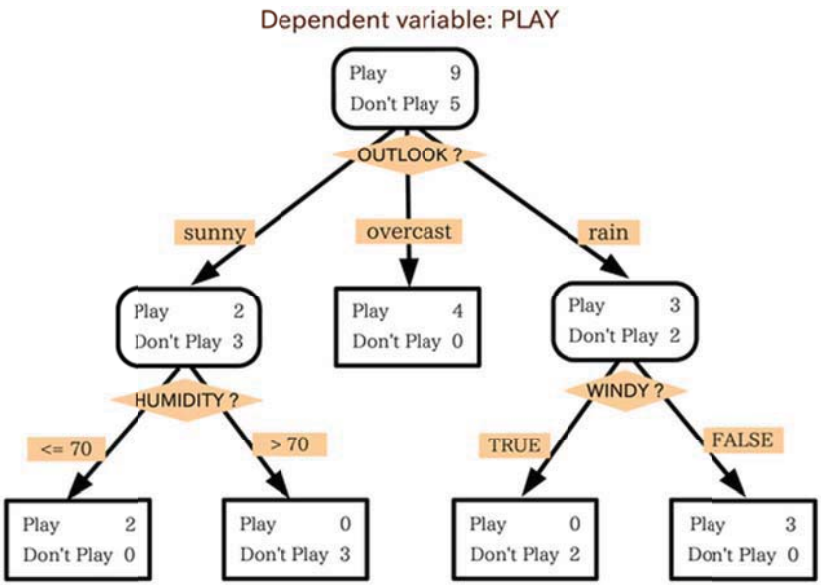


Εικόνα 5: ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ[5]

[5] *ΠΗΓΗ:* http://www.gamedev.net/reference/programming/features/vehiclenn/figure1.png

- *Δέντρα Αποφάσεων – Decision Trees*

Τα Δέντρα Αποφάσεων είναι μοντέλα στήριξης της λήψης αποφάσεων, τα οποία δημιουργούν κανόνες ώστε να ταξινομούν ένα σύνολο δεδομένων [3]. Κάθε κόμβος του Δένδρου ελέγχει την τιμή ενός χαρακτηριστικού, κάθε κλαδί αναπαριστά ένα πιθανό αποτέλεσμα του ελέγχου και κάθε φύλλο αναπαριστά κάποια ταξινόμηση ή κάποιο σύνολο ταξινομήσεων. Ένα άγνωστης τάξεως υπόδειγμα ακολουθεί πορεία από την αρχή έως κάποιο φύλλο του Δένδρου για την ταξινόμηση του [09].



Εικόνα 6: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ[6]

- *Επαγωγή Κανόνων – Rule Induction*

Ο όρος «Επαγωγή Κανόνων» αναφέρεται στην χρήση if-then κανόνων σε σύνολα από δεδομένα, οι οποίοι βασίζονται κυρίως σε στατιστικά μοντέλα [3]. Η προϋπόθεση του Κανόνα είναι συνήθως ένα σύνολο από ελέγχους, οι οποίοι συμπλέκονται με λογικές συζεύξεις ενώ το αποτέλεσμα είναι η εκχώρηση κάποιας τιμής ή κάποιας ταξινόμησης ή συνόλου ταξινομήσεων [09].

- *Γενετικοί Αλγόριθμοι – Genetic Algorithms*

Οι Γενετικοί Αλγόριθμοι είναι μια τεχνική που βασίζεται σε στοιχεία/συστατικά φυσικής εξέλιξης (*natural evolution*), χρησιμοποιώντας γενετικούς συνδυασμούς [2].

[6] *ΠΗΓΗ:* http://gautam.lis.illinois.edu/monkmiddleware/public/analytics/decision_tree.png

- *Μέθοδος Γειτνίασης – Nearest Neighbor Method*

Είναι μια τεχνική ταξινόμησης δεδομένων όπου το κάθε νέο υπόδειγμα συγκρίνεται με τα ήδη υπάρχοντα υποδείγματα με χρήση κατάλληλου μέτρου απόστασης και τελικά εκχωρείται στην τάξη του πλησιέστερου υποδείγματος [09].

- *Απεικόνιση Δεδομένων – Data Visualization*

Είναι μια εικονική αναπαράσταση των σχέσεων μεταξύ των δεδομένων

2.5. Μέθοδοι Data Mining

Στο *Data Mining* χρησιμοποιούνται πολλά είδη μεθόδων που χρησιμοποιούν κανόνες Μηχανικής Μάθησης, όπως:

- Ταξινόμηση - Classification
- Συσχέτιση - Association
- Ομαδοποίηση - Clustering
- Στατιστική Ανάλυση – Statistical Analysis
- Αριθμητική Πρόβλεψη – Numerical Prediction
- Παλινδρόμηση – Regression Analysis

και άλλες. Θα αναλύσουμε τις τρεις πρώτες παρακάτω καθώς είναι οι πιο σημαντικές και πιο συχνά χρησιμοποιούμενες

2.5.1. Ταξινόμηση

Αποτελεί την πιο δημοφιλή και αποτελεσματική μέθοδο. Οι αλγόριθμοι ταξινόμησης εφαρμόζονται σε δεδομένα, τα οποία έχουν πρώτα ταξινομηθεί σε συγκεκριμένες κλάσεις, με σκοπό την εξαγωγή κανόνων οι οποίοι μπορούν μετέπειτα να χρησιμοποιηθούν για την ταξινόμηση νέων υποδειγμάτων στις ίδιες κλάσεις. Ένα σύνολο εξαγόμενων κανόνων ονομάζεται ταξινομητής (*classifier*). Η λειτουργία ενός αλγορίθμου ταξινόμησης έχει ως εξής:

- Τροφοδοτούμε με ένα σύνολο από δεδομένα (*dataset*) έναν αλγόριθμο ταξινόμησης
- Ο αλγόριθμος έπειτα «κατανοεί» τους κανόνες βάσει των οποίων ταξινομήθηκαν τα δεδομένα
- Στη συνέχεια, βάσει αυτών των κανόνων, ο αλγόριθμος έχει την ικανότητα να ταξινομεί νέα δεδομένα

Ανάλογα το είδος του ταξινομητή, οι αλγόριθμοι ταξινόμησης χωρίζονται σε αυτούς που παράγουν λίστες αποφάσεων και σε αυτούς που παράγουν δέντρα αποφάσεων. Οι τελευταίοι αποτελούν την πιο παλιά μορφή και έκφραση του Data Mining [3].

2.5.2. Συσχέτιση

Σκοπός της είναι η εύρεση των σημαντικότερων αλληλεξαρτήσεων μεταξύ των διαφόρων πεδίων/χαρακτηριστικών του συνόλου εκπαίδευσης[2]. Η πιο συνηθισμένη μορφή της μεθόδου είναι η ανάλυση του «καλαθιού της αγοράς» (*market basket analysis*), όπου σκοπός είναι να αναγνωρισθούν προϊόντα που αγοράζονται μαζί[3]. Ένας κανόνας συσχέτισης είναι μια έκφραση της μορφής $X \Rightarrow Y\ [S, C]$, όπου Χ και Υ είναι σύνολα τιμών των πεδίων, πχ σύνολα οικονομικών αγαθών. Το S και το C αναφέρονται στην *υποστήριξη* (*support*) και την *εμπιστοσύνη* (*confidence*) αντίστοιχα του κάθε κανόνα. Η υποστήριξη είναι οι επιτυχείς προβλέψεις του κανόνα ως προς το σύνολο των δεδομένων και είναι καθαρός αριθμός, ενώ η εμπιστοσύνη είναι οι επιτυχείς προβλέψεις του κανόνα ως προς τα υποδείγματα στα οποία αυτός εφαρμόζεται και εκφράζεται με ποσοστό. Στις περισσότερες περιπτώσεις τίθενται ελάχιστα όρια για την υποστήριξη και την εμπιστοσύνη των υπό αναζήτηση κανόνων [09].

2.5.3. Ομαδοποίηση

Οι αλγόριθμοι ομαδοποίησης είναι ιδιαίτερα διαδεδομένοι και η λογική τους μοιάζει αρκετά με αυτή των αλγορίθμων ταξινόμησης. Η ειδοποιός διαφορά είναι πως τα δεδομένα του συνόλου εκπαίδευσης δεν είναι προταξινομημένα. Αντίθετα το σύνολο των εγγραφών χωρίζεται σε ομάδες έτσι ώστε οι εγγραφές της ίδιας ομάδας να έχουν περισσότερες ομοιότητες μεταξύ τους με βάση κάποια προκαθορισμένα κριτήρια, απ'ότι με εγγραφές άλλων ομάδων. Σε ορισμένους αλγορίθμους ομαδοποίησης υπάρχει η δυνατότητα μια εγγραφή να ανήκει σε περισσότερες της μια ομάδες ταυτόχρονα (*διάγραμμα Venn*), ενώ άλλοι αλγόριθμοι δίνουν την πιθανότητα καταχώρησης του κάθε υποδείγματος σε μια ομάδα.

Η μέθοδος της ομαδοποίησης μπορεί να είναι είτε στατιστική είτε αριθμητική (*statistical/ numerical clustering*), οπότε χρησιμοποιούνται διάφορα αριθμητικά κριτήρια ομοιότητας και οι ομάδες που προκύπτουν περιγράφονται από αριθμητικές τιμές, είτε εννοιολογική (*conceptual clustering*) οπότε σε αυτή την περίπτωση ο προσδιορισμός των ομάδων βασίζεται στο νόημα και στις έννοιες που τα διάφορα αριθμητικά στοιχεία αντιπροσωπεύουν και οι τιμές που προκύπτουν είναι κατηγορικές και όχι αριθμητικές [8].

2.6. Ηθικά Ζητήματα

Η χρήση δεδομένων – ιδιαίτερα δεδομένων που αφορούν ανθρώπους – για *Data Mining* έχει σοβαρές ηθικές επιπλοκές. Όσοι εφαρμόζουν επομένως τεχνικές DM πρέπει να δρουν υπεύθυνα και με επίγνωση των ηθικών ζητημάτων που προκύπτουν από τις εφαρμογές τους.

Όταν εφαρμόζεται σε δεδομένα που αφορούν ανθρώπους το *Data Mining* χρησιμοποιείται συνήθως για να διακριτοποίηση και ταξινόμηση πχ ποιος θα πάρει το δάνειο ή την εκπτωτική προσφορά κλπ. Κάποια είδη διακριτοποίησης όμως δεν

είναι μόνο ανήθικα αλλά και παράνομα πχ διακριτοποίηση ανάλογα με το φύλλο ή τη θρησκεία ή ακόμα και τη φυλή. Το κατά πόσο είναι ανήθικα όμως εξαρτάται κάθε φορά από την εφαρμογή για την οποία χρησιμοποιούνται τα δεδομένα. Για παράδειγμα στην περίπτωση ιατρικής διάγνωσης, η χρήση τέτοιων προσωπικών στοιχείων είναι όχι μόνο ηθική και θεμιτή αλλά και αναγκαία. Η χρήση των ίδιων στοιχείων όμως κατά τη διάρκεια μελέτης της συμπεριφοράς ως προς την αποπληρωμή δανείων είναι μάλλον ανήθικη.

Το εύρος εφαρμογών του *Data Mining* συνεπάγεται την δυνατότητα χρήσης των δεδομένων που υπάρχουν σε ένα dataset για πολλούς διαφορετικούς σκοπούς απ'αυτόν για τον οποίο είχαν αρχικά συλλεχθεί. Είναι σημαντικό λοιπόν να έχουμε επίγνωση του ποιος έχει δικαίωμα χρήσης των δεδομένων, για ποιο σκοπό συλλέχθηκαν και τι είδους συμπεράσματα νομιμοποιείται κάποιος να εξάγει από αυτά. Τα όποια αποτελέσματα προκύπτουν από την εφαρμογή της τεχνικής θα πρέπει να συνοδεύονται και από κατάλληλες προειδοποιήσεις. Άλλωστε οι αλγόριθμοι του *Data Mining* είναι απλώς ένα εργαλείο, η αξιολόγηση και χρήση των αποτελεσμάτων τους είναι ζήτημα ανθρωπίνων αποφάσεων και όχι μηχανικής μάθησης [05].

3. Αλγόριθμοι Data Mining

Οι περιοχές εφαρμογής των τεχνικών του *Data Mining* όπως ήδη έχουμε αναφέρει είναι πάρα πολλές. Εφαρμογές επιστημονικές, ερευνητικές, εμπορικές, σε όλους τους τομείς της σύγχρονης κοινωνίας των πληροφοριών. Επόμενο λοιπόν είναι και η ανάπτυξη πολλών αλγορίθμων και άλλων τόσων τροποποιήσεων τους για την ορθότερη χρήση και εφαρμογή τους σε κάθε περίπτωση. Σε μια προσπάθεια να ξεχωρίσουν οι αλγόριθμοι που έχουν επηρεάσει πιο ισχυρά την κοινότητα του Data Mining, η «*Διεθνής Διάσκεψη για το Data Mining», (IEEE International Conference on Data Mining (IEEE ICDM))*, αναγνώρισε τους κορυφαίους αλγορίθμους το 2006 στο Hong Kong. Επιλέχθηκαν τόσο για την δημοτικότητά τους στον σχετικό επιστημονικό χώρο όσο και για την απλότητα αλλά και την αποδοτικότητά τους στις εφαρμογές με τις οποίες συσχετίστηκαν. Παρακάτω παρουσιάζονται μερικοί από αυτούς με μία σύντομη περιγραφή τους.

3.1. Ο αλγόριθμος C4.5

Συστήματα που δημιουργούν ταξινομητές είναι από τα πιο συχνά χρησιμοποιούμενα εργαλεία σε εφαρμογές *Data Mining*. Τέτοια συστήματα δέχονται σαν είσοδο έναν σύνολο δεδομένων *(dataset)*, αποτελούμενο από υποδείγματα που ανήκουν σε κάποιες τάξεις, και στη συνέχεια παράγουν έναν ταξινομητή ικανό να προσδιορίσει με ακρίβεια την τάξη στην οποία ένα νέο υπόδειγμα ανήκει.

Στις αρχές του 1980 ο *J. Ross Quinlan*, ένας ερευνητής στην επιστήμη της Μηχανικής Μάθησης, ανέπτυξε ένα αλγόριθμο για δέντρα αποφάσεων *(decision tree)*, τον ID3 (*Iterative Dichotomiser)*. Στην συνέχεια παρουσίασε τον διάδοχο του ID3, τον αλγόριθμο C4.5, ο οποίος αποτελεί ακόμα και σήμερα ορόσημο με το οποίο συγκρίνονται πολλοί πρόσφατα ανεπτυγμένοι αλγόριθμοι μάθησης με επίβλεψη.

<u>Δέντρα αποφάσεων</u>

Ο C4.5 υιοθετεί μια άπληστη *(greedy)* προσέγγιση στην κατά την οποία τα δέντρα αποφάσεων δημιουργούνται από πάνω προς τα κάτω *(top to down)* με μια αναδρομική διαδικασία διαίρει και βασίλευε *(divide and conquer)*. Με δεδομένο ένα σύνολο υποδειγμάτων ο αλγόριθμος δημιουργεί ένα αρχικό δέντρο χρησιμοποιώντας την τεχνική divide and conquer ως ακολούθως:

- Αν όλα τα υποδείγματα στο σύνολο ανήκουν στην ίδια τάξη ή το σύνολο υποδειγμάτων είναι πολύ μικρό, το δέντρο αποτελείται από ένα φύλλο με όνομα την τάξη με την μεγαλύτερη συχνότητα εμφάνισης.
- Αλλιώς γίνεται ένας έλεγχος με βάση μια μεταβλητή με δύο ή περισσότερες πιθανές τιμές. Ο έλεγχος αυτός γίνεται η βάση του δέντρου και από αυτή ξεκινούν κλαδιά για κάθε πιθανό αποτέλεσμα από τον έλεγχο. Έτσι το σύνολο των υποδειγμάτων «σπάει» σε μικρότερα σύνολα ανάλογα με το

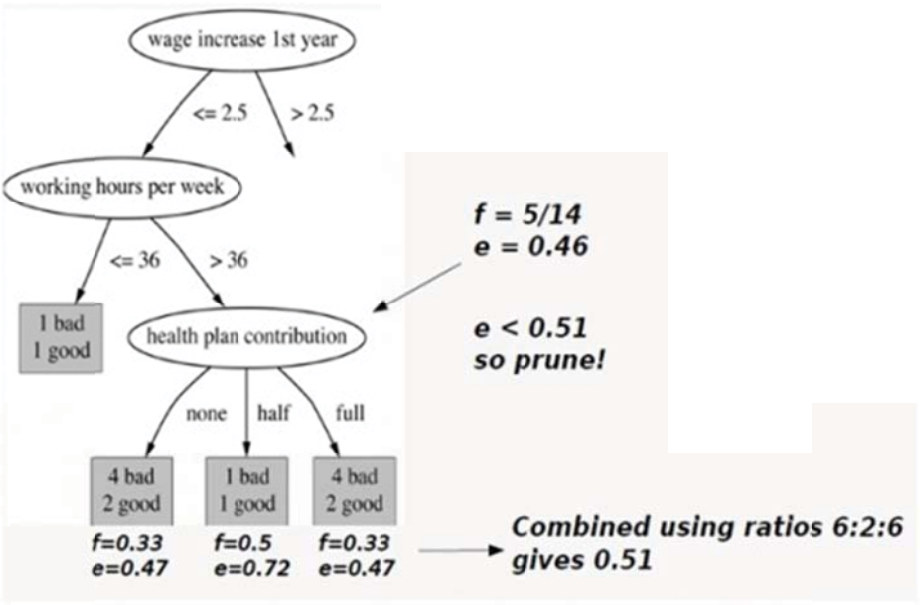
αποτέλεσμα του ελέγχου και για κάθε νέο σύνολο ακολουθείται αναδρομικά η παραπάνω διαδικασία.

Συνήθως υπάρχουν πολλοί πιθανοί έλεγχοι που μπορούν να επιλεγούν στην παραπάνω διαδικασία. Ο C4.5 χρησιμοποιεί δύο ευρετικά *(heuristic)* κριτήρια για να βαθμολογήσει αυτούς τους ελέγχους και να επιλέξει τον προτιμότερο:

- Το κέρδος πληροφορίας, το οποίο ελαχιστοποιεί την συνολική εντροπία των παραγόμενων υποσυνόλων (ωστόσο είναι θετικά προκατειλημμένο προς ελέγχους με πολλές πιθανές εξόδους)
- Και τον λόγο κέρδους, ο οποίος προκύπτει από τη διαίρεση του κέρδους πληροφορίας προς την πληροφορία που προκύπτει από τις πιθανές τιμές του ελέγχου.

Οι μεταβλητές που χρησιμοποιούνται μπορεί να είναι ονομαστικές ή αριθμητικές και αυτό καθορίζει την μορφή του ελέγχου. Για τις αριθμητικές μεταβλητές οι πιθανές μορφές είναι $\{A \le h, A \succ h\}$, όπου ***A*** η μεταβλητή και ***h*** ένα όριο που προκύπτει ταξινομώντας το σύνολο των υποδειγμάτων με βάση τις τιμές του ***A*** και επιλέγοντας το όριο μεταξύ δύο διαδοχικών τιμών που μεγιστοποιεί τα παραπάνω κριτήρια. Για τις μεταβλητές με διακριτές τιμές, υπάρχει εξ'ορισμού ένα αποτέλεσμα από τον έλεγχο για κάθε διαφορετική τιμή του ***A.*** Ωστόσο υπάρχει και η δυνατότητα ομαδοποίησης των διαφορετικών τιμών του ***A*** σε υποσύνολα, επιτρέποντας έτσι έναν έλεγχο για κάθε υποσύνολο.

Το αρχικό δέντρο που δημιουργείται στη συνέχεια υπόκειται σε κλάδεμα για να αποφευχθεί η υπερπροσαρμογή. Το κλάδεμα γίνεται από τα φύλλα προς τη ρίζα του δέντρου. Για ένα τμήμα του δέντρου, ο C4.5 βρίσκει το σταθμισμένο άθροισμα του εκτιμώμενου σφάλματος των κλαδιών και το συγκρίνει με το εκτιμώμενο σφάλμα που θα προέκυπτε αν το τμήμα αυτό του δέντρου το αντικαθιστούσε ένα φύλλο. Αν το δεύτερο είναι δεν είναι μεγαλύτερο από το πρώτο, το τμήμα αυτό του δέντρου κλαδεύεται. Με ανάλογο τρόπο ο αλγόριθμος ελέγχει αν ένα τμήμα του δέντρου θα μπορούσε να αντικατασταθεί από ένα κλαδί αυτού του τμήματος, συγκρίνοντας το εκτιμώμενο σφάλμα. Η διαδικασία του κλαδέματος ολοκληρώνεται με ένα πέρασμα από ολόκληρο το δέντρο.



Εικόνα 7: ΕΦΑΡΜΟΓΗ ΤΟΥ C4.5 ΣΕ ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ[7]

Κανόνες ταξινόμησης

Πολύπλοκα δέντρα αποφάσεων μπορεί να είναι αρκετά δυσνόητα π.χ. επειδή πληροφορίες για μια τάξη συνήθως είναι διανεμημένες σε ολόκληρο το δέντρο. Ο C4.5 εισήγαγε μια εναλλακτική διατύπωση αποτελούμενη από λίστες κανόνων της μορφής

- If A and B and C… then class X

όπου κανόνες για κάθε τάξη ομαδοποιούνται. Ένα υπόδειγμα ταξινομείται βρίσκοντας τον πρώτο κανόνα που οι συνθήκες του ικανοποιούνται. Αν δεν βρεθεί κανένας κανόνας, τότε το υπόδειγμα τοποθετείται στην προεπιλεγμένη τάξη.

Οι κανόνες του C4.5 προκύπτουν από το αρχικό δέντρο αποφάσεων, πριν υποστεί κλάδεμα. Κάθε διαδρομή από τη βάση του δέντρου ως κάποιο φύλλο γίνεται ένας πρωτότυπος κανόνας του οποίου οι συνθήκες είναι τα αποτελέσματα ελέγχων κατά μήκος της διαδρομής και του οποίου η τάξη είναι το όνομα του φύλλου. Στην συνέχεια ο κανόνας απλοποιείται, καθορίζοντας την επίδραση της απόρριψης κάθε συνθήκης κατά σειρά. Η απόρριψη μιας συνθήκης μπορεί να αυξήσει τον αριθμό των υποδειγμάτων που καλύπτονται από τον κανόνα ή τον αριθμό των υποδειγμάτων που δεν ανήκουν στην τάξη που υποδεικνύεται από τον κανόνα ή να μειώσει το εκτιμώμενο σφάλμα του κανόνα .Ένας αλγόριθμος αναρρίχησης λόφων *(hill-climbing)* χρησιμοποιείται για την απόρριψη συνθηκών από τον κανόνα μέχρι να βρεθεί το μικρότερο δυνατό σφάλμα. Για την ολοκλήρωση της διαδικασίας, ένα υποσύνολο απλοποιημένων κανόνων επιλέγεται για κάθε τάξη κάθε φορά. Αυτά τα υποσύνολα κατατάσσονται για την ελαχιστοποίηση του σφάλματος στα δεδομένα

[7] *ΠΗΓΗ:* http://octaviansima.files.wordpress.com/2011/03/c45-sample1.jpg

εκπαίδευσης και επιλέγετε μια τάξη που ορίζετε ως η προεπιλεγμένη. Το τελικό σύνολο περιέχει πολύ λιγότερους κανόνες από τον αριθμό των φύλλων στο δέντρο μετά το κλάδεμα.

Το κυριότερο μειονέκτημα των κανόνων ταξινόμησης του C4.5 είναι η ποσότητα του υπολογιστικού χρόνου και της χωρητικότητας που απαιτούνται. Σε πείραμα, επιλέχθηκαν δείγματα που αποτελούνταν από 10.000 έως 100.000 υποδείγματα. Για τα δέντρα αποφάσεων, η μετακίνηση από τα δείγματα των 10.000 σε αυτά των 100.000 υποδειγμάτων προκάλεσε αύξηση του υπολογιστικού χρόνου από 1,4 δευτερόλεπτα σε 61 δευτερόλεπτα, ένα συντελεστή της τάξης του 44. Ο χρόνος για την αντίστοιχη διαδικασία με σύνολα κανόνων αυξήθηκε από 32 δευτερόλεπτα σε 9.715 δευτερόλεπτα, συντελεστής της τάξης του 300.

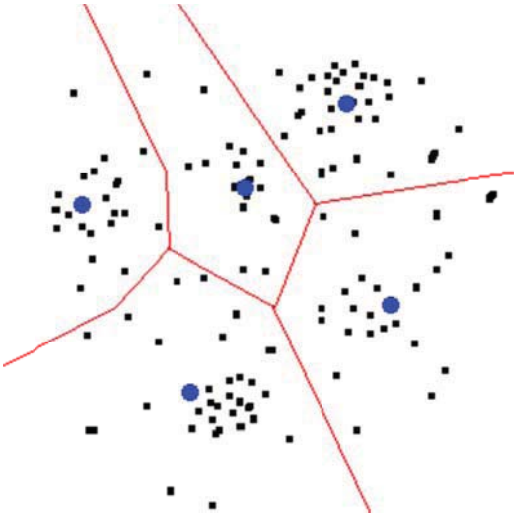
Το 1997 τον C4.5 διαδέχθηκε ο C5. Η τεχνική που ακολουθεί αυτός είναι αρκετά ταχύτερη και ελαφρά πιο ακριβής, ωστόσο αποτελεί εμπορικό αλγόριθμο και ως εκ τούτου δεν περιγράφεται στην ανοιχτή βιβλιογραφία.

3.2. Ο αλγόριθμος k-μέσων

Ο αλγόριθμος k-μέσων είναι μία απλή επαναληπτική μέθοδος για τον διαχωρισμό ενός σετ δεδομένων σε έναν καθορισμένο από το χρήστη αριθμό ομάδων, k. Ο αλγόριθμος αυτός ανακαλύφθηκε από ερευνητές από διάφορους επιστημονικούς κλάδους με πιο σημαντικούς τους Lloyd *(1957,1982)*, Forgey *(1965)*, Friedman και Rubin *(1967)* και McQueen *(1967)*.

Ο αλγόριθμος ξεκινά επιλέγοντας k σημεία στο χώρο των δεδομένων, τα οποία θα αποτελούν τα αρχικά σημεία αναφοράς της κάθε ομάδας, ονομαζόμενα και κεντροειδή *(centroids)*. Τεχνικές για την αρχική αυτή επιλογή των σημείων περιλαμβάνουν την τυχαία δειγματοληψία από το σύνολο των δεδομένων ή την μικρή μεταβολή του ολικού μέσου των δεδομένων k φορές. Στην συνέχεια ο αλγόριθμος επαναλαμβάνεται μεταξύ δύο βημάτων έως ότου να βρει σημεία σύγκλισης:

- *Βήμα 1: Ανάθεση των δεδομένων*. Κάθε σημείο από τα δεδομένα ανατίθεται στο πλησιέστερο κεντροειδές. Σε περίπτωση που ισαπέχει λαμβάνεται αυθαίρετα απόφαση. Αυτό οδηγεί στον διαχωρισμό των δεδομένων.
- *Βήμα 2*: *Επανατοποθέτηση των μέσων.* Το σημείο αναφοράς κάθε ομάδας επανατοποθετείται στο κέντρο των σημείων που την αποτελούν. Αν τα σημεία έχουν μέτρα πιθανότητας *(βάρη)* τότε η επανατοποθέτηση γίνεται στον σταθμισμένο μέσο των δεδομένων της ομάδας.



Εικόνα 8: ΔΙΑΧΩΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ ΜΕΣΩ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ Κ-ΜΕΣΩΝ[8]

Ο αλγόριθμος συγκλίνει όταν παύουν να μεταβάλλονται οι αναθέσεις των δεδομένων (και άρα και τα κεντροειδή). Σε κάθε επανάληψη απαιτούνται k επί το σύνολο των δεδομένων συγκρίσεις, κάτι που καθορίζει την πολυπλοκότητα σε χρόνο της κάθε επανάληψης. Ο αριθμός των επαναλήψεων που απαιτούνται για την σύγκλιση ποικίλλει και εξαρτάται σχεδόν γραμμικά από το όγκο των δεδομένων.

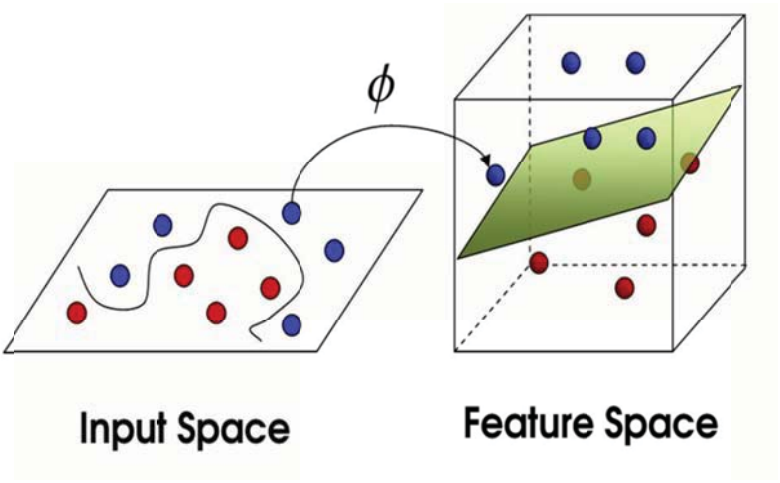
Κρίσιμο σημείο που πρέπει κανείς να έχει υπόψη του είναι η μέτρηση της «απόστασης» κατά την διαδικασία της ανάθεσης. Εξ'ορισμού η μέτρηση μπορεί να γίνει με την χρήση της *Ευκλείδειας απόστασης* όπου μπορεί κανείς να δει πως η μη αρνητική συνάρτηση κόστους $\sum_{i=1}^{N}(\arg\min_{j}\left\|x_i - c_j\right\|_2^2)$ φθίνει κάθε φορά που υπάρχει μεταβολή στην ανάθεση σημείου σε ομάδα ή στην επανατοποθέτηση τους κεντροειδούς, καθιστώντας επομένως την σύγκλιση δυνατή σε πεπερασμένο πλήθος επαναλήψεων. Ένα ακόμα ζήτημα με τον συγκεκριμένο αλγόριθμο είναι ότι συχνά παγιδεύεται σε τοπικά ακρότατα καθιστώντας τον ιδιαίτερα ευαίσθητο στην αρχική επιλογή των κεντροειδών. Το πρόβλημα αυτό είναι δυνατό να ξεπεραστεί με την επανάληψη της διαδικασίας αρκετές φορές, με διαφορετικά αρχικά κεντροειδή κάθε φορά.

3.3. Μηχανές διανυσμάτων υποστήριξης (Support vector machines - SVM)

Στα σύγχρονα συστήματα μηχανικής μάθησης, οι μηχανές διανυσμάτων υποστήριξης είναι ένα μέσο που πάντα δοκιμάζεται. Είναι ίσως η πιο στιβαρή και ακριβέστερη μέθοδος μεταξύ των γνωστών αλγορίθμων. Έχει γερή θεωρητική βάση, απαιτεί μικρό αριθμό υποδειγμάτων για εκπαίδευση και δεν επηρεάζεται από τον αριθμό των διαστάσεων του σετ δεδομένων. Επιπλέον αναπτύσσονται ταχύτατα αποδοτικές μέθοδοι για την εκπαίδευση τους.

[8] *ΠΗΓΗ:* http://mnemstudio.org/clustering-k-means-introduction.htm

Σε ένα πρόβλημα δύο τάξεων, ο στόχος των μηχανών διανυσμάτων υποστήριξης είναι να βρεθεί η καλύτερη συνάρτηση ταξινόμησης που να ξεχωρίζει τα μέλη των δύο τάξεων στα δεδομένα εκπαίδευσης. Ο τρόπος μέτρησης ώστε να γίνει κατανοητή η έννοια του καλύτερου αλγορίθμου ταξινόμησης γίνεται αντιληπτός γεωμετρικά. Για ένα σετ δεδομένων που είναι γραμμικά διαχωρίσιμο, μια γραμμική συνάρτηση ταξινόμησης αντιστοιχεί στο υπερεπίπεδο διαχωρισμού $f(x)$ που περνά μεταξύ των δύο τάξεων και τις χωρίζει στα δύο. Όταν αυτή η συνάρτηση προσδιοριστεί, κάθε νέο υπόδειγμα x_n μπορεί να ταξινομηθεί απλά βρίσκοντας το πρόσημο της συνάρτησης. Αν το $f(x_n)$ >0 , τότε το x_n ανήκει στην θετική τάξη.



Εικόνα 9: ΠΑΡΑΓΟΜΕΝΟ ΥΠΕΡΕΠΙΠΕΔΟ ΔΙΑΧΩΡΙΣΜΟΥ ΑΠΟ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ[9]

Όμως υπάρχουν πολλά τέτοια γραμμικά υπερεπίπεδα. Αυτό που προσφέρουν επιπλέον η μηχανές διανυσμάτων υποστήριξης είναι η εύρεση της συνάρτησης εκείνης που μεγιστοποιεί το περιθώριο μεταξύ των δύο τάξεων, τον κενό χώρο ανάμεσα τους κατά τον διαχωρισμό τους. Γεωμετρικά, αυτό το περιθώριο αντιστοιχεί στην ελάχιστη απόσταση που μπορεί να έχει ένα σημείο των δεδομένων από το υπερεπίπεδο. Έχοντας αυτό τον γεωμετρικό ορισμό υπόψη, μπορούμε να βρούμε τρόπους μεγιστοποίησης του περιθωρίου, ώστε παρά τον μεγάλο αριθμό πιθανών υπερεπιπέδων, μόνο λίγα να προκύπτουν σαν αποτέλεσμα των μηχανών διανυσμάτων υποστήριξης.

Ο λόγος για τον οποίο γίνεται τέτοια προσπάθεια για την εύρεση του υπερεπιπέδου μέγιστου περιθωρίου είναι ότι προσφέρει πολύ μεγάλη δυνατότητα γενίκευσης. Προσφέρει όχι μόνο την καλύτερη απόδοση κατά την ταξινόμηση (π.χ. ακρίβεια) των δεδομένων εκπαίδευσης αλλά αφήνει και χώρο για την σωστή ταξινόμηση των μελλοντικών δεδομένων.

[9] *ΠΗΓΗ:* http://imtech.res.in/raghava/rbpred/svm.jpg

Για να σιγουρευτεί ότι το υπερεπίπεδο που έχει βρεθεί είναι όντως μέγιστου περιθωρίου, ο ταξινομητής των μηχανών διανυσμάτων υποστήριξης προσπαθεί να μεγιστοποιήσει την ακόλουθη συνάρτηση ως προς τα $\vec{w}$ και b

$$L_p = \frac{1}{2}\|\vec{w}\| - \sum_{i=1}^{t} a_i y_i (\vec{w} \cdot \vec{x}_i + b) + \sum_{i=1}^{t} a_i$$

όπου t είναι ο αριθμός των υποδειγμάτων εκπαίδευσης, και a_i, $i = 1...t$, είναι μη αρνητικοί αριθμοί τέτοιοι ώστε οι παράγωγοι της L_p ως προς a_i να μηδενίζονται. Οι a_i είναι οι Λανγκραντζιανοί πολλαπλασιαστές και η L_p η Λανγκραντζιανή συνάρτηση. Στην παραπάνω συνάρτηση, το διάνυσμα $\vec{w}$ και η σταθερά b είναι αυτά που ορίζουν το υπερεπίπεδο.

3.4. Ο αλγόριθμος Apriori

Μια από τις πιο συχνές εφαρμογές του *Data Mining* είναι η αναγνώριση συνόλων στοιχείων που συναντώνται μαζί σε δεδομένα συναλλαγών και η εξαγωγή αντίστοιχων κανόνων συσχέτισης. Η εύρεση τέτοιων συνόλων στοιχείων με μεγάλη συχνότητα (με τον όρο «συχνότητα» εννοούμε το να πληρούν ένα κριτήριο ελάχιστης υποστήριξης που θα έχει θέσει ο χρήστης) δεν είναι καθόλου ασήμαντη λόγω της συνδυαστικής της δυναμικής. Μόλις βρεθούν τέτοια σύνολα στοιχείων, είναι άμεση η μετάβαση στη δημιουργία κανόνων συσχέτισης με βαθμό εμπιστοσύνης ίσο ή μεγαλύτερο από αυτό που έχει θέσει ο χρήστης.

Ο *Apriori* είναι ένας πρωτότυπος αλγόριθμος για την εύρεση συνόλων στοιχείων με μεγάλη συχνότητα χρησιμοποιώντας μία γεννήτρια υποψηφίων. Χρησιμοποιεί κατά την εξερεύνηση των υποδειγμάτων την ιδέα πως αν ένα σύνολο στοιχείων δεν είναι «συχνό» τότε και κανένα υπερσύνολο αυτού δεν θα είναι. Συμβατικά ο *Apriori* υποθέτει ότι τα υποδείγματα σε ένα σετ δεδομένων είναι ταξινομημένα αλφαβητικά. Έστω πως το συχνό σύνολο k στοιχείων είναι F_k και τα υποψήφια C_k. Ο αλγόριθμος πρώτα εξερευνά την βάση δεδομένων και αναζητά συχνά σύνολα στοιχείων μεγέθους 1 (1 στοιχείο δηλαδή αποτελεί το σύνολο) αθροίζοντας τις εμφανίσεις κάθε στοιχείο και συγκεντρώνοντας εκείνα που ικανοποιούν μια ελάχιστη υποστήριξη. Στην συνέχεια επαναλαμβάνει τα επόμενα 3 βήματα και εξάγει όλα τα συχνά σύνολα στοιχείων

1. Παράγει C_{k+1} υποψήφια συχνά σύνολα στοιχείων μεγέθους $k+1$ από τα συχνά σύνολα στοιχείων μεγέθους k.
2. Εξερευνά την βάση δεδομένων και υπολογίζει την υποστήριξη κάθε υποψηφίου συχνού συνόλου στοιχείων.
3. Προσθέτει αυτά τα σύνολα στοιχείων που ικανοποιούν την απαιτούμενη ελάχιστη υποστήριξη στο F_{k+1}

Input: A database D of transactions and *minsup*.
Output: A set F that contains all frequent itemsets of D.
Algorithm:

```
(1)   F_1 ← {frequent 1-itemsets};
(2)   k ← 2;
(3)   while F_{k-1} ≠ φ do begin
(4)     C_k ← AprioriGen(F_{k-1});
(5)     forall transactions t ∈ D do begin
(6)       C_t ← Subset(C_k, t);
(7)       forall candidates c ∈ C_t do
(8)         c.count ← c.count + 1;
(9)       end;
(10)      F_k ← {c ∈ C_k | c.count ≥ minsup};
(11)      k ← k + 1;
(12)    end;
(13)    F ← ∪_{k≥1} F_k;
```

Εικόνα 10: Ο ΑΛΓΟΡΙΘΜΟΣ Apriori[10]

Ο αλγόριθμος φαίνεται στην παραπάνω εικόνα. Η συνάρτηση AprioriGen που εμφανίζεται στην 4η γραμμή παράγει C_{k+1} υποψήφιους συνδυασμούς από το σύνολο F_k μέσω των επόμενων δύο βημάτων:

1. *Βήμα σύζευξης*: Παραγωγή R_{k+1}, τα αρχικά υποψήφια συχνά σύνολα στοιχείων μεγέθους $k+1$, μέσω της ένωσης δύο συχνών συνόλων στοιχείων μεγέθους k, P_k και Q_k, που έχουν τα πρώτα $k-1$ στοιχεία κοινά.

$$R_{k+1} = P_k \cup Q_k = \left\{ στοιχείο_1, στοιχείο_2, ..., στοιχείο_{κ-1}, στοχείο_κ, στοιχείο_{κ'} \right\}$$

$$P_k = \left\{ στοιχείο_1, ..., στοιχείο_{κ-1}, στοιχείο_κ \right\}$$

$$Q_k = \left\{ στοιχείο_1, ..., στοιχείο_{κ-1}, στοιχείο_{κ'} \right\}$$

   όπου στοιχείο 1 < στοιχείο 2 < … < στοιχείο k < στοιχείο k′

2. *Βήμα κλαδέματος:* Έλεγχος αν όλα τα σύνολα στοιχείων μεγέθους k στο R_{k+1} μπορούν να θεωρηθούν συχνά και παραγωγή του C_{k+1} αφαιρώντας όσα από αυτά δεν ικανοποιούν τις απαιτήσεις από το R_{k+1}. Η λογική πίσω από αυτό είναι ότι όποιο σύνολο μεγέθους k στο C_{k+1} δεν θεωρείται συχνό, δεν μπορεί και να αποτελεί υποσύνολο ενός συχνού συνόλου στοιχείων μεγέθους $k+1$

[10] *ΠΗΓΗ:* http://magna.cs.ucla.edu/~hxwang/axl_manual/img40.gif

Ο αλγόριθμος σαρώνει την βάση δεδομένων το πολύ $k_{\max+1}$ φορές όταν το μέγιστο μέγεθος των συχνών συνόλων στοιχείων έχει οριστεί ως $k_{\max}$. Μέσω της μείωσης του μεγέθους των υποψηφίων συνόλων ο *Apriori* πετυχαίνει καλή απόδοση. Ωστόσο σε περιπτώσεις με πάρα πολλά σύνολα συχνών στοιχείων, σύνολα μεγάλου μεγέθους ή με ορισμό πολύ χαμηλής ελάχιστης υποστήριξης, υποφέρει από το τίμημα της παραγωγής πολύ μεγάλου αριθμού υποψηφίων συνόλων και της επανειλημμένης σάρωσης της βάσης δεδομένων για τον έλεγχος όλων αυτών. Στην πράξη, η παραγωγή ενός συνόλου συχνών στοιχείων μεγέθους 100 συνεπάγεται την δημιουργία 2^{100} υποψήφιων συνόλων στοιχείων.

3.5. Ο αλγόριθμος PageRank

Ο αλγόριθμος *PageRank* παρουσιάστηκε από τους Sergey Brin και Larry Page στο *Έβδομο Διεθνές Συνέδριο του Παγκόσμιου Ιστού* (WWW7) τον Απρίλιο του 1998. Είναι ένας αλγόριθμος αναζήτησης ταξινόμησης χρησιμοποιώντας υπερσυνδέσεις από το διαδίκτυο. Με βάση τον αλγόριθμο δημιουργήθηκε η μηχανή αναζήτησης Google, που έχει γνωρίσει τρομακτική επιτυχία.

Ο *PageRank* δημιουργεί μια στατική ταξινόμηση των ιστοσελίδων υπό την έννοια ότι δίνει μια τιμή *PageRank* για κάθε ιστοσελίδα, εκτός δικτύου, που δεν επηρεάζεται από τα αντικείμενα αναζήτησης. Ο αλγόριθμος στηρίζεται στην δημοκρατική φύση του παγκόσμιου ιστού χρησιμοποιώντας την αχανή δομή διαφόρων συνδέσεων σαν δείκτη της ποιότητας κάθε ιστοσελίδας. Στην ουσία ερμηνεύει κάθε υπερσύνδεση από την ιστοσελίδα x στην ιστοσελίδα y σαν μια ψήφο της ιστοσελίδας x για την ιστοσελίδα y. Ωστόσο ο αλγόριθμος εξετάζει περισσότερα από απλά τον αριθμό των ψήφων ή των συνδέσεων που δέχεται μια ιστοσελίδα. Αναλύει και την ιστοσελίδα που δίνει την ψήφο. Ψήφοι από ιστοσελίδες που είναι «σημαντικές», έχουν μεγαλύτερη βαρύτητα και κάνουν κι άλλες σελίδες «σημαντικές».

Η λειτουργία του *PageRank* βασίζεται στις ακόλουθες ιδέες που αφορούν το κύρος κάθε ιστοσελίδας:

- Μια υπερσύνδεση από μια ιστοσελίδα προς μια άλλη προσδίδει μια μορφή αυθεντίας στην ιστοσελίδα που καταλήγει. Επομένως όσο περισσότερες υπερσυνδέσεις καταλήγουν σε μια ιστοσελίδα, τόσο μεγαλύτερο θεωρείται το κύρος της.
- Ιστοσελίδες που συνδέονται προς άλλες, έχουν και αυτές το δικό τους κύρος. Μια ιστοσελίδα με μεγαλύτερη «βαθμολογία» σε ότι αφορά το κύρος της, έχει μεγαλύτερη σημασία από μια ιστοσελίδα με μικρότερη βαθμολογία. Με άλλα λόγια μια ιστοσελίδα είναι «σημαντική» αν συνδέεται με άλλες σημαντικές ιστοσελίδες.

Η «σημαντικότητα» μιας ιστοσελίδας (η βαθμολογία της δηλαδή από τον PageRank) καθορίζεται αθροίζοντας τις βαθμολογίες των ιστοσελίδων που συνδέονται προς αυτή. Εφόσον μια ιστοσελίδα δύναται να συνδέεται με διάφορες άλλες, η βαθμολογία της σε ότι αφορά το κύρος θα πρέπει να διαμοιράζεται εξίσου σε αυτές.

Για την διατύπωση των παραπάνω ιδεών, θεωρούμε το διαδίκτυο σαν ένα προσανατολισμένο γράφημα $G=(V,E)$, όπου V το σύνολο των κόμβων (για παράδειγμα το σύνολο των σελίδων) και E το σύνολο των κατευθυνόμενων ακμών (για παράδειγμα οι υπερσυνδέσεις μεταξύ των ιστοσελίδων). Έστω ότι ο συνολικός αριθμός των ιστοσελίδων στο διαδίκτυο είναι n (άρα $n=|V|$)

Η βαθμολογία *PageRank* της σελίδας i,συμβολίζοντας την $P(i)$ ορίζεται ως:

$$P(i)=\sum_{j,i\in E}\frac{P(j)}{O_j}$$

όπου O_j ο αριθμός των υπερσυνδέσεων που ξεκινούν από την ιστοσελίδα j. Έχουμε λοιπόν ένα σύστημα n-γραμμικών εξισώσεων με n αγνώστους. Μπορούμε να χρησιμοποιήσουμε ένα πίνακα για την παράσταση όλων των εξισώσεων. Θεωρούμε ως P το διάνυσμα στήλη μεγέθους n που παριστά τις τιμές που δίνει ο *PageRank*.

$$P=(P(1),P(2),...,P(n))^T$$

Επίσης θεωρούμε A τον συμπληρωματικό πίνακα του γραφήματος μας με:

$$A_{ij}=\begin{cases}\frac{1}{O_j}\,\alpha\nu\ (i,j)\in E\\ 0\end{cases}$$

Το σύστημα των n εξισώσεων γράφεται πλέον ως:

$$\mathbf{P}=\mathbf{A}^T\mathbf{P}$$

Αυτή είναι μια χαρακτηριστική εξίσωση ιδιοσυστήματος όπου η λύση του P είναι μια ιδιοδιάνυσμα με αντίστοιχη ιδιοτιμή 1. Αφού πρόκειται για ένα κυκλικό ορισμό, λογικό είναι να απαιτείται κι ένας επαναληπτικός αλγόριθμος για να επιλυθεί. Τελικά αποδεικνύεται ότι αν ικανοποιηθούν κάποιες προϋποθέσεις, η μεγαλύτερη ιδιοτιμή είναι 1 και το PageRank διάνυσμα P είναι το βασικό ιδιοδιάνυσμα.

Όμως το πρόβλημα είναι ότι η παραπάνω εξίσωση δεν είναι ικανοποιητική γιατί το γράφημα του διαδικτύου δεν ικανοποιεί τις προϋποθέσεις που απαιτούνται. Στην πραγματικότητα η παραπάνω εξίσωση μπορεί να παραχθεί με βάση μια Μαρκοβιανή αλυσίδα. Τότε, εφαρμόζοντας κάποια αποτελέσματα από τη θεωρία των Μαρκοβιανών αλυσίδων και αυξάνοντας το γράφημα του διαδικτύου, προκύπτει η ακόλουθη εξίσωση για τον *PageRank*

$$\mathbf{P} = (1-d)e + d\mathbf{A}^T\mathbf{P}$$

όπου e είναι ένα διάνυσμα στήλη που όλα του τα στοιχεία είναι 1. Άρα για κάθε ιστοσελίδα i , ο τύπος γίνεται:

$$P(i) = (1-d) + d\sum_{j=1}^{n} A_{ji}P(j)$$

που είναι ισοδύναμος με τον αρχικό τύπο PageRank

$$P(i) = (1-d) + d\sum_{(j,i)\in E}\frac{P(j)}{Q_j}$$

Η παράμετρος d ονομάζεται συντελεστής απόρριψης και μπορεί να έχει τιμή από 0 έως 1.

Ο υπολογισμός της τιμής *PageRank* των ιστοσελίδων του διαδικτύου γίνεται χρησιμοποιώντας τον παρακάτω δυναμικό επαναληπτικό αλγόριθμο που παράγει το βασικό ιδιοδιάνυσμα με ιδιοτιμή 1.

PageRank-Iterate(G)

$P_0 \leftarrow e/n$
$k \leftarrow 1$
$\varepsilon\pi\alpha\nu\acute{\varepsilon}\lambda\alpha\beta\varepsilon$
$\quad P_k \leftarrow (1-d)e + dA^T P_{k-1}$
$\quad k \leftarrow k+1$
$\mu\acute{\varepsilon}\chi\rho\iota \ \|P_k - P_{k-1}\| < \varepsilon$
$\varepsilon\pi\acute{\varepsilon}\sigma\tau\rho\varepsilon\psi\varepsilon \quad P_k$

Η αρχή μπορεί να γίνει με οποιεσδήποτε αρχικές τιμές στον *PageRank*. Οι επαναλήψεις τελειώνουν όταν η τιμή του *PageRank* μένει σχετικά σταθερή. Στον

παραπάνω αλγόριθμο οι επαναλήψεις σταματούν όταν η απόλυτη μεταβολή της τιμής του *PageRank* από την επανάληψη είναι μικρότερη από ένα προκαθορισμένο όριο ε.

Σε ότι αφορά την αναζήτηση στο διαδίκτυο, ενδιαφερόμαστε μόνο για την βαθμολόγηση και ταξινόμηση των σελίδων, οπότε η απόλυτη σύγκλιση του αλγορίθμου δεν είναι απαραίτητη. Επομένως απαιτούνται λιγότερες επαναλήψεις. Αναφέρεται πως για μια βάση δεδομένων με 322 εκατομμύρια συνδέσεις, ο αλγόριθμος συγκλίνει σε αποδεκτό επίπεδο μετά από περίπου 52 επαναλήψεις.

3.6. Ο αλγόριθμος AdaBoost

Η ολιστική μάθηση πραγματεύεται μεθόδους που χρησιμοποιούν πολλούς αλγορίθμους εκμάθησης για την επίλυση ενός προβλήματος. Η ικανότητα γενίκευσης από έναν ολιστικό αλγόριθμο είναι συνήθως πολύ καλύτερη από αυτή ενός απλού αλγορίθμου, οπότε οι ολιστικές μέθοδοι είναι πιο ελκυστικές. Ο αλγόριθμος *AdaBoost* που προτάθηκε από τους *Yoav Freund* και *Robert Schapire* είναι ένας από τους πιο σημαντικούς ολιστικούς αλγόριθμους, καθώς έχει γερή θεωρητική βάση, μεγάλη ακρίβεια στις προβλέψεις του, είναι πολύ απλός και έχει εφαρμοστεί επιτυχώς σε πληθώρα περιπτώσεων.

Έστω πως το X αντιπροσωπεύει το χώρο των υποδειγμάτων και το Y το σύνολο των ονομάτων των διαφόρων τάξεων. Υποθέτουμε πως $Y = \{-1, +1\}$. Με δεδομένο έναν απλό βασικό αλγόριθμο (έστω Α ο αλγόριθμος αυτός) και ένα σύνολο υποδειγμάτων εκπαίδευσης $\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ όπου $x_i \in X$ και $y_i \in Y$ $(i = 1, ..., m)$, ο AdaBoost λειτουργεί ως εξής. Πρώτα θέτει ίδια βαρύτητα σε όλα τα υποδείγματα εκπαίδευσης (x_i, y_i) $(i \in \{1, ..., m\})$. Έπειτα καταχωρεί την κατανομή της βαρύτητας των υποδειγμάτων κατά τον t γύρο εκμάθησης ως D_t. Από τα υποδείγματα εκπαίδευσης και το D_t ο αλγόριθμος παράγει έναν «αδύναμο» σχήμα εκμάθησης $h_t : X \to Y$ καλώντας τον Α. Στη συνέχεια χρησιμοποιεί τα υποδείγματα εκπαίδευσης για τον έλεγχο του h_t, και αυξάνει την βαρύτητα των λάθος ταξινομημένων υποδειγμάτων. Προκύπτει έτσι μια αναθεωρημένη κατανομή βαρύτητας στα υποδείγματα D_{t+1}. Από τα υποδείγματα εκπαίδευσης και το D_{t+1}, ο *AdaBoost* παράγει ένα νέο σχήμα εκμάθησης καλώντας τον βασικό αλγόριθμο Α ξανά. Αυτή η διαδικασία επαναλαμβάνεται για Τ φορές και το τελικό μοντέλο παράγεται από τη βαρύτητα που προκύπτει από τα σχήματα εκμάθησης από τους Τ βασικούς αλγόριθμους Α κατά τη διαδικασία της εκπαίδευσης. Κατά την εφαρμογή, ο βασικός αλγόριθμος Α μπορεί να είναι ένας αλγόριθμος που χρησιμοποιεί ο ίδιος σταθμισμένα υποδείγματα εκπαίδευσης απευθείας, αλλιώς η βαρύτητα μπορεί να προκύψει χρησιμοποιώντας δείγματα των υποδειγμάτων εκπαίδευσης σύμφωνα με την κατανομή D_t.

Input: Data set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$;

Base learning algorithm $\mathcal{L}$;

Number of learning rounds T.

Process:

$D_1(i) = 1/m.$ % Initialize the weight distribution

for $t = 1, \cdots, T$:

$h_t = \mathcal{L}(\mathcal{D}, D_t)$; % Train a weak learner h_t from $\mathcal{D}$ using distribution D_t

$\epsilon_t = \Pr_{i \sim D_t}[h_t(\boldsymbol{x}_i \neq y_i)]$; % Measure the error of h_t

$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$; % Determine the weight of h_t

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & \text{if } h_t(\boldsymbol{x}_i) = y_i \\ \exp(\alpha_t) & \text{if } h_t(\boldsymbol{x}_i) \neq y_i \end{cases}$$

$$= \frac{D_t(i)\exp(-\alpha_t y_i h_t(\boldsymbol{x}_i))}{Z_t}$$ % Update the distribution, where Z_t is

% a normalization factor which enables D_{t+1} be a distribution

end.

Output: $H(\boldsymbol{x}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})\right)$

Εικόνα 11: Ο ΑΛΓΟΡΙΘΜΟΣ AdaBoost[11]

Για να αντιμετωπίσει προβλήματα με πάνω από δύο τάξεις, οι *Freund* και *Schapire* παρουσίασαν τον αλγόριθμο *AdaBoost.M1* που απαιτεί οι «αδύναμοι» βασικοί αλγόριθμοι να είναι αρκετά δυνατοί ώστε να αντιμετωπίσουν δύσκολες κατανομές βαρύτητας που δημιουργούνται κατά την εφαρμογή του *AdaBoost*. Άλλη μια διάσημη παραλλαγή του *AdaBoost* είναι ο *AdaBoost.MH* ο οποίος διασπά προβλήματα πολλών τάξεων σε μια σειρά από δυαδικά προβλήματα. Μελέτη έχει γίνει και για την εφαρμογή του αλγορίθμου σε προβλήματα παλινδρόμησης. Με την ανάπτυξη αρκετών παραλλαγών του *AdaBoost*, η ενδυνάμωση (*Boosting*) αποτελεί την πιο σημαντική περιοχή των ολιστικών μεθόδων.

3.7. Η μέθοδος kNN : k- nearest neighbor classification (ταξινόμηση k πλησιέστερων γειτόνων)

Ένας από τους πιο απλούς και μάλλον κοινούς ταξινομητές είναι ο ταξινομητής *Rote*, ο οποίος απομνημονεύει όλα τα υποδείγματα εκπαίδευσης και ταξινομεί μόνο αν τα χαρακτηριστικά ενός νέου υποδείγματος ταιριάζουν ακριβώς με κάποιο από τα υποδείγματα εκπαίδευσης. Ένα προφανές μειονέκτημα αυτής της προσέγγισης είναι πως πολλά υποδείγματα δεν θα ταξινομηθούν γιατί δεν ταιριάζουν ακριβώς με κανένα από τα δεδομένα εκπαίδευσης. Μια πιο εξελιγμένη προσέγγιση, η ταξινόμηση k πλησιέστερων γειτόνων, βρίσκει μια ομάδα k υποδειγμάτων που βρίσκονται πιο κοντά στο υπόδειγμα ελέγχου και βασίζει την ανάθεση ονόματος στην κυριαρχία κάποιας συγκεκριμένης τάξης σε αυτή την περιοχή. Τρία είναι τα απαραίτητα στοιχεία αυτής της προσέγγισης: ένα σύνολο υποδειγμάτων, για παράδειγμα κάποια αποθηκευμένα δεδομένα, ένα μέτρο απόστασης για τον

[11] *ΠΗΓΗ:* http://what-when-how.com/wp-content/uploads/2011/07/tmp1820_thumb.png?9d7bd4..jpg

υπολογισμό της απόστασης μεταξύ των υποδειγμάτων, και η τιμή του k, του αριθμού των πλησιέστερων γειτόνων. Για την ταξινόμηση ενός υποδείγματος, υπολογίζεται η απόσταση του από τα ήδη ταξινομημένα, βρίσκονται οι k πλησιέστεροι γείτονες του και οι τάξεις αυτών των υποδειγμάτων χρησιμοποιούνται για τον καθορισμό της τάξης του υποδείγματος.

Input: D, the set of k training objects, and test object $z = (\mathbf{x}', y')$

Process:

Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every object, $(\mathbf{x}, y) \in D$.

Select $D_z \subseteq D$, the set of k closest training objects to z.

Output: $y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$

Εικόνα 12: Η ΜΕΘΟΔΟΣ kNN[12]

Η παραπάνω εικόνα δείχνει μια περίληψη υψηλού επιπέδου της μεθόδου ταξινόμησης πλησιέστερων γειτόνων. Δεδομένου του συνόλου εκπαίδευσης, D, και ενός υποδείγματος ελέγχου $z = (x', y')$, ο αλγόριθμος υπολογίζει την απόσταση μεταξύ του z και των υποδειγμάτων εκπαίδευσης $(x, y) \in D$ για να καθορίσει την λίστα, D_z, με τους πλησιέστερους γείτονες. Ως x θεωρούνται τα στοιχεία του υποδείγματος εκπαίδευσης ενώ y η τάξη του. Ομοίως x' τα στοιχεία του υποδείγματος ελέγχου και y' η τάξη του.

Όταν δημιουργηθεί η λίστα πλησιέστερων γειτόνων, το υπόδειγμα ελέγχου ταξινομείται με βάση την κυρίαρχη τάξη των πλησιέστερων γειτόνων τους:

$$\text{Κυρίαρχη τάξη}: y' = \arg\max_{\upsilon} \sum_{(x_i, y_i) \in D_t} I(\upsilon = y_i)$$

όπου υ είναι το όνομα μιας τάξης, y_i η τάξη του i πλησιέστερου γείτονα και $I(...)$ μια χαρακτηριστική συνάρτηση που επιστρέφει τιμή 1 αν το όρισμα της συνάρτησης είναι αληθές (τιμή TRUE) και 0 σε κάθε άλλη περίπτωση.

Υπάρχουν ορισμένα ζητήματα που μπορεί να επηρεάσουν την απόδοση του kNN. Ένα από αυτά είναι η επιλογή του k. Αν αυτό είναι πολύ μικρό, τότε το αποτέλεσμα ενδέχεται να είναι ευαίσθητο σε σημεία θορύβου. Αντίστοιχα, αν είναι πολύ μεγάλο, τότε η «γειτονιά» μπορεί να περιέχει πάρα πολλά σημεία από άλλες τάξεις. Ένα άλλο σημαντικό ζήτημα είναι η επιλογή του μέτρου απόστασης. Αν και υπάρχουν διάφορα μέτρα απόστασης που μπορούν να χρησιμοποιηθούν, το πιο επιθυμητό είναι αυτό στο οποίο μικρότερη απόσταση μεταξύ των υποδειγμάτων υποδηλώνει μεγαλύτερη πιθανότητα να ανήκουν στην ίδια τάξη. Για παράδειγμα,

[12] *ΠΗΓΗ:* http://www.godoro.com/Divisions/Ehil/Meca/Magazines/Artcles/img/jpg/knnAlgorithm.jpg

αν ο kNN εφαρμόζεται για την ταξινόμηση εγγράφων, ίσως να είναι καλύτερα να χρησιμοποιηθεί το μέτρο του συνημίτονου παρά η Ευκλείδεια απόσταση. Κάποια μέτρα απόστασης μπορεί να επηρεάζονται από το μεγάλο πλήθος διαστάσεων των υποδειγμάτων. Είναι γνωστό ότι η Ευκλείδεια απόσταση είναι πιο δύσχρηστη στην διακριτοποίηση μεταξύ των υποδειγμάτων, όσο μεγαλώνει ο αριθμός των χαρακτηριστικών τους. Επίσης τα χαρακτηριστικά θα πρέπει να είναι κλιμακωτά για να αποφευχθεί η κυριαρχία κάποιου χαρακτηριστικού πάνω στα μέτρα απόστασης. Για παράδειγμα, υποθέτουμε ένα σύνολο υποδειγμάτων όπου το ύψος των ανθρώπων ποικίλει μεταξύ 1,5 μέτρα και 1,8 μέτρα, το βάρος του μεταξύ 50 και 120 κιλά και το εισόδημά τους μεταξύ 600€ και 60000€. Αν κάποιο μέτρο απόστασης χρησιμοποιηθεί χωρίς να έχει προηγηθεί κάποια κλιμάκωση στα υποδείγματα, το χαρακτηριστικό «εισόδημα» θα κυριαρχήσει στον υπολογισμό της απόστασης και επομένως και στην ταξινόμηση των υποδειγμάτων. Έχουν αναπτυχθεί βέβαια διάφορα σχήματα για τον υπολογισμό της βαρύτητας κάθε ξεχωριστού χαρακτηριστικού βασισμένα πάνω στα υποδείγματα εκπαίδευσης.

Οι ταξινομητές kNN ανήκουν στην κατηγορία των οκνηρών μαθησιακών σχημάτων (lazy learners), κάτι που σημαίνει ότι τα αντίστοιχα μοντέλα είναι εύκολο να παραχθούν αλλά η ταξινόμηση άγνωστων νέων υποδειγμάτων είναι μια διαδικασία με σχετικά μεγάλο υπολογιστικό κόστος καθώς απαιτεί τον υπολογισμό των k πλησιέστερων γειτόνων του κάθε υποδείγματος προς ταξινόμηση. Αυτό, γενικότερα, απαιτεί τον υπολογισμό της απόστασης των υποδειγμάτων που δεν έχουν ακόμα τεθεί σε κάποια τάξη από όλα τα ταξινομημένα υποδείγματα, κάτι που ενδεχομένως να έχει μεγάλο κόστος, ιδίως σε μεγάλα σύνολα υποδειγμάτων. Έχουν αναπτυχθεί αρκετές τεχνικές για τον αποδοτικό υπολογισμό της απόστασης των k πλησιέστερων γειτόνων, οι οποίες χρησιμοποιούν τη δομή των δεδομένων για να αποφευχθεί ο υπολογισμός της απόστασης όλων των υποδειγμάτων. Τέτοιες τεχνικές είναι συχνά εφαρμόσιμες σε δεδομένα μικρού αριθμού διαστάσεων και μπορούν να μειώσουν το υπολογιστικό κόστος χωρίς να επηρεάζουν την ακρίβεια της ταξινόμησης.

Ο αλγόριθμος kNN είναι αρκετά απλός στην κατανόηση του και στην ανάπτυξή του για τεχνικές ταξινόμησης. Παρά την απλότητά του, αποδίδει αρκετά ικανοποιητικά σε πολλές περιπτώσεις. Συγκεκριμένα οι Cover και Hart έδειξαν πως το σφάλμα του αλγορίθμου πλησιέστερων γειτόνων έχει άνω όριο το διπλάσιο του σφάλματος Bayes κάτω από συγκεκριμένες λογικές παραδοχές. Επιπλέον το σφάλμα γενικά της μεθόδου kNN προσεγγίζει ασυμπτωτικά αυτό της μεθόδου Bayes και μπορεί να χρησιμοποιηθεί επομένως για το εκτιμήσει. Ο kNN ενδείκνυται για εφαρμογές που ένα υπόδειγμα μπορεί να ανήκει σε διάφορες τάξεις και μπορεί να επιτύχει, σε τέτοια θέματα, αποτελέσματα καλύτερα από πιο προηγμένα σχήματα ταξινόμησης.

3.8. Η μέθοδος Naïve Bayes

Δεδομένου ενός συνόλου υποδειγμάτων, το καθένα από τα οποία ανήκει σε μια γνωστή τάξη και έχει ένα γνωστό διάνυσμα χαρακτηριστικών, σκοπός μας είναι η δημιουργία ενός κανόνα που θα μας επιτρέπει την ανάθεση μελλοντικών υποδειγμάτων σε κάποια τάξη, γνωρίζοντας μόνο το διάνυσμα των χαρακτηριστών του μελλοντικού υποδείγματος. Αυτού του είδους τα προβλήματα χαρακτηρίζονται προβλήματα ταξινόμησης υπό επίβλεψη, συναντώνται πολύ συχνά και έχουν οδηγήσει στην κατασκευή πολλών μεθόδων για την δημιουργία τέτοιων κανόνων. Μια πολύ σημαντική τέτοια μέθοδος είναι και ο απλοϊκός (naïve) Bayes. Η μέθοδος αυτή είναι πολύ εύκολη στην κατασκευή, καθώς δεν απαιτεί ιδιαίτερα εξελιγμένα σχήματα επαναληπτικής εκτίμησης παραμέτρων. Αυτό σημαίνει πως είναι σχεδόν έτοιμη για την εφαρμογή σε πολύ μεγάλα σύνολα υποδειγμάτων. Είναι έυκολο να ερμηνευτεί, ώστε χρήστες χωρίς ιδιαίτερη γνώση σε ότι αφορά τους ταξινομητές να είναι εύκολο να κατανοήσουν για ποιο λόγο γίνεται μια συγκεκριμένη ταξινόμηση. Τέλος, συχνά έχει απρόσμενα καλή απόδοση. Μπορεί να μην είναι ο καλύτερος ταξινομητής για μια συγκεκριμένη εφαρμογή αλλά μπορεί να βασιστεί κάποιος πάνω της λόγω της σταθερότητας της.

Θεωρούμε την ύπαρξη δύο τάξεων που ονομάζονται $i = 0,1$. Στόχος μας είναι η χρήση του αρχικού συνόλου υποδειγμάτων με γνώση των τάξεων στις οποίες ανήκουν (το σύνολο εκπαίδευσης) για την δημιουργία μιας βαθμονόμησης ώστε υποδείγματα με μεγάλη βαθμολογία να σχετίζονται με τα υποδείγματα της τάξης 1 και υποδείγματα με μικρή βαθμολογία να σχετίζονται με τα υποδείγματα της τάξης 0. Η ταξινόμηση στην συνέχεια επιτυγχάνεται συγκρίνοντας τη βαθμολογία με ένα όριο t. Αν ορίσουμε $P(i \mid x)$ την πιθανότητα ένα υπόδειγμα με διάνυσμα μέτρησης $x = (x_1, x_2, ..., x_p)$ να ανήκει στην τάξη i, τότε οποιαδήποτε μονοτονική συνάρτηση του $P(i \mid x)$ θα ήταν μια κατάλληλη βαθμονόμηση. Συγκεκριμένα ο λόγος $P(1 \mid x) / P(0 \mid x)$ θα ήταν κατάλληλος. Από τη θεωρία πιθανοτήτων γνωρίζουμε πως μπορούμε να αποσυνθέσουμε το $P(i \mid x)$ στο γινόμενο $f(x \mid i)P(i)$, όπου $f(x \mid i)$ είναι η κατανομή υπό συνθήκες του x στα υποδείγματα της τάξης i και $P(i)$ η πιθανότητα ένα υπόδειγμα να ανήκει στην τάξη i αν δεν γνωρίζουμε τίποτα παραπάνω για αυτό (η εκ τον προτέρων πιθανότητα – a priori). Άρα ο λόγος γίνεται:

$$\frac{P(1 \mid x)}{P(0 \mid x)} = \frac{f(x \mid 1)P(1)}{f(x \mid 0)P(0)}$$

Για να το χρησιμοποιήσουμε αυτό για την ταξινόμηση, πρέπει να εκτιμήσουμε τα $f(x \mid i)$ και $P(i)$. Αν το σύνολο εκπαίδευσης ήταν ένα τυχαίο δείγμα από το συνολικό πληθυσμό, το $P(i)$ θα μπορούσε να εκτιμηθεί άμεσα από το ποσοστό των υποδειγμάτων της τάξης i στο σύνολο εκπαίδευσης. Για να εκτιμήσει το $f(x \mid i)$, η

μέθοδος Naïve Bayes υποθέτει ότι τα στοιχεία του x είναι ανεξάρτητα, με $f(x \mid i) = \Pi_{j=1}^{p} f(x_j \mid i)$, και στην συνέχεια εκτιμά κάθε μια από τις κατανομές μιας μεταβλητής $f(x_j \mid i)$, $j = 1,..., p$, $i = 0,1$, ξεχωριστά. Άρα το πρόβλημα P διαστάσεων πολλών μεταβλητών έχει αναχθεί σε P προβλήματα εκτίμησης μιας μεταβλητής. Η εκτίμηση μιας μεταβλητής είναι πιο απλή και απαιτεί μικρότερου μεγέθους σύνολα εκπαίδευσης για να επιτύχει ακριβείς εκτιμήσεις. Αυτό είναι ένα από τα μοναδικά χαρακτηριστικά της μεθόδου Naïve Bayes: η εκτίμηση είναι απλή, γρήγορη και δεν απαιτεί εξελιγμένα σχήματα επαναληπτικής εκτίμησης.

Αν οι κατανομές περιθωρίου $f(x_j \mid i)$ είναι διακριτές, με κάθε x_j να παίρνει μόνο ορισμένες τιμές, τότε η εκτίμηση $\hat{f}(x_j \mid i)$ είναι ένας πολυωνυμικός εκτιμητής με τη μορφή ιστογράμματος. Απλά μετρά το ποσοστό των υποδειγμάτων της τάξης i που αντιστοιχούν σε κάθε κελί. Αν οι $f(x_j \mid i)$ είναι συνεχείς, τότε μια κοινή στρατηγική είναι ο καταμερισμός της κάθε μιας σε ένα μικρό αριθμό διαστημάτων και η χρήση ξανά πολυωνυμικών εκτιμητών, αλλά και πιο εξελιγμένες εκδόσεις βασισμένες σε συνεχείς εκτιμητές (π.χ οι εκτιμητές πυκνότητας πυρήνα – kernel estimates) χρησιμοποιούνται επίσης.

Με βάση την υπόθεση περί ανεξαρτησίας των χαρακτηριστικών, ο προηγούμενος λόγος γίνεται:

$$\frac{P(1 \mid x)}{P(0 \mid x)} = \frac{\Pi_{j=1}^{p} f(x_j \mid 1) P(1)}{\Pi_{j=1}^{p} f(x_j \mid 0) P(0)} = \frac{P(1)}{P(0)} \prod_{j=1}^{p} \frac{f(x_j \mid 1)}{f(x_j \mid 0)}$$

Θυμούμενοι ότι στόχος απλά ήταν η δημιουργία μιας βαθμονόμησης η οποία θα ήταν μονότονα σχετιζόμενη με την $P(i \mid x)$, μπορούμε να χρησιμοποιήσουμε τους λογαρίθμους που είναι μια μονότονη αύξουσα συνάρτηση. Επομένως προκύπτει ο εναλλακτικός λόγος:

$$\ln \frac{P(1 \mid x)}{P(0 \mid x)} = \ln \frac{P(1)}{P(0)} + \sum_{j=1}^{p} \ln \frac{f(x_j \mid 1)}{f(x_j \mid 0)}$$

Αν ορίσουμε $w_j = \ln(f(x_j \mid 1) / f(x_j \mid 0))$ και μια σταθερά $k = \ln(P(1) / P(0))$ βλέπουμε ότι ο λόγος παίρνει την μορφή του απλού αθροίσματος

$$\ln \frac{P(1 \mid x)}{P(0 \mid x)} = k + \sum_{j=1}^{p} w_j$$

ώστε ο ταξινομητής να αποκτήσει μια ιδιαίτερα απλή μορφή.

Η υπόθεση περί ανεξαρτησίας των χαρακτηριστικών των υποδειγμάτων για την ανάπτυξη του μοντέλου του *Naïve Bayes* μπορεί να μοιάζει αδικαιολόγητα περιοριστική. Στην πραγματικότητα ωστόσο, πολλοί παράγοντες μπορεί να υπεισέρθουν, κάτι που σημαίνει πως δεν είναι τόσο καταστροφική όσο φαίνεται. Καταρχάς, συνήθως στην αρχή γίνεται επιλογή των χαρακτηριστικών ώστε τα χαρακτηριστικά με υψηλό βαθμό συσχέτισης να αποκλειστούν εξ'αρχής με βάση το ότι είναι πολύ πιθανό να συνεισφέρουν στον ίδιο βαθμό στο διαχωρισμό μεταξύ των τάξεων. Αυτό σημαίνει πως τα χαρακτηριστικά που μένουν μπορούν να θεωρηθούν ως σχεδόν ανεξάρτητα. Επιπλέον η υπόθεση πως η συσχέτιση των χαρακτηριστικών πρέπει να είναι μηδενική, παράγει ένα βήμα έμμεσης κανονικοποίησης, μειώνοντας τον αριθμό των διαστάσεων και οδηγώντας σε πιο ακριβή ταξινόμηση. Ακόμα, σε κάποιες περιπτώσεις που τα χαρακτηριστικά έχουν κάποια συσχέτιση, η βέλτιστη επιφάνεια αποφάσεως βρίσκεται στον ίδιο τόπο με αυτή που παράγεται υπό την υπόθεση ανεξαρτησίας, οπότε κάνοντας την υπόθεση αυτή δεν υπάρχει κάποια αρνητική επιρροή σε ότι αφορά την απόδοση. Τέλος, η επιφάνεια αποφάσεως που προκύπτει από τα μοντέλα του *Naïve Bayes* μπορεί να έχει περίπλοκη μη-γραμμική μορφή: η επιφάνεια είναι γραμμική στο w_j αλλά πολύ μη-γραμμική στα αρχικά χαρακτηριστικά x_j, έτσι ώστε να προσαρμοστεί με λεπτομέρεια στις διάφορες επιφάνειες.

Τα μοντέλα που προκύπτουν από τον *Naïve Bayes* είναι πολύ ελκυστικά λόγω της απλότητας, της κομψότητας αλλά και της σταθερότητας τους. Πρόκειται για έναν από τους πιο παλιούς αλγορίθμους ταξινόμησης, που ακόμα και στην πιο απλή του μορφή παρουσιάζει αναπάντεχα υψηλή απόδοση. Χρησιμοποιείται ευρέως σε περιπτώσεις ταξινόμησης κειμένου και φιλτραρίσματος κακόβουλης αλληλογραφίας.

4. Παρουσίαση του Weka

4.1. Εισαγωγή στο Weka

Η ύπαρξη ενός αλγορίθμου εξόρυξης καθολικής εφαρμογής αποτελεί μια ουτοπία[010]. Η εμπειρία έχει δείξει ότι δεν υπάρχει κάποιο σχήμα μηχανικής μάθησης που να είναι κατάλληλο για όλες τις περιπτώσεις *data mining*. Τα σύνολα δεδομένων που συναντώνται στην πράξη ποικίλουν ευρέως και για να εξάγει κανείς ακριβή μοντέλα απαιτείται τα χαρακτηριστικά του αλγορίθμου εκμάθησης να ταιριάζουν με τη δομή του πεδίου εφαρμογής. Άλλωστε η εξόρυξη γνώσης από δεδομένα αποτελεί μια κατ'εξοχήν πειραματική επιστήμη και στηρίζεται στην εφαρμογή της μεθόδου *trial and error*[010].

Στη συγκεκριμένη εργασία θα χρησιμοποιηθεί σαν εργαλείο το Weka. Το συγκεκριμένο λογισμικό δημιουργήθηκε στο πανεπιστήμιο του Waikato στη Νέα Ζηλανδία. Η πλήρης ονομασία του είναι *Waikato Environment for Knowledge Analysis* και αποτελεί μια συλλογή από τους πλέον σύγχρονους αλγορίθμους μηχανικής μάθησης. Μεταξύ άλλων περιέχει μεθόδους για:

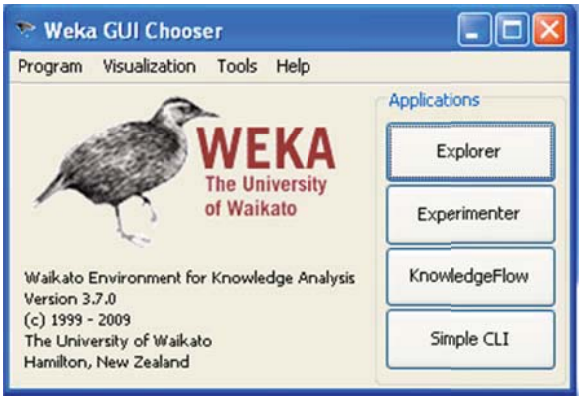
- Προεπεξεργασία δεδομένων
- Οπτικοποίηση
- Ταξινόμηση
- Ομαδοποίηση
- Εύρεση κανόνων συσχέτισης
- Παλινδρόμηση

Το λογισμικό είναι γραμμένο σε Java και είναι «ανοιχτής πηγής» (open source) και ελεύθερης διανομής.

4.2. Περιγραφή του Περιβάλλοντος του Weka

Όταν ξεκινά η εκτέλεση του Weka, ο χρήστης καλείται να διαλέξει ένα από τα τέσσερα πιθανά περιβάλλοντα εργασίας που το Weka του παρέχει με τις ονομασίες:

- Explorer
- Knowledge Flow
- Experimenter
- Simple C(ommand) L(ine) I(nterface)



Εικόνα 13: Η ΑΡΧΙΚΗ ΟΘΟΝΗ ΤΟΥ WEKA

Ακολουθεί μια συνοπτική παρουσίαση των κυρίων στοιχείων του κάθε περιβάλλοντος.

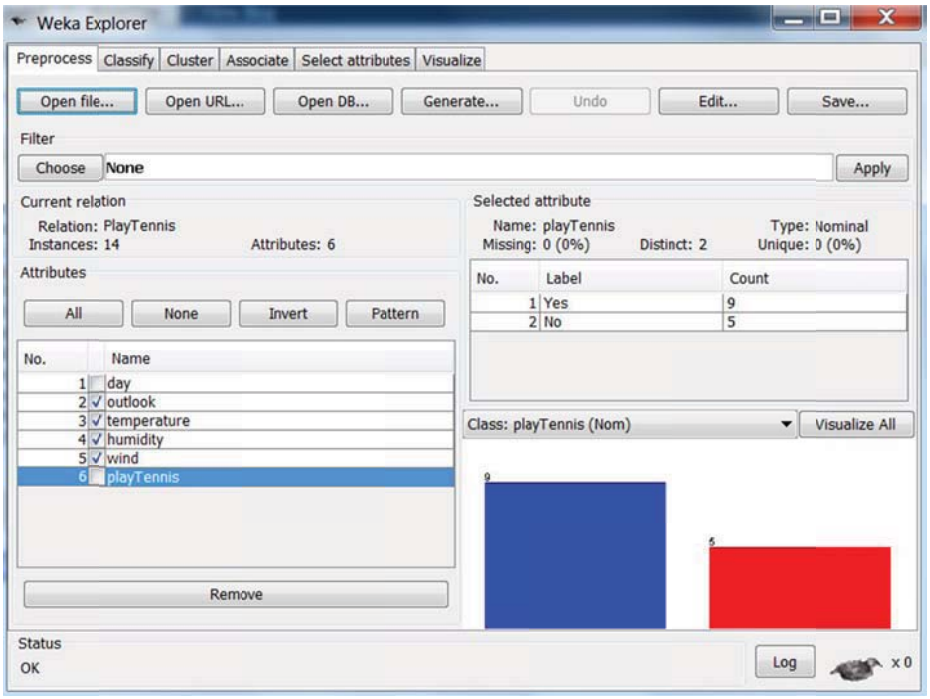
4.2.1. Explorer

Ο πιο εύχρηστος τρόπος για να χρησιμοποιήσει κανείς το Weka είναι μέσω αυτού του γραφικού περιβάλλοντος. Παρέχει πρόσβαση σε όλες τις δυνατότητες που έχει το Weka, παρουσιάζοντας του τις επιλογές του μέσα από κατάλληλα οργανωμένες λίστες. Επιλογές που δεν είναι συμβατές με την εκάστοτε διαδικασία που ακολουθεί ο χρήστης, παρουσιάζονται γκριζαρισμένες. Ακόμα χρησιμοποιούνται λογικές προεπιλεγμένες τιμές σε διάφορες επιλογές ώστε ο χρήστης να είναι σε θέση να έχει κάποια αποτελέσματα με την ελάχιστη δυνατή προσπάθεια. Μέσα από τον *Explorer* μπορεί κανείς να εξετάσει ότι αποτελέσματα έχουν προκύψει από την εφαρμογή των διάφορων αλγορίθμων, να αξιολογήσει και να συγκρίνει διαφορετικά μοντέλα που έχει δημιουργήσει από διάφορα σύνολα δεδομένων, και να οπτικοποιήσει τόσο τα μοντέλα όσο και το σύνολα δεδομένων καθαυτά.

Ο *Explorer* είναι οργανωμένος σε έξι μεγάλες κατηγορίες λειτουργιών με τις αντίστοιχες ετικέτες:

- Preprocess – Εδώ μπορεί κανείς να βρει εργαλεία και αλγορίθμους που αφορούν την επιλογή ή την τροποποίηση του συνόλου δεδομένων που τυγχάνει επεξεργασίας
- Classify – Κατηγορία που περιέχει αλγορίθμους κατάλληλους για προβλήματα ταξινόμησης ή παλινδρόμησης
- Cluster – Περιέχονται αλγόριθμοι για την εύρεση υποομάδων μέσα από το σύνολο δεδομένων
- Associate – Περιέχονται αλγόριθμοι κατάλληλοι για την εύρεση κανόνων συσχέτισης μέσα στο σύνολο δεδομένων και την αξιολόγησή τους
- Select Attributes – Αλγόριθμοι για την επιλογή των πιο σχετικών χαρακτηριστικών μέσα από το σύνολο δεδομένων

- Visualize – Εδώ μπορεί κανείς να βρει εργαλεία για την οπτικοποίηση των δεδομένων ή των μοντέλων που δημιουργεί σε δισδιάστατα γραφήματα

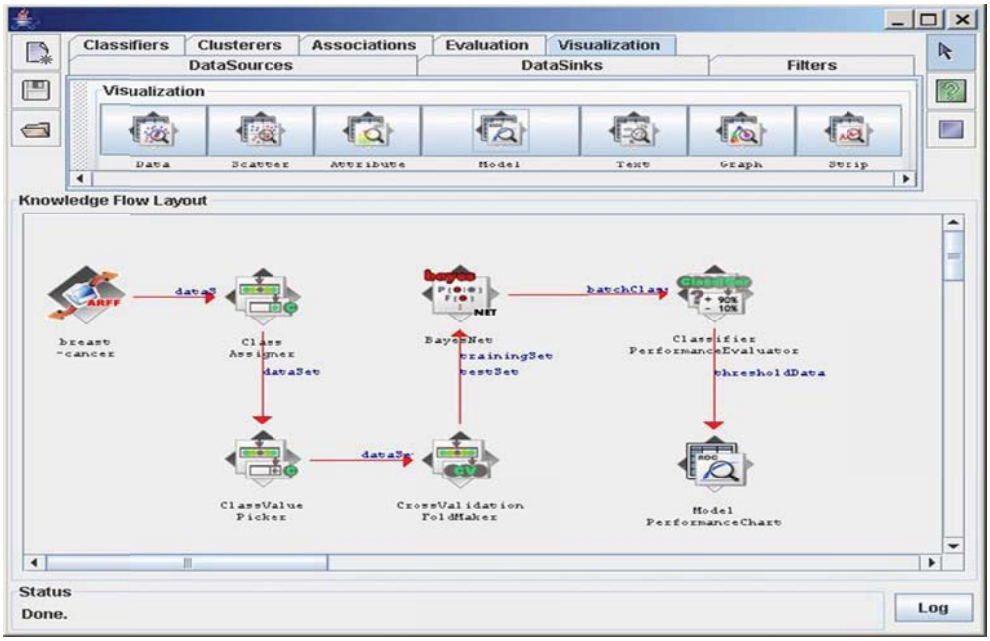


Εικόνα 14: Ο Explorer ΤΟΥ WEKA

Φορτώνοντας ένα σύνολο δεδομένων στο πρόγραμμα, εμφανίζονται γραφικά τα δεδομένα για καθένα από τα γνωρίσματα ξεχωριστά, καθώς και στατιστικές πληροφορίες για καθένα από αυτά. Ακόμα, αν στο σύνολο δεδομένων δίνεται και κάποια κλάση στην οποία ταξινομούνται, τα δεδομένα που βρίσκονται στην ίδια κλάση ταξινομούνται με το ίδιο χρώμα [3].

4.2.2. Knowledge Flow

Το περιβάλλον *Knowledge Flow* απευθύνεται σε πιο προχωρημένους χρήστες του Weka. Πιο συγκεκριμένα απευθύνεται σε όσους θέλουν να έχουν επίγνωση του πως τα δεδομένα και οι πληροφορίες που παράγονται από αυτά «κυλούν» μέσα στο σύστημα. Ο χρήστης επιλέγει τα διάφορα συστατικά κομμάτια του Weka, από μια μπάρα εργαλείων, τα τοποθετεί σε έναν πίνακα και τα συνδέει σε ένα κατευθυνόμενο γράφημα που υποδεικνύει πως γίνεται η ανάλυση και η επεξεργασία των δεδομένων.



Εικόνα 15: ΤΟ ΠΕΡΙΒΑΛΛΟΝ Knowledge Flow

Λειτουργικά, το *Knowledge Flow* περιβάλλον μοιάζει πολύ με τον *Explorer*. Μπορεί κανείς να εκτελέσεις αντίστοιχες εργασίες και στα δύο. Το *Knowledge Flow* παρέχει μια παραπάνω ευελιξία από την άποψη ότι μπορείς να εξετάσεις όλη την διαδικασία λεπτομερώς και όχι μόνο το αποτέλεσμα που προκύπτει από αυτή. Ωστόσο το στοιχείο που το ξεχωρίζει από τον *Explorer* και το κάνει να υπερέχει είναι η δυνατότητα για αυξητική λειτουργία (*incremental operation*) [5].

Αν όλα τα στοιχεία που έχουν συνδεθεί στον πίνακα έχουν τη δυνατότητα να λειτουργήσουν αυξητικά, τότε έτσι λειτουργεί ολόκληρο το μαθησιακό σχήμα. Δεν διαβάζει ολόκληρο το σετ δεδομένων που του δίνεται σαν input πριν αρχίσει η «μάθηση», όπως θα έκανε ο Explorer, αλλά διαβάζει κάθε υπόδειγμα ξεχωριστά και το προωθεί στην διαδικασία που έχει σχηματιστεί στο *Knowledge Flow* πριν πάει στο επόμενο. Μια τέτοια διάταξη μπορεί επομένως να επεξεργαστεί αρχεία οποιουδήποτε μεγέθους, ακόμα και μεγαλύτερου της κύριας μνήμης του συστήματος, καθώς δεν χρειάζεται να τα αποθηκεύσει εσωτερικά για να ξεκινήσει την διαδικασία.

4.2.3. Experimenter

Τα περιβάλλοντα *Explorer* και *Knowledge Flow* εξυπηρετούν τους χρήστες που θέλουν να διαπιστώσουν πόσο καλή απόδοση έχουν τα μαθησιακά σχήματα σε συγκεκριμένα σετ δεδομένων. Ωστόσο πιο σοβαρές ερευνητικές εργασίες απαιτούν πειράματα μεγαλύτερου εύρους, κάτι που μεταφράζεται στην εφαρμογή διαφόρων μαθησιακών σχημάτων σε πολλά διαφορετικά σετ δεδομένων και συχνά με διαφορετικές παραμέτρους. Το περιβάλλον *Experimenter* υπερέχει σε τέτοιου είδους εργασίες έναντι των δύο προηγούμενων.

Ο *Experimenter* αυτοματοποιεί την πειραματική διαδικασία. Οι πληροφορίες που προκύπτουν για τα διάφορα μαθησιακά σχήματα και τα διάφορα σετ δεδομένων μπορούν να αποθηκευθούν και να αποτελέσουν με τη σειρά τους αντικείμενο περαιτέρω μελέτης και εφαρμογής data mining. Επιπλέον ο *Experimenter* έχει ένα χαρακτηριστικό υπεροχής ανάλογο αυτού που είχε το περιβάλλον *Knowledge Flow*.

Ενώ το *Knowledge Flow* ξεπερνούσε τους περιορισμούς σχετικά με το μέγεθος του αρχείου που μπορούσε να επεξεργαστεί, εξετάζοντας κάθε υπόδειγμα από το σετ δεδομένων χωριστά χωρίς να χρειάζεται να φορτώσει ολόκληρο το σετ δεδομένων, ο *Experimenter* ξεπερνά τους χρονικούς περιορισμούς. Περιέχει υποδομές για προχωρημένους χρήστες, ώστε να διαμοιράσουν το υπολογιστικό φορτίο που απαιτείται από μεγάλου εύρους πειράματα, σε διάφορους υπολογιστές. Ευνόητο είναι ότι το συγκεκριμένο χαρακτηριστικό απαιτεί ένα επίπεδο γνώσεων από τη μεριά του χρήστη και απευθύνεται σε αρκετά προχωρημένους χρήστες.

4.2.4. Command Line Interface

Τελευταίο περιβάλλον εργασίας που μπορεί να συναντήσει κανείς στο Weka είναι το *Command Line Interface*. Επιλέγοντας το, έρχεται στην επιφάνεια ένας κενός χώρος με μια γραμμή εισαγωγής εντολών στο κάτω μέρος. Πρόκειται για το πιο απλό και χωρίς γραφικά βοηθήματα περιβάλλον εργασίας και απευθύνεται σε χρήστες που γνωρίζουν εις βάθος το Weka και τις εντολές του.

4.3. Φόρτωση Δεδομένων στο Weka

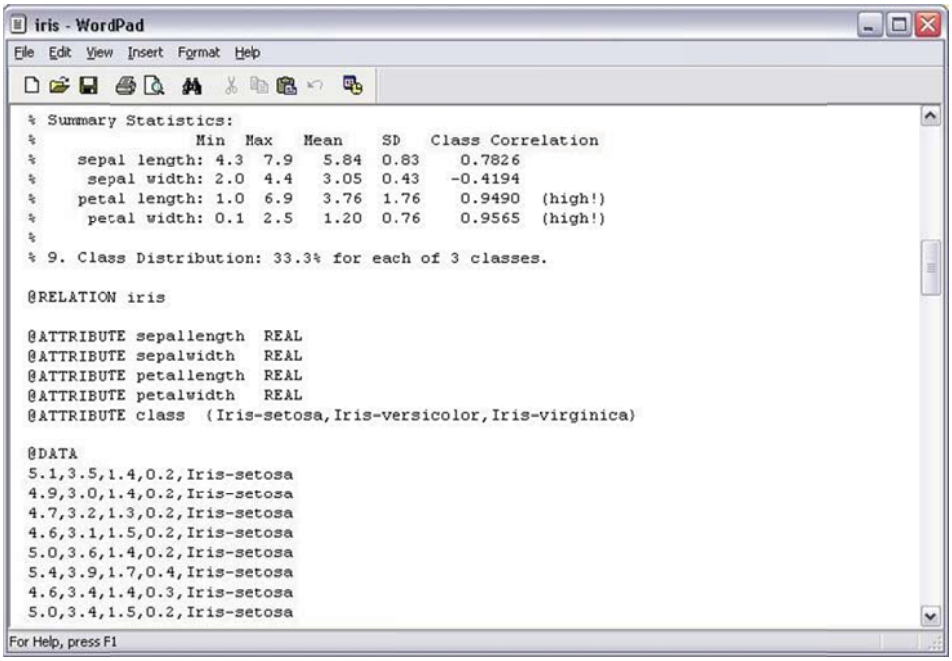
Η φόρτωση δεδομένων στο Weka μπορεί να γίνει με ποικίλους τρόπους. Στον Explorer, το περιβάλλον εργασίας που κυρίως χρησιμοποιούμε, παρατηρούμε ότι υπάρχουν στην καρτέλα preprocess οι επιλογές:

- Open file – για την εύρεση του αρχείου στον υπολογιστή μας, το οποίο θέλουμε να φορτώσουμε
- Open URL – για την φόρτωση δεδομένων κατευθείαν από κάποια ιστοσελίδα του διαδικτύου
- Open DB – για να φορτώσουμε όποια δεδομένα θέλουμε από μια βάση δεδομένων (database)
- Generate – για την δημιουργία τυχαίων δεδομένων μέσα από διάφορους αλγορίθμους, σε περίπτωση που δεν έχουμε δεδομένα διαθέσιμα και θέλουμε να πειραματιστούμε με το Weka

Το δεδομένα αυτά βέβαια μπορεί να βρίσκονται σε διάφορες μορφές και τύπους αρχείων. Το Weka περιέχει ενσωματωμένους μετατροπείς για τους πιο κοινούς τύπους αρχείων για να μετατρέψει στην τυποποίηση με την οποία μπορεί να τα χειριστεί: την τυποποίηση *.arff*

4.3.1. Η Τυποποίηση *.arff*

Η τυποποίηση *.arff* αποτελεί τη φυσική μέθοδο αποθήκευσης δεδομένων του Weka [010]. Υποστηρίζει τόσο αριθμητικά (numeric) όσο και ονομαστικά (nominal) χαρακτηριστικά (attributes) [3]. Τα δεδομένα πολύ συχνά βρίσκονται σε υπολογιστικά φύλλα ή σε βάσεις δεδομένων. Τα προγράμματα που τα χειρίζονται συνήθως επιτρέπουν την εξαγωγή των δεδομένων σε τυποποίηση *.csv*. Το Weka έχει ενσωματωμένο μετατροπέα αρχείων από *.csv* σε *.arff*. Παρόλα αυτά η διαδικασία μετατροπής είναι αρκετά απλή γι'αυτό και παρουσιάζεται παρακάτω.



Εικόνα 16: Η ΤΥΠΟΠΟΙΗΣΗ *.arff*[13]

Πρώτο βήμα είναι να ανοίξουμε το αρχείο με κατάληξη *.csv* που θέλουμε να μετατρέψουμε, με έναν επεξεργαστή κειμένου (text editor). Έπειτα προσθέτουμε το όνομα του σετ δεδομένων σε μια γραμμή στην αρχή, η οποία θα αρχίζει με την έκφραση @relation, τις πληροφορίες για τα κάθε χαρακτηριστικό χρησιμοποιώντας την έκφραση @attribute (κάθε χαρακτηριστικό σε νέα γραμμή) και μια σειρά με την έκφραση @data, που δείχνει ότι από εκεί και κάτω βρίσκονται τα δεδομένα. Τέλος κάνουμε αποθήκευση του αρχείου στην μορφή *.arff*. Στο παράρτημα στο τέλος της εργασίας παρατηρούμε ένα αρχείο *.arff* όπως αυτό προκύπτει από τον μετατροπέα του Weka. Παρότι είναι σχετικά δυσνόητο λόγω του μεγάλου όγκου των δεδομένων, είναι εμφανή τα σημεία όπου ξεκινούν με @relation , @attribute και @data.

[13] *ΠΗΓΗ:* http://research.cs.queensu.ca/home/cisc333/tutorial/wekaData.JPG

5. Παρουσίαση της εταιρείας

Για την πρακτική εφαρμογή των παραπάνω τεχνικών και μεθόδων του *Data Mining* σε υφιστάμενο πρόβλημα συνεργαστήκαμε με τη *Νότα Μασσέλος Α.Ε*. Η εταιρεία διέθεσε τόσο τα δεδομένα που χρειάστηκαν για την εργασία όσο και τον εξοπλισμό (γραφείο, ηλεκτρονικό υπολογιστή, κλπ) για την περαίωση της.

5.1. Η Νότα Μασσέλος Α.Ε.

Η Νότα ιδρύθηκε από τον Θοδωρή και τη Νότα Μασσέλου το 1962 και εξελίχθηκε σε μια από τις κυρίαρχες και μακροβιότερες εταιρείες παραγωγής και εμπορίας εσωρούχων στην Ελλάδα. Από τα πρώτα της βήματα η *Νότα* στηρίχθηκε στο design και στην καλή ποιότητα των προϊόντων της χάρη στην οποία καθιερώθηκε στην αγορά. Σήμερα απασχολεί 100 άτομα σε δύο εγκαταστάσεις της στην Αθήνα και στην Ουγγαρία. Εκτός από τα προϊόντα Νότα, τα οποία σχεδιάζονται στην Ελλάδα, η Μασσέλος ΑΕ διανέμει και τις γαλλικές συλλογές εσωρούχων *Simone Perele* και *Implicite*. Επίσης αντιπροσωπεύει τις συλλογές των *Princesse Tam Tam, Le Chat, Wacoal, Exilia* καθώς και τη σειρά εσωρούχων *Hugo Boss.*

Τα προϊόντα της Νότα στην Ελλάδα μπορούν να βρεθούν σε είκοσι ομώνυμα καταστήματα και corners στην Αθήνα, Ψυχικό, Λαμία, Θεσσαλονίκη, Πάτρα, Ηράκλειο Κρήτης και σε καλά καταστήματα εσωρούχων σε όλη την Ελλάδα. Τα προϊόντα Νότα πωλούνται επίσης σε εξειδικευμένα καταστήματα (*boutiques*) και πολυκαταστήματα (*department stores*) στην Γαλλία, Ελβετία, ΗΠΑ, Καναδά, Ιαπωνία, Κορέα, Χονγκ Κονγκ, Ουγγαρία, Πολωνία, Νορβηγία, Ρωσία, Νότιο Αφρική, Αραβικές χώρες κλπ.

Η Μασσέλος ΑΕ είναι γνωστή στην Ελληνική αγορά για την προσήλωσή της στην ποιότητα. Το σύστημα διασφάλισης ποιότητος της εταιρείας είναι συμβατό και έχει πιστοποιηθεί με βάση το πρότυπο ISO9001 από την TUV-CERT Γερμανίας από το 1999.

Η εταιρεία θεωρείται από τις πρωτοπόρες σε θέματα τεχνολογίας πληροφορικής. Είναι από τις πρώτες εταιρείες που εφάρμοσαν τεχνολογία *bar code* (1985) και ηλεκτρονικής ανταλλαγής δεδομένων-*EDI* (1994) σε ευρεία κλίμακα. Σήμερα η εταιρεία διαθέτει σύστημα ERP τεχνολογίας αιχμής με το οποίο διασυνδέονται όλες τις οι δραστηριότητές της.

Η ΝΟΤΑ συμμετέχει ενεργά σε σημαντικό αριθμό προγραμμάτων ανάπτυξης όπως: COMETT, STAR (Τηλεματική ΙΙ) , ΕΚΤ/ΟΑΕΔ, Μέντωρ, FORCE, ΠΕΠ Αττικής (ΟΠΕ), Leonardo da Vinci, ESPRIT, Πρωτοβουλία ΜΜΕ, Τεχνομεσιτεία, Κλαδικά έργα EDI, Equal, ΚΕΥΔ κλπ, επενδύοντας συστηματικά σε τεχνολογίες πληροφορικής, ανάπτυξη ανθρωπίνου δυναμικού, και συστήματα διασφάλισης ποιότητας. Έχει συμπεριληφθεί ως μελέτη περίπτωσης σε έκδοση της Ευρωπαϊκής Ένωσης με τίτλο

"Νέες μορφές οργάνωσης της εργασίας"[14]. Το 2002 η ΝΟΤΑ βραβεύθηκε από την ΕΕΔΕ (Ελληνική Εταιρεία Διοικήσεως Επιχειρήσεων) για τις σημαντικές επενδύσεις που έχει πραγματοποιήσει σε δράσεις εκπαίδευσης και ανάπτυξης ανθρώπινου δυναμικού.
Η Νότα είναι μέλος του Συνδέσμου Κατασκευαστών Ετοίμων Ενδυμάτων, του Συνδέσμου Ένδυσης Πλεκτικής και συναφών Ειδών Ελλάδος και της Διεθνούς Ομοσπονδίας Ένδυσης (IAF). [012]

Ακολουθεί μια ιστορική ανασκόπηση των σημαντικότερων στιγμών στην ιστορία της Νότα όπως αυτές καταγράφονται στην ιστοσελίδα της εταιρείας

- 1962 - Ιδρύεται η "Νότα" από τον Θοδωρή και τη Νότα Μασσέλου. Για τον πρώτο χρόνο τα προϊόντα της εταιρείας πωλούνται με το brand "Αμαρυλλίς" το οποίο γρήγορα εγκαταλείπεται και υιοθετείται το "Νότα"

- 1975 - Ανοίγει η πρώτη boutique της εταιρείας στην συμβολή των οδών Ακαδημίας και Βουκουρεστίου, η οποία παραμένει μέχρι και σήμερα το γνωστότερο κατάστημα εσωρούχων της Αθήνας.

- 1977 - Συμφωνία με την Simone Perele για παραγωγή και διάθεση των προϊόντων της κορυφαίας αυτής Γαλλικής εταιρείας στην Ελλάδα.

- 1983 - Η εταιρεία εγκαθιστά μηχανογραφικό σύστημα έκδοσης τιμολογίων και παρακολούθησης αποθήκης σε IBM PC με λογισμικό της εταιρείας Unisoft, όντας μία από τις πρώτες εταιρείες του κλάδου που αγκάλιασαν την πληροφορική.

- 1987 - Η ΝΟΤΑ συμμετέχει για πρώτη φορά στο Salon International de la Lingerie στο Παρίσι, αρχίζοντας για πρώτη φορά να εξάγει συστηματικά.

- 1989 - Η ΝΟΤΑ ιδρύει μικτή εταιρεία παραγωγής στην πόλη Pecs της Νοτίου Ουγγαρίας

- 1994 - Η ΝΟΤΑ υλοποιεί πρόγραμμα Τηλεματικής ύψους 100 εκ δραχμών και εγκαθιστά το πρώτο δίκτυο EDI στο χώρο της ένδυσης στην Ελλάδα.

- 1997 - Η εταιρική ταυτότητα και η αρχιτεκτονική των καταστημάτων ΝΟΤΑ επανασχεδιάζονται στη Γαλλία από την Argile. Το κατάστημα

[14] http://www.nota.gr/files/Corporate/casestudy_new_forms.pdf

της οδού Ακαδημίας είναι το πρώτο που ανακαινίζεται με βάση το νέο concept.

- 2003 - Η Νότα αναλαμβάνει την αποκλειστική διανομή στην Ελλάδα των συλλογών εσωρούχων και μαγιό της γαλλικής εταιρείας Princesse Tam Tam.

- 2009 - Ανακαινίζεται το κατάστημα Νότα στον Φάρο Ψυχικού από την δυναμική ομάδα αρχιτεκτόνων www.lowfat.gr. Η Νότα αναλαμβάνει την αποκλειστική αντιπροσωπεία στην Ελλάδα και στην Κύπρο της εταιρείας που εφηύρε το ανδρικό σλιπ, της Αμερικανικής Jockey (www.jockey.com). [013]

5.2. Περιγραφή των στόχων της εργασίας

Όπως είναι φυσικό μια εταιρεία σαν τη Νότα, που θεωρείται πρωτοπόρος σε θέματα τεχνολογίας πληροφορικής, έχει στη διάθεσή της μεγάλη ποσότητα αποθηκευμένων δεδομένων. Δεδομένων που είναι δυνατό να αφορούν κάθε λογής δραστηριότητα της εταιρείας: οικονομικές συναλλαγές, πελατολόγιο, ιστορικά στοιχεία κλπ. Σε τέτοιες περιπτώσεις, που η μεγάλη πληθώρα δεδομένων καθιστά την επεξεργασία τους μια αρκετά πολύπλοκη και χρονοβόρα διαδικασία, οι τεχνικές *Data Mining* μπορούν να απλουστεύσουν αρκετά τα πράγματα και να εμφανίσουν ενδεχομένως πληροφορίες που θα ήταν αδύνατο αλλιώς να εξαχθούν. Έτσι και η εταιρεία αποφάσισε να χρησιμοποιήσει τις τεχνικές αυτές για την επίτευξη δύο κύριων στόχων:

a) Την επιλογή ατόμων από το πελατολόγιο στα οποία να αποσταλούν οι κατάλογοι της νέας κολεξιόν. Το πελατολόγιο της Νότα αριθμεί μερικές χιλιάδες πελάτες. Η αποστολή του νέου καταλόγου σε όλους αυτούς τους πελάτες είναι μια διαδικασία με ένα αρκετά σοβαρό κόστος και χαμηλό (αλλά προσοδοφόρο) βαθμό απόκρισης. Κάθε τεχνική που επιτρέπει σε μια διαφημιστική καμπάνια να είναι πιο στοχευμένη, να επιτυγχάνει δηλαδή τον ίδιο αριθμό αποκρίσεων αλλά από αρκετά μικρότερο δείγμα, είναι μεγάλης αξίας για κάθε εταιρεία. Εξοικονομούνται χρήματα από την αποστολή των διαφημιστικών μέσων (καταλόγων, φυλλαδίων) σε μικρότερο αριθμό πελατών κρατώντας σταθερό τον αριθμό που ανταποκρίνονται σε αυτές, αυξάνοντας έτσι τον βαθμό απόκρισης. Έτσι και στη Νότα, σκοπός είναι η επιλογή των ατόμων που θα αποσταλούν οι νέοι κατάλογοι, ώστε το όφελος από αυτή την διαφημιστική εκστρατεία να είναι το μεγαλύτερο δυνατό με ταυτόχρονα το μικρότερο δυνατό κόστος.
b) Την εξαγωγή διάφορων άλλων πληροφοριών που μπορεί να προκύψουν και την ανάλυση του καλαθιού αγοράς (*market basket analysis)*. Η τεχνικές του *Data Mining* είναι τεχνικές δοκιμής και λάθους (*trial and error)*. Συχνά απαιτούν διάφορες δοκιμές με διάφορους συνδυασμούς χαρακτηριστικών και αλγορίθμων για την εξαγωγή αποτελεσμάτων. Αυτό έχει ως αποτέλεσμα

το εντοπισμό προτύπων διαφορετικών από αυτά που αναζητούμε εξ'αρχής, τα οποία όμως μπορεί να μεταφέρουν σημαντική πληροφορία για τα δεδομένα. Ακόμα κατά τη διαδικασία εξερεύνησης των δεδομένων προκύπτουν νέες ιδέες προς μελέτη οι οποίες μπορεί να μην είχαν τεθεί ως σκοπός στην αρχή. Όσον αφορά το *market basket analysis,* πρόκειται για τη χρήση τεχνικών συσχέτισης για την εύρεση ομάδων προϊόντων που τείνουν να εμφανίζονται μαζί στις συναλλαγές. Σε πολλά είδη εμπορικών εταιρειών αποτελεί τη μόνη δυνατή εφαρμογή τεχνικών *Data Mining* καθώς δεν απαιτεί άλλα δεδομένα πέρα από τα δεδομένα συναλλαγών. Στόχος της Νότα είναι η εύρεση τέτοιων ομάδων προϊόντων που ίσως να υπάρχουν, η γνώση των οποίων θα μπορούσε να αυξήσει της πωλήσεις μέσω κατάλληλων συνδυαστικών προσφορών ή ακόμα και μέσω της αλλαγής διαρρύθμισης των καταστημάτων ώστε να φέρνουν τις ομάδες προϊόντων πιο κοντά τη μια στην άλλη.

Οι παραπάνω είναι οι στόχοι της εταιρείας μέσα από την εργασία αυτή. Για την επίτευξή τους και την εφαρμογή των τεχνικών *Data Mining* παραχωρήθηκε ένα αρχείο δεδομένων της εταιρείας το οποίο περιείχε το σύνολο των δεδομένων συναλλαγών από το 2008 μέχρι και το 2011.

6. Προετοιμασία Δεδομένων

6.1. Εξερεύνηση των δεδομένων

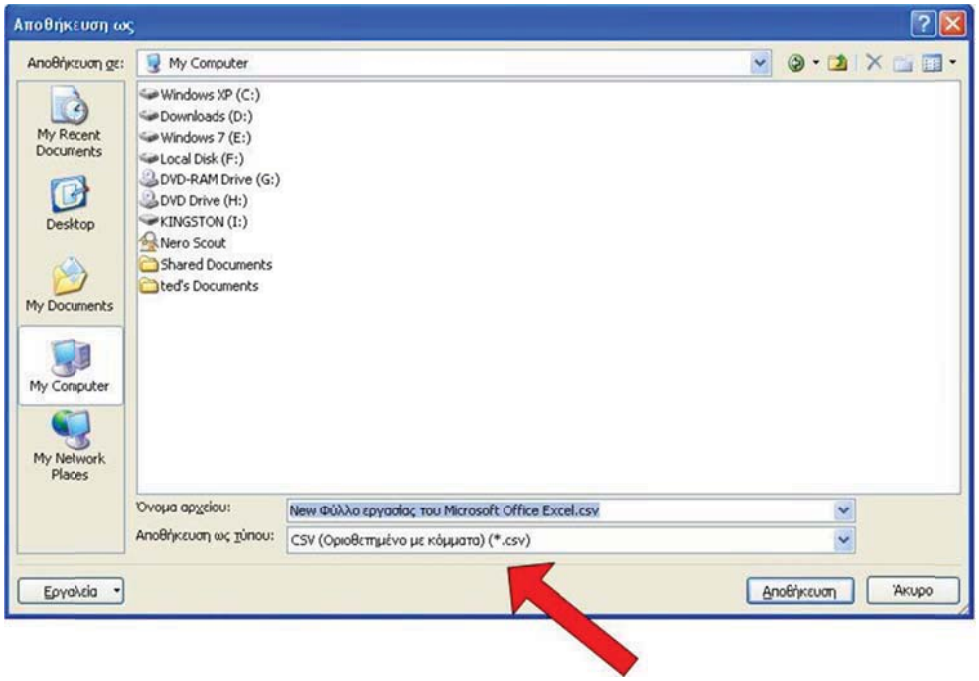
Το αρχικό dataset (αρχείο excel) περιέχει το σύνολο των δεδομένων από το αρχείο των πωλήσεων αλλά και το πελατολόγιο από την ΝΟΤΑ Α.Ε, ΝΟΤΑ Ο.Ε, το CRM της εταιρείας και την ΜΑΛΚΟΤΣΟΓΛΟΥ – ΧΕΡΕΡΑ Ο.Ε. Τα στοιχεία προέρχονται από τα καταστήματα της Ακαδημίας και του Ψυχικού, αλλά και τις πωλήσεις που είχαν γίνει στο ΑΤΤΙΚΑ και στα HONDOS της Ομόνοιας και της Πυλαίας. Κάθε υπόδειγμα (*instance*), δηλαδή κάθε σειρά του αρχείου, αφορά μια συγκεκριμένη αγορά ή επιστροφή ενός συγκεκριμένου προϊόντος. Έτσι η καταγραφή μιας επίσκεψης ενός πελάτη στο αρχείο για την αγορά πχ τεσσάρων διαφορετικών προϊόντων και την επιστροφή ενός άλλου μεταφράζετε σε πέντε υποδείγματα στο αρχείο. Στο σύνολο τους υπάρχουν 67379 υποδείγματα. Τα αρχικά attributes (οι στήλες του αρχείου) είχαν τίτλο:

- ΔΙΑΣΠΑΣΗ - Διψήφιος κωδικός εσωτερικής χρήσης της εταιρίας που χρησιμοποιείται σε περίπτωση διάσπασης των εμπορευμάτων σε διαφορετικές οικογένειες λόγω του ότι πχ. είναι είδη διαφορετικής κατηγορίας, όπως ένδυση, υπόδηση, ή τιμολογούνται με διαφορετικά ποσοστά, ή είναι διαφορετικού ΦΠΑ.
- ΚΩΔΙΚΟΣ ΑΠΟΘΗΚΗΣ ΧΩΡΟΥ - Δηλώνει τον τόπο παράδοσης στο Πολυκατάστημα.
- ΕΙΔΟΣ ΠΑΡΑΣΤΑΤΙΚΟΥ – Παίρνει την τιμή «101» εάν το συνοδευτικό παραστατικό είναι Απόδειξη Λιανικής Πώλησης ή «102» εάν είναι Απόδειξη Λιανικής Επιστροφής
- ΑΡΙΘΜΟΣ ΠΑΡΑΣΤΑΤΙΚΟΥ – Ο κωδικός αριθμός του παραστατικού
- ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΣΤΑΤΙΚΟΥ
- ΚΩΔΙΚΟΣ ΑΝΤΙΣΤΟΙΧΙΑΣ (Ο-Χ-Μ) - Συμφωνημένος κωδικός αντιστοιχίας για κάθε προϊόν που περιέχει τον Οδηγό, το Χρώμα και το Μέγεθος του προϊόντος
- ΠΟΣΟΤΗΤΑ – Η παραδοτέα ποσότητα που αναφέρεται στο παραστατικό
- ΤΕΛΙΚΗ ΚΑΘΑΡΗ ΑΞΙΑ ΚΟΣΤΟΥΣ – Η τελική καθαρή αξία που χρεώθηκε, συνυπολογισμένης οιασδήποτε έκπτωσης. Το σύνολο αυτής της στήλης πρέπει να συμφωνεί με το τελικό σύνολο του παραστατικού προ ΦΠΑ
- ΚΩΔΙΚΟΣ ΠΕΛΑΤΗ – 7ψήφιος κωδικός, μοναδικός για κάθε πελάτη
- ΠΕΡΙΓΡΑΦΗ ΕΙΔΟΥΣ – Αναλυτική ελληνική περιγραφή του προϊόντος
- ΞΕΝΟΓΛΩΣΣΗ ΠΕΡΙΓΡΑΦΗ ΕΙΔΟΥΣ – Η ξενόγλωσση περιγραφή αν υπάρχει
- ΚΩΔΙΚΟΣ ΕΙΔΟΥΣ ΑΠΟΘΗΚΗΣ – Ο κωδικός είδους που φαίνεται στο παραστατικού
- ΟΔΗΓΟΣ – Ο οδηγός του είδους
- ΧΡΩΜΑ – Το χρώμα εφόσον υπάρχει

- ΜΕΓΕΘΟΣ – Το μέγεθος εφόσον υπάρχει
- ΣΥΝΘΕΣΗ/ΠΟΙΟΤΗΤΑ ΕΙΔΟΥΣ – Η σύνθεση του είδους
- ΠΡΟΕΛΕΥΣΗ – Η προέλευση του είδους
- ΚΩΔΙΚΟΣ EAN 13 – Το EAN 13 barcode στην περίπτωση του FULL-EDI (13 χαρακτήρες)
- ΤΙΜΗ ΑΓΟΡΑΣ - Η αρχική τιμή κόστους μονάδας του παραστατικού που τιμολογείτε (χωρίς εκπτώσεις και προ ΦΠΑ)
- ΛΙΑΝΙΚΗ ΤΙΜΗ – Η τιμή πώλησης του είδους
- ΣΥΜΦΩΝΙΑ – Τιμή"01" για αγορά μας-εμπορία και τιμή "02" για άτυπη παρακαταθήκη
- ΕΠΟΧΗ – Τιμή "1" για χειμερινά, τιμή "3" για καλοκαιρινά, "0" για είδη άνευ εποχής
- ΕΤΟΣ ΕΙΔΟΥΣ - Το έτος της εμπορικής περιόδου (σαιζόν) που ανήκουν τα εμπορεύματα
- ΕΠΩΝΥΜΙΑ ΠΕΛΑΤΗ – Το επώνυμο που δηλώνει ο πελάτης
- ΠΕΡΙΟΧΗ – Η περιοχή που δηλώνει ο πελάτης ότι διαμένει
- ΤΚ – Ο ταχυδρομικός κώδικας που αντιστοιχεί στην κατοικία του πελάτη
- ΚΩΔΙΚΟΣ ΕΜΠΟΡΙΚΗΣ ΚΑΤΗΓΟΡΙΑΣ – Ο κωδικός αριθμός {1,2,3,4} που αντιστοιχεί στην εμπορική κατηγορία του είδους
- ΚΩΔΙΚΟΣ ΜΑΡΚΑΣ – Αριθμητικός κωδικός αντιστοιχίας της επωνυμίας του είδους
- ΜΑΡΚΑ – Η επωνυμία του είδους
- ΕΜΠΟΡΙΚΗ ΚΑΤΗΓΟΡΙΑ – Η εμπορική κατηγορία στην οποία ανήκει

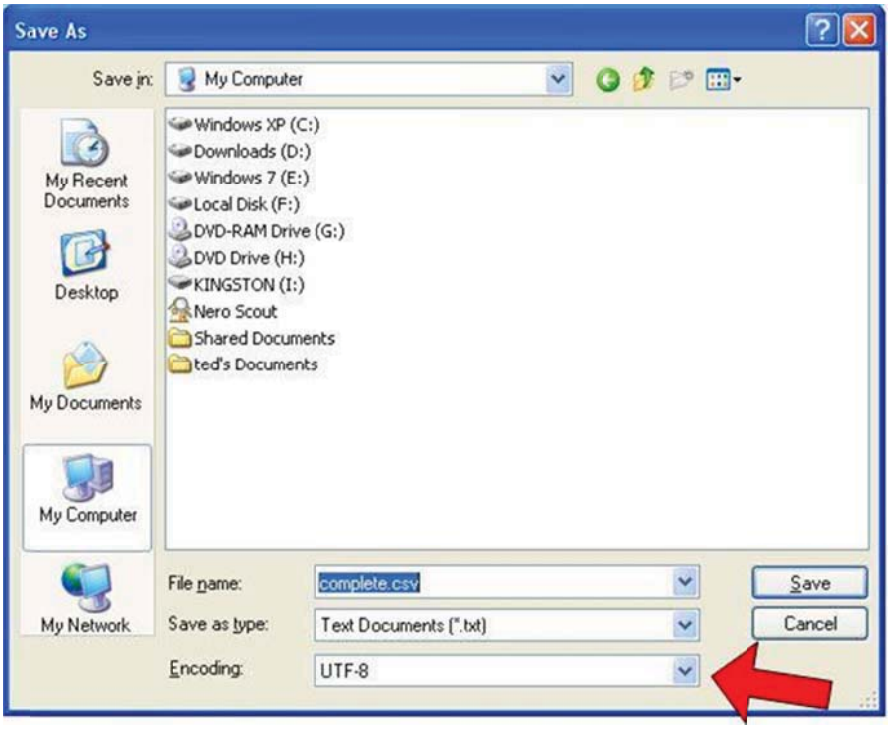
6.2. Τεχνικά Ζητήματα

Το weka επεξεργάζεται αρχεία τα οποία έχουν κατάληξη *.arff.* Επειδή όμως ο συγκεκριμένος τύπος αρχείου δεν είναι ιδιαίτερα δημοφιλής, διαθέτει ενσωματωμένο μετατροπέα αρχείων από *.csv* σε *.arff*. Τα αρχεία .csv (*comma-separated values*) αποθηκεύουν δεδομένα σε μορφή απλού κειμένου. Είναι ένας απλός τύπος αρχείου που είναι ευρέως διαδεδομένος και χρησιμοποιείται σε πολλές επιστημονικές και επιχειρηματικές εφαρμογές. [*http://en.wikipedia.org/wiki/.csv*] Στην περίπτωση μας είναι εύκολο να εξάγουμε ένα αρχείο .csv μέσα από το MS Excel καθώς διατίθεται η αντίστοιχη επιλογή



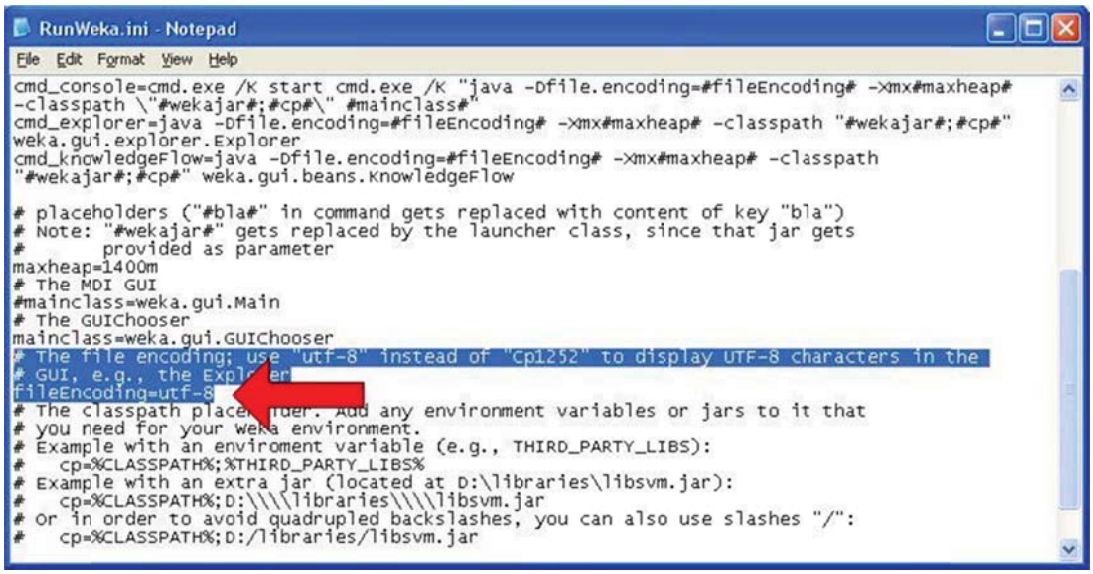
Πίνακας 1: ΑΠΟΘΗΚΕΥΣΗ ΑΡΧΕΙΟΥ ΣΕ ΜΟΡΦΗ *.csv*

Στη συνέχεια, για να μπορεί το weka να διαβάσει τους ελληνικούς χαρακτήρες που βρίσκονται στο αρχείο μας, απαιτείται η μετατροπή της κωδικοποίησης του από *Cp1252* σε *UTF-8*. Αυτό επιτυγχάνεται ανοίγοντας το αρχείο *.csv* με ένα απλό επεξεργαστή κειμένου (στην περίπτωσή μας με το notepad των windows) και αποθηκεύοντάς το ξανά αλλά με αλλαγμένη κωδικοποίηση (*encoding*). Επίσης τοποθετούμε την αντίστοιχη τιμή και στο αρχείο *weka.ini*



Πίνακας 2: ΜΕΤΑΤΡΟΠΗ ΚΩΔΙΚΟΠΟΙΗΣΗΣ ΣΕ *utf-8*

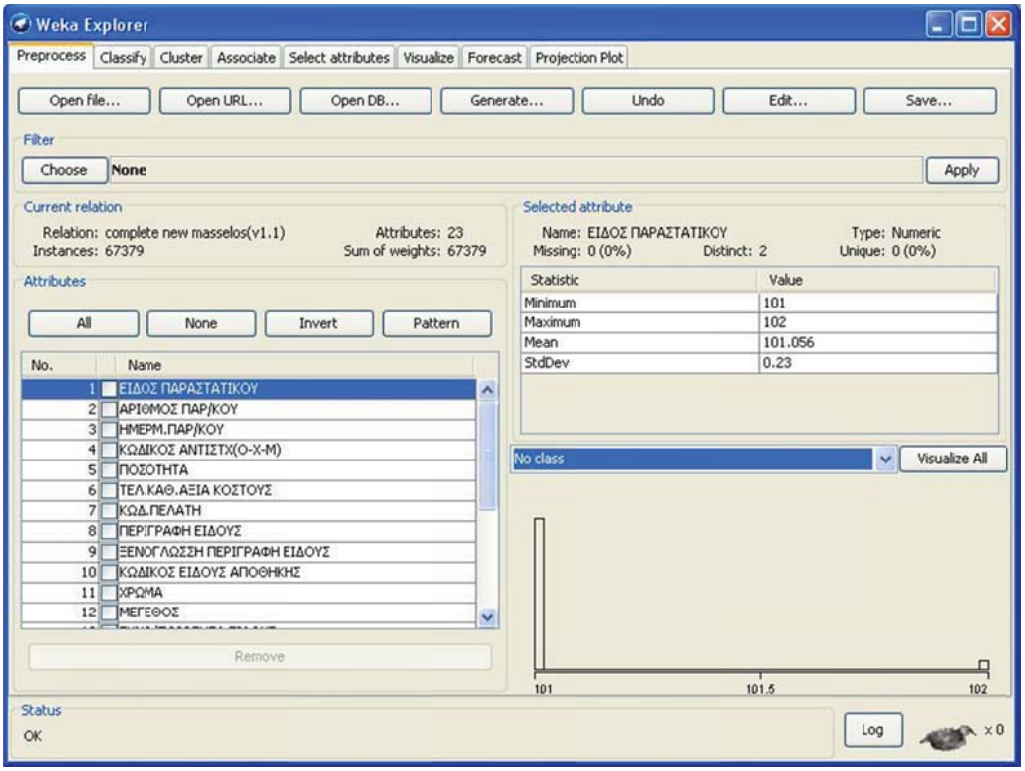
Επίσης τοποθετούμε την αντίστοιχη τιμή και στο αρχείο weka.ini



Πίνακας 3: ΤΡΟΠΟΠΟΙΗΣΗ ΣΤΟ ΑΡΧΕΙΟ weka.ini

Επόμενο βήμα είναι να αντικατασταθούν στο αρχείο του excel οι χαρακτήρες «%» και «‘», καθώς στα αρχεία .csv θεωρούνται ειδικοί χαρακτήρες και χρησιμοποιούνται για να υποδηλώσουν συνήθως αλλαγή γραμμής ή ύπαρξη σχολίων, επηρεάζοντας έτσι την σωστή εισαγωγή των δεδομένων. Το σύμβολο «%» όπου υπήρχε αντικαταστάθηκε από την συντόμευση (pct) από τη λέξη percentage. Το σύμβολο «‘» συνήθως βρισκόταν σε κείμενο ακολουθώντας το γράμμα Κ, πχ «*ΝΥΧΤΙΚΟ ΜΕΤΑΞΩΤΟ ΜΕ ΦΑΔΡΙΑ ΜΠΡΑΤΕΛΛΑ Κ' ΔΑΝΤΕΛΑ*». Σε τέτοιες περιπτώσεις αντικαταστάθηκαν οι χαρακτήρες «Κ’» από το σύμβολο «&». Στις υπόλοιπες περιπτώσεις, το σύμβολο «’» χρησιμοποιούταν απλά ως απόστροφος πχ «*7-JADE 501VERT D'EAU*», οπότε και αντικαταστάθηκε με ένα κενό.

Έπειτα από αυτές τις μετατροπές το αρχείο ήταν δυνατό να διαβαστεί από το weka και να κάνει την εξερεύνηση των δεδομένων πολύ πιο γρήγορη.



Πίνακας 4: ΑΝΟΙΓΜΑ ΑΡΧΕΙΟΥ ΜΕ ΤΟ weka

6.3. Καθαρισμός των δεδομένων

Ανοίγοντας τον weka explorer και παρακολουθώντας τα διάφορα attributes φαίνεται άμεσα ότι:

- Ο ΚΩΔΙΚΟΣ ΑΠΟΘΗΚΗΣ ΧΩΡΟΥ και η ΔΙΑΣΠΑΣΗ έχουν την τιμή 1 σε όλες τις εγγραφές. Δεν προσφέρουν επομένως κάποια πληροφορία στην ανάλυση που θα ακολουθούσε οπότε μπορούν να παραληφθούν.
- Ο ΚΩΔΙΚΟΣ ΑΝΤΙΣΤΟΙΧΙΑΣ (Ο-Χ-Μ) είναι μια σειρά χαρακτήρων που αποτελούν τη βάση για τις τιμές των ΟΔΗΓΟΣ, ΧΡΩΜΑ και ΜΕΓΕΘΟΣ. Δεν εμπεριέχει επομένως κάποια ανεξάρτητη πληροφορία σε σχέση με τα υπόλοιπα χαρακτηριστικά και μπορεί να παραληφθεί
- Ο ΚΩΔΙΚΟΣ ΕΙΔΟΥΣ ΑΠΟΘΗΚΗΣ και ο ΟΔΗΓΟΣ είναι στην ουσία το ίδιο χαρακτηριστικό, οπότε ένα από τα δύο μπορούσε να παραληφθεί
- Ο ΚΩΔΙΚΟΣ EAN 13 αποτελεί το EAN13 barcode του είδους (και οι 13 χαρακτήρες) στην περίπτωση του FULL-EDI και δεν προσφέρει κάποια πληροφορία στην ανάλυση που θα ακολουθήσει οπότε μπορεί να παραληφθεί. Το ίδιο και η ΠΡΟΕΛΕΥΣΗ.

- Η ΕΠΟΧΗ έχει σε όλες τις εγγραφές την τιμή 0 καθώς πρόκειται για προϊόντα άνευ εποχικότητας. Μπορεί επομένως να παραληφθεί.
- Η ΣΥΜΦΩΝΙΑ έχει σε όλες τις εγγραφές την τιμή «02» (άτυπη παρακαταθήκη) επομένως δεν προσφέρει κάποια πληροφορία στην ανάλυση και μπορεί να παραληφθεί.

Επιτυγχάνεται με την απλή αυτή αρχική παρατήρηση των δεδομένων μια μείωση των διαστάσεων τους κατά περίπου 26%.

Στη συνέχεια παρατηρώντας πιο προσεκτικά τις τιμές των διαφόρων χαρακτηριστικών από τον explorer του weka παρατηρούμε τα εξής:

- Το χαρακτηριστικό ΤΚ έχει πληθώρα προτύπων για την εγγραφή του ταχυδρομικού κώδικα (πχ 10433,10-433,104-33…). Κάνουμε επομένως χρήση των συναρτήσεων RIGHT, LEFT και CONCATENATE για την μετατροπή τους σε ένα κοινό format που θα είναι μια ακολουθία ψηφίων χωρίς παύλα (-) ή κάποιο άλλο σύμβολο ή κενό ανάμεσα τους. Επιπλέον υπάρχουν κάποιες εγγραφές με τιμή που έχει 4 ή 6 ψηφία. Έπειτα από έλεγχο, τα 6 ψηφία αντιστοιχούν σε ΤΚ της Κύπρου. Οι τιμές με τα 4 διορθώνονται με βάση την περιοχή που έχει δηλωθεί.
- Υπάρχουν 4 εγγραφές όπου το ΤΚ έχει εισαχθεί στη θέση της ΠΕΡΙΟΧΗΣ και αντιστρόφως. Διόρθωση των εγγραφών με αντιμετάθεση των τιμών.
- Υπάρχουν 24 εγγραφές με αρνητική τιμή στις στήλες ΠΟΣΟΤΗΤΑ και ΤΕΛΙΚΗ ΤΙΜΗ. Πρόκειται για εγγραφές που έχουν περαστεί με το χέρι, πιθανότατα λόγω προσωρινής βλάβης του συστήματος και αδυναμίας αυτόματης εισαγωγής. Διόρθωση των τιμών σε θετικές και τοποθέτηση της τιμής 102 στο ΕΙΔΟΣ ΠΑΡΑΣΤΑΤΙΚΟΥ καθώς πρόκειται για επιστροφές.
- Παρουσία της ίδιας επωνυμίας πελάτη σε διαφορετικές πιθανές τιμές της τιμής του αντίστοιχου χαρακτηριστικού. Για παράδειγμα η τιμή «ΑΝΤΩΝΑΚΟΥ» εμφανίζεται δύο φορές στις πιθανές τιμές του χαρακτηριστικού ΕΠΩΝΥΜΙΑ ΠΕΛΑΤΗ. Έπειτα από μια ταξινόμηση στο excel του χαρακτηριστικού αυτού φάνηκε για ποιο λόγο συμβαίνει αυτό. Παρότι φαίνεται να είναι η ίδια τιμή εκ πρώτης όψεως, στη μια από τις δύο περιπτώσεις τα πρώτα γράμματα είναι με λατινικούς χαρακτήρες. Προφανώς ο χρήστης του συστήματος ξεκίνησε την εισαγωγή του ονόματος με το πληκτρολόγιο να είναι στα αγγλικά αλλά δεν το κατάλαβε παρά μόνο όταν έφτασε στον χαρακτήρα «Ω» που δεν είναι ίδιος στο λατινικό αλφάβητο. Εκεί άλλαξε το πληκτρολόγιο σε ελληνικούς χαρακτήρες αλλά αμέλησε

να διορθώσει και τους προηγούμενους. Έτσι το σύστημα καταχώρησε το επώνυμο ως μια νέα τιμή. Διόρθωση των τιμών με το ίδιο πρόβλημα με την σωστή εισαγωγή των αντίστοιχων ελληνικών χαρακτήρων. Πρόκειται για συνολικά 123 λάθος γραμμένα ονόματα (Περίπου το 2.8% των συνολικών τιμών).

- Υπάρχουν ΕΠΩΝΥΜΑ ΠΕΛΑΤΩΝ που περιέχουν ορθογραφικά λάθη με αποτέλεσμα την καταχώρησή τους με νέο κωδικό πελάτη, παρότι ενδέχεται να πρόκειται για ήδη καταχωρημένους πελάτες. Ακολούθησε ορθογραφική διόρθωση 83 τιμών του χαρακτηριστικού ΕΠΩΝΥΜΙΑ ΠΕΛΑΤΗ (Περίπου το 2% των συνολικών τιμών)
- Υπάρχουν ΕΠΩΝΥΜΑ ΠΕΛΑΤΩΝ που έχουν διαφορετικό ΚΩΔΙΚΟ ΠΕΛΑΤΗ επειδή ανοίχθηκαν σε διαφορετικά καταστήματα, παρότι πρόκειται για το ίδιο φυσικό πρόσωπο. Το τελευταίο διαπιστώνεται τόσο από το ΤΚ όσο και από την σύγκριση, όσο αυτή είναι εφικτή, των μεγεθών των προϊόντων τα οποία αγόρασαν.
- Για τους ΠΕΛΑΤΕΣ που έχουν ανοιχθεί πάνω από ένας ΚΩΔΙΚΟΙ ΠΕΛΑΤΩΝ, είτε λόγω των διαφόρων προβλημάτων που προαναφέραμε πως μπορεί να υπήρχαν κατά την εισαγωγή των δεδομένων είτε λόγω απροσεξίας του χρήστη, διόρθωση των αντίστοιχων τιμών του ΚΩΔΙΚΟΥ ΠΕΛΑΤΗ. Με τη βοήθεια της ταξινόμησης του excel και τη χρήση φίλτρων διορθώνουμε τους πελάτες ώστε να έχουν ένα μοναδικό κωδικό (Αφαιρέθηκαν 199 κωδικοί – περίπου το 3.3% του συνόλου).
- Σε μερικές περιπτώσεις τα στοιχεία των πελατών ήταν ελλειπή ή ημιτελή. Για παράδειγμα μπορεί να ήταν συμπληρωμένο το πεδίο στο ΕΠΩΝΥΜΟ ΠΕΛΑΤΗ, χωρίς όμως να έχει συμπληρωθεί το αντίστοιχο πεδίο στο ΤΚ ή στην ΠΕΡΙΟΧΗ. Φυσικά αυτό μπορεί να συνέβη γιατί ενδέχεται ο πελάτης να αρνήθηκε να δώσει όλα τα στοιχεία του για λόγους προσωπικών δεδομένων. Έγινε συμπλήρωση ορισμένων εγγραφών όπου τα στοιχεία που έλειπαν ήταν γνωστά από άλλες επισκέψεις. Για τον έλεγχο και την επιβεβαίωση ότι επρόκειτο για το ίδιο πρόσωπο χρησιμοποιήθηκε εκτός από το ΕΠΩΝΥΜΟ ΠΕΛΑΤΗ, το ΤΚ και το ΜΕΓΕΘΟΣ των προϊόντων που αγόρασε όπου αυτό ήταν εφικτό.

6.4. Μετασχηματισμός των δεδομένων

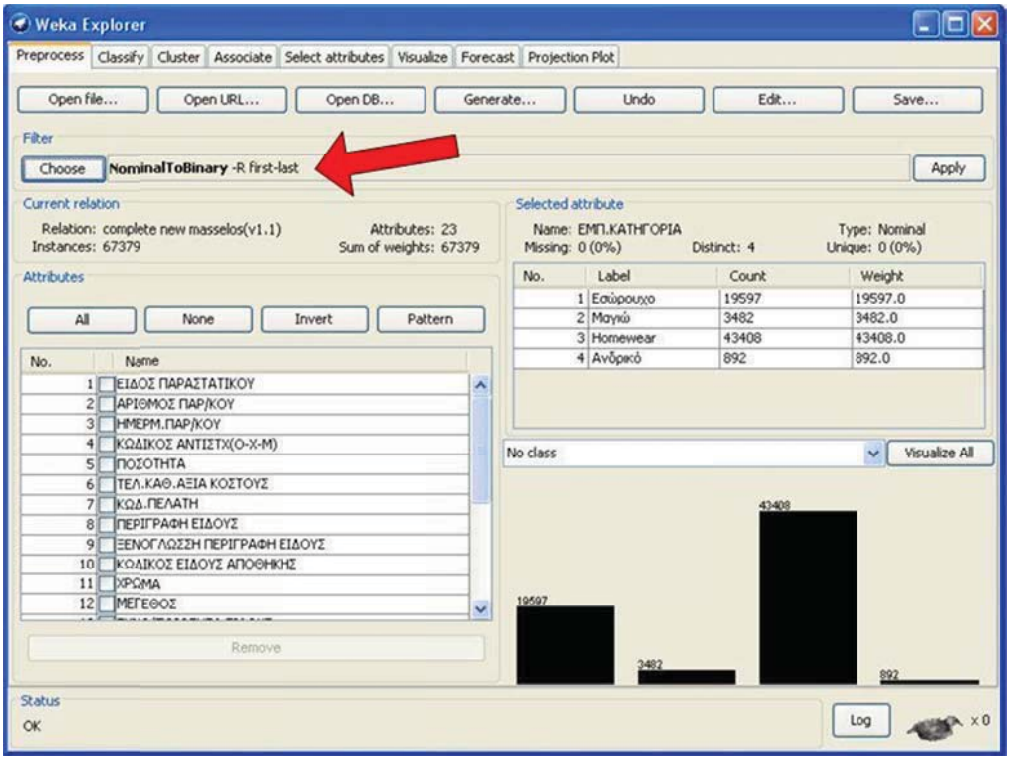
Μετά τον καθαρισμό των δεδομένων σειρά έχει ο μετασχηματισμός τους. Πρέπει να τα φέρουμε σε μορφή τέτοια ώστε να είναι πιο εύκολο για το λογισμικό που χρησιμοποιούμε να τα επεξεργαστεί αλλά και για μας να κατανοήσουμε τις πληροφορίες που θα μας δώσει. Αυτό περιλαμβάνει τόσο την μετατροπή ορισμένων

χαρακτηριστικών που χρησιμοποιούμε όσο και την δημιουργία νέων βοηθητικών χαρακτηριστικών. Ο τύπος των χαρακτηριστικών που κάνουν πιο εύκολη και ακριβή την λειτουργία κάθε αλγορίθμου είναι διαφορετικός. Ορισμένοι αλγόριθμοι λειτουργούν καλύτερα με ονομαστικά χαρακτηριστικά ενώ άλλοι με αριθμητικά χαρακτηριστικά. Γενικότερα ωστόσο οι περισσότεροι αλγόριθμοι που χρησιμοποιούνται έχουν ευχέρεια στην χρήση δυαδικών χαρακτηριστικών, attributes δηλαδή που παίρνουν δυαδικές τιμές (0 και 1). Σε αυτή την μορφή θα φέρουμε τα περισσότερα από τα χαρακτηριστικά που θα χρησιμοποιήσουμε.

6.4.1. Μετατροπή ήδη υπαρχόντων χαρακτηριστικών

Καταρχάς απαιτείται η μετατροπή μερικών από των ήδη υπαρχόντων χαρακτηριστικών σε δυαδικά. Αυτό μπορεί να γίνει είτε μέσω του weka είτε κατευθείαν στο αρχείο excel.

Μέσω του weka μπορεί κανείς να χρησιμοποιήσει το αντίστοιχο filter που υπάρχει. Αυτό μπορεί να βρεθεί ανοίγοντας τη λίστα των filters και ακολουθώντας την διαδρομή *weka.filters.attribute.NominalToBinary*



Πίνακας 5: ΦΙΛΤΡΟ Nominal to Binary

Αυτό το φίλτρο μετατρέπει ένα ονομαστικό χαρακτηριστικό σε δυαδικό.

Η ίδια διαδικασία μπορεί να γίνει και από εμάς στο excel και είναι αυτή που θα χρησιμοποιηθεί στην παρούσα ανάλυση.

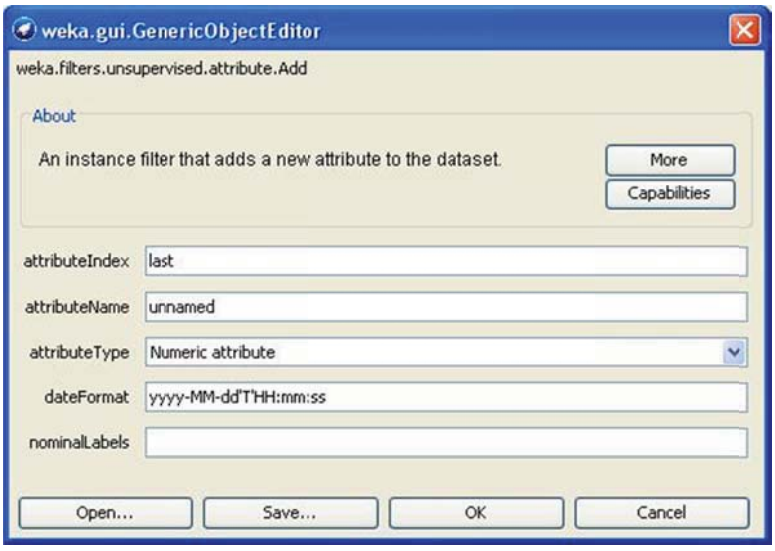
Αρχικά μετατρέπουμε το χαρακτηριστικό *ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΣΤΑΤΙΚΟΥ* σε τέσσερα δυαδικά χαρακτηριστικά που θα αφορούν την εποχή που καταχωρήθηκε το παραστατικό. Χρησιμοποιώντας την συνάρτηση if του excel δημιουργούμε τα χαρακτηριστικά *ΧΕΙΜΩΝΑΣ, ΑΝΟΙΞΗ, ΚΑΛΟΚΑΙΡΙ, ΦΘΙΝΟΠΩΡΟ*. Αυτά παίρνουν την τιμή «1» αν ημερομηνία του χαρακτηριστικού εμπίπτει στην αντίστοιχη εποχή και «0» σε αντίθετη περίπτωση.

Έπειτα δημιουργούμε το δυαδικό χαρακτηριστικό *«ΕΠΙΣΤΡΟΦΕΣ (1 καταγραφή ανά επίσκεψη)»* το οποίο θα ελέγχει κάθε επίσκεψη κάθε κωδικού πελάτη ξεχωριστά και θα παίρνει την τιμή «1» αν υπάρχει έστω και μια επιστροφή είδους (αν επομένως υπαρχει έστω και ένα *ΕΙΔΟΣ ΠΑΡΑΣΤΑΤΙΚΟΥ* με τιμή «102») και «0» σε αντίθετη περίπτωση. Αυτός ο έλεγχος θα γίνεται για κάθε διαφορετική επίσκεψη του πελάτη (δηλαδή σε κάθε διαφορετική *ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΣΤΑΤΙΚΟΥ*).

6.4.2. Δημιουργία νέων βοηθητικών χαρακτηριστικών

Μετά την μετατροπή των ήδη υπαρχόντων, προχωράμε στην δημιουργία νέων βοηθητικών χαρακτηριστικών. Και αυτά μπορούν να δημιουργηθούν είτε μέσω των filters του weka είτε απευθείας στο excel.

Στο weka, μπορεί κανείς να ορίσει ένα νέο χαρακτηριστικό με το φίλτρο *weka.filters.unsupervised.attribute.Add.*



Πίνακας 6: ΦΙΛΤΡΟ ΓΙΑ ΤΗΝ ΠΡΟΣΘΗΚΗ ΧΑΡΑΑΚΤΗΡΙΣΤΙΚΩΝ

Έχει επιλογές να ορίσει τόσο το όνομα του χαρακτηριστικού όσο και τον τύπο του ή της τιμές που θα μπορεί να πάρει.

Ακόμα, μεταξύ άλλων, άξια αναφοράς είναι το *weka.filters.unsupervised.attribute.AddExpression* όπου δημιουργεί ένα νέο χαρακτηριστικό εφαρμόζοντας μια μαθηματική έκφραση πάνω σε ήδη υπάρχον

χαρακτηριστικό και το *weka.filters.unsupervised.attribute.AddID* όπου δημιουργεί ένα νέο χαρακτηριστικό με μοναδική τιμή για κάθε εγγραφή.

Στην περίπτωση μας, για την δημιουργία των νέων βοηθητικών χαρακτηριστικών, θα καταφύγουμε απευθείας στο excel. Δημιουργούμε αρχικά δύο χαρακτηριστικά με τις ονομασίες *Temporary Count 1* και *Temporary Count 2*. Αυτά θα είναι προσωρινοί καταμετρητές που θα υποδεικνύουν, έπειτα από ταξινόμηση των δεδομένων κατά κωδικό πελάτη, ο πρώτος το τέλος μιας επίσκεψης πελάτη και την αρχή μιας άλλης, και ο δεύτερος την αλλαγή του κωδικού πελάτη που εξετάζουμε. Η χρήση τους θα γίνεται κυρίως κατά την λειτουργία των επόμενων χαρακτηριστικών.

Έπειτα δημιουργούμε το χαρακτηριστικό *ΕΠΙΣΚΕΨΕΙΣ*, το οποίο θα υπολογίζει το πλήθος των επισκέψεων κάθε κωδικού χρησιμοποιώντας το *Temporary Count 1*. Από αυτό προκύπτει το χαρακτηριστικό *ΚΟΣΤΟΣ / ΕΠΙΣΚΕΨΗ* που αθροίζει τη χρηματική συναλλαγή που έλαβε χώρα σε κάθε επίσκεψη του πελάτη ξεχωριστά. Να σημειώσουμε ότι οι επιστροφές λογίζονται ως αρνητικό μέγεθος στην παραπάνω άθροιση. Ακόμα δημιουργούμε και το *ΣΥΝΟΛΙΚΗ ΑΞΙΑ ΑΓΟΡΩΝ* το οποίο αθροίζει τη αξία των αγορών ενός πελάτη από όλες τις επισκέψεις που έχει κάνει και το *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ* που υπολογίζει το μέσο όρο της αξίας των αγορών ανά επίσκεψη του πελάτη.

Τέλος δημιουργούμε βοηθητικά χαρακτηριστικά σχετικά με την χρονική περίοδο που έγιναν αγορές από τον κάθε πελάτη. Πρώτα δημιουργούμε το *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008* το οποίο υπολογίζει το ποσοστό των καταγεγραμμένων επισκέψεων του πελάτη που αφορούν το έτος 2008 προς το σύνολο των καταγεγραμμένων επισκέψεων του. Αντίστοιχα και για τα έτη 2009, 2010, 2011 για τα οποία έχουμε στοιχεία. Έπειτα με ανάλογο τρόπο σκέψης δημιουργούμε το *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ* το οποίο υπολογίζει το ποσοστό των καταγεγραμμένων επισκέψεων του πελάτη τους χειμερινούς μήνες προς το σύνολο των επισκέψεών του. Αντίστοιχα και για την άνοιξη, το καλοκαίρι και το φθινόπωρο.

7. Επεξεργασία Δεδομένων

Με τα δεδομένα καθαρισμένα πλέον και τη δημιουργία κάποιων βασικών βοηθητικών χαρακτηριστικών πραγματοποιημένη, είμαστε έτοιμοι να ξεκινήσουμε τη διαδικασία του *data mining*.

7.1. Ομαδοποίηση

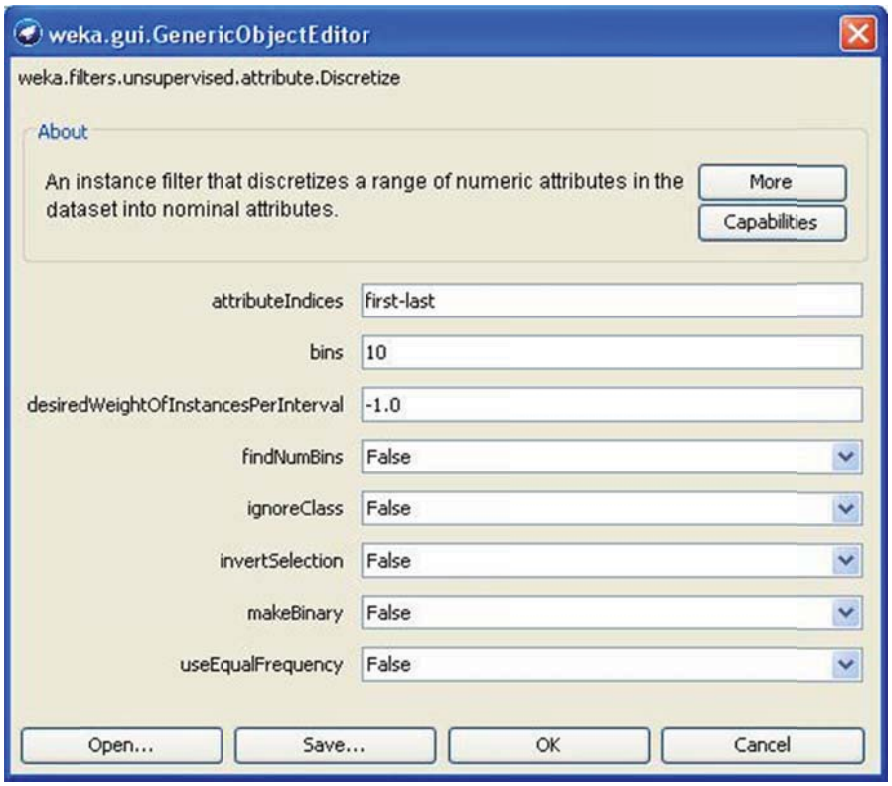
Το πρώτο ζητούμενο της ανάλυσής μας είναι η δημιουργία μιας mailing list, μιας λίστας πελατών δηλαδή, στην οποία θα αποσταλούν ενημερωτικοί κατάλογοι με τα νέα προϊόντα της εταιρείας. Στόχος μας είναι η μείωση του κόστους με την επιλογή των πελατών που είναι πιο πιθανό να ανταποκριθούν στην συγκεκριμένη προώθηση. Προσπαθούμε επομένως να εξάγουμε πρότυπα από τα στοιχεία που έχουμε για τους πελάτες και να τους κατατάξουμε σε ομάδες ανάλογα με την αγοραστική τους συμπεριφορά. Τα στοιχεία που θα χρησιμοποιήσουμε για την δημιουργία των ομάδων αυτών αφορούν:

a) Την αξία των αγορών του κάθε πελάτη
b) Τις καταγεγραμμένες επισκέψεις που έχει ένας πελάτης στην εταιρεία
c) Το έτος που έκαναν αγορές
d) Την εποχή που πραγματοποιήθηκαν οι αγορές αυτές

Καταρχάς, για την ομαδοποίηση, χρειαζόμαστε ένα αρχείο το οποίο θα έχει σε κάθε του γραμμή και ένα μοναδικό κωδικό πελάτη. Δεν θέλουμε οι αγορές του ενός πελάτη να αναλύονται σε 5, για παράδειγμα, υποδείγματα και ενός άλλου μόνο σε 1, γιατί έτσι ο αλγόριθμος μεροληπτεί εις βάρος των πελατών που έχουν μικρό αριθμό καταγραφών. Χρησιμοποιούμε λοιπόν την επιλογή του excel για κατάργηση διπλοτύπων με βάση τον κωδικό πελάτη. Έτσι έχουμε μόνο ένα υπόδειγμα για κάθε κωδικό και συγκεντρωμένα στα βοηθητικά χαρακτηριστικά τα περισσότερα στοιχεία που αφορούν την αγοραστική του συμπεριφορά. Επίσης χρησιμοποιούμε το *weka.filters.unsupervised.instance.RemoveWithValues* για να αφαιρέσουμε όσους κωδικούς πελατών δεν αναφέρονται σε πελάτες (πχ ο κωδικός που αντιστοιχεί στο επώνυμο πελάτη *ΜΑΣΣΕΛΟΣ* προορίζεται για εσωτερική χρήση και δεν αντιστοιχεί σε κάποιο πραγματικό πελάτη) ή δεν μπορούν να προσφέρουν κάποια πληροφορία (πχ ο κωδικός που αντιστοιχεί στο επώνυμο πελάτη *ΠΕΛΑΤΗΣ ΛΙΑΝΙΚΗΣ* αφορά ένα σύνολο πελατών για τους οποίους δεν υπάρχουν στοιχεία, επομένως δεν μπορεί να προσφέρει κάποια πραγματική πληροφορία στην συγκεκριμένη ανάλυση) ή αποτελούν πολύ ακραίες τιμές οι οποίες επηρεάζουν αρνητικά την ακρίβεια του μοντέλου.

a) Για την αξία των αγορών του κάθε πελάτη θα χρησιμοποιήσουμε τη *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ*. Ωστόσο, όπως προαναφέραμε, προτιμάμε να φέρουμε τα χαρακτηριστικά μας σε δυαδικά ή σε αριθμητικά με ελάχιστη τιμή το 0 και μέγιστη το 1. Για την μετατροπή των αριθμητικών χαρακτηριστικών σε δυαδικά απαιτείται

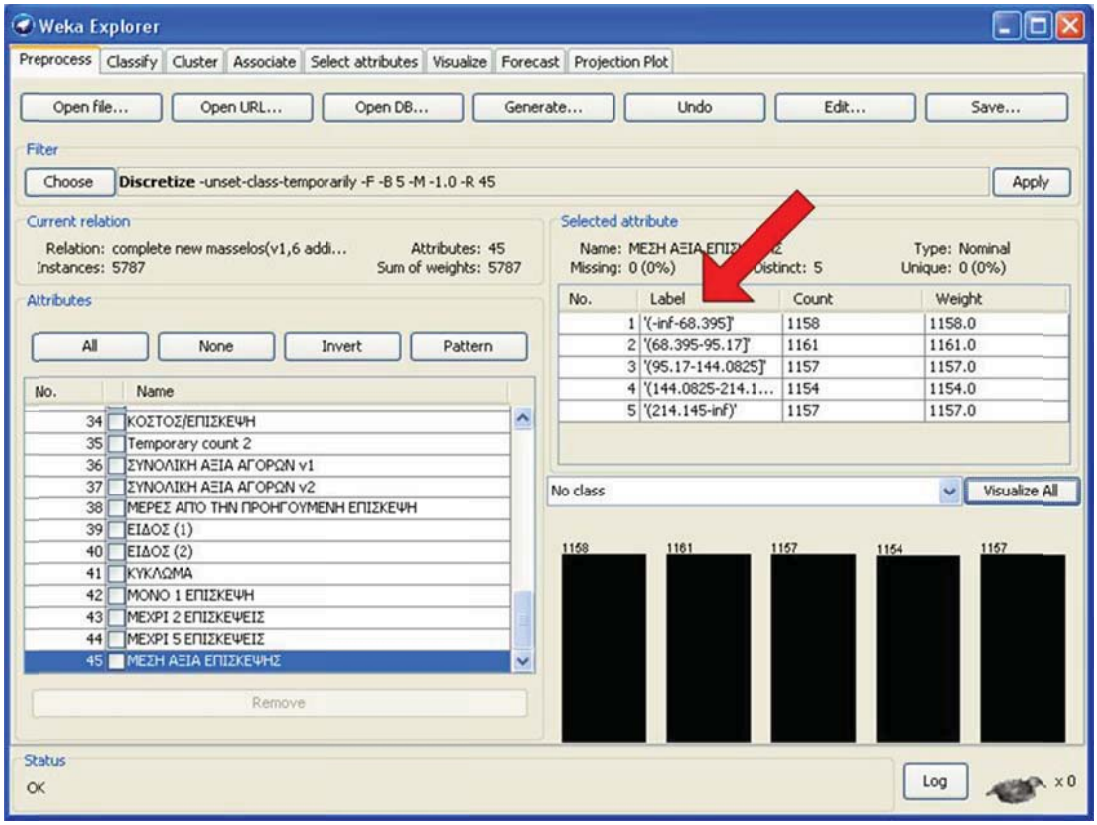
διακριτοποίηση. Αυτή γίνεται με τη βοήθεια του *weka.filters.unsupervised.attribute.Discretize*



Πίνακας 7: ΦΙΛΤΡΟ ΓΙΑ ΤΗ ΔΙΑΚΡΙΤΟΠΟΙΗΣΗ ΑΡΙΘΜΗΤΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Με αυτό έχουμε τη δυνατότητα εκτός των άλλων να επιλέξουμε σε ποιό χαρακτηριστικό θα κάνουμε διακριτοποίηση και αν θα είναι σταθερού εύρους ή σταθερής συχνότητας. Για την ανάλυση μας επιλέγουμε τη σταθερή συχνότητα, δηλαδή η δημιουργία των υποσυνόλων θα γίνει με κριτήριο την ισοκατανομή των υποδειγμάτων μέσα σε αυτά. Η επιλογή της αυτόματης εύρεσης του βέλτιστου αριθμού υποσυνόλων λειτουργεί μόνο για την περίπτωση του σταθερού εύρους, οπότε στην περίπτωσή μας απαιτούνται δοκιμές για να βρούμε τα υποσύνολα που μας φαίνονται πιο λογικοφανή και λειτουργικά.

Δοκιμάζοντας την διακριτοποίηση με διάφορους αριθμούς υποσυνόλων, επιλέγουμε ως βέλτιστο το νούμερο 5. Αυτό γιατί και τα υποσύνολα που προκύπτουν είναι ισόποσα χωρισμένα αλλά και γιατί τα όρια των υποσυνόλων συνάδουν με τον πραγματικό κόσμο. Για παράδειγμα ένα υποσύνολο με *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ* «125-135€», μπορεί να προκύπτει από τον αλγόριθμο και να περιέχει ίδιο αριθμό υποδειγμάτων με τα άλλα υποσύνολα, αλλά δεν περιγράφει κάποια ομάδα πελατών που στην πραγματικότητα αποτελεί από μόνη της μια κατηγορία λόγω του πολύ μικρού εύρους του.



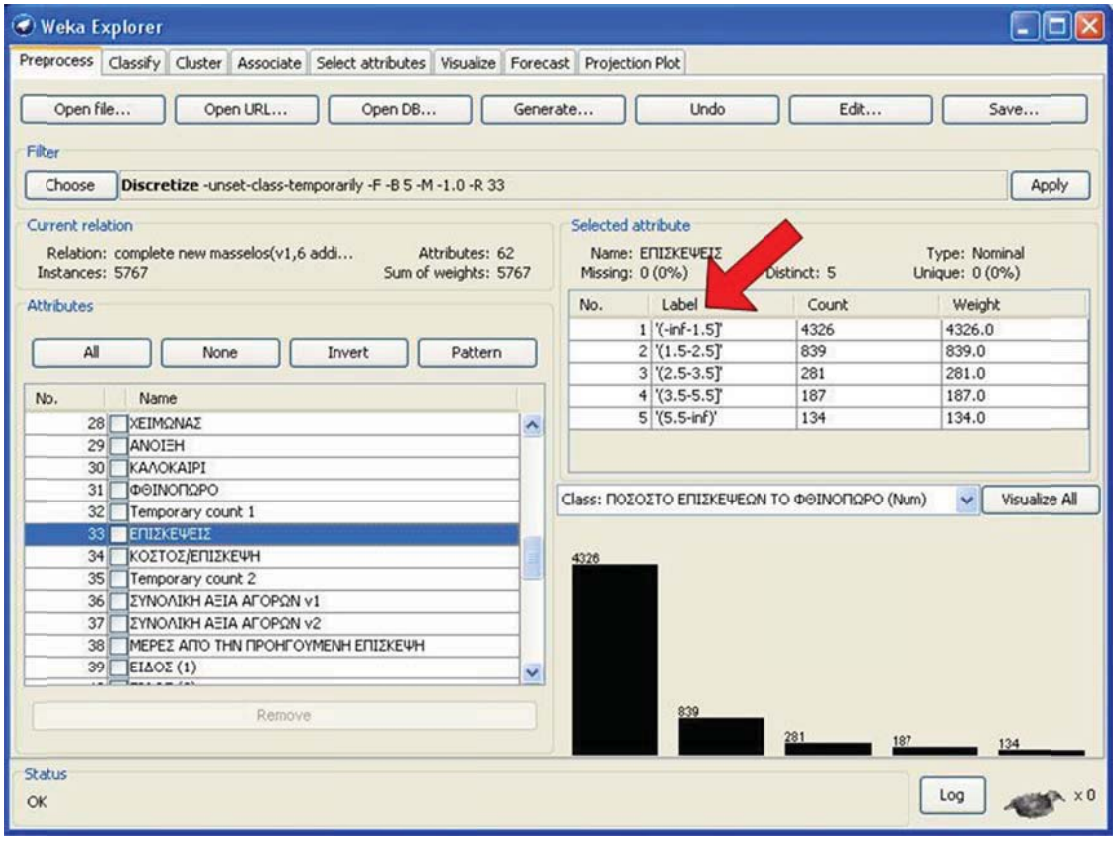
Πίνακας 8: ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΙΑΚΡΙΤΟΠΟΙΗΣΗΣ ΩΣ ΠΡΟΣ ΤΗΝ ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ

Δημιουργούμε τα δυαδικά χαρακτηριστικά:

- *ΜΕΣΗ ΑΞΙΑ < 68€*
- *ΜΕΣΗ ΑΞΙΑ ΜΕΧΡΙ 95€*
- *ΜΕΣΗ ΑΞΙΑ ΜΕΧΡΙ 143€*
- *ΜΕΣΗ ΑΞΙΑ 212€*

Τα χαρακτηριστικά αυτά παίρνουν την τιμή 0 αν η μέση αξία αγορών του πελάτη είναι μεγαλύτερη από το αναγραφόμενο όριο (είναι δηλαδή ψευδή) και την τιμή 1 αν η μέση αξία αγορών είναι μικρότερη από το αντίστοιχο όριο, είναι δηλαδή αληθής η πρόταση.

b) Την αντίστοιχη διαδικασία κάνουμε και για να περιγράψουμε το ποσό των καταγεγραμμένων επισκέψεων. Χρησιμοποιούμε το χαρακτηριστικό ΕΠΙΣΚΕΨΕΙΣ και με το φίλτρο του weka το διακριτοποιούμε. Πειραματιζόμενοι με τον αριθμό των υποσυνόλων καταλήγουμε στα 5 υποσύνολα ως βέλτιστο σημείο συμβιβασμού μεταξύ της ισοκατανομής των υποσυνόλων και της αντιστοιχίας στην πραγματικότητα.



Πίνακας 9: ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΙΑΚΡΙΤΟΠΟΙΗΣΗΣ ΩΣ ΠΡΟΣ ΤΟΝ ΑΡΙΘΜΟ ΕΠΙΣΚΕΨΕΩΝ

Βλέπουμε όμως ότι το υποσύνολο νο.3 {2.5 – 3.5} αφορά μόνο τους πελάτες που έχουν κάνει 3 επισκέψεις και δεν είναι πολύ μεγάλο σε μέγεθος. Επίσης θεωρούμε πως δεν έχει κάποιο ιδιαίτερο γνώρισμα ή χαρακτηριστικό σε σχέση με τους πελάτες των 4 επισκέψεων. Με την κατανομή λοιπόν να παραμένει της ίδιας περίπου μορφής δημιουργούμε στο excel τα δυαδικά χαρακτηριστικά που εκφράζουν τα αντίστοιχα υποσύνολα:

- *ΜΕΧΡΙ 1 ΕΠΙΣΚΕΨΗ (<1.5)*
- *ΜΕΧΡΙ 2 ΕΠΙΣΚΕΨΕΙΣ (1.5 – 2.5)*
- *ΜΕΧΡΙ 5 ΕΠΙΣΚΕΨΕΙΣ (2.5 – 5.5)*

Κάποιος μπορεί να αναρωτηθεί γιατί δεν χρησιμοποιούσαμε τότε την διακριτοποίηση με 4 υποσύνολα. Ο λόγος είναι ότι στα 4 υποσύνολα, ο αλγόριθμος έκανε την τελευταία διάκριση στις 3 επισκέψεις. Όμως οι 3 επισκέψεις δεν θεωρήθηκαν αριθμός ικανός για να χαρακτηρίσει τον πελάτη ως «πιστό». Κάτι που μπορεί να το ισχυριστεί κάποιος για πελάτες που έχουν άνω των 5 επισκέψεων. Επίσης βλέπουμε ότι η διαφορά στην κατανομή των υποδειγμάτων θα ήταν αμελητέα, καθώς αφορά ένα πλήθος κωδικών πελατών μικρότερο του 5% του συνολικού.

Και σε αυτά τα χαρακτηριστικά χρησιμοποιούμε τις τιμές 0 και 1 όπως στην παραπάνω περίπτωση. Για τα υποδείγματα που η συνθήκη του χαρακτηριστικού

είναι αληθής 1 και για αυτές που είναι αναληθείς 0. Έτσι για παράδειγμα ένας πελάτης με 8 καταγεγραμμένες επισκέψεις θα εκφραζόταν με την τιμή 0 σε όλα τα παραπάνω χαρακτηριστικά.

c) Τα δεδομένα που έχουμε τοποθετούνται χρονικά από το 2008 μέχρι και το 2011. Εμείς θέλουμε σε ένα υπόδειγμα να δείξουμε σε ποια έτη και πως κατανεμήθηκαν οι επισκέψεις που έχει κάνει κάθε πελάτης. Για το σκοπό αυτό δημιουργούμε χαρακτηριστικά που περιέχουν το ποσοστό των επισκέψεων κάθε πελάτη που έγινε ένα συγκεκριμένο έτος προς το σύνολο των καταγεγραμμένων επισκέψεων του. Δημιουργούμε επομένως τα:

- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008*
- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009*
- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010*
- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2011*

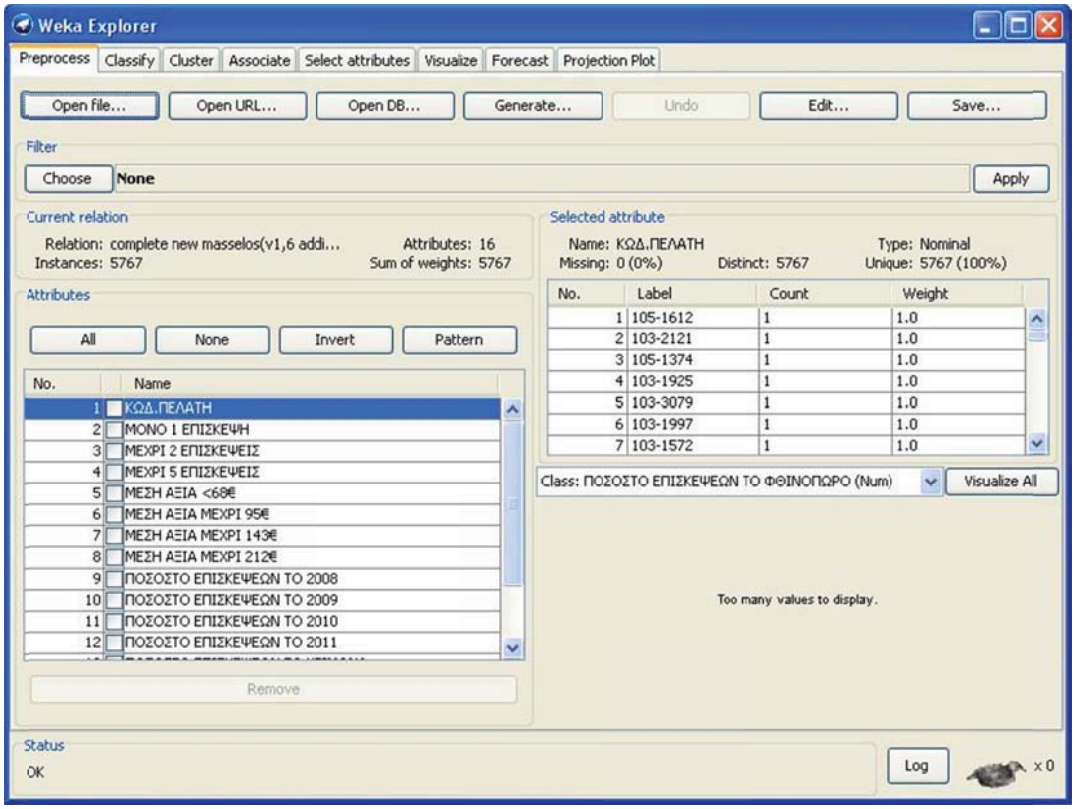
Η τιμή του καθενός από αυτά προκύπτει ως ο λόγος των επισκέψεων του πελάτη με ημερομηνία στο έτος κάθε χαρακτηριστικού προς τις συνολικές του επισκέψεις. Τα χαρακτηριστικά αυτά επομένως παρότι δεν είναι δυαδικά αλλά συνεχή με ελάχιστη τιμή 0 και μέγιστη 1, μοιάζουν αρκετά στον τρόπο έκφρασης της πληροφορίας με τα δυαδικά και μας βολεύουν για την περιγραφή των ομάδων.

d) Τέλος για την εποχή που πραγματοποιήθηκαν οι αγορές δημιουργούμε χαρακτηριστικά αντίστοιχα με τα παραπάνω:

- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ*
- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΗΝ ΑΝΟΙΞΗ*
- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΚΑΛΟΚΑΙΡΙ*
- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΦΘΙΝΟΠΩΡΟ*

Η λειτουργία τους, η δομή τους και η πληροφορία που εκφράζουν είναι το αντίστοιχο των παραπάνω χαρακτηριστικών σε ότι αφορά τις εποχές.

Έτοιμοι πλέον, ανοίγουμε το αρχείο μας με τον explorer του weka, αφαιρούμε τα υπόλοιπα χαρακτηριστικά και κρατάμε μόνο τα προαναφερθέντα μαζί με το χαρακτηριστικό ΚΩΔΙΚΟΣ ΠΕΛΑΤΗ. Στην παρακάτω εικόνα βλέπουμε τα χαρακτηριστικά και στον δίπλα πίνακα επιβεβαιώνεται ότι έχουμε ένα υπόδειγμα για κάθε κωδικό πελάτη.



Πίνακας 10: ΑΡΧΕΙΟ ΕΤΟΙΜΟ ΓΙΑ ΤΗ ΔΙΑΔΙΚΑΣΙΑ ΟΜΑΔΟΠΟΙΗΣΗΣ

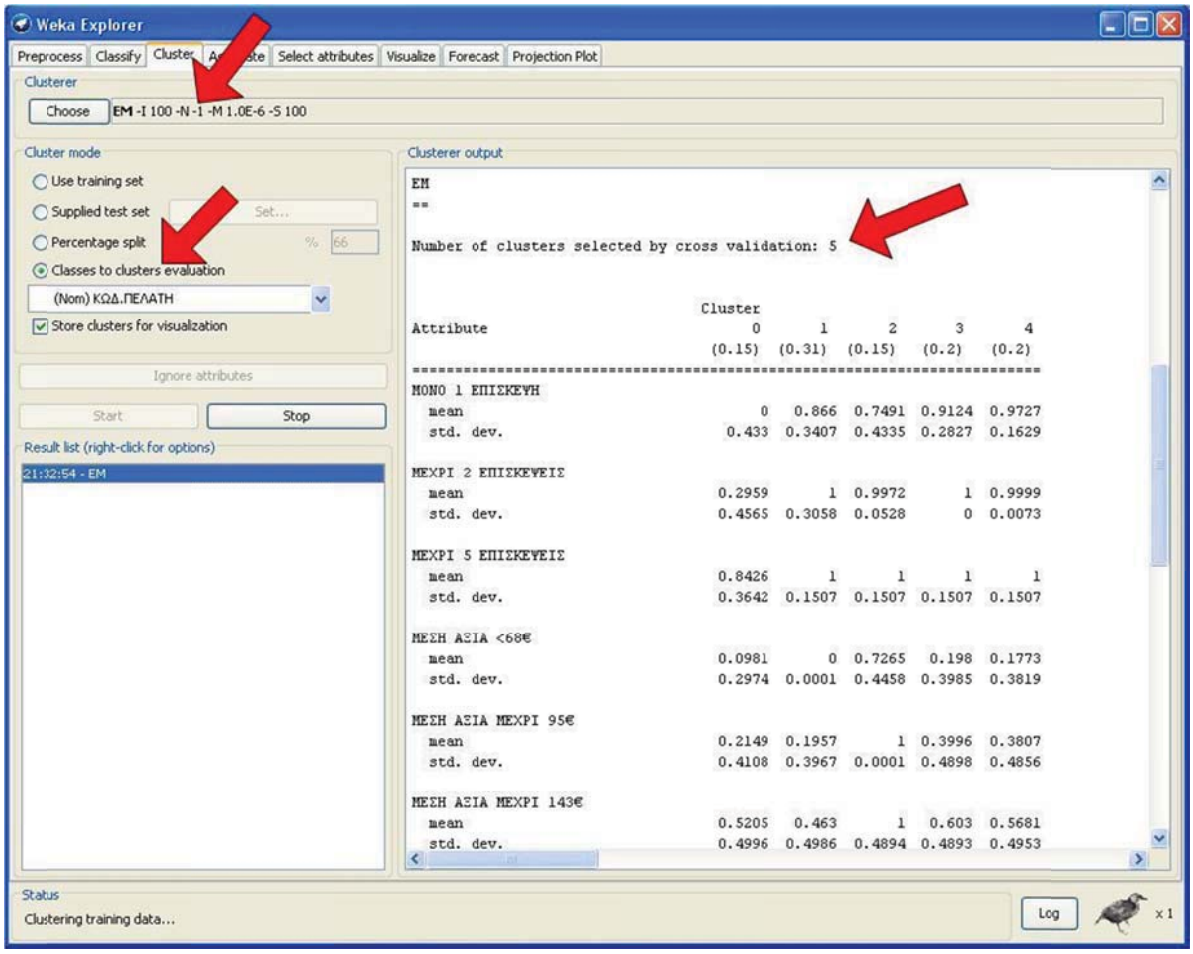
Για την ομαδοποίηση θα χρησιμοποιήσουμε τους δύο πιο γνωστούς και ευρέως χρησιμοποιημένους αλγορίθμους ομαδοποίησης. Τον EM και τον KMeans.

Ο KMeans (k μέσων) είναι ο πιο γνωστός αλγόριθμος ομαδοποίησης. Χρησιμοποιείται για τον διαχωρισμό των υποδειγμάτων σε k ομάδες. Στην αρχή επιλέγονται στην αρχή με τυχαίο τρόπο τα κέντρα των ομάδων, εκχωρούνται τα υποδείγματα στη συνέχεια στις ομάδες με κριτήριο την απόσταση τους από τα κέντρα και στη συνέχεια επαναϋπολογίζονται τα κέντρα μέχρι να επέλθει σύγκλιση.

Ο EM (Expectation – maximization) (Μεγιστοποίησης εκτίμησης) αποτελεί μια γενίκευση του KMeans σε πιθανοκρατικό πλαίσιο. Σε αυτόν τον αλγόριθμο υπολογίζονται οι πιθανότητες κάθε υποδείγματος να ανήκει σε μια ομάδα και στην συνέχεια γίνεται εκτίμηση των παραμέτρων κατανομών από τις πιθανότητες των ομάδων και αποθηκεύονται οι πιθανότητες των ομάδων ως βαρύτητες σε κάθε υπόδειγμα. Όταν η βελτίωση είναι αμελητέα ο αλγόριθμος διακόπτεται.

Αρχικά χρησιμοποιούμε τον explorer του weka και στην καρτέλα Cluster επιλέγουμε τον αλγόριθμο EM. Στην επιλογή numClusters σχετικά με τον αριθμό των ομάδων που θα δημιουργηθούν, επιλέγουμε «-1». Αυτό δίνει την εντολή στον αλγόριθμο να υπολογίσει μόνος του τον βέλτιστο αριθμό των ομάδων. Ο αριθμός αυτός προκύπτει έπειτα από διασταυρωμένη επικύρωση (cross validation) με κριτήριο το λογάριθμο πιθανότητας (*loglikelihood*). Επίσης επιλέγουμε το *Classes to clusters evaluation* το οποίο αφαιρεί το χαρακτηριστικό ΚΩΔΙΚΟΣ ΠΕΛΑΤΗ κατά την

ομαδοποίηση και τον επανεισάγει κατά την διάρκεια της αξιολόγησης του αποτελέσματος. Το αποτέλεσμα φαίνεται στην εικόνα που ακολουθεί.

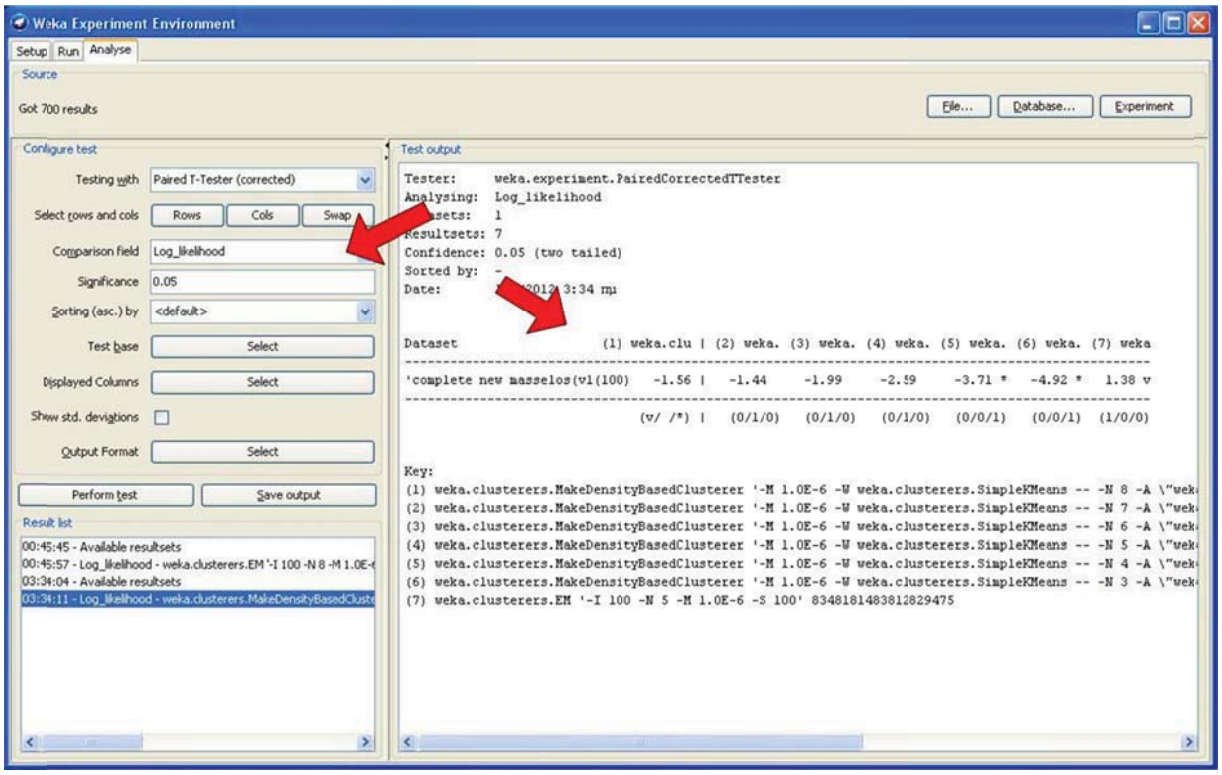


Πίνακας 11: ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΟΥ ΕΜ

Παρατηρούμε ότι ο αλγόριθμος επέλεξε ως βέλτιστη διαίρεση τις 5 ομάδες υποδειγμάτων. Επίσης μπορούμε να δούμε τους μέσους όρους και τις τυπικές αποκλίσεις των υποδειγμάτων κάθε χαρακτηριστικού για κάθε ομάδα. Γενικά το πόσες ομάδες πρέπει να δημιουργηθούν σε ένα πρόβλημα ομαδοποίησης είναι ένα ζήτημα που δεν έχει θεωρητικά τεκμηριωμένη απάντηση. Είναι επομένως ανοικτό σε παραδοχές και συχνά επιλέγεται μια λύση που να μπορεί εύκολα να γίνει αντιληπτός ο διαχωρισμός και τα όρια μεταξύ των ομάδων που προτείνει χωρίς να περιπλέκει μεγάλο αριθμό από αυτές. Υπάρχουν κάποια μέτρα που μπορούν σε γενικές γραμμές να αποτελέσουν μέτρα σύγκρισης (ο λογάριθμος πιθανότητας loglikelihood είναι ένα από αυτά) αλλά δεν είναι κριτήρια ικανά να αποδείξουν ότι ένας αριθμός ομάδων είναι καλύτερος από έναν άλλον.

Στην συνέχεια δοκιμάζουμε στον weka experimenter το μοντέλο ΕΜ των 5 ομάδων, που μας πρότεινε ο αλγόριθμος, απέναντι σε μοντέλα παραγόμενα από τον αλγόριθμο KMeans. Ο KMeans δεν έχει διαδικασία αυτόματης εύρεσης του

βέλτιστου αριθμού ομάδων οπότε θα πειραματιστούμε με διάφορες τιμές. Στο πείραμα μας τοποθετούμε τις τιμές {3,4,5,6,7,8}.



Πίνακας 12: ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ ΤΟΥ KMeans ΜΕ ΤΟΝ ΕΜ

Θέτουμε σαν κριτήριο σύγκρισης το λογάριθμο πιθανότητας. Όσο μεγαλύτερη η τιμή του, τόσο πιο ομοιογενή είναι τα σύνολα υποδειγμάτων που προκύπτουν. Βλέπουμε πως το μοντέλο του ΕΜ είναι αυτό που επιτυγχάνει την μεγαλύτερη τιμή του οπότε και είναι αυτό που θα χρησιμοποιήσουμε.

Επιστρέφουμε στον explorer και αφού φορτώσουμε το αρχείο που έχουμε με τα μοναδικά υποδείγματα για κάθε κωδικό, επιλέγουμε την καρτέλα Cluster. Εκεί επιλέγουμε τον ΕΜ. Ίδια διαδικασία με προηγουμένως λοιπόν μόνο που πλέον γνωρίζουμε πόσες ομάδες θέλουμε να δημιουργηθούν. Αλλάζουμε την αρχική τιμή του numClusters σε «5», επιλέγουμε το *Classes to clusters evaluation* για το χαρακτηριστικό *ΚΩΔΙΚΟΣ ΠΕΛΑΤΗ* και ξεκινάμε τον αλγόριθμο. Προκύπτουν οι παρακάτω ομάδες:

EM -1 (5) clusters.txt - Notepad

File Edit Format View Help

```
EM
==

Number of clusters selected by cross validation: 5
```

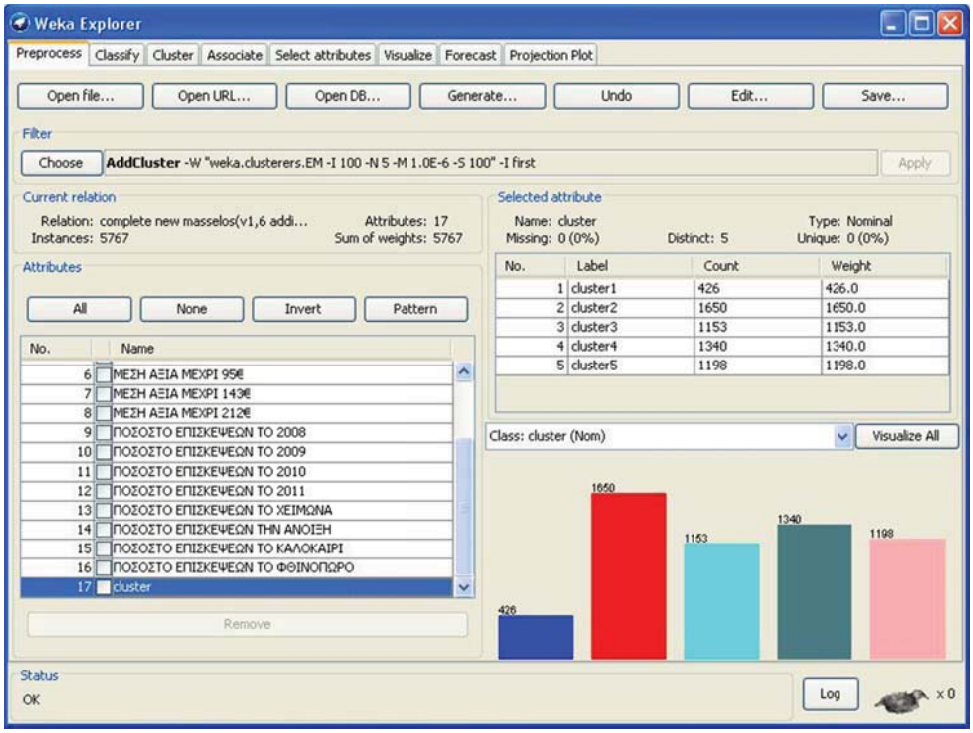
Attribute	Cluster 0 (0.15)	1 (0.31)	2 (0.15)	3 (0.2)	4 (0.2)
ΜΟΝΟ 1 ΕΠΙΣΚΕΨΗ					
mean	0	0.866	0.7491	0.9124	0.9727
std. dev.	0.433	0.3407	0.4335	0.2827	0.1629
ΜΕΧΡΙ 2 ΕΠΙΣΚΕΨΕΙΣ					
mean	0.2959	1	0.9972	1	0.9999
std. dev.	0.4565	0.3058	0.0528	0	0.0073
ΜΕΧΡΙ 5 ΕΠΙΣΚΕΨΕΙΣ					
mean	0.8426	1	1	1	1
std. dev.	0.3642	0.1507	0.1507	0.1507	0.1507
ΜΕΣΗ ΑΞΙΑ <68€					
mean	0.0981	0	0.7265	0.198	0.1773
std. dev.	0.2974	0.0001	0.4458	0.3985	0.3819
ΜΕΣΗ ΑΞΙΑ ΜΕΧΡΙ 95€					
mean	0.2149	0.1957	1	0.3996	0.3807
std. dev.	0.4108	0.3967	0.0001	0.4898	0.4856
ΜΕΣΗ ΑΞΙΑ ΜΕΧΡΙ 143€					
mean	0.5205	0.463	1	0.603	0.5681
std. dev.	0.4996	0.4986	0.4894	0.4893	0.4953
ΜΕΣΗ ΑΞΙΑ ΜΕΧΡΙ 212€					
mean	0.7782	0.7264	1	0.8072	0.7669
std. dev.	0.4155	0.4458	0.4	0.3945	0.4228
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008					
mean	0.1838	0	0.0753	0	1
std. dev.	0.2828	0.4081	0.2339	0.4081	0.0031
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009					
mean	0.3405	0.7155	0.6145	0	0
std. dev.	0.2945	0.4442	0.4639	0.4536	0.0019
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010					
mean	0.2936	0	0.0435	1	0
std. dev.	0.2652	0.4036	0.141	0.4036	0.002
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2011					
mean	0.1821	0.2845	0.2668	0	0
std. dev.	0.25	0.4442	0.4277	0.3381	0.0014
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ					
mean	0.2293	0.1116	0.2993	0.2695	0.2902
std. dev.	0.2699	0.3048	0.4377	0.4405	0.453
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΗΝ ΑΝΟΙΞΗ					
mean	0.2104	0.2601	0.2475	0.1731	0.2801
std. dev.	0.252	0.4298	0.4127	0.3751	0.4481
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΚΑΛΟΚΑΙΡΙ					
mean	0.3011	0.4327	0.2879	0.3795	0.2746
std. dev.	0.2991	0.4846	0.4293	0.4812	0.4457
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΦΘΙΝΟΠΩΡΟ					
mean	0.2592	0.1956	0.1653	0.1779	0.1551
std. dev.	0.2813	0.3845	0.3503	0.3784	0.3615

Πίνακας 13: ΤΙΜΕΣ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΤΩΝ ΟΜΑΔΩΝ ΠΟΥ ΠΡΟΚΥΠΤΟΥΝ ΑΠΟ ΤΟΝ ΕΜ

Επιχειρώντας μια πρώτη πρόχειρη ανάλυση μπορούμε γρήγορα να βγάλουμε κάποια αποτελέσματα για τις ομάδες.

- Η πρώτη ομάδα όπως φαίνεται από τα χαρακτηριστικά που αφορούν τις επισκέψεις, περιέχει πελάτες με πιο συχνές επισκέψεις σε σχέσεις με τις άλλες ομάδες και που κρίνοντας από το ποσοστό επισκέψεων ανά τα διάφορα έτη, συνεχίζουν και την επισκέπτονται. Μπορεί να χαρακτηριστεί ως η ομάδα των πιστών πελατών.
- Η πέμπτη ομάδα αποτελείται από πελάτες που όπως φαίνεται έχουν επισκέψεις καταγεγραμμένες το 2008 και, κρίνοντας και από την πολύ χαμηλή τυπική απόκλιση (η τιμή του std. dev. κάτω από το μέσο όρο), σχεδόν καμία τα πιο πρόσφατα έτη.
- Η δεύτερη ομάδα περιέχει τους πελάτες με την μεγαλύτερη μέση αξία αγορών. Αν και δεν έχουν τον αριθμό επισκέψεων της ομάδας 0, έχουν αρκετές αγορές μέσα στο 2011, το πιο πρόσφατο έτος στα στοιχεία μας και επίσης μεγάλο ποσοστό των αγορών τους έχει γίνει το καλοκαίρι, που είναι ο στόχος της λίστας αποστολής.

Χρησιμοποιούμε το *weka.filters.unsupervised.attribute.AddCluster* για να δημιουργήσουμε ένα νέο χαρακτηριστικό στο αρχείο μας που θα περιέχει την ομάδα στην οποία ανήκει κάθε υπόδειγμα, κάθε κωδικός πελάτη δηλαδή. Προσοχή χρειάζεται στο γεγονός ότι, ενώ κατά την ομαδοποίηση στην καρτέλα των clusters, η αρίθμηση των ομάδων ξεκινούσε από το 0, με τη χρήση των φίλτρων ξεκινά από το 1. Επίσης έχουμε και μια εικόνα για το ποσό των υποδειγμάτων που βρίσκεται σε κάθε ομάδα.



Πίνακας 14: ΣΥΝΟΛΟ ΥΠΟΔΕΙΓΜΑΤΩΝ ΠΟΥ ΑΝΗΚΟΥΝ ΣΕ ΚΑΘΕ ΟΜΑΔΑ ΤΟΥ ΕΜ

7.2. Ταξινόμηση

Πλέον έχουμε δημιουργήσει ένα χαρακτηριστικό, το *ΟΜΑΔΑ*, το οποίο μας δείχνει σε ποια ομάδα ανήκει το κάθε υπόδειγμα κατά συνέπεια ο κάθε *ΚΩΔΙΚΟΣ ΠΕΛΑΤΗ*. Να σημειώσουμε ότι κάποιοι κωδικοί πελατών δεν έχουν τιμή σε αυτό το χαρακτηριστικό γιατί είχαν εξαιρεθεί από τη διαδικασία της ομαδοποίησης για τους λόγους που έχουμε προαναφέρει. Αποτελεί επομένως το χαρακτηριστικό αυτό την τάξη του κάθε υποδείγματος και μας επιτρέπει να δοκιμάσουμε στα δεδομένα μας την ταξινόμηση (*classification*). Σκοπός μας είναι να μπορέσουμε να περιγράψουμε καλύτερα την δομή των σχηματισμένων ομάδων.

Πρώτα θα δοκιμάσουμε έναν απλό αλγόριθμο ταξινόμησης Naïve Bayes. Ο ταξινομητής Naïve Bayes γενικά παράγει υποθέσεις πιθανοτήτων αντί για προβλέψεις όπως κάνουν οι υπόλοιποι ταξινομητές. Αντί δηλαδή να προβλέπει σε ποια τάξη ανήκει κάθε υπόδειγμα, υπολογίζει την πιθανότητα που έχει το κάθε υπόδειγμα να ανήκει σε κάθε μια τάξη. Στην περίπτωσή μας χρησιμοποιείται για να δείξει τους μέσους όρους των υποδειγμάτων που ανήκουν στην κάθε ομάδα. Να σημειώσουμε ότι έχουμε χρησιμοποιήσει δύο ακόμα χαρακτηριστικά: το χαρακτηριστικό *ΚΥΚΛΩΜΑ*, που παράγεται από τα τρία πρώτα ψηφία του κωδικού πελάτη και μας δείχνει σε ποιό κατάστημα ανοίχτηκε ο κωδικός πελάτη, και το χαρακτηριστικό «*ΤΚ 3 πρώτα ψηφία*», που χρησιμοποιείται σε μια προσπάθεια να προσδιοριστεί γεωγραφικά η κατανομή των πελατών της κάθε ομάδας. Επίσης να σημειώσουμε ότι δεν χρησιμοποιούμε τα χαρακτηριστικά μας σε διακριτή μορφή αλλά αφήνουμε τον ταξινομητή να κάνει την διάκριση στο σημείο που αυτός

υπολογίζει. Τέλος να σημειωθεί ότι το ποσοστό των σωστών εκτιμήσεων του συγκεκριμένου ταξινομητή φτάνει σχεδόν το 86.5 %.

Αντιγράφουμε τα αποτελέσματα στον επεξεργαστή κειμένου για να είναι πιο εύκολη η παράθεσή τους εδώ. Αναλύοντας τα παρατηρούμε ότι:

- Στο χαρακτηριστικό *ΕΠΙΣΚΕΨΕΙΣ* φαίνεται ξεκάθαρα ότι η πρώτη ομάδα αποτελείται από πελάτες με μεγάλο αριθμό επισκέψεων, κάτι που είχε φανεί και από την ομαδοποίηση.
- Στο χαρακτηριστικό *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ* φαίνεται πως η τρίτη ομάδα περιέχει το χαμηλότερο μέσο όρο αξίας αγορών. Αυτό εξηγείται εν μέρει από τον μεγάλο συγκριτικά αριθμό επιστροφών που έχουν οι πελάτες της σε σχέση με αυτούς των άλλων ομάδων. Αντίθετα η δεύτερη ομάδα έχει τους πελάτες με την μεγαλύτερη μέση αξία επίσκεψης.
- Η τέταρτη ομάδα φαίνεται να έχει την μεγάλη πλειοψηφία των επισκέψεων της το 2010, ενώ η δεύτερη και η τρίτη το 2009. Παρόλα αυτά οι δύο τελευταίες έχουν σημαντικό ποσοστό επισκέψεων και το 2011.
- Σχεδόν στο σύνολό τους, οι πελάτες που ανήκουν στο κύκλωμα «201» και το «104», εμπεριέχονται στην πέμπτη ομάδα. Να σημειωθεί ότι οι κωδικοί του «201» που στην συνέχεια είχαν επισκεφτεί το κατάστημα στο Ψυχικό, «105», έχουν διορθωθεί σε «105» ώστε να υπάρχει ένας μοναδικός κωδικός για την επεξεργασία του συγκεκριμένου πελάτη.

NaiveBayes.txt - Notepad

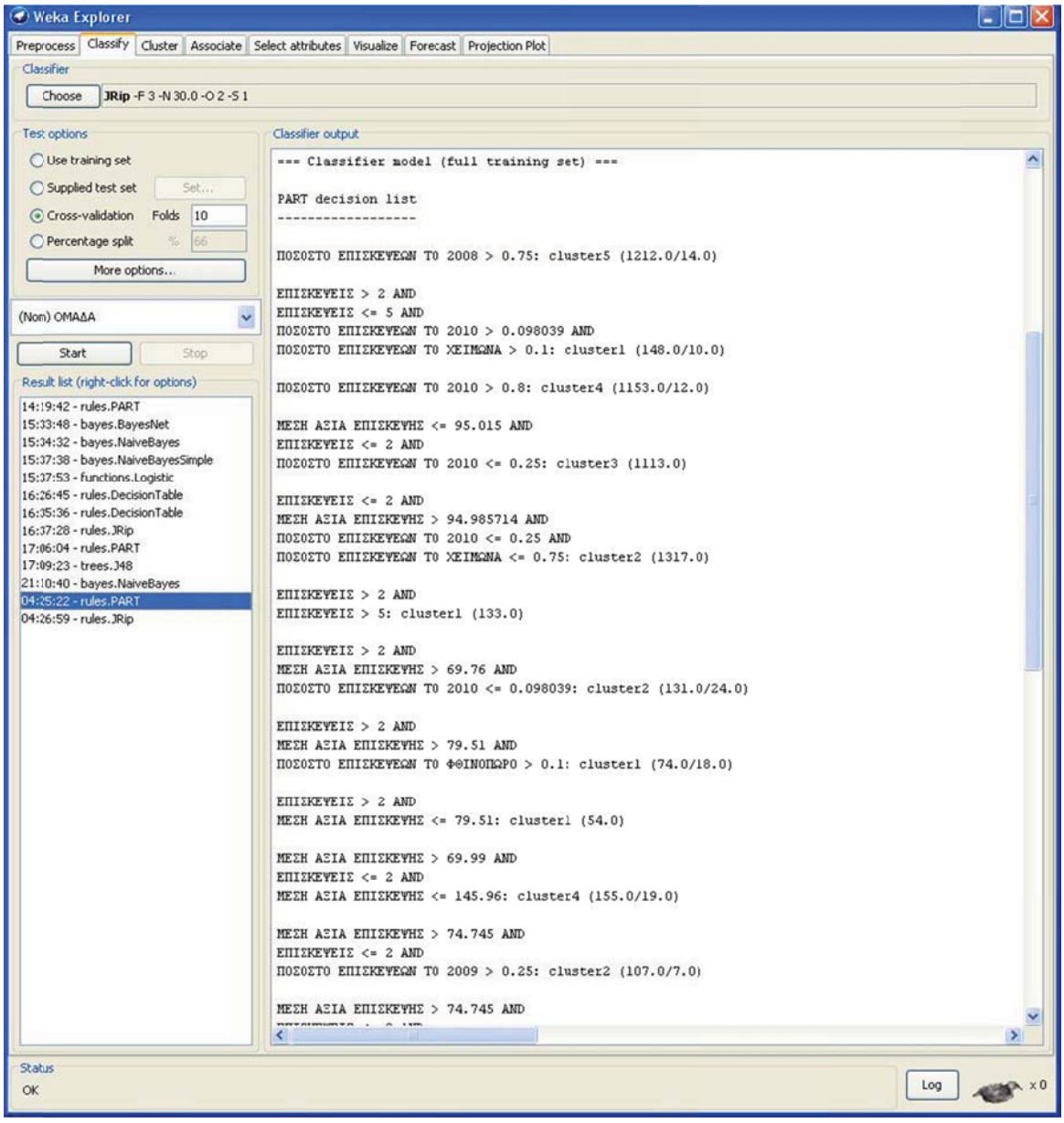
File Edit Format View Help

Naive Bayes Classifier

Attribute	Class cluster4 (0.23)	cluster2 (0.29)	cluster5 (0.21)	cluster3 (0.2)	cluster1 (0.07)
ΕΠΙΣΚΕΨΕΙΣ					
mean	0.505	1.006	0.1976	0.4737	6.2257
std. dev.	1.1302	1.4784	0.7487	1.1014	7.683
weight sum	1340	1650	1198	1153	426
precision	3.0345	3.0345	3.0345	3.0345	3.0345
ΚΥΚΛΩΜΑ					
105.0	673.0	659.0	1.0	449.0	207.0
103.0	666.0	961.0	985.0	699.0	218.0
201.0	2.0	6.0	152.0	1.0	1.0
104.0	1.0	3.0	64.0	1.0	4.0
310.0	1.0	10.0	1.0	1.0	1.0
11-	1.0	1.0	1.0	1.0	1.0
311.0	2.0	15.0	1.0	8.0	1.0
199.0	2.0	3.0	1.0	1.0	1.0
301.0	1.0	1.0	1.0	1.0	1.0
102.0	1.0	1.0	1.0	1.0	1.0
0.0	1.0	1.0	1.0	1.0	1.0
[total]	1351.0	1661.0	1209.0	1164.0	437.0
ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ					
mean	140.3037	211.7468	155.6529	44.736	132.6791
std. dev.	121.4383	122.9107	141.2627	56.0319	89.2362
weight sum	1340	1650	1198	1153	426
precision	0.5285	0.5285	0.5285	0.5285	0.5285
ΕΠΙΣΚΕΨΕΙΣ ΜΕ ΕΠΙΣΤΡΟΦΕΣ / ΕΠΙΣΚΕΨΕΙΣ					
mean	0.141	0.0429	0.1632	0.3148	0.1606
std. dev.	0.3344	0.1633	0.3624	0.4454	0.2051
weight sum	1340	1650	1198	1153	426
precision	0.0238	0.0238	0.0238	0.0238	0.0238
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008					
mean	0.0149	0.0379	1	0.0178	0.1095
std. dev.	0.0851	0.1418	0.0064	0.0926	0.1951
weight sum	1340	1650	1198	1153	426
precision	0.0385	0.0385	0.0385	0.0385	0.0385
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009					
mean	0.041	0.6594	0	0.7056	0.3277
std. dev.	0.1754	0.4357	0.0033	0.441	0.2666
weight sum	1340	1650	1198	1153	426
precision	0.02	0.02	0.02	0.02	0.02
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010					
mean	0.897	0.0341	0	0.0173	0.3538
std. dev.	0.2639	0.1318	0.0032	0.0915	0.2566
weight sum	1340	1650	1198	1153	426
precision	0.0192	0.0192	0.0192	0.0192	0.0192
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2011					
mean	0.047	0.2685	0	0.2593	0.2075
std. dev.	0.1942	0.4218	0.0035	0.4283	0.2379
weight sum	1340	1650	1198	1153	426
precision	0.0208	0.0208	0.0208	0.0208	0.0208
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ					
mean	0.3207	0.1004	0.2915	0.209	0.2566
std. dev.	0.4534	0.2653	0.4507	0.3944	0.2461
weight sum	1340	1650	1198	1153	426
precision	0.0213	0.0213	0.0213	0.0213	0.0213
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΗΝ ΑΝΟΙΞΗ					
mean	0.1623	0.271	0.276	0.2541	0.196
std. dev.	0.3579	0.4124	0.443	0.4245	0.2149
weight sum	1340	1650	1198	1153	426
precision	0.0204	0.0204	0.0204	0.0204	0.0204

Πίνακας 15: ΤΙΜΕΣ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΤΩΝ ΟΜΑΔΩΝ ΟΠΩΣ ΤΙΣ ΕΚΤΙΜΑ Ο Naive Bayes

Στη συνέχεια δοκιμάζουμε τον ταξινομητή JRip και τον PART. Και οι δύο είναι ταξινομητές που δημιουργούν λίστες κανόνων της μορφής IF…THEN…ELSE… για να περιγράψουν τη δομή των παραγόμενων μοντέλων. Η διαδικασία που ακολουθείται από τους αλγορίθμους αυτούς είναι επαναληπτική. Το σύνολο των δεδομένων ελέγχεται και παράγεται ο πρώτος κανόνας ο οποίος περιγράφει με συγκεκριμένη ελάχιστη κάλυψη και ακρίβεια ένα μέρος από τα δεδομένα. Έπειτα αυτό το κομμάτι αποκόπτεται και επαναλαμβάνεται η διαδικασία για το υπόλοιπο πλήθος των δεδομένων. Θέτουμε στους αλγορίθμους στην επιλογή της ελάχιστης κάλυψης τον αριθμό 30 (κάθε κανόνας δηλαδή να εκφράζει τουλάχιστον 30 υποδείγματα – πελάτες) ώστε να αποφύγουμε την υπερπροσαρμογή. Ο αλγόριθμος PART στο παραδειγμά μας επιτυγχάνει ποσοστό σωστά ταξινομημένων υποδειγμάτων 97%.



Πίνακας 16: ΚΑΝΟΝΕΣ ΠΟΥ ΠΡΟΚΥΠΤΟΥΝ ΑΠΟ ΤΟΝ ΤΑΞΙΝΟΜΗΤΗ PART

Οι πρώτοι από τους παραγόμενους κανόνες του PART είναι αυτοί που φαίνονται στην εικόνα. Οι αριθμοί στην παρένθεση δείχνουν πόσα υποδείγματα καλύπτει ο κανόνας, δηλαδή σε πόσα η υπόθεση του κανόνα ισχύει, και πόσα λάθος ταξινομημένα υποδείγματα προκύπτουν από τον κανόνα. Ο δεύτερος αριθμός αν υπάρχει χωρίζεται από τον πρώτο με το σύμβολο «/» αλλιώς παραλείπεται. Το AND ενώνει τις υποθέσεις του κανόνα ενώ το αποτέλεσμα του κανόνα βρίσκεται μετά το σύμβολο «:». Έτσι για παράδειγμα ο πρώτος κανόνας στην εικόνα εκφράζεται ως:

Αν το *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008* είναι μεγαλύτερο του 0,75 τότε το υπόδειγμα ανήκει στην πέμπτη ομάδα. Ο κανόνας ισχύει για 1212 υποδείγματα εκ των οποίων τα 14 είναι λάθος ταξινομημένα.

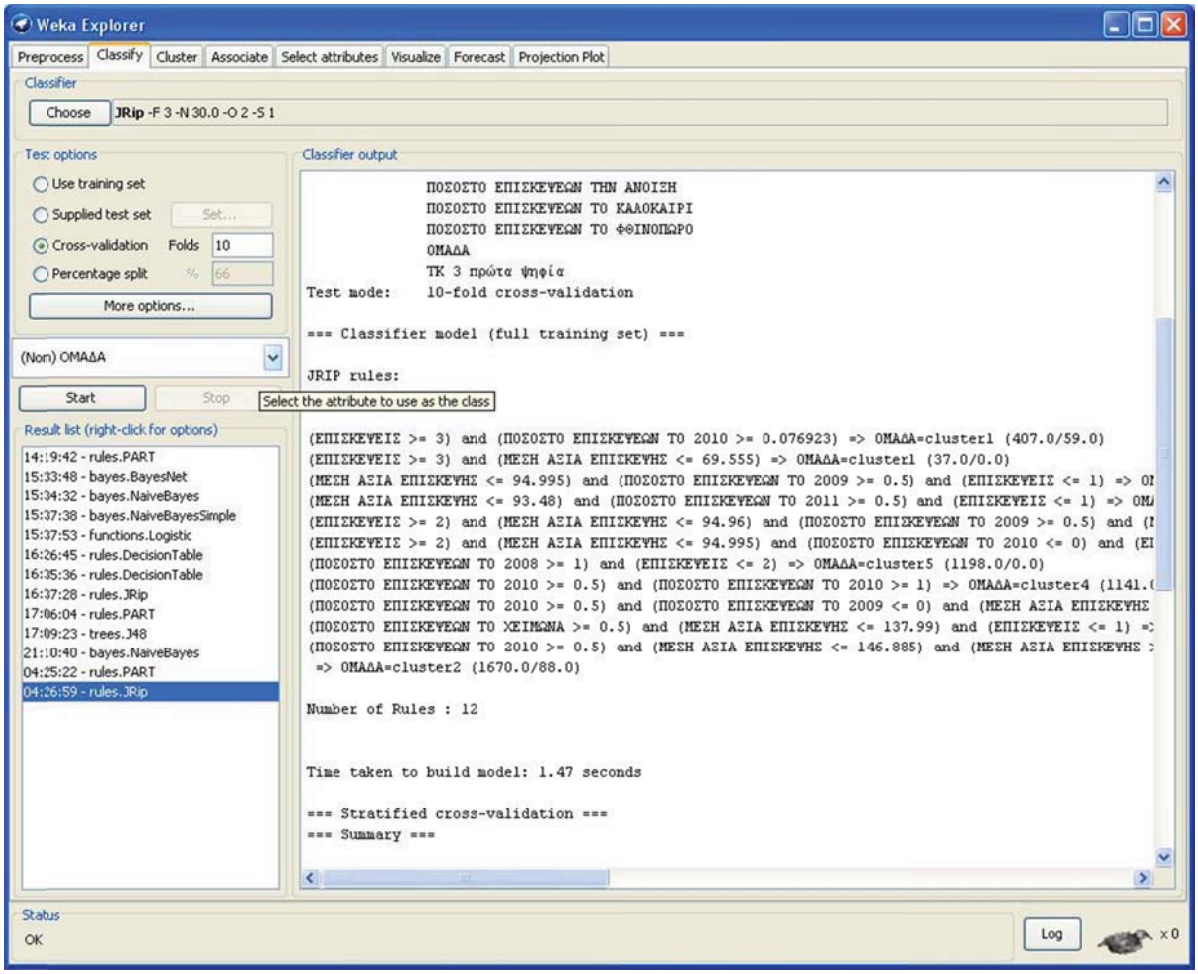
Στη συνέχεια παραθέτουμε τους κανόνες:

- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 > 0.75: cluster5 (1212.0/14.0)*
- *ΕΠΙΣΚΕΨΕΙΣ > 2 AND*
  ΕΠΙΣΚΕΨΕΙΣ <= 5 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 > 0.098039 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ > 0.1: cluster1 (148.0/10.0)
- *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 > 0.8: cluster4 (1153.0/12.0)*
- *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 95.015 AND*
  ΕΠΙΣΚΕΨΕΙΣ <= 2 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 <= 0.25: cluster3 (1113.0)
- *ΕΠΙΣΚΕΨΕΙΣ <= 2 AND*
  ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ > 94.985714 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 <= 0.25 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ <= 0.75: cluster2 (1317.0)
- *ΕΠΙΣΚΕΨΕΙΣ > 2 AND*
  ΕΠΙΣΚΕΨΕΙΣ > 5: cluster1 (133.0)
- *ΕΠΙΣΚΕΨΕΙΣ > 2 AND*
  ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ > 69.76 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 <= 0.098039: cluster2 (131.0/24.0)
- *ΕΠΙΣΚΕΨΕΙΣ > 2 AND*
  ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ > 79.51 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΦΘΙΝΟΠΩΡΟ > 0.1: cluster1 (74.0/18.0)
- *ΕΠΙΣΚΕΨΕΙΣ > 2 AND*
  ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 79.51: cluster1 (54.0)
- *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ > 69.99 AND*
  ΕΠΙΣΚΕΨΕΙΣ <= 2 AND
  ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 145.96: cluster4 (155.0/19.0)
- *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ > 74.745 AND*
  ΕΠΙΣΚΕΨΕΙΣ <= 2 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009 > 0.25: cluster2 (107.0/7.0)

- *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ > 74.745 AND*
  ΕΠΙΣΚΕΨΕΙΣ <= 2 AND
  ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 > 0.25: cluster4 (47.0/8.0)
- *ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ > 74.745: cluster2 (81.0/23.0)*
- *: cluster3 (42.0/5.0)*

Να τονιστεί όπως ήδη προαναφέρθηκε πως έπειτα από κάθε κανόνα, το πλήθος των δεδομένων που αυτός εκφράζει αφαιρείται από το σύνολο το δεδομένων και ο αλγόριθμος εξετάζει το υπόλοιπο πλήθος για νέους κανόνες. Έτσι για παράδειγμα μετά τον τελευταίο κανόνα, το πλήθος των δεδομένων που έχει μείνει ανατίθεται χωρίς κανόνα στην τρίτη ομάδα.

Οι παραγόμενοι κανόνες του JRip αν και είναι της ίδιας μορφής IF…THEN…ELSE…, έχουν μια μικρή διαφορά στον τρόπο που παρουσιάζονται από το weka. Η τοποθέτηση των κανόνων γίνεται σε μία σειρά. Ο αλγόριθμος αυτός επιτυγχάνει ποσοστό σωστά ταξινομημένων υποδειγμάτων 96%.

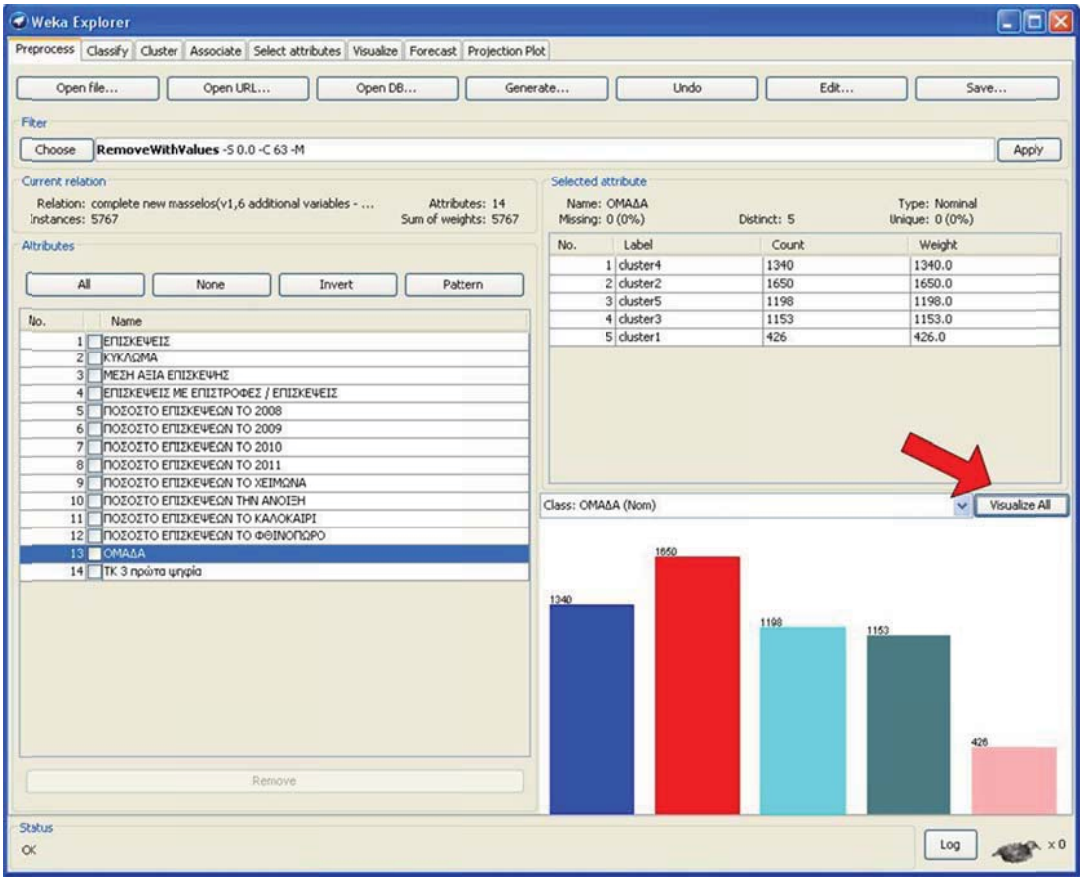


Πίνακας 17: ΚΑΝΟΝΕΣ ΠΟΥ ΠΡΟΚΥΠΤΟΥΝ ΑΠΟ ΤΟΝ ΤΑΞΙΝΟΜΗΤΗ JRip

Παραθέτουμε τους κανόνες:

- *(ΕΠΙΣΚΕΨΕΙΣ >= 3) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 >= 0.076923) => ΟΜΑΔΑ=cluster1 (407.0/59.0)*
- *(ΕΠΙΣΚΕΨΕΙΣ >= 3) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 69.555) => ΟΜΑΔΑ=cluster1 (37.0/0.0)*
- *(ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 94.995) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009 >= 0.5) and (ΕΠΙΣΚΕΨΕΙΣ <= 1) => ΟΜΑΔΑ=cluster3 (718.0/0.0)*
- *(ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 93.48) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2011 >= 0.5) and (ΕΠΙΣΚΕΨΕΙΣ <= 1) => ΟΜΑΔΑ=cluster3 (255.0/0.0)*
- *(ΕΠΙΣΚΕΨΕΙΣ >= 2) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 94.96) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009 >= 0.5) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 69.855) => ΟΜΑΔΑ=cluster3 (78.0/0.0)*
- *(ΕΠΙΣΚΕΨΕΙΣ >= 2) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 94.995) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 <= 0) and (ΕΠΙΣΚΕΨΕΙΣ <= 2) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 <= 0.5) => ΟΜΑΔΑ=cluster3 (81.0/0.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 >= 1) and (ΕΠΙΣΚΕΨΕΙΣ <= 2) => ΟΜΑΔΑ=cluster5 (1198.0/0.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 >= 0.5) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 >= 1) => ΟΜΑΔΑ=cluster4 (1141.0/0.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 >= 0.5) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009 <= 0) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ >= 72.14) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΗΝ ΑΝΟΙΞΗ <= 0) => ΟΜΑΔΑ=cluster4 (52.0/1.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ >= 0.5) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 137.99) and (ΕΠΙΣΚΕΨΕΙΣ <= 1) => ΟΜΑΔΑ=cluster4 (64.0/0.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 >= 0.5) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 146.885) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ >= 69.495) => ΟΜΑΔΑ=cluster4 (66.0/11.0)*
- *=> ΟΜΑΔΑ=cluster2 (1670.0/88.0)*

Τέλος χρησιμοποιούμε την δυνατότητα που μας παρέχει το weka για εποπτεία του συνόλου των δεδομένων μέσω της επιλογής Visualize All.

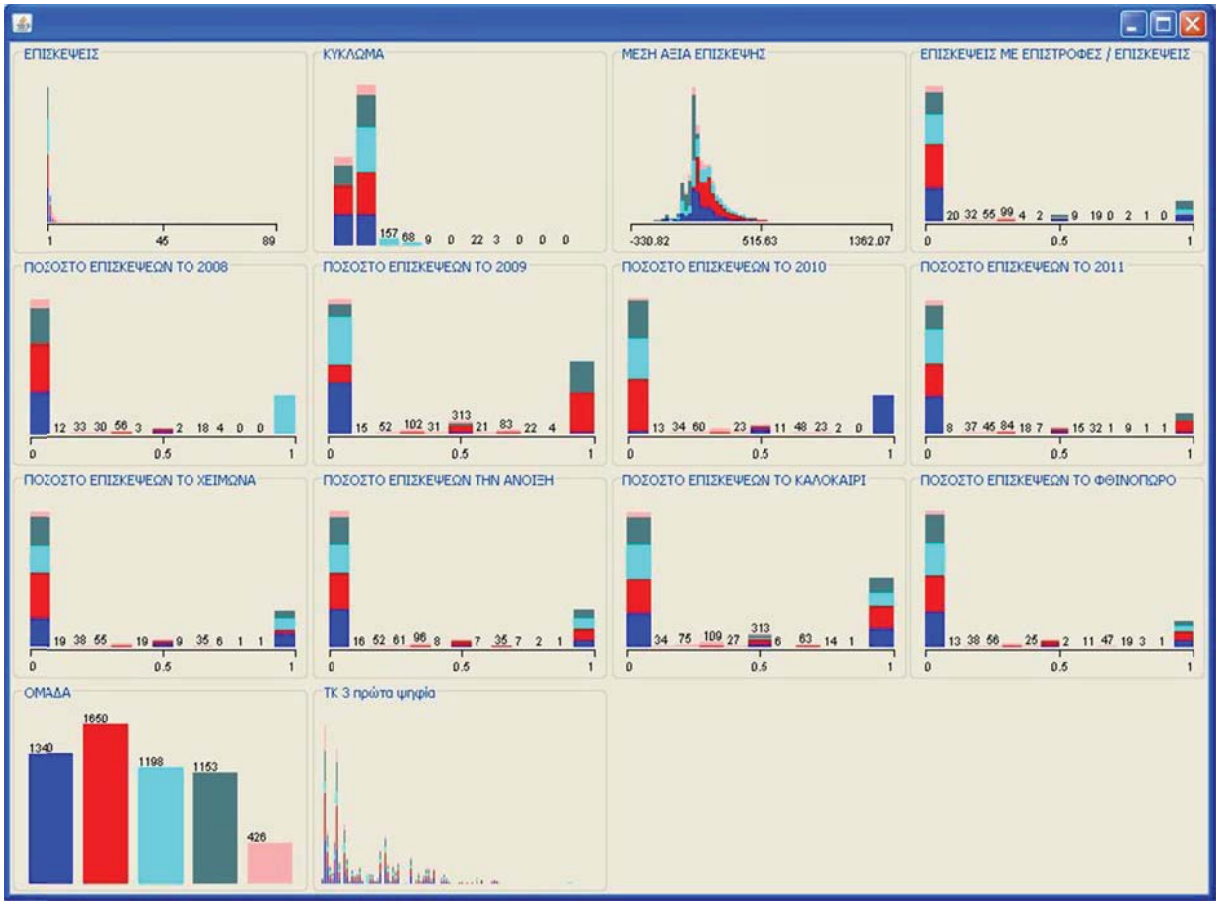


Πίνακας 18: ΟΠΤΙΚΟΠΟΙΗΣΗ ΜΕ ΔΙΑΦΟΡΕΤΙΚΟ ΧΡΩΜΑ ΤΩΝ ΠΑΡΑΓΟΜΕΝΩΝ ΟΜΑΔΩΝ

Η κάθε τάξη των υποδειγμάτων αναφέρεται με διαφορετικό χρώμα. Έτσι έχουμε:

- Ροζ για την πρώτη ομάδα (cluster1)
- Κόκκινο για την δεύτερη ομάδα
- Γκρι για την τρίτη ομάδα
- Βαθύ μπλε για την τέταρτη ομάδα
- Γαλάζιο για την πέμπτη ομάδα

Με τη χρωματική αυτή κωδικοποίηση είναι δυνατό να γίνει μια εξέταση και εξερεύνηση των δεδομένων. Φαίνεται άμεσα λοιπόν το μεγάλο ποσοστό της πέμπτης ομάδας για τις επισκέψεις το 2008 και αντίστοιχα της τέταρτης ομάδας για το 2010. Επίσης εύκολα διαπιστώνει κανείς μια συνολική αύξηση των επισκέψεων το καλοκαίρι σε σχέση με τις υπόλοιπες εποχές.



Πίνακας 19: ΟΠΤΙΚΟΠΟΙΗΣΗ ΣΤΟ ΣΥΝΟΛΟ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Από τους κανόνες επιβεβαιώνεται η εικόνα που ήδη έχουμε σχηματίσει για τον διαχωρισμό των ομάδων πελατών.

1. Η πρώτη ομάδα περιέχει τους πελάτες με τον μεγαλύτερο αριθμό επισκέψεων. Έχουν παρόμοια ποσοστά επισκέψεων για τα διάφορα έτη αλλά και τις εποχές και η μέση αξία αγορών τους κυμαίνεται σε μέτρια επίπεδα συγκριτικά με τις άλλες ομάδες. Μπορούν να θεωρηθούν ως οι «πιστοί» πελάτες της εταιρείας. (426 υποδείγματα – 7.4%)
2. Η δεύτερη ομάδα περιέχει και αυτή πελάτες με συγκριτικά αρκετές επισκέψεις σε σχέση με τους υπόλοιπους. Η μέση αξία επίσκεψης είναι η υψηλότερη που συναντάμε και το ποσοστό των επιστροφών που έχουν το χαμηλότερο. Οι επισκέψεις τους έγιναν κυρίως το 2009 και το 2011 και το ποσοστό επισκέψεών τους για το καλοκαίρι αγγίζει το 40% ενώ το ποσοστό τους για το χειμώνα είναι το χαμηλότερο από τις υπόλοιπες ομάδες. (1650 υποδείγματα – 28.61%)
3. Η τρίτη ομάδα έχει χαμηλό αριθμό επισκέψεων. Η μέση αξία επίσκεψης είναι η χαμηλότερη που συναντάμε και οι πελάτες αυτής της ομάδας έχουν αρκετές επιστροφές. Οι επισκέψεις τους έγιναν κατά ένα ποσοστό περίπου 70% το 2009 και γύρω στο 25% το 2011. (1153 υποδείγματα – 20%)

4. Η τέταρτη ομάδα έχει και αυτή χαμηλό αριθμό επισκέψεων κατά μέσο όρο. Η μέση αξία επίσκεψης είναι στο μέτρια συγκριτικά με τις υπόλοιπες ομάδες. Κύρια διαφορά της είναι το πολύ μεγάλο ποσοστό επισκέψεων το 2010, με ένα μέσο όρο σχεδόν 90%. Επίσης φαίνεται να έχει μια προτίμηση στις αγορές το χειμώνα και το καλοκαίρι. (1340 υποδείγματα – 23.2%)
5. Τέλος η πέμπτη ομάδα έχει και αυτή μικρό αριθμό επισκέψεων. Το ποσοστό επισκέψεων το 2008 αγγίζει το 100% και στα επόμενα έτη δεν υπάρχει σχεδόν καμία επίσκεψη. Πρόκειται για πελάτες που δείχνουν πως μάλλον έχουν χάσει το ενδιαφέρον τους για την εταιρεία. (1198 υποδείγματα – 20.77%)

Η πέμπτη ομάδα, και ειδικότερα η μη ύπαρξη εγγραφών μετά το 2008 για αυτούς του πελάτες, είναι ένα κομμάτι που απαιτεί λίγο περισσότερη διερεύνηση. Και αυτό γιατί έπειτα από διασταύρωση με την βάση δεδομένων της εταιρείας, αποτελεί ένα πολύ μεγάλο ποσοστό (περίπου το 80%) των καταγεγραμμένων επώνυμων αγορών του συγκεκριμένου έτους. Έχει ενδιαφέρον να εξετάσουμε αν υπήρξε μια γενικότερη αύξηση των ανώνυμων αγορών ή αν κάποιο υποκατάστημα είχε μια διαφορετική ποσότητα ως προς το μέσο όρο εισαγωγής επώνυμων και ανώνυμων εγγραφών. Δημιουργούμε λοιπόν τον παρακάτω πίνακα.

ΚΥΚΛΩΜΑ	ΕΤΟΣ ΑΓΟΡΩΝ							
	2008		2009		2010		2011	
	ΕΠΩΝΥΜΕΣ	ΑΝΩΝΥΜΕΣ	ΕΠΩΝΥΜΕΣ	ΑΝΩΝΥΜΕΣ	ΕΠΩΝΥΜΕΣ	ΑΝΩΝΥΜΕΣ	ΕΠΩΝΥΜΕΣ	ΑΝΩΝΥΜΕΣ
102	0	4564	0	3099	0	3400	0	2212
103	1437	0	2181	0	1123	0	903	0
104	88	0	3	0	1	0	0	0
105	75	0	1389	211	1278	0	779	0
11	1	0	0	0	0	0	0	0
199	0	0	1	0	1	0	2	0
201	177	1237	3	67	0	0	0	0
301	0	0	0	0	197	154	1543	1341
310	0	0	0	0	0	209	12	1860
311	0	0	0	2806	0	3240	24	0
ΣΥΝΟΛΟ	1778	5801	3577	6183	2600	7003	3263	5413

Πίνακας 20: ΚΑΤΑΓΡΑΦΗ ΕΠΩΝΥΜΩΝ ΚΑΙ ΑΝΩΝΥΜΩΝ ΑΓΟΡΩΝ ΑΝΑ ΕΤΟΣ ΚΑΙ ΚΥΚΛΩΜΑ ΚΑΤΑΓΡΑΦΗΣ

Ο πίνακας αυτός μας δείχνει το σύνολο των παραστατικών που κόπηκαν ανά κύκλωμα και ανά έτος καθώς και πόσα από αυτά περιείχαν επώνυμα στοιχεία του πελάτη και πόσα ανώνυμα. Στην πρώτη στήλη έχουμε τα κυκλώματα και έπειτα οι αγορές, τα αντίστοιχα παραστατικά, χωρίζονται στα τέσσερα έτη που μελετάμε και

σε επώνυμες και ανώνυμες αγορές. Τέλος στην κατώτερη γραμμή βρίσκουμε το σύνολο κάθε στήλης.

Παρατηρούμε πως το κύκλωμα 102 περιέχει μόνο ανώνυμα παραστατικά. Αυτός φαίνεται να είναι και ο λόγος δημιουργίας του αρχικά. Να καταγράφει τα ανώνυμα παραστατικά, όσα εισάγονταν για τον γενικό ανώνυμο πελάτη λιανικής. Αντίστοιχα το κύκλωμα 103 περιέχει μόνο επώνυμα παραστατικά, παραστατικά στα οποία ο πελάτης έχει δώσει το όνομά του. Επίσης μπορούμε να διακρίνουμε και άλλα στοιχεία όπως την διακοπή της χρήσης του κυκλώματος 201 μετά το έτος 2009 ή την μη χρήση των κυκλωμάτων 310 και 311 κατά το έτος 2008. Να σημειωθεί πως το κύκλωμα 11 υφίσταται για εσωτερική χρήση στην εταιρεία.

Ένα αξιοπρόσεκτο στοιχείο είναι η μείωση των επώνυμων παραστατικών του κυκλώματος 103 από το έτος 2009 στο έτος 2010, χωρίς να υπάρχει ανάλογη μείωση για το κύκλωμα 105. Τα δύο αυτά κυκλώματα είναι τα κύρια της εταιρείας και λογικά θα περίμενε κανείς να έχουν παρόμοια συμπεριφορά. Ωστόσο ενώ οι επώνυμες εγγραφές του κυκλώματος 103 μειώνονται κατά σχεδόν 50% στα δύο αυτά έτη, οι αντίστοιχες του 105 μειώνονται μόνο κατά 8%.

Στη συνέχεια προσπαθούμε να εξάγουμε κάποιες πληροφορίες σχετικά με την κατανομή των πελατών ως προς την εμπορική κατηγορία των προϊόντων τα οποία αγόρασαν. Για αυτή την ανάλυση δημιουργούμε τέσσερα βοηθητικά χαρακτηριστικά αντίστοιχα αυτών που έχουμε δημιουργήσει σε προηγούμενες περιπτώσεις:

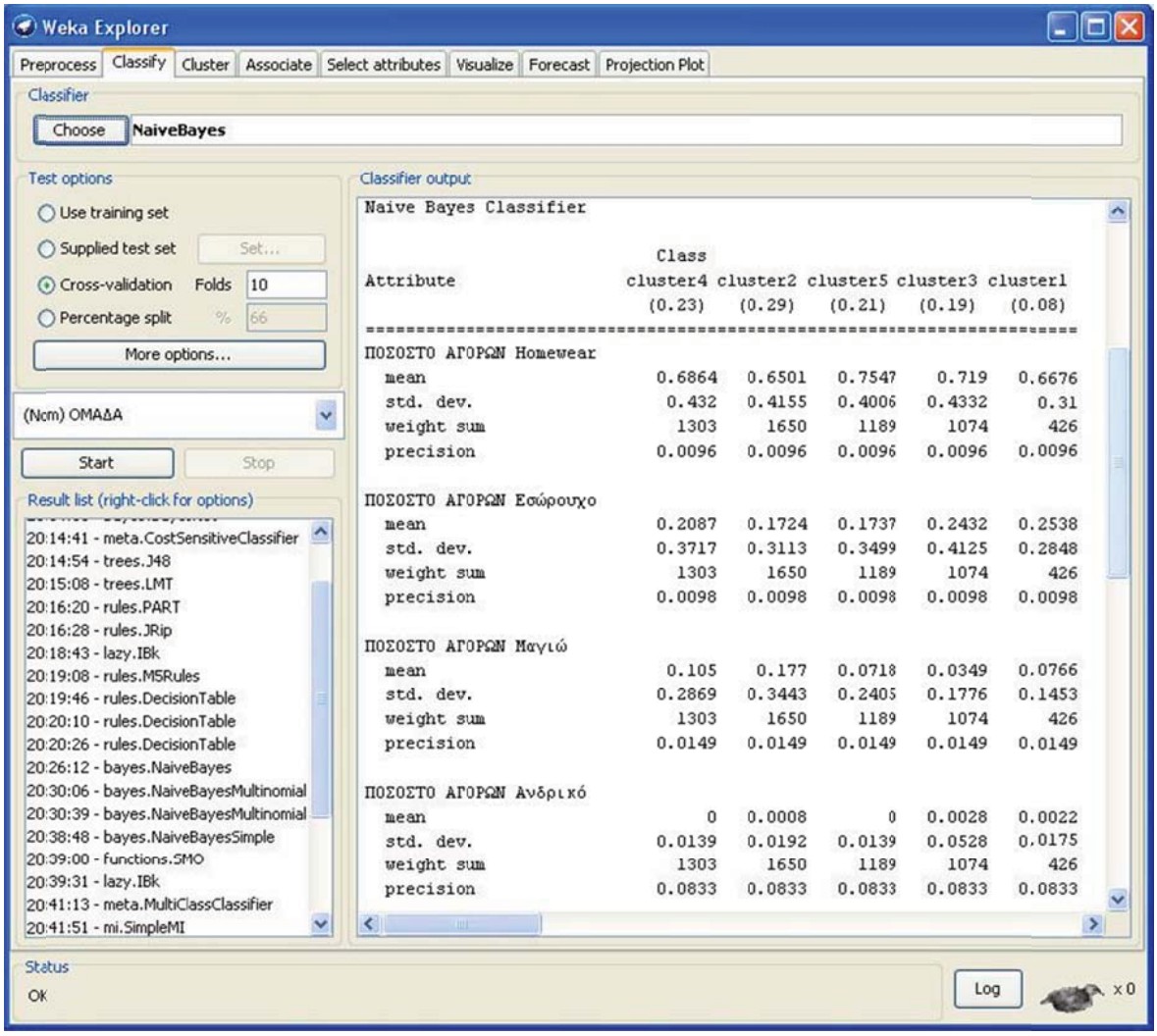
- *ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Homewear*
- *ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Εσώρουχο*
- *ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Μαγιό*
- *ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Ανδρικό*

Όπως και σε αντίστοιχες προηγούμενες περιπτώσεις, τα χαρακτηριστικά αυτά δείχνουν το ποσοστό των αγορών του πελάτη που αντιστοιχούν στην αναφερόμενη εμπορική κατηγορία. Η κατηγοριοποίηση αυτή είναι αυτή που χρησιμοποιεί η ίδια η εταιρεία για τα προϊόντα της. Για την δημιουργία αυτών των χαρακτηριστικών χρησιμοποιούμε το excel. Ωστόσο υπάρχουν 125 περιπτώσεις πελατών, για τους οποίους έχουν καταγραφεί επισκέψεις που αφορούν μόνο επιστροφές και όχι αγορές, με αποτέλεσμα να προκύπτει σφάλμα στην τιμή του αντίστοιχου χαρακτηριστικού. Οι περιπτώσεις αυτές εξαιρούνται προσωρινά για την συγκεκριμένη ανάλυση.

Χρησιμοποιούμε αρχικά στο weka τον Naïve Bayes από τους ταξινομητές. Σκοπός μας είναι να πάρουμε μια εικόνα της κατανομής των διαφόρων ανωτέρω ποσοστών ανά τις δημιουργηθείσες ομάδες. Προκύπτει το παρακάτω μοντέλο το οποίο

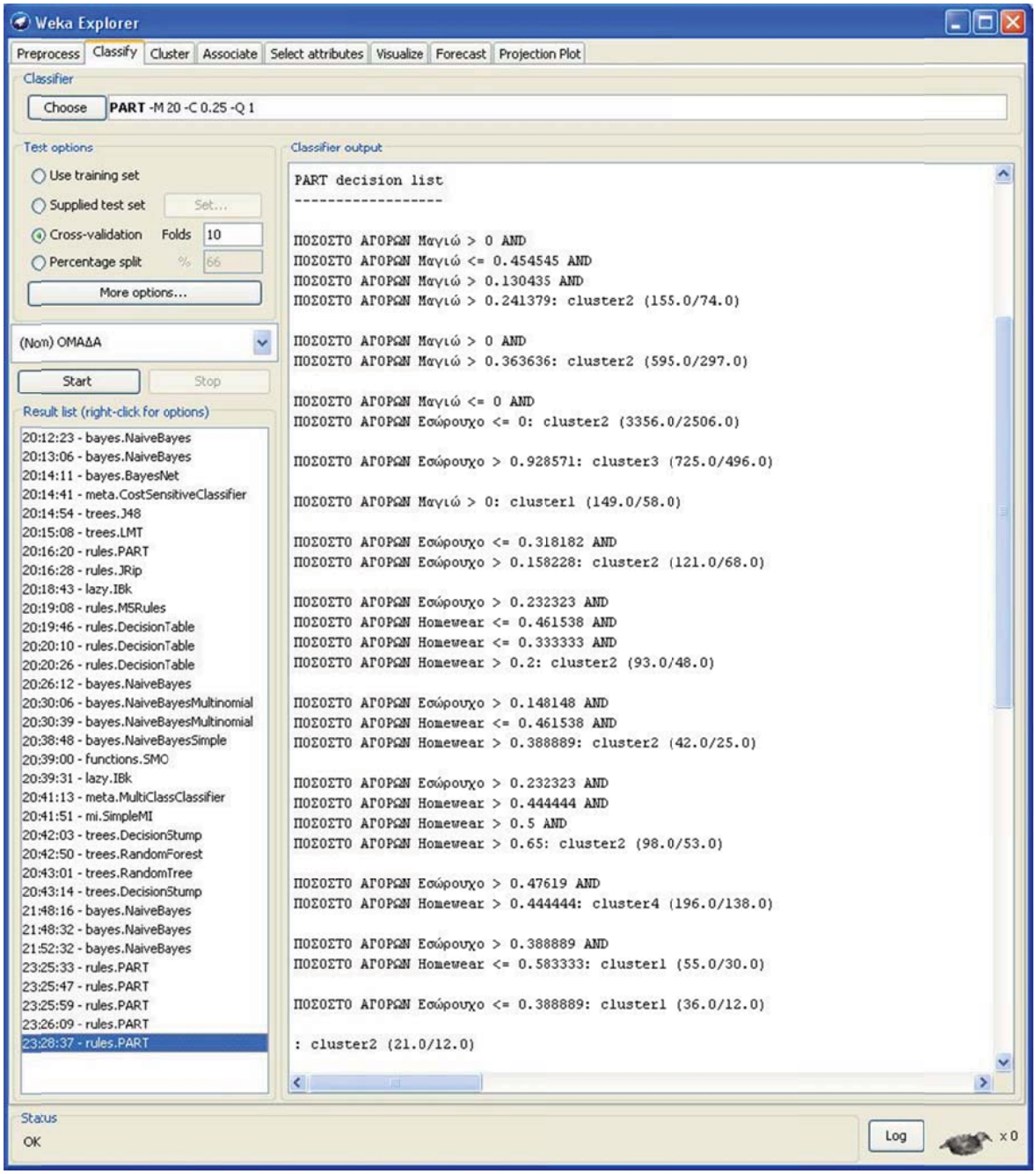
χαρακτηρίζεται από χαμηλή όμως ακρίβεια με ποσοστό επιτυχημένων ταξινομήσεων που αντιστοιχεί σε 27.56%. Αυτό σημαίνει πως το μοντέλο αυτό δεν μπορεί να μας δώσει ασφαλή πρόβλεψη ταξινόμησης με βάση τα χαρακτηριστικά αυτά ωστόσο μπορεί να μας δώσει περιγραφή της δομής των χαρακτηριστικών. Το μοντέλο μας δίνει μια τιμή του μέσου όρου και της τυπικής απόκλισης που έχουν τα παραπάνω χαρακτηριστικά ανά ομάδα. Άμεσα μπορεί κάποιος να εξάγει τα εξής συμπεράσματα:

- Η ομάδα των «πιστών» πελατών είναι αυτή με το μεγαλύτερο μέσο όρο σε ποσοστό αγορών σε εσώρουχα και ταυτόχρονα την μικρότερη τυπική απόκλιση.
- Η δεύτερη ομάδα έχει αισθητά μεγαλύτερο ποσοστό αγορών από τις υπόλοιπες σε ότι αφορά τα προϊόντα της κατηγορίας μαγιό. Κάτι που συνάδει άμεσα με την μεγάλη συχνότητα επισκέψεων που έχει η συγκεκριμένη ομάδα την περίοδο του καλοκαιριού.
- Η πέμπτη ομάδα πελατών είναι πρώτη στο μέσο όρο σε ποσοστό αγορών Homewear και από τις τελευταίες σε ποσοστό αγορών της κατηγορίας εσώρουχο.



Πίνακας 21: ΑΠΟΤΕΛΕΣΜΑΤΑ Naive Bayes ΣΧΕΤΙΚΑ ΜΕ ΤΗΝ ΚΑΤΗΓΟΡΙΑ ΠΡΟΪΟΝΤΩΝ

Δοκιμάζουμε στην συνέχεια και τον αλγόριθμο PART καθώς και τον JRip για την προβολή κανόνων, με τις επιτυχημένες ωστόσο ταξινομήσεις να παραμένουν στα ποσοστά της τάξης του 30%. Παραθέτουμε τα αποτελέσματα του PART, ωστόσο οι κανόνες που εξάγονται δεν φαίνεται να έχουν αρκετά μεγάλη ακρίβεια και αναλογία σωστών προς λανθασμένες προβλέψεις για να μπορούν να χρησιμοποιηθούν ασφαλώς για την εξαγωγή συμπερασμάτων.



Πίνακας 22: ΑΠΟΤΕΛΕΣΜΑΤΑ PART ΣΧΕΤΙΚΑ ΜΕ ΤΗΝ ΚΑΤΗΓΟΡΙΑ ΠΡΟΪΟΝΤΩΝ

Μια ακόμα ενδιαφέρουσα ανάλυση αφορά τα κυκλώματα. Ταξινόμηση δηλαδή όχι με βάση τις ομάδες πελατών που έχουμε δημιουργήσει αλλά με βάση τα κυκλώματα, τα οποία αντιστοιχούν στα διάφορα υποκαταστήματα της εταιρείας. Χρησιμοποιούμε την καρτέλα με την προεπεξεργασία του weka για να ρυθμίσουμε ποια χαρακτηριστικά χρειαζόμαστε για την συγκεκριμένη ανάλυση. Αφαιρούμε τα υποδείγματα που αφορούν τα κυκλώματα 11 και 199 καθώς περιέχουν πολύ μικρό αριθμό υποδειγμάτων και δεν μπορούν να μας προσφέρουν κάποια πληροφορία. Αντιθέτως με την αφαίρεση τους βελτιώνεται η ακρίβεια του μοντέλου. Κύριο σημείο ενδιαφέροντος είναι τυχόν διαφορές μεταξύ των κυκλωμάτων 103 και 105, αλλά και ο εντοπισμός σημαντικών στοιχείων και για τα υπόλοιπα κυκλώματα.

Χρησιμοποιούμε και εδώ τον Naïve Bayes. Το μοντέλο μας επιτυγχάνει ένα ποσοστό επιτυχημένων προβλέψεων της τάξης του 46%. Ωστόσο όπως και σε προηγούμενες αναλύσεις σκοπός μας είναι περισσότερο η περιγραφή της δομής των δεδομένων και λιγότερο η ακρίβεια του μοντέλου. Θα μπορούσαμε να δοκιμάσουμε πολυπλοκότερα μοντέλα και αλγορίθμους που θα επιτύγχαναν καλύτερη ακρίβεια ωστόσο δεν θα είχαν την απλότητα και την ευκρίνεια των δομών που προσφέρει ο συγκεκριμένος αλγόριθμος.

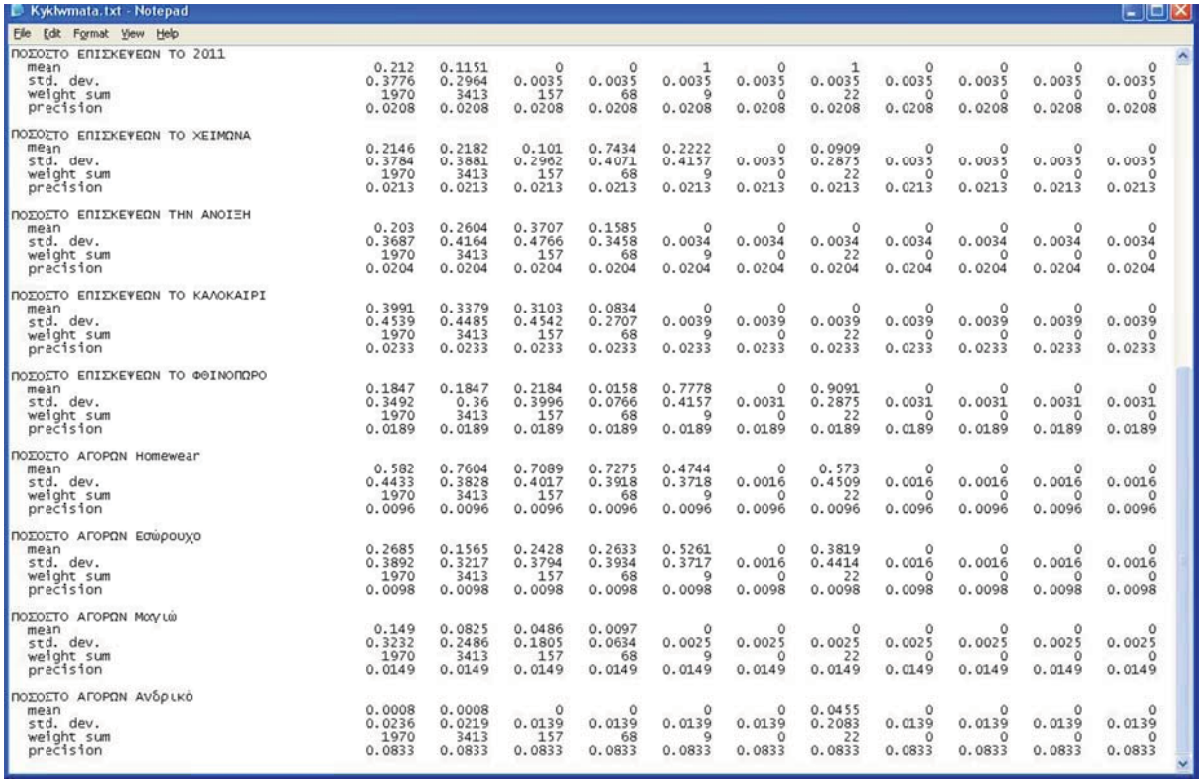
Kykloma.txt - Notepad

File Edit Format View Help

Naive Bayes Classifier

Attribute	Class 105.0 (0.35)	103.0 (0.6)	201.0 (0.03)	104.0 (0.01)	310.0 (0)	11- (0)	311.0 (0)	199.0 (0)	301.0 (0)	102.0 (0)	0.0 (0)
ΕΠΙΣΚΕΨΕΙΣ											
mean	1.2292	0.9469	0.3479	0.6694	0.6743	0	0.2759	0	0	0	0
std. dev.	2.3293	3.1692	0.9668	1.2582	1.2616	0.5057	0.8724	0.5057	0.5057	0.5057	0.5057
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	3.0345	3.0345	3.0345	3.0345	3.0345	3.0345	3.0345	3.0345	3.0345	3.0345	3.0345
ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ											
mean	146.4586	147.7915	220.0661	172.2044	316.3613	0	182.3153	0	0	0	0
std. dev.	116.3128	123.2092	167.4956	133.6067	128.6006	0.084	107.7105	0.084	0.084	0.084	0.084
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.5038	0.5038	0.5038	0.5038	0.5038	0.5038	0.5038	0.5038	0.5038	0.5038	0.5038
ΕΠΙΣΚΕΨΕΙΣ ΜΕ ΕΠΙΣΤΡΟΦΕΣ / ΕΠΙΣΚΕΨΕΙΣ											
mean	0.1005	0.1638	0.0032	0.0648	0	0	0	0	0	0	0
std. dev.	0.2679	0.3441	0.0398	0.1839	0.004	0.004	0.004	0.004	0.004	0.004	0.004
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0238	0.0238	0.0238	0.0238	0.0238	0.0238	0.0238	0.0238	0.0238	0.0238	0.0238
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008											
mean	0.0114	0.321	0.9809	0.9819	0	0	0	0	0	0	0
std. dev.	0.0681	0.448	0.1369	0.0922	0.0064	0.0064	0.0064	0.0064	0.0064	0.0064	0.0064
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0385	0.0385	0.0385	0.0385	0.0385	0.0385	0.0385	0.0385	0.0385	0.0385	0.0385
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009											
mean	0.4087	0.3686	0.0191	0.0147	0	0	0	0	0	0	0
std. dev.	0.4564	0.457	0.1369	0.0888	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010											
mean	0.3677	0.1953	0	0.0037	0	0	0	0	0	0	0
std. dev.	0.4428	0.3721	0.0032	0.0301	0.0032	0.0032	0.0032	0.0032	0.0032	0.0032	0.0032
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0192	0.0192	0.0192	0.0192	0.0192	0.0192	0.0192	0.0192	0.0192	0.0192	0.0192
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2011											
mean	0.212	0.1151	0	0	1	0	1	0	0	0	0
std. dev.	0.3776	0.2964	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208

Πίνακας 23: ΑΠΟΤΕΛΕΣΜΑΤΑ Naive Bayes ΣΧΕΤΙΚΑ ΜΕ ΤΑ ΚΥΚΛΩΜΑΤΑ ΚΑΤΑΓΡΑΦΗΣ (1)



Kykloumata.txt - Notepad

File Edit Format View Help

ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2011											
mean	0.212	0.1151	0	0	1	0	1	0	0	0	0
std. dev.	0.3776	0.2964	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ											
mean	0.2146	0.2182	0.101	0.7434	0.2222	0	0.0909	0	0	0	0
std. dev.	0.3784	0.3881	0.2962	0.4071	0.4157	0.0035	0.2875	0.0035	0.0035	0.0035	0.0035
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0213	0.0213	0.0213	0.0213	0.0213	0.0213	0.0213	0.0213	0.0213	0.0213	0.0213
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΗΝ ΑΝΟΙΞΗ											
mean	0.203	0.2604	0.3707	0.1585	0	0	0	0	0	0	0
std. dev.	0.3687	0.4164	0.4766	0.3458	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΚΑΛΟΚΑΙΡΙ											
mean	0.3991	0.3379	0.3103	0.0834	0	0	0	0	0	0	0
std. dev.	0.4539	0.4485	0.4542	0.2707	0.0039	0.0039	0.0039	0.0039	0.0039	0.0039	0.0039
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0233	0.0233	0.0233	0.0233	0.0233	0.0233	0.0233	0.0233	0.0233	0.0233	0.0233
ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΦΘΙΝΟΠΩΡΟ											
mean	0.1847	0.1847	0.2184	0.0158	0.7778	0	0.9091	0	0	0	0
std. dev.	0.3492	0.36	0.3996	0.0766	0.4157	0.0031	0.2875	0.0031	0.0031	0.0031	0.0031
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0189	0.0189	0.0189	0.0189	0.0189	0.0189	0.0189	0.0189	0.0189	0.0189	0.0189
ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Homewear											
mean	0.582	0.7604	0.7089	0.7275	0.4744	0	0.573	0	0	0	0
std. dev.	0.4433	0.3828	0.4017	0.3918	0.3718	0.0016	0.4509	0.0016	0.0016	0.0016	0.0016
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0096	0.0096	0.0096	0.0096	0.0096	0.0096	0.0096	0.0096	0.0096	0.0096	0.0096
ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Εσώρουχο											
mean	0.2685	0.1565	0.2428	0.2633	0.5261	0	0.3819	0	0	0	0
std. dev.	0.3892	0.3217	0.3794	0.3934	0.3717	0.0016	0.4414	0.0016	0.0016	0.0016	0.0016
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0098	0.0098	0.0098	0.0098	0.0098	0.0098	0.0098	0.0098	0.0098	0.0098	0.0098
ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Μαγιώ											
mean	0.149	0.0825	0.0486	0.0097	0	0	0	0	0	0	0
std. dev.	0.3232	0.2486	0.1805	0.0634	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0149	0.0149	0.0149	0.0149	0.0149	0.0149	0.0149	0.0149	0.0149	0.0149	0.0149
ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Ανδρικό											
mean	0.0008	0.0008	0	0	0	0	0.0455	0	0	0	0
std. dev.	0.0236	0.0219	0.0139	0.0139	0.0139	0.0139	0.2083	0.0139	0.0139	0.0139	0.0139
weight sum	1970	3413	157	68	9	0	22	0	0	0	0
precision	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833

Πίνακας 24: ΑΠΟΤΕΛΕΣΜΑΤΑ Naive Bayes ΣΧΕΤΙΚΑ ΜΕ ΤΑ ΚΥΚΛΩΜΑΤΑ ΚΑΤΑΓΡΑΦΗΣ (2)

Βλέπουμε ότι τα κυκλώματα «11», «199», «301», «102» και «0» δεν περιέχουν υποδείγματα, απλά τοποθετήθηκαν για λόγους πληρότητας. Η προσοχή μας εστιάζεται στα κυκλώματα «103» και «105» καθώς περιέχουν συντριπτικά περισσότερα υποδείγματα σε σχέση με τα υπόλοιπα. Ωστόσο κάποια εντυπωσιακά στοιχεία των υπόλοιπων κυκλωμάτων τονίζονται και αυτά.

Κάποια ενδιαφέροντα στοιχεία που μπορεί κάποιος άμεσα να παρατηρήσει:

- Οι πελάτες του κυκλώματος 105 έχουν ελαφρώς περισσότερες επισκέψεις κατά μέσο όρο από αυτούς του 103.
- Αν και φαίνεται να υπάρχει μια σχέση μεταξύ των υποδειγμάτων που περιέχει ένα κύκλωμα και της *ΜΕΣΗΣ ΑΞΙΑΣ ΕΠΙΣΚΕΨΗΣ*, το κύκλωμα 201 έχει αρκετά μεγάλη τιμή στο συγκεκριμένο χαρακτηριστικό. Πιθανότατα σχετίζεται με την πολύ μικρή τιμή του *ΕΠΙΣΚΕΨΕΙΣ ΜΕ ΕΠΙΣΤΡΟΦΕΣ / ΕΠΙΣΚΕΨΕΙΣ*. Το κύκλωμα 310 φαίνεται να έχει και αυτό πολύ μέση αξία επίσκεψης ωστόσο αναφέρεται σε μόνο 9 υποδείγματα.
- Το κύκλωμα 105 φαίνεται να έχει μικρότερη τιμή στο *ΕΠΙΣΚΕΨΕΙΣ ΜΕ ΕΠΙΣΤΡΟΦΕΣ / ΕΠΙΣΚΕΨΕΙΣ* από το 103
- Το κύκλωμα 104 παρότι αφορά μικρό αριθμό πελατών, έχει εντυπωσιακά υψηλή τιμή κατά μέσο όρο στο χαρακτηριστικό *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ*

- Το κύκλωμα 105 περιέχει πελάτες με μεγαλύτερο *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΚΑΛΟΚΑΙΡΙ* σε σχέση με το 103, το οποίο με τη σειρά του υπερέχει στο *ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΗΝ ΑΝΟΙΞΗ*
- Το κύκλωμα 103 έχει υποδείγματα με πελάτες με αισθητά μεγαλύτερο *ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Homewear* σε σχέση με το 105, το οποίο όμως υπερέχει σε *ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Εσώρουχο* και *ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Μαγιό*.

Δοκιμάζουμε και τον ταξινομητή JRip για να εξάγουμε κάποιους ποσοτικούς κανόνες σχετικά με την δομή των χαρακτηριστικών ανά κύκλωμα. Θέτουμε ως ελάχιστο όριο κάλυψης στους κανόνες τα 20 υποδείγματα και προκύπτουν οι εξής 7 κανόνες:

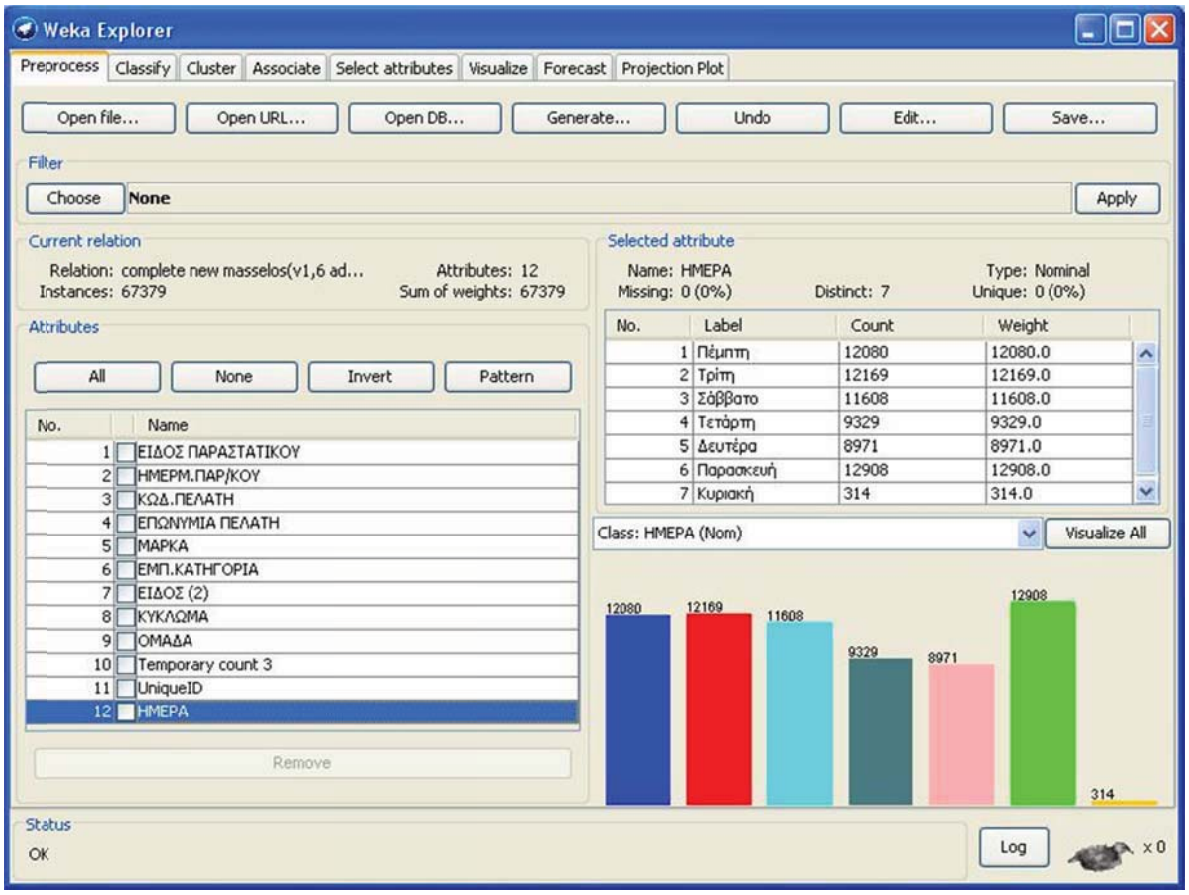
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 <= 0.055556) and (ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Homewear <= 0.425) and (ΕΠΙΣΚΕΨΕΙΣ >= 2) => ΚΥΚΛΩΜΑ=105.0 (281.0/65.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 <= 0) and (ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Homewear <= 0.571429) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 172.105) => ΚΥΚΛΩΜΑ=105.0 (741.0/298.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 <= 0.083333) and (ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Homewear <= 0.6) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 215.8675) => ΚΥΚΛΩΜΑ=105.0 (162.0/74.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 <= 0.25) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009 <= 0.375) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ ΧΕΙΜΩΝΑ >= 0.25) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 <= 0.333333) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ >= 72.57) => ΚΥΚΛΩΜΑ=105.0 (130.0/47.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 <= 0.055556) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2009 <= 0.833333) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 154.015) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ >= 43.43) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 78.53) => ΚΥΚΛΩΜΑ=105.0 (141.0/52.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 <= 0.375) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 >= 0.333333) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 186.34) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ >= 125.406667) and (ΜΕΣΗ ΑΞΙΑ ΕΠΙΣΚΕΨΗΣ <= 153.13) => ΚΥΚΛΩΜΑ=105.0 (77.0/26.0)*
- *(ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2008 <= 0.055556) and (ΠΟΣΟΣΤΟ ΕΠΙΣΚΕΨΕΩΝ ΤΟ 2010 >= 0.333333) and (ΠΟΣΟΣΤΟ ΑΓΟΡΩΝ Homewear <= 0.888889) => ΚΥΚΛΩΜΑ=105.0 (190.0/84.0)*

7.3. Συσχέτιση

Στη συνέχεια επιχειρούμε να εφαρμόσουμε κάποια μέθοδο συσχέτισης για να βρούμε προϊόντα ή άλλα χαρακτηριστικά τα οποία ενδέχεται να συσχετίζονται μεταξύ τους. Σκοπός μας είναι να κάνουμε μια «ανάλυση του καλαθιού αγοράς» (*market basket analysis*). Αποτέλεσμα αυτής της ανάλυσης είναι η εύρεση προϊόντων τα οποία ενδέχεται να συνδυάζονται συχνά στις καταγεγραμμένες πωλήσεις ή η εύρεση σχέσεων μεταξύ προϊόντων και άλλων χαρακτηριστικών.

Χρησιμοποιούμε το αρχικό μας αρχείο που περιείχε όλες τις εγγραφές από τις επισκέψεις αναλυτικά. Μας ενδιαφέρει το περιεχόμενο των πωλήσεων που έγιναν σε όλους τους πελάτες, ανεξαρτήτως του αν είχαν δώσει τα στοιχεία τους ή όχι. Ωστόσο απαιτεί κάποιους μετασχηματισμούς για να το φέρουμε σε μορφή κατάλληλη για επεξεργασία.

Αρχικά δημιουργούμε στο excel ένα επιπλέον χαρακτηριστικό, το χαρακτηριστικό *ΗΜΕΡΑ*. Αυτό μας αναγράφει την ημέρα της εβδομάδος που προκύπτει από το χαρακτηριστικό *ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΣΤΑΤΙΚΟΥ*.



Πίνακας 25: ΟΠΤΙΚΟΠΟΙΗΣΗ ΤΩΝ ΑΓΟΡΩΝ ΑΝΑ ΗΜΕΡΑ ΤΗΣ ΕΒΔΟΜΑΔΟΣ

Άμεσα παρατηρούμε ένα δυνητικά ενδιαφέρον στοιχείο: τον μειωμένο αριθμό αγορών τις ημέρες Τετάρτη και Δευτέρα. Εξηγείται βέβαια από το μειωμένο ωράριο λειτουργίας των καταστημάτων τις ημέρες αυτές. Παρατηρούμε όμως και μια μικρή αύξηση των επισκέψεων για την ημέρα Παρασκευή. Βλέπουμε και κάποιες επισκέψεις την Κυριακή. Αυτές αφορούν τις Κυριακές πριν τα Χριστούγεννα και την Πρωτοχρονιά.

Στην συνέχεια δημιουργούμε ακόμα ένα χαρακτηριστικό, το *UniqueID*, το οποίο φαίνεται στη λίστα με τα χαρακτηριστικά στην παραπάνω εικόνα. Αυτό αποδίδει ένα μοναδικό αριθμό - κλειδί σε κάθε επίσκεψη και χρησιμοποιεί ως βάση τόσο τον *ΑΡΙΘΜΟ ΠΑΡΑΣΤΑΤΙΚΟΥ* όσο και την *ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΣΤΑΤΙΚΟΥ* για να είναι

αυτός ο αριθμός μοναδικός για κάθε ομάδα υποδειγμάτων που αφορούν μια ξεχωριστή επίσκεψη. Υπό άλλες συνθήκες θα μπορούσε να χρησιμοποιηθεί ο *ΑΡΙΘΜΟΣ ΠΑΡΑΣΤΑΤΙΚΟΥ* σαν *UniqueID* ωστόσο ο αριθμός αυτός είναι δημιουργημένος στο σύστημα ώστε να ανακυκλώνεται έπειτα από συγκεκριμένες χρονικές περιόδους.

Έπειτα, αφού μελετούμε μόνο τις αγορές στη συγκεκριμένη ανάλυση, αφαιρούμε από το αρχείο τα υποδείγματα που αφορούν επιστροφές μέσω του *weka.filters.unsupervised.instance.RemoveWithValues*. Στο φίλτρο επιλέγουμε τις υποδείγματα που έχουν τιμή «102» στο χαρακτηριστικό *ΕΙΔΟΣ ΠΑΡΑΣΤΑΤΙΚΟΥ*.

Ο αλγόριθμος που θα χρησιμοποιήσουμε είναι ο *Apriori*. Πρόκειται για έναν αλγόριθμο που μέσω επαναλήψεων μειώνει την ελάχιστη κάλυψη (support) στα υποδείγματα μέχρι να βρει τον αριθμό κανόνων και σχέσεων που έχει ορίσει ο χρήστης με την απαιτούμενη ελάχιστη εμπιστοσύνη (confidence). Ο *Apriori* επεξεργάζεται ονομαστικά ή δυαδικά χαρακτηριστικά αλλά όχι αριθμητικά. Άλλωστε πολλά από τα αριθμητικά χαρακτηριστικά που έχουμε δημιουργήσει αφορούν ποσότητες και μεγέθη που δεν μας χρειάζονται στην παρούσα φάση. Απομακρύνουμε λοιπόν τα παραπάνω χαρακτηριστικά και κρατάμε μόνο τα ονομαστικά χαρακτηριστικά που θα χρησιμοποιήσουμε:

- *ΜΑΡΚΑ*
- *ΕΙΔΟΣ*
- *ΟΜΑΔΑ*
- *UniqueID*
- *ΗΜΕΡΑ*

Το χαρακτηριστικό *UniqueID* δεν είναι ονομαστικό αλλά το χρειαζόμαστε για να αλλάξουμε τη μορφή του αρχείου μας.

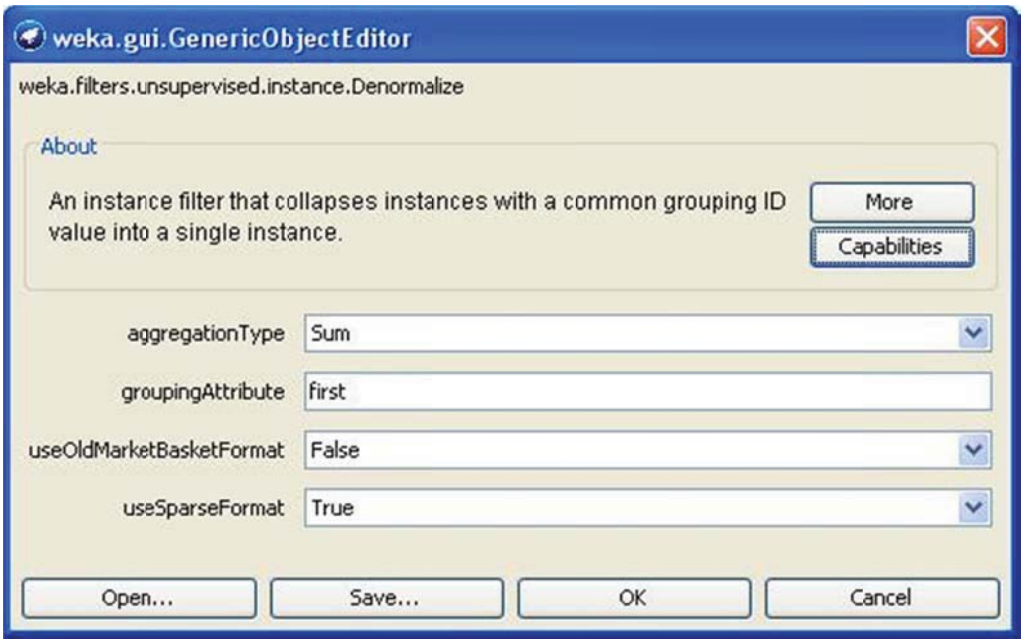
Το αρχείο στην αρχική του μορφή έχει ένα υπόδειγμα (μια σειρά) για κάθε διαφορετικό είδος σε κάθε συναλλαγή. Είναι της μορφής δηλαδή:

ΑΡ. ΠΑΡΑΣΤΑΤΙΚΟΥ	**ΗΜΕΡΟΜΗΝΙΑ**	**ΕΙΔΟΣ**
0152	07/07/10	ΠΑΝΤΟΦΛΕΣ
0152	07/07/10	ΣΟΥΤΙΕΝ
0152	07/07/10	ΜΠΛΟΥΖΑ
0153	08/07/10	ΦΟΥΣΤΑ
0153	08/07/10	ΦΟΡΕΜΑ
0154	09/07/10	ΝΥΧΤΙΚΟ

Ωστόσο για να το επεξεργαστούμε με τον *Apriori* θα πρέπει να το μετατρέψουμε σε μορφή που θα έχει ένα υπόδειγμα (μια γραμμή) για κάθε συναλλαγή. Ο παραπάνω πίνακας σε μια τέτοια μετατροπή θα γινόταν

ΑΡ. ΠΑΡΑΣΤΑΤΙΚΟΥ	ΗΜΕΡΟΜΗΝΙΑ	ΠΑΤΝΟΦΛΕΣ	ΣΟΥΤΙΕΝ	ΜΠΛΟΥΖΑ	ΦΟΥΣΤΑ	ΦΟΡΕΜΑ	ΝΥΧΤΙΚΟ
0152	07/07/10	True	True	True	False	False	False
0153	08/07/10	False	False	False	True	True	False
0154	09/07/10	False	False	False	False	False	True

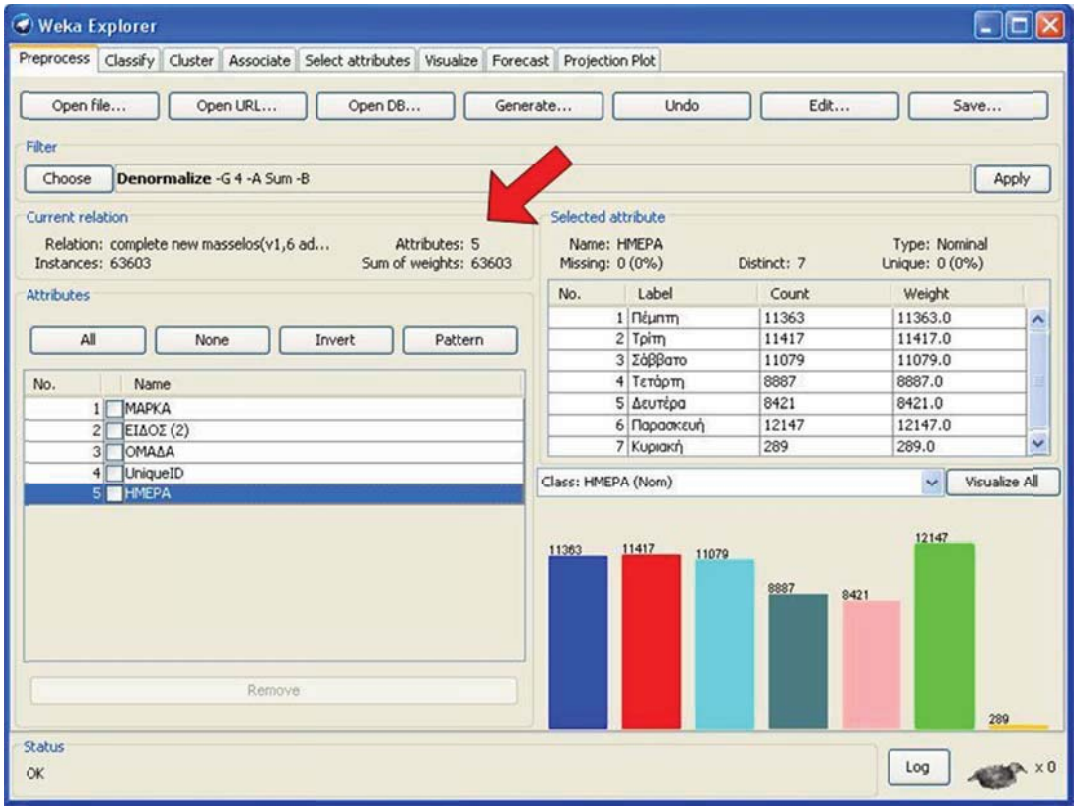
Αυτή η μετατροπή επιτυγχάνεται μέσω του weka.filters.unsupervised.instance.Denormalize. Το φίλτρο αυτό συγχωνεύει υποδείγματα με ένα κοινό χαρακτηριστικό-κλειδί, και τα μετατρέπει σε ένα υπόδειγμα.



Πίνακας 26: ΦΙΛΤΡΟ Denormalize

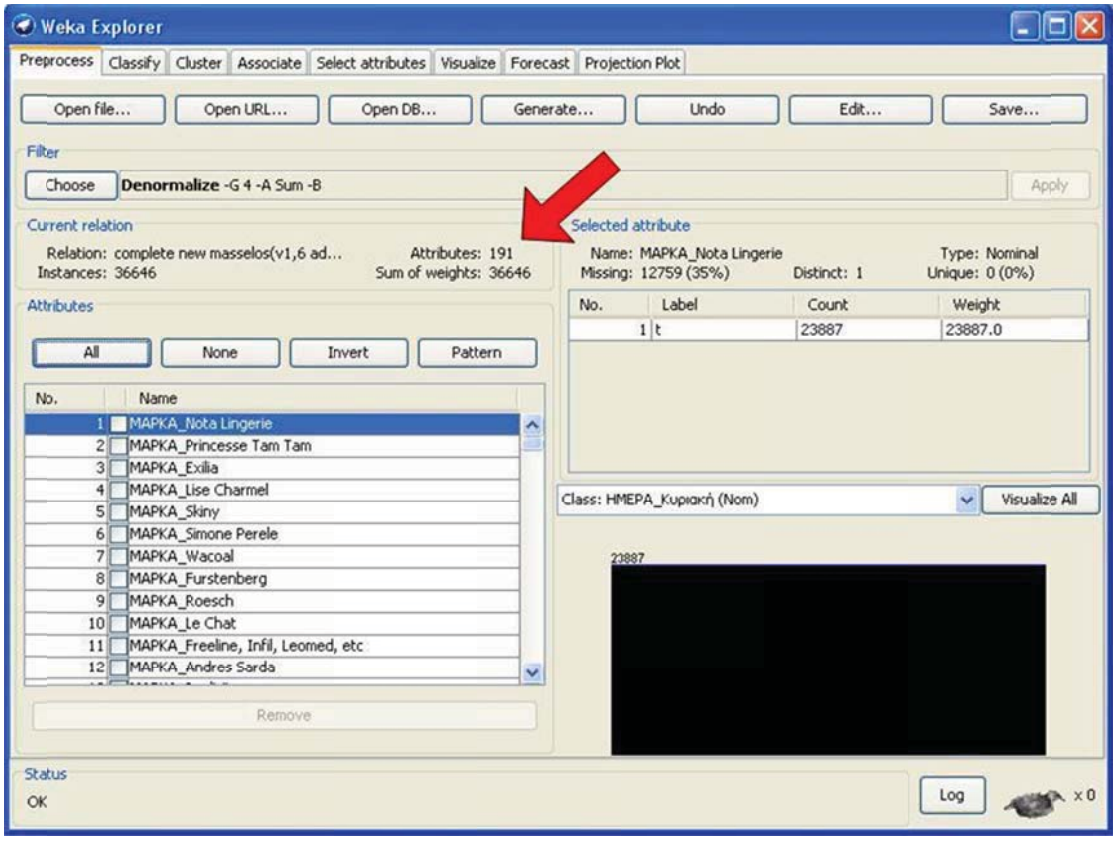
Στην επιλογή *groupingAttribute* τοποθετούμε το χαρακτηριστικό-κλειδί ενώ υπάρχουν ακόμα επιλογές σχετικά με τον χειρισμό των αριθμητικών δεδομένων κατά τη συγχώνευση των υποδειγμάτων και με τον χειρισμό αραιών πινάκων. Να τονιστεί πως λειτουργεί με την υπόθεση ότι τα υποδείγματα είναι εκ των προτέρων ταξινομημένα ως προς το χαρακτηριστικό-κλειδί.

Εκτελώντας το φίλτρο από τα αρχικά πέντε χαρακτηριστικά



Πίνακας 27: ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΠΡΙΝ ΤΗΝ ΕΚΤΕΛΕΣΗ ΤΟΥ ΦΙΛΤΡΟΥ Denormalize

δημιουργούνται 191 δυαδικά χαρακτηριστικά

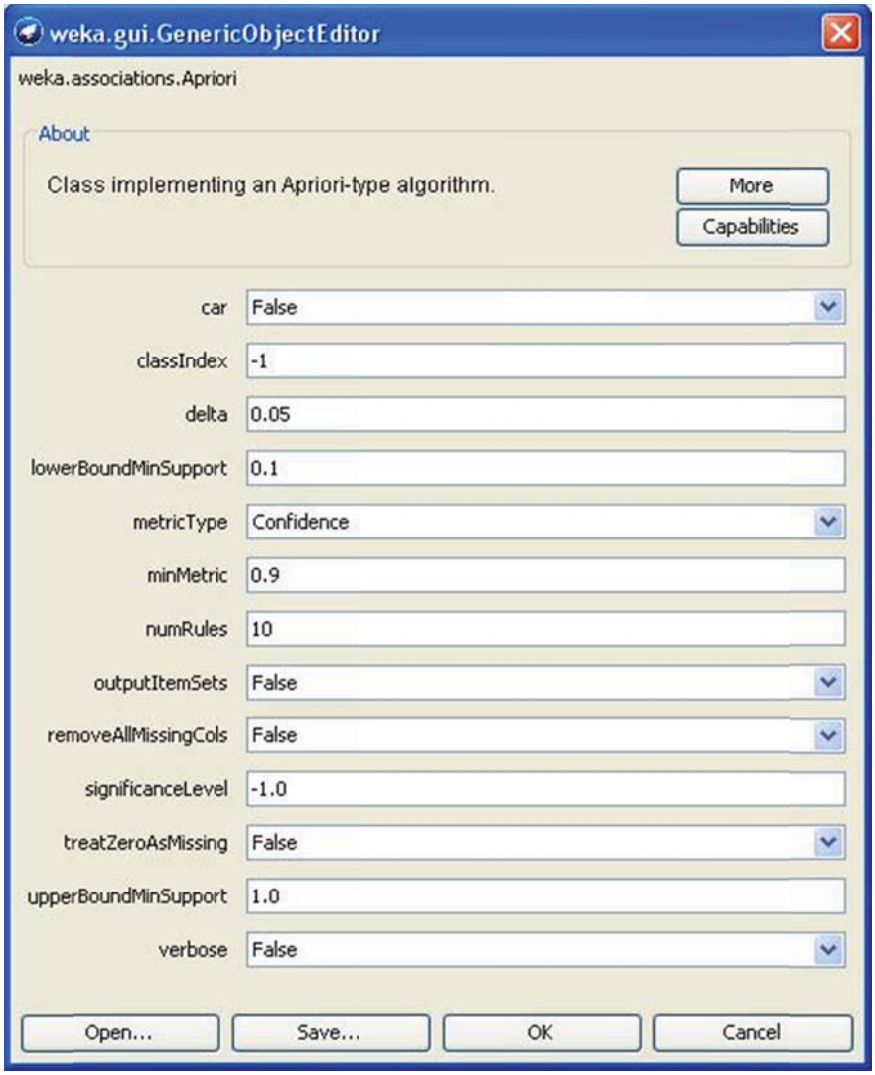


Πίνακας 28: ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕΤΑ ΤΗΝ ΕΚΤΕΛΕΣΗ ΤΟΥ ΦΙΛΤΡΟΥ Denormalize

Τα δυαδικά χαρακτηριστικά παρατηρούμε πως έχουν πάρει την τιμή «t» που αντιστοιχεί στο *true* ενώ αντί για *false* χαρακτηρίζονται ως *missing.* Αυτό συμβαίνει γιατί ενεργοποιήσαμε την επιλογή του φίλτρου *useOldMarketBasketFormat* η οποία παράγει δεδομένα σε μορφή που είναι ταχύτερη επεξεργάσιμη από τον *Apriori.* Στην συνέχεια αφαιρούμε το χαρακτηριστικό *UniqueID* και είμαστε έτοιμοι να εφαρμόσουμε τον αλγόριθμο *Apriori*.

Όπως βλέπουμε οι επιλογές που υπάρχουν για τον *Apriori* είναι πολλές. Οι πιο σημαντικές από αυτές:

- *car* - εξάγει συσχετίσεις με την τάξη των υποδειγμάτων αν είναι true και όχι γενικά μεταξύ των χαρακτηριστικών
- *metricType* - θέτει τον κριτήριο με βάσει το οποίο θα ταξινομηθούν οι κανόνες. Τα πιθανά κριτήρια είναι το confidence, το lift, το support, και το leverage.
- *minMetric* - θέτει την ελάχιστη τιμή του ανωτέρω κριτηρίου
- *numRules* - ο αριθμός των κανόνων που εξάγει ο αλγόριθμος



Πίνακας 29: ΟΙ ΔΙΑΘΕΣΙΜΕΣ ΕΠΙΛΟΓΕΣ ΣΤΟΝ Apriori

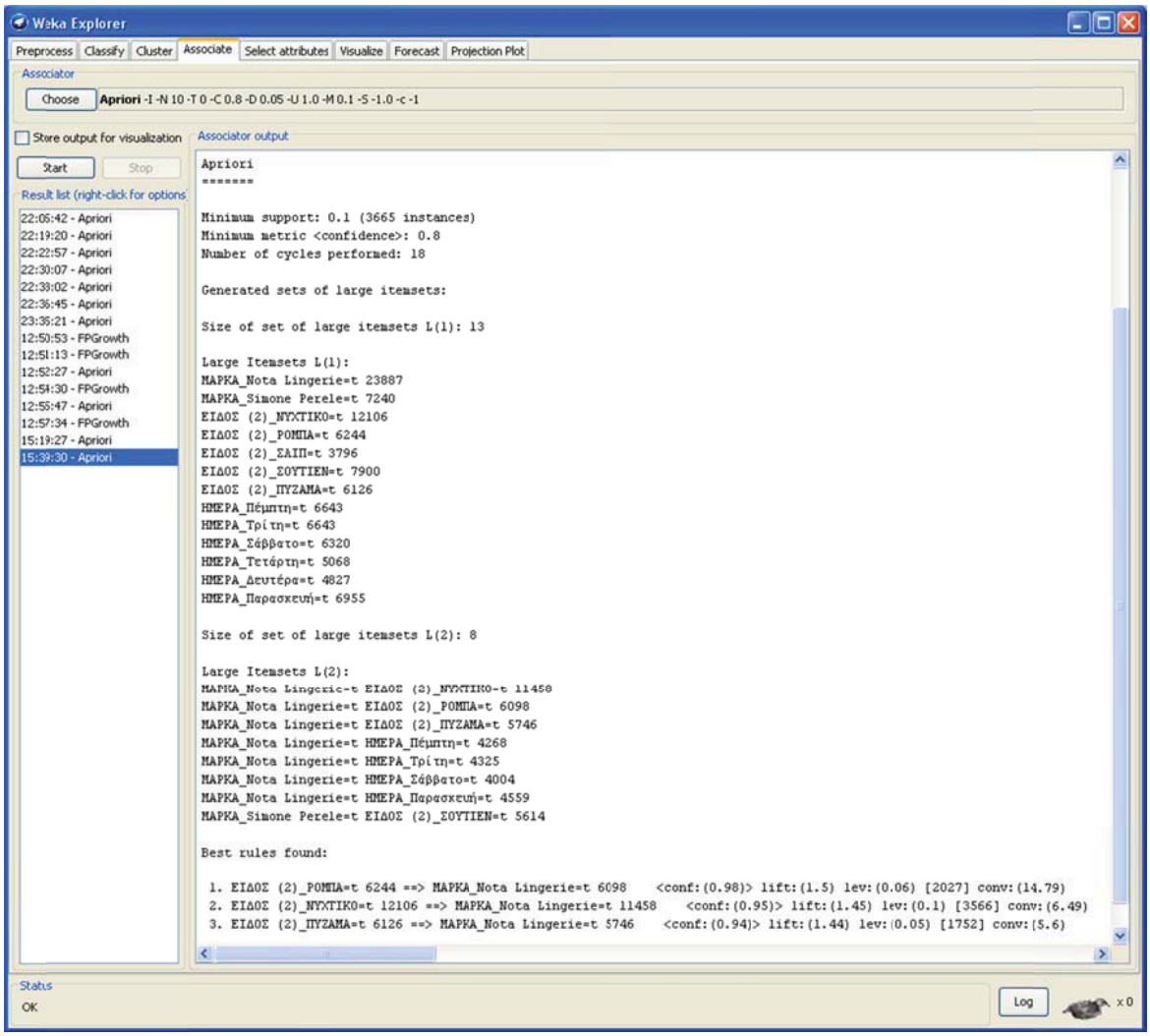
Στην περίπτωσή μας θα χρησιμοποιήσουμε για κριτήρια το *confidence* και το *lift.*

Το *confidence* (εμπιστοσύνη) είναι το πιο συχνά χρησιμοποιούμενο κριτήριο. Προκύπτει ως ο λόγος των υποδειγμάτων στα οποία ένας κανόνας επαληθεύεται ως προς το σύνολο των υποδειγμάτων τους οποίους καλύπτει. Για παράδειγμα η εμπιστοσύνη σε έναν κανόνα της μορφής $[x] \Rightarrow [z]$προκύπτει ως ο λόγος του συνόλου των υποδειγμάτων που συνυπάρχουν τα $[x],[z]$ως προς το σύνολο των υποδειγμάτων που ο κανόνας καλύπτει, δηλαδή τα υποδείγματα που περιέχουν το $[x]$. Οι τιμές που παίρνει το κριτήριο αυτό είναι από 0 έως 1 και όσο πιο κοντά στο 1 η τιμή του τόσο πιο ισχυρός είναι ο κανόνας αυτός.

Το *lift* είναι επίσης ένα πολύ συχνά χρησιμοποιούμενο κριτήριο. Υπολογίζει πόσο πιο συχνά τα $[x],[z]$του ανωτέρω παραδείγματος εμφανίζονται σε ένα αρχείο απ' ότι θα εμφανιζόντουσαν αν ήταν στατιστικά ανεξάρτητα. Μετρά και αυτό την αξία ενός κανόνα και οι τιμές που μπορεί να πάρει είναι από 0 έως θεωρητικά άπειρο. Τιμή μικρότερη του 1 σημαίνει πως το αριστερό στοιχείο του κανόνα έχει αρνητική

επίδραση ως προς την εμφάνιση του δεξιού στοιχείου το κανόνα. Τιμή ίση με το 1 σημαίνει πως τα δύο στοιχεία είναι στατιστικά ανεξάρτητα οπότε η εμφάνιση του αριστερού στοιχείου σε κάποιο υπόδειγμα δεν επηρεάζει τις πιθανότητες εμφάνισης του δεξιού. Τιμή μεγαλύτερη του 1 σημαίνει πως το αριστερό στοιχείο του κανόνα έχει θετική επίδραση ως προς την εμφάνιση του δεξιού στοιχείου του κανόνα, αυξάνοντας τις πιθανότητες αυτή να συμβεί.

Εφαρμόζουμε λοιπόν αρχικά τον *Apriori* με κριτήριο το *confidence* και ελάχιστη τιμή 0,8.



Πίνακας 30: ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΕΚΤΕΛΕΣΗΣ ΤΟΥ Apriori ΜΕ ΚΡΙΤΗΡΙΟ ΤΟ *Confidence*

Παρατηρούμε ότι ο αλγόριθμος δημιούργησε ένα σύνολο με 13 χαρακτηριστικά τα οποία είναι τα πιο συχνά συναντόμενα στο αρχείο μας. Από αυτά προκύπτουν 8 σχέσεις συχνά συνυπαρχόντων χαρακτηριστικών.

- ΜΑΡΚΑ_Nota Lingerie=t ΕΙΔΟΣ (2)_ΝΥΧΤΙΚΟ=t 11458
- ΜΑΡΚΑ_Nota Lingerie=t ΕΙΔΟΣ (2)_ΡΟΜΠΑ=t 6098
- ΜΑΡΚΑ_Nota Lingerie=t ΕΙΔΟΣ (2)_ΠΥΖΑΜΑ=t 5746

- ΜΑΡΚΑ_Nota Lingerie=t ΗΜΕΡΑ_Πέμπτη=t 4268
- ΜΑΡΚΑ_Nota Lingerie=t ΗΜΕΡΑ_Τρίτη=t 4325
- ΜΑΡΚΑ_Nota Lingerie=t ΗΜΕΡΑ_Σάββατο=t 4004
- ΜΑΡΚΑ_Nota Lingerie=t ΗΜΕΡΑ_Παρασκευή=t 4559
- ΜΑΡΚΑ_Simone Perele=t ΕΙΔΟΣ (2)_ΣΟΥΤΙΕΝ=t 5614

Για την κατανόηση των αποτελεσμάτων η πρώτη σχέση διαβάζεται ως: η μάρκα *Nota Lingerie* σε συνδυασμό με το χαρακτηριστικό *ΝΥΧΤΙΚΟ* συναντώνται στο αρχείο μας 11458 φορές

Τέλος ο αλγόριθμος εξάγει του παρακάτω 3 κανόνες:

1. ΕΙΔΟΣ (2)_ΡΟΜΠΑ=t 6244 ==> ΜΑΡΚΑ_Nota Lingerie=t 6098 <conf:(0.98)> lift:(1.5) lev:(0.06) [2027] conv:(14.79)
2. ΕΙΔΟΣ (2)_ΝΥΧΤΙΚΟ=t 12106 ==> ΜΑΡΚΑ_Nota Lingerie=t 11458 <conf:(0.95)> lift:(1.45) lev:(0.1) [3566] conv:(6.49)
3. ΕΙΔΟΣ (2)_ΠΥΖΑΜΑ=t 6126 ==> ΜΑΡΚΑ_Nota Lingerie=t 5746 <conf:(0.94)> lift:(1.44) lev:(0.05) [1752] conv:(5.6)

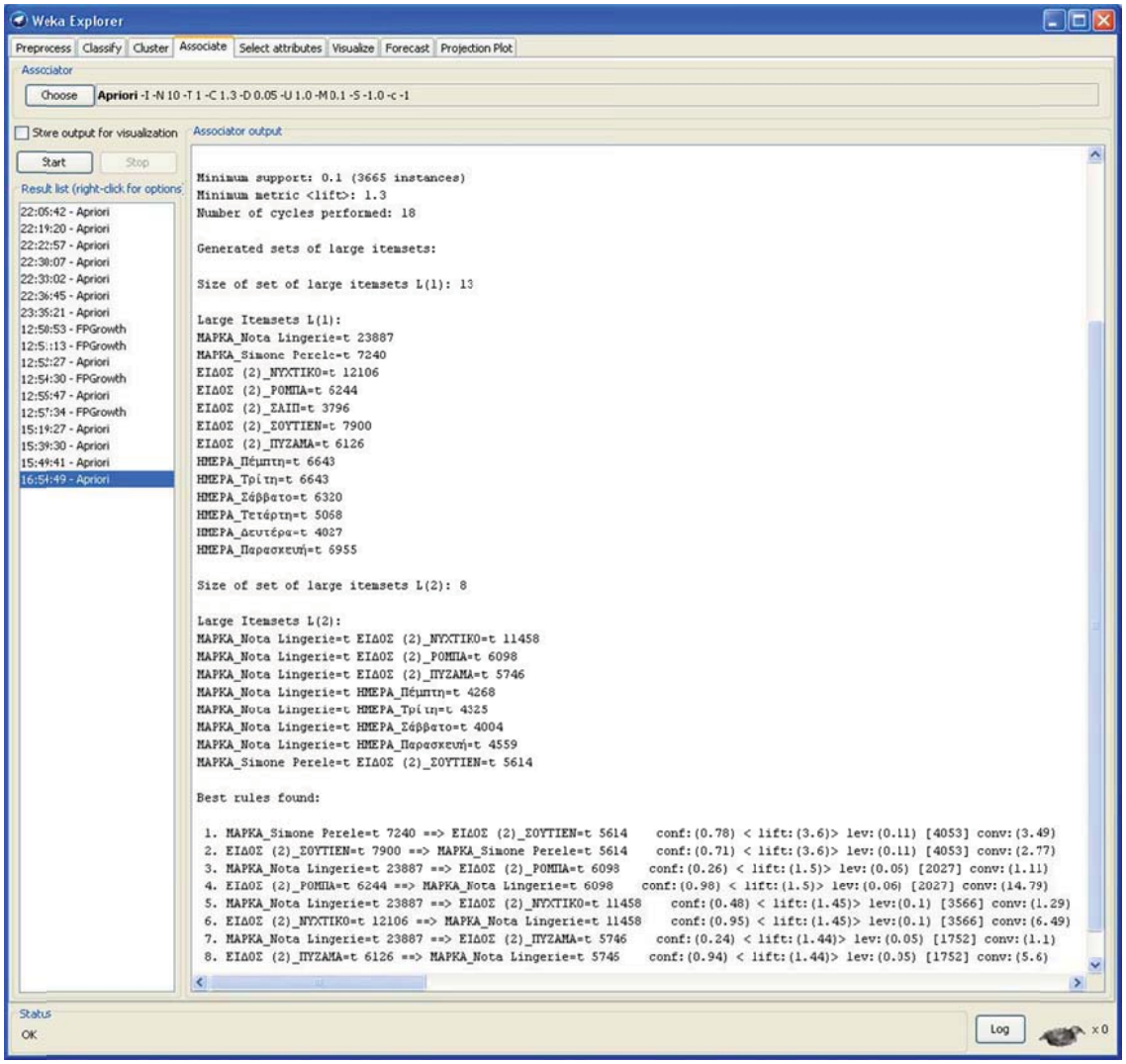
Αυτοί είναι οι κανόνες που είναι εντός των ορίων που έχουμε ορίσει. Σε κάθε κανόνα βλέπουμε τις εμφανίσεις του αριστερά σκέλους του (6244 στον πρώτο κανόνα – η κάλυψη του κανόνα), τις εμφανίσεις που επαληθεύουν τον κανόνα, ταυτόχρονη δηλαδή εμφάνιση και των δύο σκελών του κανόνα (6098 στον πρώτο), την τιμή του *confidence* (0.98), την τιμή του *lift* (1.5) και τις τιμές των άλλων δύο κριτηρίων (*leverage* και *departure*). Το *departure* είναι ένα ακόμα κριτήριο για την στατιστική ανεξαρτησία μεταξύ των προς μελέτη χαρακτηριστικών.

Στη συνέχεια εφαρμόζουμε τον *Apriori* με κριτήριο το *lift* και ελάχιστη τιμή 1,3. Σε αυτή την περίπτωση προκύπτουν 8 κανόνες:

1. ΜΑΡΚΑ_Simone Perele=t 7240 ==> ΕΙΔΟΣ (2)_ΣΟΥΤΙΕΝ=t 5614 conf:(0.78) < lift:(3.6)> lev:(0.11) [4053] conv:(3.49)
2. ΕΙΔΟΣ (2)_ΣΟΥΤΙΕΝ=t 7900 ==> ΜΑΡΚΑ_Simone Perele=t 5614 conf:(0.71) < lift:(3.6)> lev:(0.11) [4053] conv:(2.77)
3. ΜΑΡΚΑ_Nota Lingerie=t 23887 ==> ΕΙΔΟΣ (2)_ΡΟΜΠΑ=t 6098 conf:(0.26) < lift:(1.5)> lev:(0.06) [2027] conv:(1.11)
4. ΕΙΔΟΣ (2)_ΡΟΜΠΑ=t 6244 ==> ΜΑΡΚΑ_Nota Lingerie=t 6098 conf:(0.98) < lift:(1.5)> lev:(0.06) [2027] conv:(14.79)
5. ΜΑΡΚΑ_Nota Lingerie=t 23887 ==> ΕΙΔΟΣ (2)_ΝΥΧΤΙΚΟ=t 11458 conf:(0.48) < lift:(1.45)> lev:(0.1) [3566] conv:(1.29)
6. ΕΙΔΟΣ (2)_ΝΥΧΤΙΚΟ=t 12106 ==> ΜΑΡΚΑ_Nota Lingerie=t 11458 conf:(0.95) < lift:(1.45)> lev:(0.1) [3566] conv:(6.49)
7. ΜΑΡΚΑ_Nota Lingerie=t 23887 ==> ΕΙΔΟΣ (2)_ΠΥΖΑΜΑ=t 5746 conf:(0.24) < lift:(1.44)> lev:(0.05) [1752] conv:(1.1)

8. ΕΙΔΟΣ (2)_ΠΥΖΑΜΑ=t 6126 ==> ΜΑΡΚΑ_Nota Lingerie=t 5746 conf:(0.94) < lift:(1.44)> lev:(0.05) [1752] conv:(5.6)

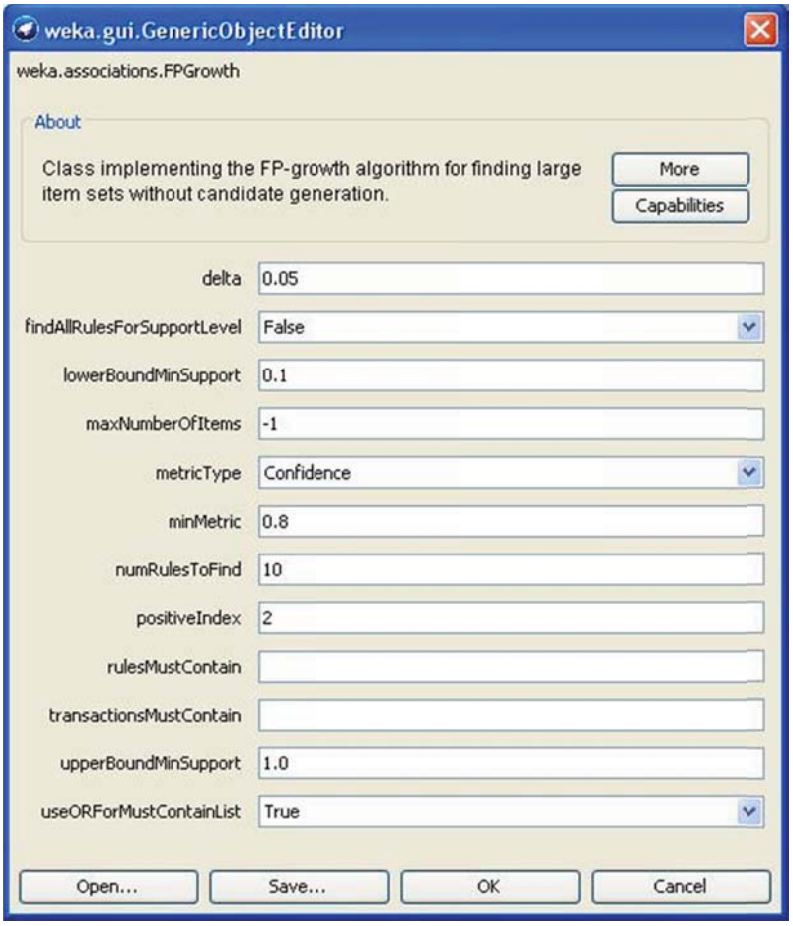
Έχουν προστεθεί κανόνες που είναι εντός του ορίου για το *lift* σε σχέση με την προηγούμενη περίπτωση αλλά με πολλή χαμηλή εμπιστοσύνη όπως ο 3ος ή ο 5ος ή ο 7ος κανόνας.



Πίνακας 31: ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΕΚΤΕΛΕΣΗΣ ΤΟΥ Apriori ΜΕ ΚΡΙΤΗΡΙΟ ΤΟ *Lift*

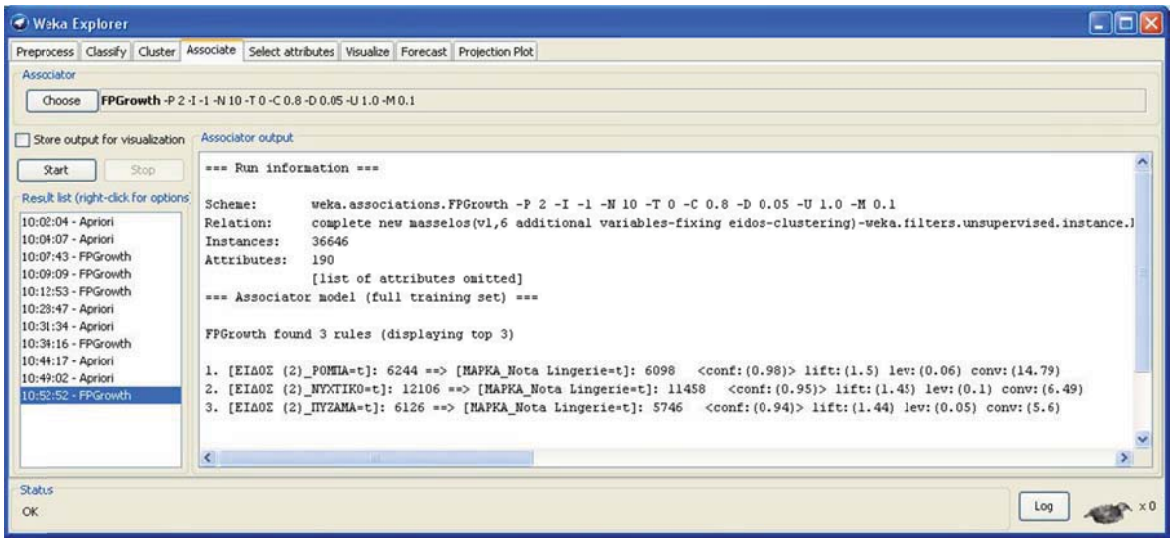
Παρατηρούμε ότι και στις δύο περιπτώσεις ο αλγόριθμος δεν μας έδωσε κάποιο κανόνα για σχέση μεταξύ ειδών ο οποίος δυνητικά ίσως να είχε μεγαλύτερη αξία αλλά για σχέσεις μεταξύ ειδών και μάρκας. Αυτό κυρίως οφείλεται στην ποικιλία εγγραφών για προϊόντα ίδιας κατηγορίας με αποτέλεσμα τη διάσπασή τους σε μικρά υποσύνολα τα οποία ο αλγόριθμος δεν εντοπίζει ως σημαντικά.

Δοκιμάζουμε να χρησιμοποιήσουμε και τον αλγόριθμο *FPGrowth*. Είναι και αυτός ένας αλγόριθμος συσχετίσεων που βρίσκεται στην καρτέλα *Associate* του *Weka*.

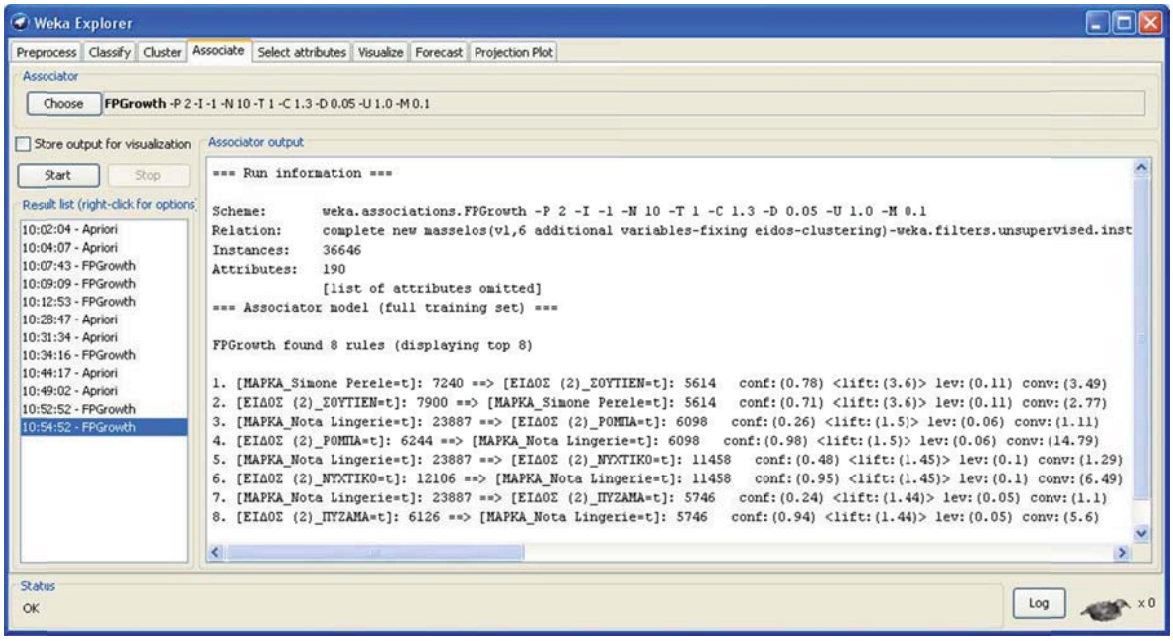


Πίνακας 32: ΟΙ ΔΙΑΘΕΣΙΜΕΣ ΕΠΙΛΟΓΕΣ ΣΤΟΝ FPGrowth

Η κύρια διαφορά του με τον *Apriori* είναι ότι δεν δημιουργεί αρχικά μια λίστα υποψηφίων συνόλων στοιχείων που εμφανίζονται πιο συχνά αλλά καταφεύγει απευθείας στην δημιουργία κανόνων. Εφαρμόζουμε τον *FPGrowth* στο αρχείο μας δύο φορές με τα ίδια κριτήρια όπως παραπάνω (*confidence* 0.8 στη πρώτη και *lift* 1.3 στην δεύτερη).



Πίνακας 33: ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΕΚΤΕΛΕΣΗΣ ΤΟΥ FPGrowth ΜΕ ΚΡΙΤΗΡΙΟ ΤΟ *Confidence*



Πίνακας 34: ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΕΚΤΕΛΕΣΗΣ ΤΟΥ FPGrowth ΜΕ ΚΡΙΤΗΡΙΟ ΤΟ *Lift*

Οι κανόνες που παράγονται στην πρώτη περίπτωση (*confidence>0.8*):

1. [ΕΙΔΟΣ (2)_ΡΟΜΠΑ=t]: 6244 ==> [ΜΑΡΚΑ_Nota Lingerie=t]: 6098 <conf:(0.98)> lift:(1.5) lev:(0.06) conv:(14.79)
2. [ΕΙΔΟΣ (2)_ΝΥΧΤΙΚΟ=t]: 12106 ==> [ΜΑΡΚΑ_Nota Lingerie=t]: 11458 <conf:(0.95)> lift:(1.45) lev:(0.1) conv:(6.49)
3. [ΕΙΔΟΣ (2)_ΠΥΖΑΜΑ=t]: 6126 ==> [ΜΑΡΚΑ_Nota Lingerie=t]: 5746 <conf:(0.94)> lift:(1.44) lev:(0.05) conv:(5.6)

Και στη δεύτερη (*lift>1.3*):

1. [ΜΑΡΚΑ_Simone Perele=t]: 7240 ==> [ΕΙΔΟΣ (2)_ΣΟΥΤΙΕΝ=t]: 5614 conf:(0.78) <lift:(3.6)> lev:(0.11) conv:(3.49)
2. [ΕΙΔΟΣ (2)_ΣΟΥΤΙΕΝ=t]: 7900 ==> [ΜΑΡΚΑ_Simone Perele=t]: 5614 conf:(0.71) <lift:(3.6)> lev:(0.11) conv:(2.77)
3. [ΜΑΡΚΑ_Nota Lingerie=t]: 23887 ==> [ΕΙΔΟΣ (2)_ΡΟΜΠΑ=t]: 6098 conf:(0.26) <lift:(1.5)> lev:(0.06) conv:(1.11)
4. [ΕΙΔΟΣ (2)_ΡΟΜΠΑ=t]: 6244 ==> [ΜΑΡΚΑ_Nota Lingerie=t]: 6098 conf:(0.98) <lift:(1.5)> lev:(0.06) conv:(14.79)
5. [ΜΑΡΚΑ_Nota Lingerie=t]: 23887 ==> [ΕΙΔΟΣ (2)_ΝΥΧΤΙΚΟ=t]: 11458 conf:(0.48) <lift:(1.45)> lev:(0.1) conv:(1.29)
6. [ΕΙΔΟΣ (2)_ΝΥΧΤΙΚΟ=t]: 12106 ==> [ΜΑΡΚΑ_Nota Lingerie=t]: 11458 conf:(0.95) <lift:(1.45)> lev:(0.1) conv:(6.49)

7. [ΜΑΡΚΑ_Nota Lingerie=t]: 23887 ==> [ΕΙΔΟΣ (2)_ΠΥΖΑΜΑ=t]: 5746 conf:(0.24) <lift:(1.44)> lev:(0.05) conv:(1.1)
8. [ΕΙΔΟΣ (2)_ΠΥΖΑΜΑ=t]: 6126 ==> [ΜΑΡΚΑ_Nota Lingerie=t]: 5746 conf:(0.94) <lift:(1.44)> lev:(0.05) conv:(5.6)

Βλέπουμε ότι οι κανόνες που προκύπτουν είναι ίδιοι με αυτούς που προκύπτουν από τον *Apriori* και τον επιβεβαιώνουν.

8. Συμπεράσματα – Προτάσεις

Στο 5^{ο} κεφάλαιο περιγράψαμε τους στόχους που είχε η εταιρεία μέσα από αυτή την εργασία. Επίσης τα αποτελέσματα των εκάστοτε αλγορίθμων παρατέθηκαν στην ενότητα με την περιγραφή τους. Ορισμένα από αυτά παρήγαγαν γνώση πολλή ενδιαφέρουσα για την εταιρεία που δεν ήταν εύκολο να εξαχθεί με άλλη διαδικασία και άλλα παρήγαγαν αποτελέσματα και κανόνες που ήταν ήδη γνωστοί και επιβεβαίωναν τις ήδη υπάρχουσες γνώσεις. Παρακάτω συνοψίζουμε τα αποτελέσματα της μελέτης σε σχέση με τους αρχικούς στόχους που είχαν τεθεί:

Πρώτος και κύριος στόχος μέσα από αυτή την εργασία ήταν η επιλογή των υποψήφιων πελατών για την αποστολή ενημερωτικών μέσων για τη νέα κολεξιόν. Σκοπός ήταν η μείωση του αριθμού των πελατών προς τους οποίους θα στέλνονταν ενημερωτικά μέσα, σε αυτούς από τους οποίους θα υπήρχε πραγματικό όφελος για την εταιρεία, μειώνοντας έτσι το κόστος την ενημερωτικής καμπάνιας χωρίς να θυσιάζονται τα οφέλη από αυτή. Για τον σκοπό αυτό χρησιμοποιήσαμε όπως είδαμε αλγορίθμους ομαδοποίησης, τον EM και τον KMeans, ώστε να χωρίσουμε το πλήθος των πελατών σε ομάδες με κοινά χαρακτηριστικά και στη συνέχεια αλγορίθμους ταξινόμησης για περιγράψουμε ακόμα καλύτερα την δομή αυτών των ομάδων. Αυτό είχε ως αποτέλεσμα την δημιουργία 5 ομάδων πελατών με κοινά χαρακτηριστικά. Κατά την αξιολόγησή τους προέκυψαν οι ακόλουθες διαπιστώσεις:

I. Η πρώτη ομάδα, η ομάδα των πιστών πελατών, απορρίπτεται από την επιλογή για την *mailing list*. Πρόκειται για πιστούς πελάτες που, ούτως ή άλλως, θα επανέλθουν για αγορές στην εταιρεία, οπότε θα ήταν μάλλον περιττή η αποστολή σε αυτούς ενημερωτικών μέσων και δεν θα ταίριαζε με την επιθετική πολιτική της καμπάνιας για την αύξηση των πωλήσεων.
II. Η δεύτερη ομάδα, αποτελείται από πελάτες που είναι και αυτοί σχετικά συνεπείς στις αγορές τους από την εταιρεία. Εμφανίζονται δε περισσότερο τους καλοκαιρινούς μήνες, οπότε αποτελούν και αυτοί πελάτες που πιθανότατα θα έλθουν τους μήνες που αφορά η νέα κολεξιόν ούτως ή άλλως για αγορές. Η αποστολή και σε αυτούς ενημερωτικών μέσων κρίθηκε μάλλον περιττή.
III. Οι υπόλοιπες τρεις ομάδες αποτελούνται από πελάτες που δεν έχουν τόσο συχνή εμφάνιση στα δεδομένα συναλλαγών. Πρόκειται για πελάτες που δεν τους έχει κερδίσει ακόμα η εταιρεία. Ειδικά η 4^{η} και η 5^{η} ομάδα αποτελούνται από πελάτες που εμφανίστηκαν κατά μεγάλη πλειοψηφία μόνο ένα έτος στα δεδομένα συναλλαγών. Η 3^{η} ομάδα έχει πελάτες που έχουν χαμηλό ποσοστό επισκέψεων τους καλοκαιρινούς μήνες. Και οι 3 ομάδες επομένως περιέχουν πελάτες τους οποίους θέλει η

εταιρεία να επαναδραστηριοποιήσει ως προς τις αγορές τους σε αυτή.

Βλέπουμε λοιπόν ότι οι 3 τελευταίες ομάδες πελατών είναι αυτές που θα ήταν πιο ωφέλιμο να σταλούν τα ενημερωτικά μέσα ως προς τη νέα κολεξιόν και που θα αποτελέσουν την mailing list. Συγκριτικά με το σύνολο των καταγεγραμμένων πελατών στα αρχεία της εταιρείας, η mailing list θα περιλαμβάνει μόνο το 64% αυτών και το οποίο θα είναι οι πελάτες που είναι πιο πιθανό να αυξήσουν τις αναμενόμενες πωλήσεις της. Επιτυγχάνεται επομένως μια μείωση του κόστους της τάξεως του 36% χωρίς να θυσιάζεται το πιθανό όφελος από αυτή την διαφημιστική καμπάνια.

Δεύτερος στόχος ήταν η εξαγωγή τυχόν πληροφοριών που θα προέκυπταν κατά την όλη διαδικασία και μια ανάλυση του καλαθιού αγοράς για τον εντοπισμό τυχόν συσχετίσεων που μπορεί να υπάρχουν μεταξύ των προϊόντων. Πληροφορίες σχετικά με την δομή των ομάδων που προέκυψαν από την ομαδοποίηση υπήρξαν διάφορες και είναι καταγεγραμμένες στο αντίστοιχο κεφάλαιο. Αφορούν την αγοραστική δύναμη των διαφόρων ομάδων, την συχνότητα εμφάνισής τους τα διάφορα έτη και τις διάφορες εποχές, την συχνότητα επιστροφών στις διάφορες αγορές που είχαν κάνει κλπ. Αυτές ήταν αναμενόμενο να εξαχθούν καθώς ήταν σχετικές με το αντικείμενο του πρώτου στόχου της εργασίας. Στη συνέχεια όμως και με την χρήση των αλγορίθμων ταξινόμησης, τόσο του Naïve Bayes όσο και των αλγορίθμων παραγωγής κανόνων JRip και PART, προέκυψαν και πληροφορίες που δεν είχαν άμεση σχέση με τον πρώτο στόχο της εργασίας και οι οποίες θα μπορούσαν να έχουν κάποια αξία σε ότι αφορά διαφορετικά θέματα της εταιρείας. Για παράδειγμα κατά την ανάλυση των επώνυμων και ανώνυμων εγγραφών ανά κύκλωμα καταγραφής και ανά έτος φάνηκε μια μείωση των επώνυμων εγγραφών από το έτος 2009 στο 2010 για το κύκλωμα 103 της τάξης του 50% την ώρα που αντίστοιχη μείωση στο κύκλωμα 105 ήταν μόλις 8%. Αυτό θα μπορούσε να καταδείξει μια μεταφορά της προτίμησης των επώνυμων πελατών για αγορές στο κύκλωμα 105 ή μια ασυνέπεια των υπαλλήλων στο κύκλωμα 103 ως προς τις επώνυμες καταγραφές. Ακόμα, οι προτιμήσεις των ομάδων πελατών ως προς τις κατηγορίες προϊόντων και οι διαφορές των στοιχείων των πελατών ανάλογα το κύκλωμα καταγραφής είναι δύο ακόμα παραδείγματα πληροφοριών που ενώ δεν έχουν άμεση σχέση με τον κύριο στόχο της εργασίας, εξήχθησαν κατά τη μελέτη και μπορούν να δώσουν πληροφορίες και κατευθύνσεις σε διάφορους τομείς δραστηριότητας της εταιρείας. Για παράδειγμα το ότι το κύκλωμα 105 υπερέχει σε αγορές κατά την περίοδο του καλοκαιριού έναντι του κυκλώματος 103 (τα δύο κύρια κυκλώματα της εταιρείας), θα μπορούσε να χρησιμοποιηθεί για την διαφορετική ίσως διαρρύθμιση των αντίστοιχων καταστημάτων. Το ένα με πιο καλοκαιρινό στυλ από το άλλο και με περισσότερη έμφαση σε ανάλογη κατηγορία προϊόντων όπως για παράδειγμα τα μαγιό. Σε ότι αφορά το δεύτερο κομμάτι του

στόχου, το *market basket analysis*, δεν βρέθηκε κάποιος συσχετισμός ανάμεσα σε προϊόντα. Βρέθηκε μόνο ανάμεσα σε μάρκα και είδος προϊόντος, που όπως ήταν αναμενόμενο η μάρκα Nota Lingerie συσχετίστηκε με διάφορα είδη, μιας και είναι η κυρίαρχη μάρκα της εταιρείας.

Ωστόσο από την ενασχόληση με την διαδικασία του *data mining* στη Νότα προέκυψαν χρήσιμα συμπεράσματα αλλά και εμπειρίες τα οποία συνοψίζονται παρακάτω:

Η ποιότητα των δεδομένων είναι ο σημαντικότερος συντελεστής τόσο για την ακρίβεια των αποτελεσμάτων από αλγορίθμους *data mining* όσο και για την εξαγωγή ορθών συμπερασμάτων κατά την ανάλυσή τους. Είναι απαραίτητη η σωστή διαχείριση και αποθήκευση αυτών των δεδομένων από μια κατάλληλη πλατφόρμα και η σωστή συντήρηση της εκάστοτε βάσης δεδομένων. Ωστόσο τα περισσότερα λανθασμένα ή ημιτελή δεδομένα προκύπτουν από το σημείο εισαγωγής τους, τα σημεία πωλήσεως δηλαδή. Εκεί ο εκάστοτε πωλητής λόγο του άγχους να εξυπηρετήσει όσο πιο σωστά και γρήγορα γίνεται τον πελάτη, συχνά αμελεί την σημασία της σωστής εισαγωγής των δεδομένων. Παραδείγματα τέτοιων αμελειών είναι το άνοιγμα δεύτερου και τρίτου κωδικού πελάτη για το ίδιο άτομο ή η παράλειψη εισαγωγής διαθέσιμων στοιχείων για την ταχύτερη εξυπηρέτηση ή ακόμα κάποια λάθος πληκτρολόγηση στοιχείων κατά την καταχώρηση που μπορεί να προκύπτει λόγω βιασύνης. Αυτό έχει ως αποτέλεσμα την αλλοίωση των δεδομένων και τη μείωση της ποιότητας τους και συχνά είναι δύσκολος έως αδύνατος ο εντοπισμός και η διόρθωσή τους. Για την αποφυγή τέτοιων περιπτώσεων προτείνονται δύο μέτρα:

1. Η αυστηρή τυποποίηση και ο έλεγχος κατά την εισαγωγή των δεδομένων.

Το σύστημα να είναι σε θέση να ελέγχει απευθείας τα δεδομένα κατά την εισαγωγή και να προτείνει ήδη υπάρχοντα σε περίπτωση που μοιάζουν. Για παράδειγμα κατά την εισαγωγή επωνυμίας πελάτη έπειτα από την εισαγωγή των τριών ή τεσσάρων πρώτων γραμμάτων να προτείνει το σύστημα εγγραφές που έχουν την ίδια επωνυμία. Με αυτό τον τρόπο θα μειωθεί αρκετά η ύπαρξη διπλοεγγραφών και λανθασμένων εγγραφών λόγω ορθογραφικών λαθών. Ακόμα να υπάρχει μια τυποποίηση κατά την εισαγωγή αριθμητικών στοιχείων όπως του ταχυδρομικού κωδικού, για παράδειγμα τρία ψηφία, παύλα «-», δύο ψηφία (104-33), ώστε να αποφεύγεται η καταλάθος πληκτρολόγηση λιγότερων ή περισσοτέρων χαρακτήρων. Τέλος μια ακόμα λύση θα μπορούσε να ήταν η σύνδεση του ταχυδρομικού κώδικα με την περιοχή μέσα από κατάλληλα ενημερωμένο αρχείο, έτσι ώστε μόνο με την εισαγωγή του ταχυδρομικού κώδικα να ενημερώνεται αυτόματα και η περιοχή, μειώνοντας την πιθανότητα λάθους. Αυτές οι απλές λύσεις θα βοηθούσαν σημαντικά στην βελτίωση της ποιότητας των δεδομένων της εταιρείας με

αποτέλεσμα την αμεσότερη και ταχύτερη επεξεργασία τους και παραγωγή αποτελεσμάτων χωρίς να απαιτείται τόσο χρονοβόρα προεπεξεργασία και καθαρισμός τους.

2. Η απόδοση κινήτρων στους πελάτες

Διαπιστώσαμε πως πολλά λάθη γίνονται κατά την εισαγωγή των δεδομένων λόγω πιθανής αμέλειας ή βιασύνης. Ωστόσο αυτό θα μειωνόταν δραματικά αν ο ίδιος ο πελάτης ενδιαφερόταν για την σωστή εισαγωγή των στοιχείων του. Αυτό μπορεί να γίνει με την απόδοση κατάλληλων κινήτρων στους πελάτες. Για παράδειγμα η ύπαρξη προσφοράς με το που συμπληρώσει κάποιος πελάτης ένα συγκεκριμένο ποσό αγορών ή έναν αριθμό επισκέψεων σε κάποιο κατάστημα, θα προέτρεπε τον πελάτη να τονίσει τα στοιχεία του και να ελέγξει όσο αυτό είναι δυνατό την σωστή εισαγωγή τους ώστε να επωφεληθεί της προσφοράς. Ακόμα θα μπορούσε να χρησιμοποιηθεί ειδική κάρτα με barcode, την οποία θα «φόρτωνε» την πρώτη φορά της επισκέψεώς του ο πελάτης με τα στοιχεία του και θα την χρησιμοποιούσε στις επόμενες αγορές, μειώνοντας αισθητά την πιθανότητα διπλοεγγραφών ή λοιπών λανθασμένων εγγραφών.

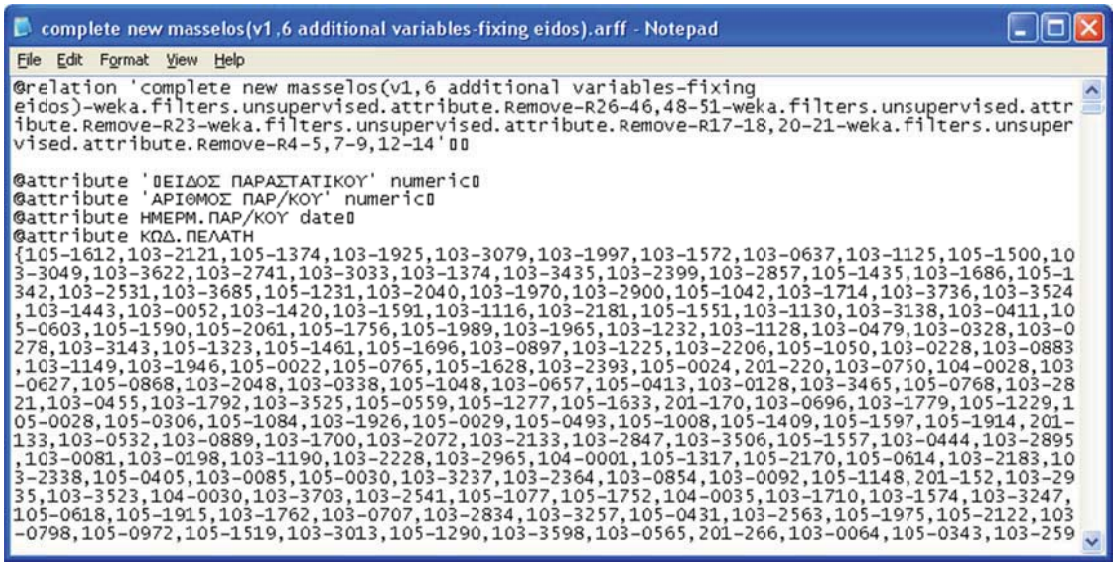
Εκτός όμως από την ποιότητα, θα ήταν προσοδοφόρο για τη διαδικασία του data mining να εισάγονται και άλλα στοιχεία που αφορούν τον πελάτη στα δεδομένα του πελατολογίου. Η εισαγωγή της ηλικίας θα ήταν ένα πολύ σημαντικό στοιχείο που θα χρησίμευε για τον εντοπισμό προτύπων αγορών στις διάφορες ηλικιακές κατηγορίες. Θα αναδείκνυε ποιά προϊόντα προτιμώνται από τις νεότερες ηλικίες και ποια από τις πιο ώριμες ώστε να είναι επιτυχημένη μια στοχευμένη διαφημιστική εκστρατεία αλλά και να είναι πιο ακριβείς τυχόν προτάσεις των πωλητών κατά τη διάρκεια της πώλησης. Ακόμα, αντίστοιχα αποτελέσματα θα ήταν δυνατό να επιτευχθούν με την προσθήκη του φύλου στα στοιχεία των πελατών, αναδεικνύοντας έτσι προϊόντα ή αγοραστικές συνήθειες και διαφορές που μπορεί να έχουν τα δύο φύλα μεταξύ τους.

9. Παράρτημα

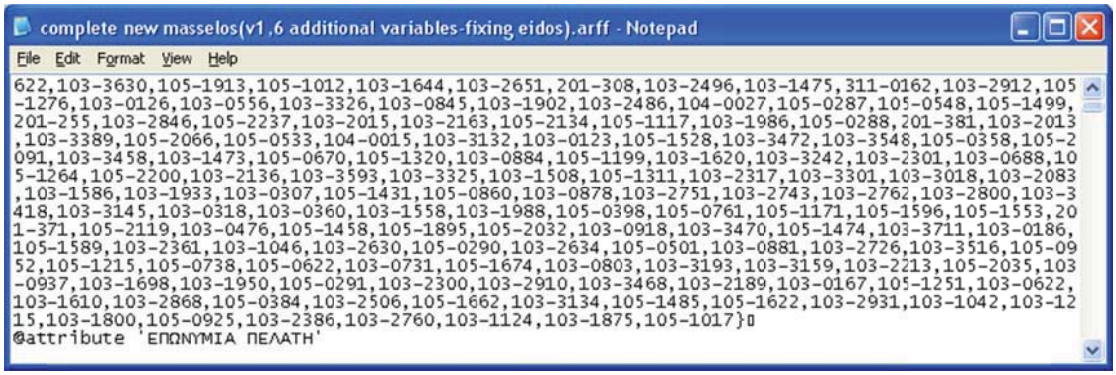
Στις παρακάτω εικόνες φαίνεται η δομή ενός από τα αρχεία *.arff* που χρησιμοποιήθηκαν κατά τη διάρκεια της εργασίας. Δεν είναι τόσο ευκρινές όσο το πρότυπο αρχείο που παρουσιάσαμε στο κεφάλαιο 4.3.1 καθώς η μετατροπή του από *.csv* σε *.arff* έχει γίνει απευθείας από το Weka και στην συνέχεια έχει υποστεί αρκετές επεξεργασίες. Ωστόσο φαίνονται τα βασικά μέρη ενός τέτοιου αρχείου. Στην πρώτη σειρά της πρώτης εικόνας διακρίνεται η χαρακτηριστική φράση «@relation» που ακολουθείται από το όνομα του αρχείου. Επίσης φαίνονται και τα φίλτρα που έχουν χρησιμοποιηθεί για τις διάφορες επεξεργασίες που έχει υποστεί το αρχείο αυτό, το Weka τα ενσωματώνει σε αυτό το σημείο.

Στην συνέχεια διακρίνουμε διάφορες γραμμές με την χαρακτηριστική φράση «@attribute». Αυτή σηματοδοτεί την ύπαρξη κάποιου χαρακτηριστικού, το όνομα του οποίου βρίσκεται αμέσως μετά. Μετά το όνομα φαίνεται ο τύπος του χαρακτηριστικού, numeric για αριθμητικά χαρακτηριστικά, date για ημερομηνίες κλπ. Για τα ονομαστικά (nominal) χαρακτηριστικά δεν βλέπουμε τον τύπο τους αλλά τις τιμές που είναι δυνατό να λάβουν.

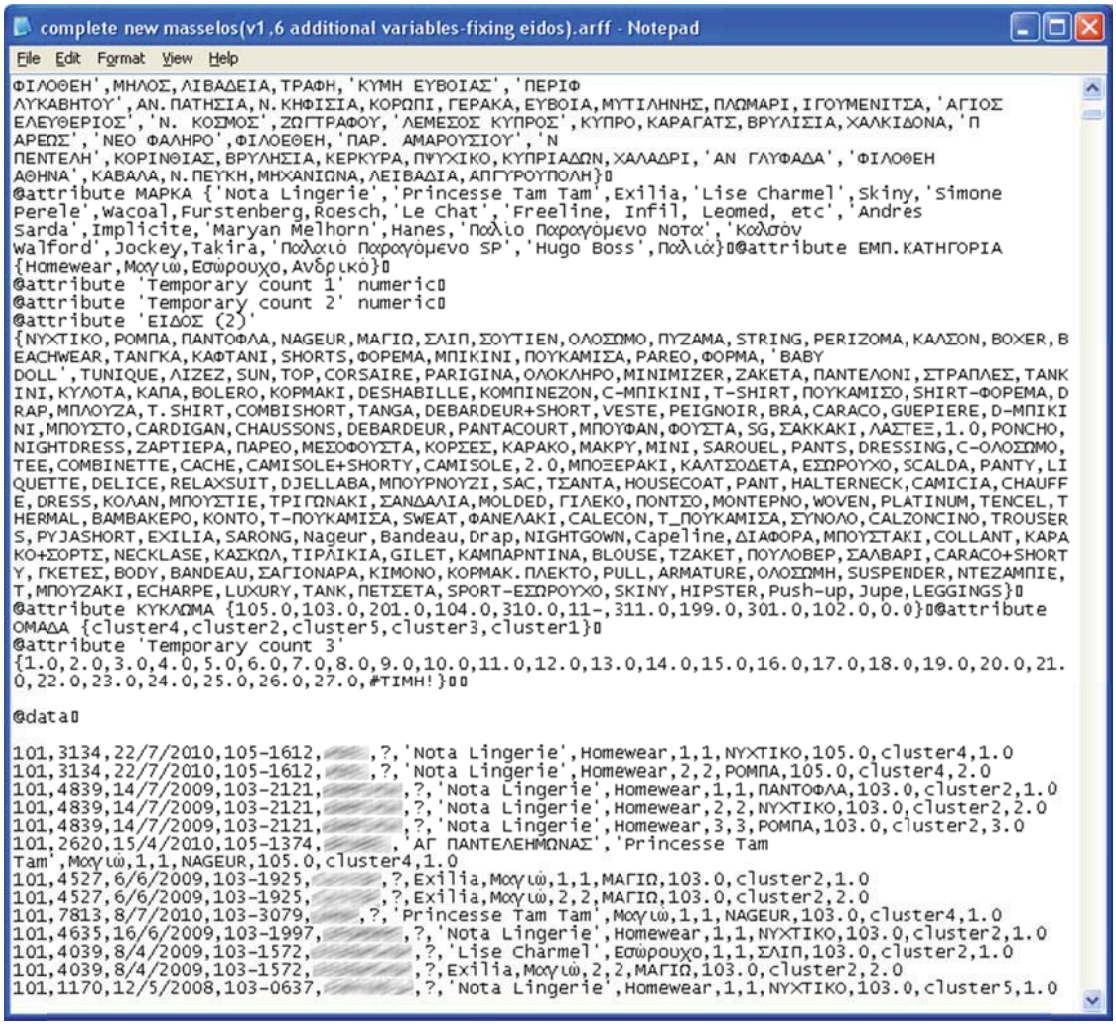
Τέλος στην τρίτη εικόνα διακρίνουμε την φράση «@data» που σηματοδοτεί την αρχή καταγραφής των διαφόρων υποδειγμάτων. Οι τιμές των χαρακτηριστικών στα υποδείγματα χωρίζονται με κόμμα «,», και ακολουθούν την σειρά με την οποία εμφανίστηκαν τα χαρακτηριστικά. Στην πρώτη εικόνα, για παράδειγμα, παρατηρούμε ότι πρώτο χαρακτηριστικό παρουσιάστηκε το «ΕΙΔΟΣ ΠΑΡΑΣΤΑΤΙΚΟΥ», οπότε η πρώτη τιμή σε κάθε υπόδειγμα θα αφορά αυτό το χαρακτηριστικό. Αναλόγως και για τα υπόλοιπα. Τα θολά τμήματα της τρίτης εικόνας στο κομμάτι «@data» αφορούν το χαρακτηριστικό «ΕΠΩΝΥΜΙΑ ΠΕΛΑΤΗ» και έχουν υποστεί επεξεργασία λόγω των προσωπικών δεδομένων που μεταφέρουν.



Εικόνα 17: Αρχείο *.arff (1)*



Εικόνα 18: Αρχείο *.arff (2)*



Εικόνα 19: Αρχείο *.arff (3)*

10. Αναφορές

Συγγράμματα

[1] : Βουτσινάς Βασίλειος, *Θέματα Επιχειρηματικής Νοημοσύνης, Θεωρητική Θεμελίωση και Εφαρμογές*, εκδ. Κωσταράκη Π. Ευρυδίκη, Αθήνα, 2006

[2] : Παγουρόπουλος Απόστολος, *Διπλωματική Εργασία με Θέμα «Data Mining στη Χρηματοοικονομική Ανάλυση»,* Πάτρα, 17/10/2006

[3]: Χαραλάμπους Κωνσταντίνος, *Διπλωματική Εργασία με Θέμα «Οικονομικές Επιπτώσεις Ολυμπιακών Αγώνων και Τεχνικές Εξόρυξης Πληροφορίας: Απόπειρες Εξαγωγής Αποτελεσμάτων και Συμπεράσματα»,* Αθήνα, Ιούλιος 2009

[4]: Jiawei Han and Micheline Kamber, *Data Mining Concepts and Techniques,* εκδ. Morgan Kaufmann, 2006

[5]: Ian H Witten and Eibe Frank, *Data Mining Practical Machine Learning Tools and Techniques,* εκδ. Morgan Kaufmann, 2005

[6]: Χαλκίδη Μ, Βαρζιγιάννης Μ, *Εξόρυξη γνώσης από βάσεις δεδομένων και τον παγκόσμιο ιστό,* εκδ. Gutenberg, 2006

[7]: Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* εκδ. AAAI press/The MIT press, 1996

[8]: Richter M Michael, *Advances in Data Mining: Theoretical Aspects and Applications,* εκδ. Springer, 2007

Ηλεκτρονικές Πηγές

[01] Artificial Intelligence – The basics

http://www.humansfuture.org/artificial_intelligence_tutorial.php.htm

[02] Artificial Intelligence (Scope, Implications & Recent Advancements)

http://www.slideshare.net/guest45deb89/artificial-intelligence-2444532?src=related_normal&rel=200065

[03] Artificial Intelligence - Wikipedia

http://en.wikipedia.org/wiki/Artificial_intelligence

[04] Θεοδωρίδης Γιάννης, Πελέκης Νίκος, «*Αποθήκες Δεδομένων και Εξόρυξη Γνώσης, Εισαγωγή»*

http://infolab.cs.unipi.gr/courses/dwdm/slides/1-DWDM-intro.pdf

[05] Τζιραλής Γεώργιος, *«Αλγόριθμοι Εξόρυξης Πληροφορίας, Διάλεξη 1, Εισαγωγή»*

http://sites.google.com/site/gtziralis/Lecture01_Introduction.pdf

[06]: European Journal of Operational Research, Volume 187, Issue 3, Sigurdur Olafsson, Xiaonan Li, Shuning Wu, «*Operations research and data mining*»

http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VCT-4MBT21C-5&_user=7351776&_coverDate=06%2F16%2F2008&_rdoc=50&_fmt=high&_orig=browse&_srch=doc-info%28%23toc%235963%232008%23998129996%23676722%23FLA%23display%23Volume%29&_cdi=5963&_sort=d&_docanchor=&_ct=57&_acct=C000059671&_version=1&_urlVersion=0&_userid=7351776&md5=1195f9f255d148e9dab0d2e9c39eeeff

[07]: Information&Management, Volume 39, Issue 3, Indranil Bose and Radha K Mahapatra, « Business data mining — a machine learning perspective»

http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VD0-4471P3X-4&_user=7351776&_coverDate=12%2F20%2F2001&_rdoc=4&_fmt=high&_orig=browse&_srch=doc-info%28%23toc%235968%232001%23999609996%23268072%23FLA%23display%23Volume%29&_cdi=5968&_sort=d&_docanchor=&_ct=6&_acct=C000059671&_version=1&_urlVersion=0&_userid=7351776&md5=0d3305f6eb91e16ffa46e6344b748881

[08]: Expert Systems with Applications, Volume 36, Issue 4, Mehdi Toloo, Babak Sohrabi, Soroosh Nalchigar, « A new method for ranking discovered rules from data mining by DEA »

http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V03-4TV2X8G-1&_user=7351776&_coverDate=05%2F31%2F2009&_rdoc=134&_fmt=high&_orig=browse&_srch=doc-info%28%23toc%235635%232009%23999639995%23931095%23FLA%23display%23Volume%29&_cdi=5635&_sort=d&_docanchor=&view=c&_ct=157&_acct=C000059671&_version=1&_urlVersion=0&_userid=7351776&md5=f177ad008aca30577f6a8cc09753b34c

[09]: Τζιραλής Γεώργιος, *«Αλγόριθμοι Εξόρυξης Πληροφορίας, Διάλεξη 4, Απεικόνιση Γνώσης, Αξιοπιστία&Αποτίμηση»*

http://sites.google.com/site/gtziralis/Lecture04_KnowledgeRepresentationCredibilityEvaluation.pdf

[010]: Τζιραλής Γεώργιος, *«Αλγόριθμοι Εξόρυξης Πληροφορίας, Διάλεξη 2, Συνιστώσες Δεδομένων - Οπτικοποίηση και Εξερεύνηση»*

http://sites.google.com/site/gtziralis/Lecture02_DataComponentsVisualizationExploration.pdf

[011]: Data Mining, Predictive Modeling, Techniques

http://www.statsoft.com/textbook/data-mining-techniques/

[012]: Ταυτότητα της Νότα Μασσέλος Α.Ε.

http://www.nota.gr/default.asp?pid=3&la=1

[013]: Ιστορία της Νότα Μασσέλος Α.Ε.

http://www.nota.gr/default.asp?pid=4&la=1